IMPLEMENTING APPROPRIATE MULTIVARIATE METHODS FOR HIGHER QUALITY RESULTS FROM GENETIC ASSOCIATION STUDIES IN SUBSTANCE ABUSE POPULATIONS

by

Derek Beaton

APPROVED BY SUPERVISORY COMMITTEE:

____
Hervé Abdi, PhD, Chair

____
Francesca M. Filbey, PhD

____
Richard M. Golden, PhD

____
Karen M. Rodrigue, PhD

Dedicated to the coffee and craft beer purveyors of the DFW metroplex.

IMPLEMENTING APPROPRIATE MULTIVARIATE METHODS FOR HIGHER QUALITY RESULTS FROM GENETIC ASSOCIATION STUDIES IN SUBSTANCE ABUSE POPULATIONS

by

DERECK BEATON, BS, MS

DISSERTATION

Presented to the Faculty of

The University of Texas at Dallas

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY IN

COGNITION AND NEUROSCIENCE

THE UNIVERSITY OF TEXAS AT DALLAS

May 2017

ACKNOWLEDGMENTS

I would like to thank my advisor Dr. Hervé Abdi and my close mentor Dr. Francesca Filbey. Their guidance, support, and advice were instrumental in the dissertation process, as well as for my NRSA (F31DA035039). Many thanks are due to Dr. Karen Rodrigue and Dr. Richard Golden for their roles in, and beyond, this dissertation. I would also like to thank two generations of Abdi lab members: (1) Dr. Anjali Krishnan and Dr. Joseph Dunlop, and (2) Cherise Chin Fatt, Michael Kriegsman, and Ju-chi Yu. Furthermore a large debt of gratitude is due to all of the BBS faculty, staff, and administration that make the transition from student to scholar possible. I would also like to thank Cara and Debz for their endless patience. Furthermore, I'd like to thank Café Silva, Cultivar Coffee, Goodfriend, and (especially) Lakewood Growler for providing the necessities to get through a dissertation. Penultimately, the dissertation became an easier process because of a small writing group: Nazly and Gérard. Finally and most importantly, this dissertation would not have gotten finished without Jenny's support and motivation.

February 2017

IMPLEMENTING APPROPRIATE MULTIVARIATE METHODS FOR HIGHER QUALITY RESULTS FROM GENETIC ASSOCIATION STUDIES IN SUBSTANCE ABUSE POPULATIONS

Derek Beaton, PhD
The University of Texas at Dallas, 2017

Supervising Professor: Hervé Abdi

For nearly a century, detecting the genetic contributions to cognitive and behavioral phenomena has been a core interest for psychological research. Recently, this interest has been reinvigorated across many related domains including and especially psychiatric research. Furthermore, genotyping technologies (e.g., microarrays) that provide genetic data, such as single nucleotide polymorphisms (SNPs), are routinely available and easily accessible to almost any researcher. These SNPs—which represent pairs of nucleotide letters (e.g., AA, AG, or GG) found at specific positions on human chromosomes—are best considered as categorical variables. However, a categorical coding scheme can make difficult the analysis of their relationships with behavioral, diagnostic, or clinical measurements because most multivariate techniques developed for the analysis between sets of variables are designed for quantitative variables. Furthermore, there are many—not just one or a few—genetic contributions to complex behaviors and disorders such as substance abuse, thus requiring multivariate techniques to fully understand the many genetic contributions. To palliate this problem, I present a generalization of partial least squares (PLS)—

a technique used to extract the information common to two different data tables measured on the same observations—called partial least squares correspondence analysis (PLS-CA)—that is specifically tailored for the analysis of categorical and mixed ("heterogeneous") data types. I further extend PLS-CA with a ridge-like regularization called Smoothed PLS-CA (SmooPLS-CA). SmooPLS-CA adjusts for overfitting and noise that can lead to the interpretation of spurious effects in high dimensional-low sample size data such as genetics and genomics. PLS-CA and SmooPLS-CA were both applied to two genetic data sets within substance use disorders (SUDs) that focused on a large number of genes: an archived set ("discovery") and an external set ("validation"). The goal of the two data sets were to discover markers of SUDs in one set, and then validate those markers in an independent and completed sequestered set. SmooPLS-CA showed no advantage over standard PLS-CA: bootstrap resampling techniques provided robust results regardless of regularization. Finally, multiple genes were identified as contributors to a broad case-control (i.e., SUDs vs. control group) effect. Some of the identified genes play key roles in the glutamatergic (e.g., GRIN2B) and dopaminergic systems (e.g., CCKBR), where other genes play complex or even undefined roles (e.g., PRKCE). In sum there are many robust, albeit small, genetic effects as opposed to only a few large effects that contribute to SUDs.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Substance use disorders (SUDs)—like many of the DSM-5's criteria for diseases and disorders—are defined by a continuum (a.k.a., spectrum) that reflects the severity of the disease. While SUDs are defined differently for various types of substances (e.g., alcohol, stimulants, food), SUDs typically share many symptoms. These symptoms include behavioral and cognitive impairments *because* of substance use, often in social and work domains (DSM5, 2015; NIDA, 2015). Other symptoms—and consequences—of SUDs also include lack of self-regulation (Baumeister & Heatherton, 1996; Beaton, Abdi, & Filbey, 2014; Khantzian, 2013), as well as increased use and craving, in general for SUDs (Everitt & Robbins, 2016; Volkow, Koob, & McLellan, 2016, Sinha, 2013; Tiffany & Wray, 2012) and for specific substances (Di Nicola et al., 2015; Filbey, Schacht, Myers, Chavez, & Hutchison, 2009b; Meule, Lutz, Vögele, & Kübler, 2014). SUDs and their corresponding behaviors have substantial impact (such as medical and financial problems) on individuals, those around them, and society at large because—as noted by the National Institute on Drug Abuse (NIDA)—drug abuse has major economic effects (caused, e.g., by spanning purchase, health care, criminal costs, and lack of productivity) that exceed $700 billion, and contributes to nearly 600,000 deaths per year in the United States alone (NIDA, 2015). But, at the individual level, SUDs have an even higher cost evaluated by their impact on family, friends, and those who actually suffer from SUDs and addiction. Individuals with SUDs often must confront social, work, and financial issues, and even—in cases of particularly dangerous substances such as opiates and heroin—are putting at risk the health and lives of themselves and, sometimes, others. For example, a rural part of southern Indiana has recently experienced an outbreak of H.I.V., due (almost exclusively) to intravenous drug use (Goodnough, 2015). However, Indiana is not an isolated case; there has been a substantial increase in drug abuse—across all substances—through out the United States over the past decade (Jones, Logan, Galdden, & Bohm, 2015).

People with SUDs have, for a very long time, been associated with the stigma of "lack of will power" (Lieberman, 2015). However, public attitude towards SUDs and drug abuse have changed substantially in recent years; to such a degree that it has been a topic of honest discussion and debate during the 2016 presidential primary, in a state particularly hard hit by addiction (New Hampshire; Keith, 2016). For example, marijuana is now seen by some as a safer substance to use than alcohol (Nutt et al., 2010; Pew Research Center, 2014). Furthermore, public attitude and policy are finally aligning, in part, with modern psychiatric science: SUDs—like other psychiatric disorders—are *diseases* that are better served by treatment, not by jail time. In fact, Massachusetts—one of the places hit the hardest, in recent years, by drug abuse—has started to address these issues: The Gloucester police department, and other agencies, now grants amnesty for individuals suffering from opiate abuse (Beck, 2015). However, while there are a number of viable treatment options, the origins of SUDs and addiction in general remain poorly understood.

SUDs are very heritable disorders (Agrawal et al., 2012). There have been numerous—often family or twin—studies illustrating that SUDs, drug abuse—and even other forms of addiction, such as gambling—are explained by more than just environmental factors (Kendler et al., 2012). In fact, SUDs are believed to be, in part, biological (i.e., genetic) in origin (Loth, Carvalho, & Schumann, 2011; Palmer et al., 2015; Uhl, 2004). Today, SUDs are often studied with candidate gene studies (i.e., one or just a few pre-selected genes), large-scale candidate gene panels (e.g., thousands of markers, as in Hodgkinson et al., 2008; Saccone et al., 2007; Beaton et al., in prep), and even genome-wide studies (Bierut et al., 2010; Wetherill et al., 2015).

Genome-wide technology—intended to provide massive, but respectively sparse, genotypic data across the entire genome—was considered a "breakthrough" (Pennisi, 2007) that promised to improve our understanding of health, personality, individual differences, and even causal effects of genes (Stranger, Stahl, & Raj, 2011; Weiner & Hudson, 2002). The accessibility and low cost of genotyping array technology have led to an abundance of genome wide studies (GWAS) in numerous domains: GWAS central (www.gwascentral.org) lists 1,831 studies, NHGRI (and EBI; http://www.ebi.ac.uk/), lists 2,223 GWA studies, and the keyword "GWAS" generated over 47,800 hits on Google Scholar[1]. A substantial amount of GWAS growth can be attributed to studies in the neuropsychiatric domains, such as the Alzheimer's Disease Neuroimaging Initiative (ADNI), Psychiatric Genetics Consortium (PGC), and the Study of Addiction: Genetics and Environment (SAGE). Despite such a range of applications, large-scale genetics and genomics studies have yet to yield any "breakthroughs," especially in substance abuse research (Hutchison, 2010). Furthermore, there have been varied results with relatively massive samples in other well-studied psychiatric disorders (such as schizophrenia: Arnedo et al., 2014; Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014; Sekar et al., 2016; The Network and Pathway Analysis Subgroup of the Psychiatric Genomics Consortium, 2015).

[1] All as of October, 2015.

Currently, there is a distinct shortage of (bio) statistically appropriate data-analytic tools designed by, and especially for, research in the genetics of SUDs. This is underscored by the somewhat recent NIDA executive summary—punctuated by a variety of funding calls since—that expressed caution on the interpretation of genome-wide—and more generally, genetic—studies for substance abuse populations (NIDA, 2010). In general, NIDA—and the field at large—have recognized that there are substantial drawbacks to current approaches and methods to detect the genetic contributions to SUDs.

One potential issue is that SUDs—like many other complex psychiatric disorders and diseases—are likely to be, at the very least, oligogenic ("a few genes") or, much more likely, polygenic ("many genes"). Yet, most studies use methods and tools that were designed around the "OGOD" (one-gene one-disorder) principle; an idea that is considerably outdated and nearly abandoned for complex traits, diseases, and disorders—see, for example Xiaolin Zhu, Need, Petrovski, and Goldstein (2014) where the authors point to many instances of one gene (at a time) that appears to contribute to *many* neuropsychiatric disorders. Often, the goals, methods, and tools still used in these studies are designed to find the best *single* marker (a.k.a., "causal variant", and therefore use a series of univariate tests). Further complicating the matter, studies—not just in SUDs, but across almost any genetic or genomic study—are often conducted with just two groups: a control group and a case group (e.g., a specific SUD). So, unfortunately, the growth in the number of large-scale genetics and genome-wide studies, with multiple groups—especially in alcohol and drug addiction research—is not paralleled by a growth of better, more appropriate, and more rigorous statistical methods or tools designed to fully exploit these complex data sets. Thus, to fully exploit the complexity of genetics in SUDs, we must first address several statistical and methodological issues. In my dissertation (as an extension of some of my current work: NIH F31DA035039), I intend to address some of these issues. Broadly, there are two aims of my dissertation: (1) develop better, more powerful, statistical techniques for genetic and genomic association studies for the behavioral and brain sciences, and (2) apply these techniques to identify a potential better set of genetic markers to characterize (general and specific contributions to) substance use disorders.

1.1 Research Questions

To address the issues of high-dimension, low sample size (HDLSS) genetic data, I extended a recently developed framework for multivariate analysis of genetic and genomic data: Partial Least Squares-Correspondence Analysis (PLS-CA; see Chapter 4; Beaton, Dunlop, et al., 2016). In Beaton, Dunlop, et al., (2016) we established a multivariate framework—based on partial least squares—designed around the categorical nature of genetic (and many other, e.g., clinical, diagnostic) data types, but that can be adapted for use of mixed-data types (i.e., any combination of categorical, ordinal, continuous; Beaton, Dunlop, et al., 2016). Currently, PLS-CA—like most other techniques— relies in part on *post-hoc* inference techniques to identify genetic markers of interest. However, the current trend in the statistical sciences—especially with respect to extremely high-dimensional data sets—is that some form of regularization (or sparsification) is not only beneficial for interpretation of high-dimensional data, but also better suited for over-fitting and ill-posed problems (Jolliffe, Trendafilov, & Uddin, 2003). Currently, there are only a few approaches that exist that could potentially enhance PLSCA with regularized (or sparsified) conditions (Allen, 2013; Josse, Chavent, Liquet, & Husson, 2012; Takane & Hwang, 2006; Verbanck, Josse, & Husson, 2013).

Given the statistical trends for HDLSS problems (e.g., regularization, sparsifcation), I have developed a new version of PLS-CA: Smoothed PLS-CA ("SmooPLS-CA"). Both PLS-CA and SmooPLS-CA are compared against one another and applied to a subset of genome-wide data from SUDs populations. The broadest aim of this dissertation is similar to some of my recent work (Beaton et al., in prep; NIH F31DA035039): With the right data and analytical tools, do we have sufficient power to detect genetic markers in (relatively) small sample sizes? In this study, I will first analyze (separately with PLS-CA SmooPLS-CA) data from a relatively small genome-wide data set (*N* = 431; from Dr. Francesca M. Filbey). The markers identified in this analysis will then be used in a completely sequestered, and much larger data set (original *N* > 3,285; Study on Addiction: Genetics and Environment; SAGE). The SUD groups of interest in this dissertation are marijuana and nicotine users. The goal here is to see if we can, in a small data set, identify adequate genetic markers that will sufficiently predict SUD diagnosis in a much larger, excluded, data set (i.e., a validation step). My dissertation studies were as follows:

1. Design and implement a theoretically sound approach to minimize the number of interpretable items, while also accounting for over-fitting with PLSCA (with respect to genetics).
2. Application to genome-wide based data:
    a. Apply standard PLSCA and SmooPLS-CA to a relatively (for genetics) small sample data set; markers identified in either will be used to classify participants (with respect to their SUD or a control group) in an excluded data set. If power happens to be sufficient in the "small" analysis, classification should be reasonable in the excluded data set.

b. Compare which approach is more powerful: standard PLSCA vs. SmooPLS-CA (with *post-hoc* inference tests); this helps answer a statistical question about which pipeline is more powerful and directly addresses analytical tradeoffs (e.g., computational time vs. accuracy).

Generally, genome-wide and large-scale genetics analyses are considered "sufficiently powered" when $N \geq 5{,}000$. However, many branches of the NIH (and other agencies) have provided funding to collect genome-wide data to many researchers; although, almost no single researcher—especially in brain and behavioral sciences—can attain a sample size "sufficient" for a genome-wide study. My dissertation will address a broader, more important question: can we analyze small sample, large-scale genetic data sets and trust those results? If so, this will provide a practical and economical set of tools that can be used to investigate the genetic nature of complex neurological and psychiatric disorders.

CHAPTER 2

SUBSTANCE USE DISORDERS AND THEIR LIKELY GENETIC CONTRIBUTIONS

Substance use disorders (SUDs)—like many of the DSM-5's criteria for disorders—are defined by severity on a scale (DSM5, 2015), where the term "addiction" is reserved for a chronic state of extreme SUD severity (Volkow et al., 2016). Severe SUD states are often characterized by a desire to quit, but self-control (or self-regulation) is very low, and compulsive substance use is very high. Symptoms of SUDs often include behavioral and cognitive changes and impairments leading to impairments often in social and work domains (NIDA, 2015).

SUDs are heritable (Agrawal et al., 2012), where the heritability of SUDs is in part biological (Uhl, 2004) and in part environmental (as assessed through family and twin studies; Kendler et al., 2012)—with upwards of 50% of SUDs explainable by genetic factors (Volkow, Wang, Fowler, & Tomasi, 2012). In recent years, the National Institute of Drug Abuse (NIDA) and the National Institute on Alcohol Abuse and Alcoholism (NIAAA) have placed a substantial amount of resources in, and have prioritized the study of human genetics and its contribution to SUDs (expressed by NIDA in **NOT-DA-12-012).** With a better understanding of the genetic mechanisms of substance abuse and addiction, researchers and clinicians could design better, more targeted treatments than those currently available (NIDA, 2010, p. iv).

In this chapter, I provide a comprehensive literature review on the genetic contributions to marijuana and nicotine use disorders—with additional material from other SUDs and psychiatric and neurological disorders. I also provide additional reviews of specific models of SUDs (as they pertain to the genetics of SUDs), as well as genes related to various aspects of behavioral and brain research, with particular emphasis on psychiatric disorders. Parts of this chapter are adapted from my qualifying thesis, my F31 grant application, and some on-going work (Beaton, Abdi, & Filbey, in prep) as well as Beaton, Filbey, & Abdi (2014). Beaton, D., Abdi, H., & Filbey, F. M. (2014) Unique aspects of impulsive traits in substance use and overeating: specific contributions of common assessments of impulsivity. The American journal of drug and alcohol abuse, 40(6), 463-475. http://dx.doi.org/10.3109/00952990.2014.937490 reprinted by permission of Taylor & Francis LLC (http://www.tandfonline.com).

2.1 Behavioral and physiological (neurological) aspects of SUDs and addiction

There are a number of behaviors, as well as physiological and neurobiological changes, associated with SUDs and addiction. Some of these behaviors include craving for a substance (Filbey et al., 2009b; Volkow, Fowler, et al., 2010), risk-taking (DeWitt, Aslan, & Filbey, 2014; Fernie, Cole, Goudie, & Field, 2010), lack of self-regulation (Cole, Logan, & Walker, 2011; Heatherton & Wagner, 2011), and disruption in executive functions such as decision making (Bickel, Koffarnus, Moody, & Wilson, 2014; Ernst et al., 2003) and cognitive control (Dalley, Everitt, & Robbins, 2011; Pharo, Sim, Graham, Gross, & Hayne, 2011). SUDs also tend to co-occur with stress and trauma-based disorders (Read et al., 2012; Tull, Gratz, Coffey, Weiss, & McDermott, 2013).

One of the most studied behavioral types associated with SUDs: impulsivity (and associated aspects): a relatively complex trait often studied across many domains including personality disorders (Witt et al., 2010), and self-regulatory failures (Baumeister & Heatherton, 1996) such as substance use and overeating disorders (Dawe & Loxton, 2004). In SUDs, high levels of impulsivity—which are believed to also be a risk factor for addiction (Kreek, Nielsen, Butelman, & LaForge, 2005)—may be associated with an increase in drug use (Perry & Carroll, 2008), alcohol use (Fernie et al., 2013), and pathologic substance abuse (Belin, Mar, Dalley, Robbins, & Everitt, 2008). High levels of impulsivity could also impact treatment strategies (Bankston et al., 2009). “Impulsivity” is a multifaceted and heterogeneous concept that includes aspects of disinhibition, inattention, sensation seeking, and deficits in decision-making (Evenden, 1999). Prior work has shown traces of unique impulsivity traits associated to different substance users. For example, Meda et al., (2009) derived a five-factor model based on state and trait measures in healthy controls vs. “at-risk/addicted” participants, where Beaton et al., (2014) showed there are multiple, non-linear—and much more nuanced—facets to impulsivity in SUDs including traits unique to marijunana + nicotine co-users.

While various concepts play a role in SUDs—either through initiation of use, or maintenance of use—such as impulsivity, stress, and dysregulation (Jentsch et al., 2014; G. Koob & Kreek, 2007), interoception (Naqvi & Bechara, 2010; Noël, Brevers, & Bechara, 2013), and exteroception (DeWitt, Ketcherside, McQueeny, Dunlop, & Filbey, 2015), there are two primary (and prominent) models of SUDs and addiction: (1) reward (“positive” reinforcement; Robinson & Berridge, 2008) and (2) stress (“negative” reinforcement; Koob & Le Moal, 2001; Koob & Kreek, 2007), which play substantial roles in (1) early and intentional use stages (i.e., initiation) and (2) maintenance of substance use (often related to withdrawal).

The reward and stress hypotheses play important theoretical roles in the genetics of SUDs, because both the reward and stress circuits (mostly, the dopaminergic and corticoid systems) are highly involved in SUDs and addiction. Furthermore, we have a generally good understanding of the biology—and thus genetics—of the reward and stress circuits. Though, it is now more generally understood that, while these two systems are primary contributors (perhaps to SUDs and addiction in general, not for specific SUDs), there are other (neural and genetic) contributors to SUDs and addiction including specific genetic systems for particular disorders (e.g., nicotinic receptors for nicotine, cannabinoid receptors for cannabis; Volkow & Muenke, 2012; Volkow, Wang, Fowler, Tomasi, & Telang, 2011). Below I provide a brief description of these circuits to provide a context of why particular genes (1) have been proposed as candidate genes, (2) are almost routinely found in a variety of SUDs and psychiatric disorders, and (3) within both of these circuits have brought about two separate polygenic (i.e., multiple gene) panels to explain SUDs (and even other psychiatric disorders).

2.1.1 The reward circuit

The "reward circuit" (Feder, Nestler, & Charney, 2009; Volkow et al., 2011) is primarily a dopaminergic system. At the seat of the reward circuit are two key regions: the nucleus accumbens (NAc) and the ventral tegmental area (VTA); these regions regulate response to reward, and are the origin of dopaminergic cells, respectively. From here, the NAc and VTA feed into the dorsal striatum (which plays a role in aspects of cognition, e.g., executive function) and motor cortex (for actions). The NAc and VTA are, in part, top-down regulated by frontal regions (e.g., anterior cingulate cortex), and more medial structures involved in emotion, memory, and learning, such as the hippocampus and amygdala. To summarize, the reward circuit is important because it is largely believed that the top-down influence of the NAc and VTA can trigger a dopaminergic (reward) response, which, in turn, elicits an action to engage in reward seeking behavior (e.g., substance use; via dorsal striatum and motor regions, in the first stage as described in Volkow et al., 2016). While this is the "core" reward circuit, other regions—such as the precuneus—may also play a role in reward processing (Filbey & DeWitt, 2012).

2.1.2 The stress circuit

The stress circuit is primarily a corticoid system (mostly regarded as a glucocorticoid, but there are also other corticoids involved, such as mineralocorticoid; Vogel, Fernández, Joëls, & Schwabe, in press, 2016). At the seat of the "stress" circuit is the hypothalamic-pituitary-adrenal (HPA) axis. The HPA axis—via the adrenal gland—releases glucocorticoids that regulate, or can be triggered by, other brain regions and "brain stress systems" (Koob & Le Moal, 2001; López, Akil, & Watson, 1999). Regions regulated by the HPA axis include, for example, the amygdala, hippocampus, prefrontal cortex, and—importantly—the NAc. This circuit is important because it is believed to be related to maintenance factors (as in stress-relief a.k.a. "self-medication"; second stage described in Volkow et al., 2016). This same "stress" circuit is also called the "resilience" circuit (Morrow & Flagel, 2015) as it may help to *avoid* further substance use as well as provide "resilience" (a.k.a. a protective factor) against other disorders (e.g., PTSD; Feder et al., 2009).

2.2 Genetics of Substance Use Disorders

The biological (including genetic) underpinnings of SUDs have been a primary concern in the field that has spanned almost four decades (Camps, 1972; Cruz-Coke, 1982; Devor & Cloninger, 1989). In 1989, the NIAAA began a study called the "Collaborative Study on the Genetics of Alcoholism" (COGA; http://grantome.com/grant/NIH/U10-AA008401-27)—related to another study included in this dissertation: "Study on Addiction: Genes and Environment" (SAGE)—which primarily addressed alcohol use disorders (at the time referred to as "alcoholism"), but also other—believed to be heritable—SUDs (Bierut, Dinwiddie, Begleiter, et al, 1998). The clear familial transmission of several SUDs (with the help of the COGA study and its individual studies), spurred some of the earliest work—in alcohol use disorders—on specific genetic contributions (Heinz, Mann, Weinberger, & Goldman, 2001; Long et al., 1998).

However, before the advent of inexpensive and widely-available genotyping platforms, some of the earliest studies—beyond alcohol use disorders—were twin and family studies designed to estimate the heritability of SUDs, as well as estimate the contributions from genetic and/or environmental factors (Kendler, Jacobson, Prescott, & Neale, 2003; Kendler, Prescott, Myers, & Neale, 2003)—a paradigm that is still employed today (Kendler, Sundquist, Ohlsson et al., 2012; Lynskey et al., 2012). Some important contributions of twin and family studies show combined genetic and environmental effects (Kendler, Schmitt, Aggen, & Prescott, 2008; Vrieze, Hicks, Iacono, & McGue, 2012).

Shortly after many of the family and twin studies (up to the mid-to-late 2000s, and even in recent years), this work was extended to also estimate contributions of, and relationships with, behavioral (sometimes referred to as “phenotypic”) and other biological, physiological, and neurological (sometimes referred to as “endophenotypic”) measures (see, e.g., Meyer-Lindenberg & Weinberger, 2006) in SUDs (Edenberg & Foroud, 2006; Palmer et al., 2015; Wetherill et al., 2015).

Around the time of these large-scale family studies, genotyping, and in particular genome-wide, platforms became much more accessible (and usable) to researchers outside the biological sciences. In brain and behavioral sciences, one of the most typical types of genetic data—called single nucleotide polymorphisms (SNPs; pronounced “snips”)—lists the possible alleles of a nucleotide pair at a given position for the corresponding chromosomes (i.e., one maternal and one paternal). In practice, SNPs are detected as their DNA nucleotide pairs, called genotypes. These genotypes are in general classified as the major homozygote, heterozygote, and minor homozygote (e.g., AA, AG, GG, respectively, assuming that AA is found more often than GG in the population of interest) where zygosity is determined by allele frequency (e.g., G is a minor allele because it is less frequent than A). The general form of SNPs is to consider *A* the major allele (most frequent), and *a* the minor allele, where *AA* is the major homozygote, *aa* the minor homozygote, and *Aa* the heterozygote.

Studies have been performed with single or multiple (small and large scale) candidate (hypothesis-driven; relatively small) gene sets, or genome-wide (exploratory; very large) arrays (which primarily use SNPs) in nearly all SUDs such as cocaine (Bi, Gelernter, Sun, & Kranzler, 2014; Gancarz et al., 2014), nicotine (Hancock, Reginsson, et al., 2015), heroin (Hancock, Levy, et al., 2015), and even comorbid disorders such as alcohol dependence and obesity (K.-S. Wang, Zuo, Pan, Xie, & Luo, 2015; L. Wang et al., 2013). In fact, a number of reviews outline the genetic (and other biological) contributions of SUDs, related disorders, and related traits (Bevilacqua & Goldman, 2011; Loth, Carvalho, & Schumann, 2011; Munafò & Flint, 2011). Though there has been a substantial amount of work to understand the genetic contributions to SUDs, results tend to be weak, inconsistent, or not replicable (Hutchison, 2010 p. 579), and sometimes originated from methods designed to work around some of the genetic complexities (e.g., “multilocus (genetic) profile scores”; as in Papiol et al., 2014). Even though there are inconsistencies within the literature—especially with respect to exact SNPs and genotypes, as well as magnitudes and directions of effects—there do appear to be a number of very likely genetic candidates for SUDs in general and even specific SUDs.

2.2.1 How do we know which genes to study?

Many of the genetic markers of interest in SUDs and addiction come from a number of other physiological, biological, and (especially) neurobiological effects associated with SUDs, and in particular: reward and stress processes associated with SUDs and addiction. Studies of substance abuse and addiction regularly find brain regions that are strongly associated with particular neurotransmitter systems—especially those related to reward (Volkow, Wang, et al., 2010; Volkow et al., 2011)—such as serotonergic, cholinergic, dopaminergic, and—to a lesser degree opioid—systems. Often, these studies show that brain regions associated with particular neurotransmitter systems work differently in substance abusing individuals (as compared to controls). Furthermore, these differences are apparent across many specific substances such as cocaine (Adinoff et al., 2010; Volkow, Fowler, et al., 2010), alcohol (Beck et al., 2009), marijuana (Filbey & DeWitt, 2012; Filbey et al., 2009b), nicotine (Bühler et al., 2010) and even food (Gearhardt et al., 2011; Horstmann et al., 2011). Given that a wide body of research implicates multiple neurotransmitter systems—across multiple substances—then it would follow that the genes that regulate these systems are likely contributors to substance abuse. Furthermore, a number of these neural systems have relatively well known genetic contributions. For example, some studies—in just alcohol abuse—revealed a diverse set of neural systems with genetic contributions across glucocorticoid (Desrivières et al., 2011), opoid (Bart et al., 2004), serotonergic (Kranzler, Hernandez-Avila, & Gelernter, 2002), and dopaminergic, as well as opioid (Filbey et al., 2008) systems.

2.2.2 Specific genes of nicotine abuse

Like most SUDs, there appears to be a transmissible (or heritable) factor for nicotine use, abuse, and/or dependence within families (Buka, Shenassa, & Niaura, 2003) or other familial and social environments (Kendler et al., 2008). Primary and meta-analytic studies mention genes related to nicotinic acetylcholine receptors (nAChRs). The most common genes associated with nAChRs use the symbol prefix of *CHRN* (e.g., *CHRNA5*, *CHRNB4*)[2]. Markers for *CHRN** genes are routinely implicated in nicotine use, abuse, and dependence. For example, both *CHRNA6* and *CHRNB3* (as candidate genes) have been associated with subjective responses (self-report of effects with respect) to tobacco use (Zeiger et al., 2008). Additionally, some of the strongest genome-wide (including meta-analytic) effects have been observed in a variety of nicotinic receptor genes (most are contained within chromosome 15): *CHRNA6*, *CHRNB3*, and *CHRNA5* (J. Z. Liu et al., 2010; TAG, 2010; Thorgeirsson et al., 2010), where *CHRNA5* has shown some of the strongest effects with (subjective) response to tobacco usage (L.-S. Chen et al., 2012); all of these genome-wide nicotinic receptor effects reinforces some of the strongest effects observed in more narrow, candidate-like approaches with the *CHRNA5/A3/B4* cluster (Greenbaum, Rigbi, Teltsh, & Lerer, 2009; Saccone et al., 2007). Importantly, SNPs in this same chromosomal region appear to play a role in nicotine addiction in both African American and European American populations (Culverhouse et al., 2014). Additionally, an earlier study showed that serotonin transporter and receptor genes (*SLC6A4*, *HTR3A*, *HTR3B*) also contribute to nicotine dependence in African American and European American cohorts (Yang et al., 2013).

[2] For the purposes of brevity, hereafter I use a wildcard symbol (*) in cases where multiple genes with the same "prefix" (e.g., CHRN*) are implicated in an effect.

In two of the most recent analyses—one of which was a genome-wide meta-analysis—some “usual suspects” (i.e., the *CHRNA5/A3/B4* group on chromosome 15) showed a strong effect on nicotine use, but markers associated with *CHRNA4* showed the strongest association with nicotine dependence (via the Faegerstrom’s Test for Nicotine Dependence; Hancock, Reginsson, et al., 2015). Furthermore, a genome-wide meta-analysis derived set of candidate markers for polygenic risk scores (in the same *CHRNA5/A3/B4* cluster, with some additional markers, e.g., *EGLN2*, *ADAMTS7*) also showed strong association with a variety of nicotine use phenotypes (e.g., initiation, dependence, family usage; see Belsky et al., 2013). *CHRN** markers also appear to play a role in the mortality rates of smokers (Kupiainen et al., 2016).

While *CHRN** genes are some of the most reliable markers for nicotine usage and dependence, other genes appear to show similar effects: *DRD4* has also been implicated as a mediating factor of subjective response to tobacco usage (after abstinence; Harrell et al., 2015). In fact, a variety of dopamine markers—including *DRD1* (W. Lee et al., 2012) and the *DRD2/ANKK1* cluster (Voisey et al., 2012)—have also been shown to either contribute to nicotine usage and dependence, or to interact with (or modulate) the effects of *CHRN** genes in nicotine administration and attention paradigms (Ahrens et al., 2015; Breckel et al., 2015). Additionally, *DRD4* genotypes appear to mediate reward to smoking cues in short-term abstinent smokers (Xu et al., 2014). Dopamine (receptor and transport) genes are critically important to neurological reward processing, and play a major role in a particular genetic panel for SUDs (see later section “The reward panel”). Beyond *CHRN** and *DRD** genes, Schlaepfer, Hoft, and Ehringer (2008) provide a review of co-addiction with nicotine, and suggest that *CHRN** and *GABA*-ergic genes play a role in nicotine use and other SUDs (e.g., alcohol use disorders).

Though *CHRN** genes typically appear to show the strongest effects and are reported most often in the literature (Ware & Munafò, 2014), some researchers have pointed to yet other—somewhat unexpected—genes. Jensen and Sofuoglu (2015) make a case for *FKBP5*—a gene, which plays a critical role in stress/resilience (especially with respect to childhood trauma and PTSD), and is an important gene in another genetic panel for SUDs (see later section "The stress/resilience panel")—as a mediator of stress response during nicotine withdrawal. Finally, other markers—particularly glutathione transferase genes—have shown increased effects of anxiety in co-morbid nicotine and mood disorder populations (Nunes et al., 2014).

In sum, research focused on nicotine use, abuse, and dependence, has yielded only a few promising candidate markers, though these markers are what would be expected for nicotine dependence (e.g., *CHRN**), behavioral and neural effects of reward/craving (e.g., *DRD**), and even the "stress" of quitting (e.g., *FKBP5*).

2.2.3 Specific genes of marijuana abuse

Marijuana misuse (including abuse and dependence) like nicotine use is known to be "transmissible" within families (Hopfer, Stallings, Hewitt, & Crowley, 2003). While environment certainly plays a role in marijuana use like it does for other substances, family and linkage studies showed specific genetic markers in the *GABA*-ergic and endocannabinoid systems (i.e., *GABRA2* and *CNR1*, respectively) related to marijuana use within family probands (Agrawal et al., 2008). There also exist moderate and small effects in African Americans and European Americans, respectively, of *NRG1* in cannabis dependence (Han et al., 2012). Furthermore, a primary endocannabinoid receptor—*CNR1*—has been shown as a likely candidate—because it is a target receptor for cannabinoids (Agrawal et al., 2009). *CNR1* has also been shown to work in conjunction with a gene that breaks down endocannabindoids—*FAAH*—to influence withdrawal and craving effects (Haughey, Marshall, Schacht, Louis, & Hutchison, 2008). Additionally, a number of markers have been suggested to play a role in chronic marijuana use (including abuse and dependence), including *CNR1* and *CNR2*, *FAAH*, *GABRA2*, *DRD2* (or *ANKK1* due to locus proximity and gene-gene/protein-protein interactions), *SLC6A3* (previously known as *DAT1*), and *OPRM1* (Agrawal & Lynskey, 2009). However, the only genome-wide association studies on cannabis dependence showed no (traditionally) significant effects, but some moderate effects of the *ANKFN1* and *FTO* genes (Agrawal et al., 2011; Agrawal et al., 2014).

Beyond these markers, there are few other candidate genes that have been studied. The *ABCB1* gene—which may play a role in tetrahydrocannabinol distribution—was shown to increase the risk of cannabis dependence (Benyamina et al., 2009). Additionally, the *DRD4* gene—a prominent gene reward processing—plays a role in marijuana use (Vaske, Boisvert, Wright, & Beaver, 2013), as well as marijuana use and depression comorbidity (Bobadilla, Vaske, & Asberg, 2013). In fact, a number of other genes studied in marijuana use typically stem from comorbidities. The *COMT* gene—which is integral in the transport and degradation of dopamine (and other transmitters)—appears to influence age of onset and marijuana usage in schizophrenia (Estrada et al., 2011; Pelayo-Terán et al., 2010). Finally, a haplotype of the *FK5BP* gene—a prominent gene in stress/resilience, and as noted plays a role in many stress-related disorders—plays a small role in marijuana dependence in adolescents that were mistreated as children (Handley, Rogosch, & Cicchetti, 2015).

In sum, research focused on marijuana use, abuse, and dependence has identified multiple candidate markers. However, the most likely candidate markers appear to be related to cannabinoids (i.e., *CNR1*, *CNR2*, *FAAH*), to both the *GABA*-ergic (i.e., *GABRA2*) and dopaminergic (i.e., *SLC6A3*, *DRD2*, *DRD4)*, and to a lesser extent the *COMT* systems.

2.2.4 Genetics of related behavioral and neural traits, phenotypes, and endophenotypes

Part of the motivation behind studying associated (with SUDs) traits, as opposed to strict categorical diagnoses/groups, is that strict categories may contain heterogeneous "diagnoses"; these ideas are inherent in the Research Domain Criteria (RDoC; https://www.nimh.nih.gov/research-priorities/rdoc/index.shtml; Yee, Javitt, & Miller, 2015). However, many DSM diagnoses are also moving towards "spectrums" and "dimensions" of severity (Weinberger, Glick, & Klein, 2015). The field of "imaging genetics" is based on the premise that, biologically, we should see a better explanation of variance from genes to brain function than we would from genes to behavior, or genes to diagnosis because brain function is "closer" to genetic function than behavior or diagnosis (Meyer-Lindenberg & Weinberger, 2006); though, it has been argued that with the correct set of behavioral instruments, it is possible to boost the genetic signal (Beaton, Dunlop, et al., 2016; Bloss, Schiabor, & Schork, 2010). However, the idea that something is "closer" to genetics is not without scrutiny (Iacono, Vaidyanathan, Vrieze, & Malone, 2014).

Before reviewing the literature about genetic contributions to traits, phenotypes, and endophenotypes in SUDs we need to clearly define these terms. First, the standard dictionary definitions (Apple OSX dictionary application, accessed January, 2016) of trait and phenotype are:

- Trait: (1) a distinguishing quality or characteristic, (2) a genetically determined characteristic.
- Phenotype: [a set of] observable characteristics of an individual resulting from the interaction of its genotype(s) with the environment.

Endophenotype does not have a standard dictionary definition; its initial use can be traced back to the study of insects (John & Lewis, 1966) in order to parse apart what was believed to be two components of phenotypes based largely on chromosomal variation: external (exogenous) and internal (endogenous) phenotypes (exophenotype and endophenotype, respectively). To quote John & Lewis (1966):

> *The endophenotype, by definition, does not affect the competitive efficiency or, therefore, the adaptedness of the individual; it affects the number and nature of the offspring and is, in consequence, the subject of retrospective selection. […] it is clearly time to examine more fully not the exophenotype but the endophenotype, not the obvious and external but the microscopic and internal, not the genic but the chromosomal.* (p. 720)

John and Lewis described endophenotypes as chromosomal variations (something that cannot be directly observed at the time) that could impact *subsequent* generations. This definition was loosely adapted for psychiatry by Gottesman and Gould (2003) to include almost anything that was not strictly a diagnostic category:

> *An endophenotype may be neurophysiological, biochemical, endocrinological, neuroanatomical, cognitive, or neuropsychological (including configured self-report data) in nature.* (p. 636)

However, Gottesman and Gould did provide some criteria for what an endophenotype is:

> *1. The endophenotype is associated with illness in the population.*
>
> *2. The endophenotype is heritable.*
>
> *3. The endophenotype is primarily state-independent (manifests in an individual whether or not illness is active).*
>
> *4. Within families, endophenotype and illness co-segregate.*
>
> *[...]*
>
> *5. The endophenotype found in affected family members is found in nonaffected family members at a higher rate than in the general population.* (p. 638)

A broader re-interpretation of Gottesman and Gould's definition is provided in (Ersche et al., 2012):

> *Endophenotypes are quantitative traits, mediating between the predisposing genes (genotypes) and the clinical symptoms (phenotypes) in complex disorders* (p. 602)

where quantitative traits tend to be some non-categorical measure used as a proxy for categories (i.e., diagnoses). Furthermore, there are other similar terms in the literature (e.g., quantitative trait, phenotype, intermediate phenotype) that are also used interchangeably where the definition has evolved into some measure believed to be "closer" to genetics than diagnostic category (e.g., Reitz & Mayeux, 2009):

> *In conclusion, given that the pathways from genotypes to end-stage phenotypes are circuitous at best, discernment of endophenotypes more proximal to the effects of genetic variation can improve statistical power and thereby be a powerful tool in the identification of genes linked to complex disorders.* (p. 186)

Finally, an even broader definition of phenotype and intermediate phenotype has found its way into, mostly, neurological disorders (Vounou et al., 2012):

> […] *we identified voxels that provide an imaging signature of the disease with high classification accuracy; then we used this multivariate biomarker as a phenotype* […] (p. 700)

This definition broadens the idea of "phenotype" with respect to the first dictionary definition of trait ("a distinguishing quality or characteristic"). Vounou and colleagues used a quantitative measure that best discriminates or identifies *a priori* groups as their "phenotype"; in this particular study the phenotype was "voxel-wise longitudinal changes" that best discriminated Alzheimer's Disease from control groups (which was done *before* any genetic association), a procedure that violates Gottesman and Gould's rules except the first ("associated with illness in the population").

While these are the (very open) definitions primarily used today, two common parts of the original definitions of a phenotype (and endophenotype) are that phenotypes must be *genetic* and *heritable*. In the literature at large, heritability often is not (or cannot be) computed. While some promising neural or behavioral patterns have been suggested as endophenotypes (implying genetic and inheritance components)—including working memory activation in the dorsolateral prefrontal cortex (DLPFC) in schizophrenia (Callicott et al., 2003; Potkin et al., 2009) and even impulsivity traits as measured by routinely used scales (Ersche, Turton, Pradhan, Bullmore, & Robbins, 2010; Robbins, Gillan, Smith, de Wit, & Ersche, 2012) or lab measures (Anokhin, Grant, Mulligan, & Heath, 2015)—these are far from confirmed as phenotypes or endophenotypes as initially defined.

Because of the inconsistency of definitions and the difficulty in estimating heritability, I will provide definitions to help distinguish these terms from one another (with respect to the brain and behavioral sciences, largely in the context of psychiatric, and to some extent neurological, disorders). Trait will be a largely umbrella term where genetics effects are either confirmed or suspected (direct or indirect) contributors to some measured effect. Thus, I provide the following two definitions:

- Phenotype: (often directly) observable behavioral traits, with presumed genetic contributions, shown to be associated with particular, or common to a variety of disorders (e.g., impulsivity in SUDs, working memory performance in Schizophrenia); often measured with subjective or objective instruments.
- Endophenotype: (often indirectly) observable neural traits, with presumed genetic contributions, shown to be associated with particular, or common to a variety of disorders (e.g., “reward circuit” activation in SUDs, ventricle size in schizophrenia); often measured with invasive and non-invasive technology (e.g., functional or structural imaging, electrophysiological recordings).

With distinct definitions of phenotype and endophenotype we can now review the literature, in the context of SUDs, in a clearer way. However, these definitions are relaxed because: (1) there is no requirement of heritability estimates (though an effect may be genetic, it is not always inherited), even though it is well known that SUDs are quite heritable, and (2) some endo/phenotypes may exist separately from SUDs and thus could co-occur with SUDs.

2.2.4.1 Phenotypic-genotypic associations in SUDs

A wide variety of genetic markers have been associated with behaviors common to SUDs. Though this dissertation focuses on marijuana and nicotine use, there are fewer phenotypic-genotypic studies on these groups than some other SUDs (e.g., cocaine, alcohol). However, the majority of studies include control groups, thus studies outside of the phenotypic-genotypic associations in marijuana or nicotine still provide an expectation of what to expect in SUDs and especially in SUDs vs. control.

"Rule breaking" is shown to be a mediating factor between *GABRA2* and substance misuse (Trucco, Villafuerte, Heitzeg, Burmeister, & Zucker, 2014). *MAOA* and *SLC6A4* (a.k.a., *5-HTTLPR*) magnify effects of depression in cocaine users (Moeller et al., 2014); *DRD4* explains co-morbid depression and marijuana use (Bobadilla et al., 2013); and *GST** genes are associated with higher anxiety in nicotine dependence (Nunes et al., 2014).

Impulsive, sensation and novelty-seeking, and risk taking behaviors are common traits in SUDs, where some SUDs show distinct aspects of impulsivity (Beaton et al., 2014). In fact, impulsivity and related traits have been suggested as possible phenotypes or endophenotypes (Congdon & Canli, 2005; Robbins et al., 2012; Whelan et al., 2012) in SUDs, though the genetic origins of these behaviors have limited study outside of disordered populations. One such study showed that multiple genetic markers (*COMT*, *DRD4*, *BDNF*, *SLC6A4*, and *ANKK1*) contributed both on the whole (polygenic), and individually to various aspects of impulsivity (Carver, LeMoult, Johnson, & Joormann, 2014). Impulsivity is increased in marijuana-related problems with *CNR1*, but not *FAAH* (Bidwell et al., 2013), and even in adolescent populations (Buchmann et al., 2014), where *OPRM1* genotypes explain an increase in both impulsivity and alcohol use

(Pfeifer et al., 2015). A complex interaction was observed between *HTR2B*, impulsivity, aggression, and alcohol use in Finnish founder populations (Tikkanen et al., 2015), and, finally, in a review, *SLC6A4* (a.k.a. *5-HTTLPR*) genotypes appear to increase the risk of both sexual risk behaviors and alcohol use disorders (Rubens et al., 2015).

Like impulsivity, there are a number of other personality traits with suspected genetic contributions, though the findings there are also inconsistent. These genes are primarily dopaminergic and serotonergic (Munafò & Flint, 2011). There is also a suspected link between personality traits and *resilience* ("constraint") against SUDs through a variety of brain circuits, which are linked to dopaminergic (receptors found in, e.g., VTA, reward circuit) and serotonergic genes (receptors found in, e.g., ACC, amygdala; Belcher, Volkow, Moeller, & Ferré, 2014)[3]. This brings up two points: (1) resilience is a likely phenotype, and (2) the brain regions involved—in SUDs, related traits, and resilience—could be likely phenotypes or endophenotypes. A number of studies have indeed shown genetic contributions to treatment options in SUDs, such as: treatment effects mediated by dopamine (Feldstein Ewing, LaChance, Bryan, & Hutchison, 2009) and serotonin in alcohol use disorders (Kenna et al., 2014), and also *CREB* (which mediates *NPY* a gene linked to the thalamus and satiation) effects on treatment efficacy in methamphetamine abusers (Heinzerling, Demirdjian, Wu, & Shoptaw, 2016).

[3] This paper offers yet another, very loose, definition of "endophenotype:" "The endophenotype concept is understood as simpler clues to genetic underpinnings than the disease syndrome itself, and involves the genetic analysis of any of a variety of biological markers (cognitive, neurophysiological, anatomical, biochemical, etc.) of the disease." (p. 211)

As noted in Volkow et al. (2011) with respect to brain circuit disruptions and differences (compared to controls) routinely found in SUDs, we should expect to find contributions—to SUDs—from a generally wide array of genetic markers in the serotonergic, cholinergic, dopaminergic, and opioid systems. Thus, we should also expect that genes in these systems would contribute to brain changes or differences (endophenotypes), which could lead to behavioral changes (phenotypes), and thus result in a disorder.

2.2.4.2 Endophenotypic-genotypic associations in SUDs

The endophenotypic-genotypic associations in SUDs vary across a wide array of SUDs and trait behaviors (possible phenotypes) often linked to SUDs. Though the focus of this dissertation is on marijuana and nicotine use, there are fewer enodphenotypic-genotypic studies on these groups than some other SUDs (e.g., cocaine, alcohol). However, the majority of studies include control groups, thus studies outside of the endophenotypic-genotypic associations in marijuana or nicotine still provide an expectation of what to expect in SUDs and especially in SUDs vs. control.

Some of the most studied associations are in cocaine users: *SCL6A3* (*DAT1*) mediates craving, and neural response to craving, in (temporarily) abstinent users (Moeller et al., 2013), wide-spread contributions from dopamine, serotonin, and *GABA*-ergic genes to white matter degradation in cocaine users (Azadeh et al., in press, 2016) as well as contributions of both *MAOA* and *SLC6A4* in aversive-related event-related potentials in cocaine users with depressive symptoms, which could impact treatment (Moeller et al., 2014). Monoamine transport genes (i.e., helpers of *MAO-A* and *-B*) also play a role in resting state connectivity in alcohol dependence (Zhu et al., 2015).

Early work on endophenotype-genotype associations in SUDs have shown that endocannabinoid markers—specifically *CNR1*—are associated with cue-elicited responses in alcohol (Hutchison et al., 2008), where particular alleles are associated with responsivity in prefrontal regions. There are also effects of *CNR1* on prefrontal connectivity in conjunction with lower working memory performance (Colizzi et al., 2015). Furthermore, additional endocannabinoid markers—*CNR1* and *FAAH*—have shown a number of effects in (as expected) marijuana use including mediating cue-elicited neural response, largely in the "reward circuit" regions (Filbey, Schacht, Myers, Chavez, & Hutchison, 2009a, p. 1), as well as differences in hippocampal and amygdala volumes associated with *CNR1* and *FAAH* (Schacht, Hutchison, & Filbey, 2012), where *FAAH* alone has shown a relationship with lower white matter integrity—in cannabis users—in forceps minor and bilateral uncinate fasciculus (frontolimbic regions, see Shollenbarger, Price, Wieser, & Lisdahl, 2015)—regions also associated with the reward circuit. Other research implicates the *DRD2* and *PENK* genes with (impulsive and neurotic) personality traits and amygdala activation (for impulsivity-related tasks) in cannabis dependence (Jutras-Aswad et al., 2012). *DRD4* repeat markers play a role in substance abuse—in particular marijuana and "hard drugs"—in psychiatrically hospitalized (for behavioral disorders) adolescents (Mallard, Doorley, Esposito-Smythers, & McGeary, 2016). A *DBH* SNP was associated with increased impulsivity, as well as decreased functional connectivity (in a reward circuit) in marijuana and cocaine users (who were under the influence of small doses; Ramaekers et al., 2015). Finally, in a recent review Moeller, London, & Northoff (2014) implicated many *GABA*-ergic related regions show substantial differences in SUDs (from controls) in resting state connectivity.

2.2.5 Expected and unexpected genetic contributions to SUDs

Though this dissertation is concerned, primarily, with marijuana and nicotine cohorts, many of the previously mentioned genes have been shown, repeatedly, in SUDs and addiction studies both generally (vs. a control group), and specifically (e.g., within marijuana or cocaine use groups). Some examples of particular (and sometimes presumed mono-) genetic effects include dopamine in alcohol (Filbey et al., 2008; Kang et al., 2014), opioid markers in a wide array of studies including opiate abuse (Clarke et al., 2014), the severity of opiate withdrawal (Jones, Luba, Vogelman, & Comer, 2016), and methamphetamine abuse (Li et al., 2004). Even zinc finger (*ZNF**) markers have been shown to have a role, separately, in comorbid alcohol use and opioid dependence (Ali et al., 2015), and again in heroin abuse (Hancock, Levy, et al., 2015; Sun et al., 2015). Furthermore, some of the most robust markers of substance abuse—across a wide array of SUDs, not just in nicotine use—tend to be nicotinic gene markers (Melroy-Greif, Stitzel, & Ehringer, 2015; Sherva et al., 2010). Interestingly, *CB1* and *CB2* (endocannabinoid markers) respectively, play roles in cocaine use in rats (McReynolds et al., 2015), and alcohol-reward behaviors in (knock out) mice (Powers, Breit, & Chester, 2015), where a related gene—*FAAH*—also showed evidence of regulation of nicotine withdrawal (Merritt, Martin, Walters, Lichtman, & Damaj, 2008), likely by breaking down endocannabinoids in a variety of brain structures (e.g., amygdala, hypothalamus; as shown in rats; Cippitelli et al., 2011).

While much research has been done with a focus on small sets or single genetic markers, there are two relatively complex findings that can be gleaned from the literature: (1) very specific genetics of multi-substance use, or associated with general substance abuse, and (2) very specific genetic markers that, according to the literature, contribute to many traits associated with

SUDs or other neuropsychiatric disorders. Furthermore, several individual genes have been suggested as general risk factors for any SUD: *OPRM1* (Schwantes-An et al., 2015), *GABRA2* and *DRD2* (Kreek et al., 2005), as well as *SLC6A4* (Belcher et al., 2014), usually because these genes are routinely found as significant contributors across a variety of SUDs.

Usually, we can expect a particular gene (e.g., *OPRM1*, an opiate receptor) *that matches the substance of interest* to contribute to a particular SUD (e.g., heroin use), to quote Volkow and Mueke (2012):

> *"Thus, the tetrahydrocannabinol (THC) in marijuana takes advantage of the cannabinoid receptor type 1 (CB1), heroin acts through mu opioid receptors and nicotine through nicotinic receptors; alcohol affects the dopaminergic system via multiple targets, including GABA, N-methyl-D-aspartate (NMDA), cannabinoid and serotonin receptors."* (p. 774-775)

However, this is not always the case, and particular genes (e.g., *OPRM1*) may contribute to SUDs without "matching" a substance (e.g., cannabis dependence). For examples, *CNR1* (an endocannabinoid receptor) and *FAAH* (that breaks down endocannabinoids) have been associated with heroin addiction (Proudnikov et al., 2010), *OPRM1* (an opiate receptor) alters binding in various brain regions (e.g., amygdala, ACC) in tobacco smokers (Ray et al., 2011), and *CHRNB3* (a nicotinic receptor) has been associated with cannabis use (Agrawal et al., 2015). Notably, Agrawal et al., (2015) also found genetic associations between *BDNF* and smoking initiation, as well as associations between *CHRNA5/A3* and nicotine/cannabis co-use. However, endocannabinoid receptors (especially *CB1*, a.k.a. *CNR1*) appear to contribute to a wide variety of reward processes and addiction (Parsons & Hurd, 2015).

To continue with unexpected genetic effects in SUDs, two well known genes outside of SUDs—*ApoE* and *DISC1*—have, quite curiously, been shown to play a role in unexpected ways: *ApoE*—the largest non-causative risk factor for Alzheimer's Disease (Karch & Goate, 2015)—has in fact been linked with smoking (Kalapatapu & Delucchi, 2013), where *DISC1*—named from "***Di***srupted in ***Sc***hizophrenia ***1***" (Millar et al., 2000)—shows effects of cocaine use in rats (Gancarz et al., in press, 2016). Similarly, *ApoB*—in a similar fashion to *ApoE*, commonly associated with lipids and cholesterol—appears to mediate the relationship between bipolar disorder and binge-eating (Winham et al., 2014).

In the same vein as Zhu et al., (2014), it would appear one gene can and often does contribute to many SUDs—and even other diseases or disorders, such as Alzheimer's or schizophrenia. However, this section also makes clear that the direction of this relationship is not just "one gene" that appears to contribute to many disorders, rather, numerous genes that have been studied (often individually) all appear to contribute to *one disorder*. In fact, it is much more likely that multiple genes contribute to multiple disorders in a variety of complex ways (see Table 2.1 for a selective illustration of multiple genes that may contribute to multiple disorders).

Table 2.1

Many genes, many disorders

Gene	Cannabis	Nicotine	Other Substances	Multiple SUDS or SUDS + Other Disorders	Other Neuropsychiatric Disorders
COMT	Tunbridge et al., 2015	Ashare et al., 2013; Johnstone et al., 2007	Ting Li et al., 2012; Schellekens et al., 2013; Tao Li et al., 2004; Mus et al., 2013	Estrada et al., 2011; Pelayo-Terán et al., 2010; C.-K. Chen, Lin, Chiang, Su, & Wang, 2014	Goenjian et al., 2015; Inoue et al., 2015; Sampaio et al., 2015
BDNF	Agrawal et al., 2015	Lang et al., 2006; M. D. Li, Lou, Chen, Ma, & Elston, 2008	Greenwald, Steinmiller, Śliwerska, Lundahl, & Burmeister, 2013; Chen et al., 2015; Corominas-Roso et al., 2015; Su et al., 2015	Cheah et al., 2014	Suchanek, Owczarek, Kowalczyk, Kucia, & Kowalski, 2011; Timpano, Schmidt, Wheaton, Wendland, & Murphy, 2011
*DRD**	Vaske, Boisvert, Wright, & Beaver, 2013	Lee et al., 2012; Voisey et al., 2012	Clarke et al., 2014; Filbey et al., 2008; Kang et al., 2014; Sullivan et al., 2013; Bousman et al., 2010	Bobadilla, Vaske, & Asberg, 2013	Watanabe, Shibuya, & Someya, 2015
CNR1	Agrawal et al., 2009; Bidwell et al., 2013; Filbey, Schacht, Myers, Chavez, & Hutchison, 2009	X. Chen et al., 2008	Proudnikov et al., 2010; Hutchison et al., 2008; Clarke, Bloch, et al., 2013; Okahisa et al., 2011	Onwuameze et al., 2013	Juhasz et al., 2009; P. Monteleone et al., 2009; Palmiero Monteleone et al., 2010; Tiwari et al., 2010
OPRM1	N/A	Ray et al., 2011; Zhang, Kendler, & Chen, 2006	Drakenberg et al., 2006; Pfeifer et al., 2015; Heinzerling, McCracken, Swanson, Ray, & Shoptaw, 2012; Clarke, Crist, et al., 2013	N/A	Davis et al., 2011
*CHRN**	Agrawal et al., 2015	Hancock et al., 2015; Thorgeirsson et al., 2010	Coon et al., 2014; Haller et al., 2014; Garcia-Ratés, Camarasa, Escubedo, & Pubill, 2007	Lubke, Stephens, Lessem, Hewitt, & Ehringer, 2012; McEachin et al., 2010	Hartz et al., 2011; Stephens et al., 2012

Note. Some of the most studied genes in SUDs also appear in other neurological and psychiatric disorders (and vice versa). This table provides a brief overview of how, essentially, some of the most popular genes to study (rows) can be found in almost any domain of interest (columns). See also volume **33** of *Neuropsychopharmocology* with a special section dedicated to *COMT*: http://www.nature.com/npp/journal/v33/n13/index.html#Special-Theme:-Catechol-O-Methyl-Transferase-(*COMT*),-Recent-Findings

2.2.6 Failures to replicate and non-significant findings.

Even though some reviews (e.g., Bevilacqua & Goldman, 2011; Loth et al., 2011) imply strong genetic factors—with estimates of genetic contributions upwards of 50% (Volkow et al., 2012, 2011)—other reviews express more caution with respect to the role of genetics in substance abuse, stating that "the predictive utility of the genetic factors studied to date is weak" (Hutchison, 2010, p. 579). In fact, there have been a number of failures in replication studies across SUDs, related disorders, and related traits. One of the earliest replication studies (Franke et al., 2001) failed to find an association in *OPRM1* with either alcohol use or opiate addiction, as well as related. While this study was a non-replication, numerous studies since then have essentially found *OPRM1* as a risk factor for SUDs, and in particular opiate use. However, these effects may not exist across a broad population. Recently, Rouvinen-Lagerström et al., (2013) found no association of *OPRM1* with alcohol dependence *in a Finnish population*, though a related study showed "non-specific liability" (i.e., general risk factor) of *OPRM1* in SUDs (Schwantes-An et al., 2015). Similarly, there are also reported failures—of *GABRA2*—in Italian alcohol use populations (Onori et al., 2010). Recently, much more emphasis has been placed on "polygenic risk scores" or "multilocus (genetic) profile scores" (Nikolova, Ferrell, Manuck, & Hariri, 2011; van Eekelen et al., 2011). However, those too failed to replicate. In an attempted replication study, Hart et al., (2013) were unable to find the same effects they initially reported across *as many as 12* (highly cited) prior studies (see Hart et al., 2013 for the list of studies). Finally, a number of weak effects have been reported through new studies and meta-analyses: *ANKFN1* and *FTO* in cannabis use (Agrawal et al., 2011), *SLC6A4* in alcohol use (Villalba et al., 2015), and *DRD2* effects in alcohol use (Munafò, Matheson, & Flint, 2007).

2.3 Several theoretical oligo- and polygenic panels for SUDs and Addiction

In recent years—perhaps because of so many genes appear to play a role in so many diseases, disorders, and traits—there have been three proposed genetic panels (a.k.a. sets of genes) that are of particular interest in SUDs and addiction: (1) the "stress/resilience" panel (Feder et al., 2009), which is related to the hypothalamus-pituitary-adrenal (HPA) axis and stress circuit, (2) the "reward deficiency syndrome" (RDS) panel (Blum, Oscar-Berman, Demetrovics, Barh, & Gold, 2014)—which is purported to be a superordinate classification of many DSM disorders (and their respective behaviors; see the following RDS section for more details), and finally, (3) a panel (that actually predates the previously mentioned reward and stress panels) that was created by a variety of collaborators with close ties to NIAAA (Hodgkinson et al., 2008), called the "Addictions Array"—with 130 candidate genes (and ~12,000 SNPs) intended to target a wide array of SUDs.

While they have been around for a considerable amount of time, none of these panels has been tested on the whole—usually just parts of each panel are used in mono- (single) or oligogenic (a few) candidate studies. Here, I will briefly discuss each of these panels, as well as one additional panel we refer to as the "substance-specific" panel, how they are related to one another, and how they are related to various aspects/models/theories of SUDs and addiction.

2.3.1 The stress/resilience panel

Feder and colleagues (2009) proposed a "psychobiolog[ical] and molecular genetic" model of "resilience." Here, resilience means the ability to cope with stressful and/or traumatic situations (a.k.a. "stress-resistance"). As previously noted, stress is a critical part of some theories of SUDs and addiction, as the inability to cope with stress (i.e., lack of resilience) and trauma could, (1) be a maintenance factor in SUDs, (2) lead to SUDs, or (3) lead to SUDs comorbid with other disorders. Feder et al., (2009) proposed these particular markers because of the neural regions largely responsible for response to stress, fear, and reward; these regions are part of three systems: the hypothalamus-pituitary-adrenal (HPA) axis, the "fear circuit," and the "reward circuit." Through these neural systems, a number of genetic markers are proposed largely due to how these systems signal within, and to one another. Broadly, Feder et al., (2009) suggest that noradrenergic, serotonergic, dopaminergic genes and even glucocorticoid receptor (*GR*) genes (e.g., *NR3C1*) play critical roles within these systems, in the context of stress response (and resilience).

While the HPA axis actually includes a wide array of expressed genes, Feder et al., (2009) propose only a few specific genes (and gene markers)—almost all of which have a varied history in the SUDs and addiction literature. The specific markers proposed are (often specific SNPs or genotypes of): *CRHR1*, *FKBP5*, a very specific serotonin transporter gene (*5-HTTLPR*, more recently known as): *SLC6A4*, *COMT* (which, amongst other roles, degrades dopamine), *NPY*, *BDNF*, *MAOA* and *EGR1* (formerly referred to as *NGFI-A*).

While the stress/resilience panel is related to reward circuitry in the brain, it has little overlap with the reward panel (see Blum et al., 2014 and the next section). However, the stress/resilience panel is actually a subset of the "Addictions Array" (Hodgkinson et al., 2008; see Section 2.3.4). In general, the stress/resilience panel is considered promising because of the stress hypotheses of substance abuse, which suggests stress is a factor in initiation and/or maintenance of SUDs (Johnston, Linden, & van den Bree, 2015; Morrow & Flagel, 2016).

2.3.2 The reward panel

Much more recently, Blum et al., (2014) have proposed a very specific genetic panel (under a trademark) titled the "Genetic Addiction Risk Score" (GARS$^{\text{TM}}$), which is based on a genetic model—based almost exclusively on the idea of "reward" via dopamine, and in particular *DRD2*—of multiple psychiatric disorders that includes substance abuse. The GARS panel is based on earlier work—referred to as "Reward Deficiency Syndrome" (RDS)—by Blum and colleagues (2000). The primary RDS behaviors (and analogous DSM disorders) are comprised of SUDs and addiction related disorders. But Blum and colleagues (2000, 2014) also include a variety of other behaviors (and disorders) supposedly subordinate to RDS. In 2000, Blum and colleagues suggested that:

> *"Therefore lack of D2 receptors causes individuals to have a high risk for multiple addictive, impulsive and compulsive behavioral propensities, such as severe alcoholism, cocaine, heroin, marijuana and nicotine use, glucose bingeing, pathological gambling, sex addiction, ADHD, Tourette's Syndrome, autism, chronic violence, posttraumatic stress disorder, schizoid/avoidant cluster, conduct disorder and antisocial behavior."* (Blum et al., 2000)

In 2014, Blum and some of the same colleagues suggested that behaviors (and disorders) subordinate to RDS include compulsive (e.g., body dysmorphia) or impulsive (e.g., Tourette's Syndrome and Autism) behaviors, and personality disorders (e.g., paranoia and schizotypy; see Table 1 of Blum et al., 2014). While GARS is one of the few proposed polygenic panels, it is a proprietary panel[4] and has little-to-no formal evaluation in the literature, it was designed almost entirely based on literature reviews.

The GARS™ is mostly based on the premise that the brain "reward circuitry" is dopaminergic, and so GARS™ focuses on the *DRD2* gene. However, other dopamine receptor genes (i.e., *DRD1*, *DRD3*, and *DRD4*) are also suggested as possible molecular markers for "reward deficiency" and thus substance abuse. Similar to the resilience panel, the GARS panel also implicates (obviously) dopaminergic, and serotonergic systems. Specifically, both panels include *MAOA*, *COMT*, and *SLC6A4*. Blum et al., (2014) provide a specific list of the GARS panel, which includes: *DRD1-4*, *DAT1* (better known as *SLC6A3*), *5-HTTLPR* (better known as *SLC6A4*), *OPRM1*, *GABRB3*[5], *MAOA*, and *COMT*. However, throughout the review in this GARS proposal, several other markers are implicated in RDS (*DBH*, *GABRA3*, *HTR1A*, and *HTR2A*) but not included in the GARS panel. Finally, some markers are mentioned (more so in passing) and not strongly implicated (e.g., *HTR3B*, *CNR1*). For the purposes of this review, the GARS panel should be considered as two panels: (1) the core GARS panel: *DRD1-4*, *SLC6A3* (a.k.a. *DAT1*), *SLC6A4* (a.k.a. *5-HTTLPR*), *OPRM1*, *GABRB3*, *MAOA*, and *COMT*, and (2) the extended GARS panel: which is the core plus *DBH*, *GABRA3*, *HTR1A*, and *HTR2A*. While other markers are mentioned in the GARS paper, they are mentioned more so in passing, and not proposed—directly—as part of the RDS/GARS model.

The GARS™/RDS panel proposes that, on the molecular level, there are broad disruptions of the reward circuitry (dopaminergic system), which predisposes individuals to "reward

[4] http://blogs.discovermagazine.com/neuroskeptic/2015/08/17/strange-world-reward-deficiency-syndrome-3/

[5] Though, through a typo, it is suggested that it is *GABRA3*. It should be pointed out that this paper has a number of typos for genetic names and markers, amongst other typos and errors.

deficiency" disorders such as SUDs (and other "reward deficiency" disorders claimed by Blum and colleagues: such as autism, schizophrenia, and Tourette's syndrome). The genes put forth in the GARS™ panel encompass genes expressed in a wide variety of reward circuit regions in the brain, which—as noted in prior sections—have shown effects in a number of SUDs, and associated (endo-)phenotypes. As previously noted in the "stress/resilience" panel section, there is little overlap between the "stress/resilience" and "reward" panels (only *SLC6A4*, *COMT*, and *MAOA*). But, the common genes between these two panels linked through a common trait: their role in dopamine transport and degradation.

2.3.3 Genes specifically associated with substances (i.e., "substance-specific" panel)

In on-going work (Beaton, Abdi, & Filbey, in prep) we propose a "substance-specific" panel that is, essentially, a selection of genetic markers related to particular substances of abuse. This is not a new idea: endogenous receptors are the usual targets of particular SUDs (see, e.g., Volkow & Muenke, 2012; Volkow et al., 2011, and the marijuana and nicotine sections with endocannabinoid and nicotinic genes). In previous work we used three substance use groups: marijuana, nicotine, and a comorbid marijuana/nicotine group. Here we selected all cannabinoid and (neuronal) nicotinic related genes (generally from a wide array of literature). We expected in this work that each substance-specific set of markers would be more related to their respective groups (e.g., endocannabinoid with the marijuana group) than to other substance-matched markers.

Endogenous receptors are obvious candidate genes—for specific SUDs. While this is not a strict panel, per se, with respect to the groups in this dissertation and our on-going work, the panel is comprised mostly of *CNR1*, *CNR2*, *CHRNA**, and *CHRNB** (nicotinic neuronal-type, not muscle-type receptor) markers. The same principles apply to other SUDs.

2.3.4 The NIAAA "Addictions Array" panel

The NIAAA panel (a.k.a., "Addictions Array") is a very large set (130) of candidate genes that was actually turned into a commercial chip (via Illumina's GoldenGate platform; page 506 of Hodgkinson et al., 2008). The "Addictions Array" panel is a superset of the previously mentioned panels: The stress/resilience panel, the reward (GARS) panel, and the review-based "substance-specific panel" are subsets of the "Addictions Array" panel. Broadly, "Addictions Array" is comprised of 130 candidate genes that are linked through a variety of molecular and neural systems. Generally, the "Addictions Array" includes markers (i.e., single nucleotide polymorphisms, a.k.a. SNPS) for genes in 13 broad domains / systems aimed at the neurobiological underpinnings of endogenous receptors for particular psychoactive substances, signaling through various pathways, or a variety of other models/theories of SUDs and addiction (e.g., stress, circadian rhythms): (1) Cholinergic (e.g., CHRM), (2) Stress (e.g., NPY), (3) Adrenergic (e.g., ADRA), (4) Metabolism (e.g., ADH), (5) Dopaminergic (e.g., DRD2), (6) Serotonergic (e.g., HTR1), (7) GABA-ergic (e.g., GABRA), (8) opioid (e.g., *OPRM1*), (9) glycine (e.g., *GLRB*), (10) NMDA (*GRIK*), (11) cannabinoid (e.g., *CNR1*), (12) "signal transduction" (e.g,. *MAPK*), and, finally (13) "other" (e.g., *BDNF*, *CLOCK*). Interestingly, circadian rhythm genes (e.g., *CLOCK*) appear to interact with stress (via the HPA-axis) and reward across a variety of SUDs (Perreau-Lenz & Spanagel, 2015). This particular array is a compromise between (very large) candidate gene studies and (very small) genome-wide association studies. In fact, the "Addictions Array" is a direct predecessor to two recent genome-wide technologies: (1) the “PsychArray” (http://www.illumina.com/products/psycharray.html), which roughly comprisses half candidate genes and half exploratory SNPs across the genome; and thus aimed at studying a wide array of psychiatric disorders (in part, commissioned by the Psychiatric Genetics Consortium; http://www.med.unc.edu/pgc/), and (2) Smokescreen©, a technology much like the “Addictions Array” with 800k SNPs across “1000 nominated candidate genes”, which was made through a NIDA SBIR (http://grants.nih.gov/grants/guide/notice-files/NOT-DA-16-013.html)[6].

The “Addictions Array” (as well as Smokescreen©) is a compromise between hypothesis (i.e., candidate) and exploratory (i.e., genome-wide) approaches to identify general and specific genetic markers most associated with SUDs. While most of these genes have been mentioned in the context of the prior panels (“stress / resilience”, “reward,” “substance-specific”), there are a number of other genes of interest (such as *CLOCK* and other genes associated with circadian rhythms; Falcón & McClung, 2009; Hasler, Smith, Cousins, & Bootzin, 2012; Parekh, Ozburn, & McClung, 2015).

2.3.5 A final note on the panels

A recent review on the “neuroscience of resilience,” with respect to addiction (Morrow & Flagel, 2016), provides a brief overview of genetic and epigenetic contributions to stress/resilience with respect to SUDs. Much more interestingly, Morrow and Flagel (2016) propose a new model of that integrates particular aspects of several existing models of substance abuse into a model that includes reward, stress, resilience (as a separate, protective factor),

[6] This is a fairly recent technology, and it does not appear that the full list of the 1031 candidate genes is published anywhere at this time.

developmental, neurobiological, and psychosocial factors. Morrow and Flagel propose an addiction “pipeline” where various genes may play a role (often through environmental interactions). These include stress (with HPA axis, as in the stress panel), reward (with respect to mostly dopamine and the ventral tegmental area), and plasticity/development (e.g., *BDNF*), where endogenous receptors (genes for receptors that bind with particular substances e.g., endocannabinoids for cannabis; nicotinic for nicotine) help mediate risk for, and resilience against, addiction. Coincidentally, a separate group has proposed that addiction can be characterized, at least in part, by combining the theoretical bases of the dopamine (reward; “positive” system) and stress (resilience; “negative” system) hypotheses, wherein both systems are influenced by genetics and environment (Johnston et al., 2015). Furthermore, a recent article has suggested the involvement of receptors much like glucocorticoid: mineralocorticoid (Vogel et al., 2016); while Vogel et al., (2016) presented cases for various stress-based disorders, which in some cases are co-morbid with or precede SUDs. In sum, while these models (Johnston et al., 2015), especially the Morrow and Flagel (2015) model, are broader than their predecessors, they do incorporate many important genetic contributions on various systems, and how these systems work together both distinct from, and with respect to, SUDs and addiction.

2.4 Some concluding remarks on the complexity of genetics in SUDs

The prior sections have largely focused on the likely, unlikely, and even controversial (e.g., not replicable) genetic contributions to SUDs, addiction, common co-morbidities, and associated traits. The genetics of SUDs—and almost any neuropsychiatric disorder—are complex, and there are few explanations—besides attempts with “missing heritability” and "endophenotypes"—to justify why so many genetic markers appear to play a role in so many different aspects of behavioral and brain sciences (or, again, why they appear to *not* play a role).

This review highlighted a number of ways in which the genetics of SUDs (and other disorders) have been studied: mostly across a wide array of hypothesis-driven and candidate gene studies, some of the more exploratory (large scale and GWAS; most of which are simple case-control) studies, and finally correlates between genetics and the phenotypes (defined here has largely behavioral) and endophenotypes (defined here as largely neural or biological) in SUD populations. However, among the "well-delineated" sections in the review, there are many genetic markers that come up across many disorders (even far outside the SUDs and addiction literature; see Table 2.1).

If one were to conclude based on the available literature, which genetic markers we are to expect as contributors to specific SUDs (e.g., marijuana or nicotine dependence), and/or SUDs in general (as compared to control), then we will find a large amount of studies to provide confirmatory, contradictory, or completely unexpected results. In fact, the review I have provided illustrates the exact nature of the genetic contributions to brains, behaviors, and their associated disorders and dysfunctions: it is extraordinarily complex. And yet, this complexity is unmatched—and at times ignored or impossible to achieve—when it comes to study design, methodological choices, and available tools.

Part of this apparent dissonance in the literature may be a result of the field taking too narrow an approach to identifying genetics of SUDs—many genes we associate with particular SUDs were identified by studies of either (1) categorical classification of individuals, or (2) the behaviors most associated with those categorical classifications. If, in fact, these approaches are too narrow to decipher the genetic contributions to SUDs and addiction, then it may help to use methods tailored to look at genetics more broadly, and/or to also include traits (phenotypes and endophenotypes) associated with SUDs.

2.4.1 Current GWAS techniques can be resource-demanding or costly.

Genome-wide technology was almost immediately considered a "breakthrough" (Pennisi, 2007) that promised to improve our understanding of health, personality, individual differences, and even causal effects of genes (Stranger, Stahl, & Raj, 2011). The accessibility of genotyping array technology has led to an abundance of genome wide data across numerous domains. Despite the enthusiasm and abundance of genome-wide data, these studies failed to delivered on the GWAS promise, largely because statistical concerns (Visscher, Brown, McCarthy, & Yang, 2012) and criticisms over lack of power (Cantor, Lange, & Sinsheimer, 2010). Consequently, the resulting trend leans towards using massive sample sizes (~200,000 for the largest one to date), ~40,000 for Alzheimer's (Hollingworth et al., 2011), and ~6,500 in Schizophrenia (Purcell et al., 2009) as massive sample sizes became one of the only believed solutions to achieving a "breakthrough." Yet, the NIMH has only (very) recently considered *any* GWAS to be such a "breakthrough" (NIMH, 2016). The study (Sekar et al., 2016)—which showed effects of C4 genes (in Chromosome 6) for schizophrenia—came about because the authors leveraged resources across multiple research groups (including Psychiatric Genetics Consortium) with access to "65,000 [genome-wide samples], 700 postmortem brains, and the precision of mouse genetic engineering […]" (NIMH, 2016). However the Sekar (2016) study was preceded by three other (related) "breakthrough" studies: in schizophrenia (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014), in schizophrenia and other psychiatric disorders (The

Network and Pathway Analysis Subgroup of the Psychiatric Genomics Consortium, 2015), and in “schizophrenias” (Arnedo et al., 2014); all of which implicated a variety of other genetic factors not necessarily common across all of these studies (including likely polygenic factors), although two of the studies (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014; The Network and Pathway Analysis Subgroup of the Psychiatric Genomics Consortium, 2015) included common data with the Serak (2016) study. Contrary to these “breakthroughs,” two recent studies have shown null results in exactly these domains: The first was in a relatively small sample (*N* > 500; Voineskos et al., 2015) and the second (Franke et al., 2016) was a massive sample that lead to an unexpected and extraordinary failure (to identify genetic correlates of brain imaging). Franke et al., (2016) amassed the largest schizophrenia samples to date (33,636 schizophrenia, 43,008 controls, with a total of 11,840 subjects providing subcortical brain imaging data) and concluded:

> *“With a comprehensive set of analyses, we did not find evidence for notable genetic correlations, either at a high level (that is, common variant genetic architecture) or for single genetic markers. [...] Similarly, we did not find evidence that common SNPs have pleiotropic effects on these MRI volumes and schizophrenia. Our results suggest alternative hypotheses that require consideration and refutation: that the volumetric differences observed in schizophrenia may be epiphenomena unrelated to its primary genetic causes, [...]”* (p. 420).

Clearly, studies with extremely large samples, or those that use cross-species techniques demand incredible resources often not available to most researchers (even through collaborative efforts, unless collaboration can be done with team sizes from a few hundred to a thousand) and yet even doing so does not necessarily reveal similar results (and in fact can yield null results). Some of the most distinct differences between these studies are not the study designs or samples (recall, at least three of the studies draw from the same data sets), but rather the methodological choices made during both quality control (preprocessing) and analytical stages of the studies.

It has been proposed that, better methods (e.g., Cantor et al., 2010; Kemper, Daetwyler, Visscher, & Goddard, 2012), better quality control (Turner et al., 2011; Zuvich et al., 2011), and systematic approaches to genotype-phenotype results (Saccone et al., 2008), will provide better and more reliable results from genetic and genomic studies. Some of the suggested methods include leveraging additional information to help detect better "genetic signal" (Schifano, Li, Christiani, & Lin, 2013). If better methods can yield better results, then this now becomes a more cost-effective way to utilize (relatively) small genetic and genomic data sets. While some work has been done in this area (see the next chapter), several of these newer methods are completely impractical for most researchers to use. For example, Hibar, Stein, Kohannim, Jahanshad, Saykin, et al. (2011) recently proposed a "voxelwise GeneWAS"—which is an extension of similar methods (Shen et al., 2010), but the Hibar et al., (2011) method reduces the size of the genomic data as compared to the Shen et al., (2010) method—a technique that requires for its implementation: "[...] approximately 13 days" and parallel computing on "[...] a cluster of 10 high performance 8-core CPU nodes [...]" because "[t]he total number of tests of association for vGeneWAS is very high (18,044 genes × 31,662 voxels)."

Thus, in summary, we are left in quite a predicament if we expect SUDs—and related traits or disorders—to originate from complex and/or polygenic contributions. If we cannot possibly design a single study to meet all the expected criteria (e.g., sample size, population diversity), then we have to rely on other ways to potentially increase power, and, perhaps this can be achieved with more sophisticated—and powerful—analytical methods. In this dissertation (see Chapter 5), I propose a new technique designed to increase power, and potentially provide higher quality results for SUDs (which are highly heritable, but the genetic results are notoriously weak).

CHAPTER 3

STATISTICAL METHODS FOR GENETICS ANALYSES

Parts of this chapter are adapted from my qualifying thesis, my F31 grant application, and some on-going work (Beaton, Abdi, & Filbey, in prep) as well as Beaton, Dunlop, ADNI, & Abdi (2016). Copyright © 2015 American Psychological Association. Reproduced with permission. The official citation that should be used in referencing this material is Beaton, D., Dunlop, J., Abdi, H., & Alzheimer's Disease Neuroimaging Initiative. (2016). Partial least squares correspondence analysis: A framework to simultaneously analyze behavioral and genetic data. Psychological Methods, 21(4), 621-651. http://dx.doi.org/10.1037/met0000053. This article may not exactly replicate the authoritative document published in the APA journal. It is not the copy of record. No further reproduction or distribution is permitted without written permission from the American Psychological Association.

Interests in the genetic contributions to psychological traits date back over a century, and with these interests came advancements in measurement and assessment of genetic contributions to behavior. Thorndike (1905) made some of the earliest efforts "to make modern statistical methods current in psychology" (Sanford, 1908, p. 142) as well as genetics and was soon followed by Fisher (1919) who favored more formal approaches. Over a decade later, Thurstone (1934) proposed a multivariate factorial analysis—akin to multidimensional scaling—to understand cognitive abilities and personality traits. Thurstone (1934) proposed these approaches, in part, because he firmly believed "that the isolation of the mental abilities will turn out to be essentially a problem in genetics."

And now again—just as in the early days of Thorndike or Fisher—new statistical methods are being developed to understand how genetics may contribute to brain structure and function, as well as to behaviors (or traits). Most of these modern statistical techniques—like Thurstone's approaches—are multivariate in nature. Further, some of these recent techniques are specifically designed for simultaneous analysis of behavioral and genetic data (Bloss et al., 2010)—mostly because doing so appears to increase the statistical power to detect genetic contributions to traits, behaviors, and phenotypes (van der Sluis et al., 2013; Schifano et al., 2013; Seoane et al., 2014). This boost in power can be quite effective (and welcome) in psychological research where sample sizes cannot reach the large "standard" sizes for genome-wide studies ($N \approx 5000$).

Here, I review the current statistical approaches to genetic and genomic association studies. First I present some standard approaches, and highlight particular multivariate and penalized (regularized or "sparsified") methods and provide introductions to traditional methods, followed by more advanced—typically multivariate—methods; Specifically, I review Principal Components Analysis (PCA), Partial Least Squares (PLS), Canonical Correlation Analysis (CCA), discriminant analyses, Correspondence Analysis (CA), and finally regularization techniques. Each section presents the mathematics of one of these techniques, and a final section shows how these techniques are connected. Finally, this chapter concludes with a major point: SNPs are not inherently quantitative data, and treating them as such likely obfuscates true inheritance patterns, as well as complex polygenic effects (e.g., epistasis); This conclusion justifies the new technique developed in this dissertation (i.e., an extension of PLSCA; Beaton, Dunlop, et al., 2016).

The core of Chapter 3—which is a review of methodological development—has many references to the Alzheimer's Disease Neuroimaging Initiative (ADNI) and other large scale (often collaborative) projects (e.g., ENIGMA). The frequent references to ADNI, and other large-scale projects, in Chapter 3 are mostly because these types of projects—a large, (mostly) open access data set analyzed by thousands of researchers—have produced a large number of new techniques for genetic analyses.

3.1 Traditional Approaches to (and Typical Problems in) Association Studies

Some of the most common approaches to testing genetic or genomic associations is through univariate tests designed to identify SNPs of interest either in case-control studies (i.e., unpaired *t*-test), or with respect to quantitative traits (i.e., simple, mass-univariate regressions via contrasts, see, e.g., Frommlet, Bogdan, & Ramsey, 2016). While one factor ANOVAs or χ^2 tests (independence, or goodness-of-fit) correspond to the actual design of SNP analyses (see Table 3.1), it is most common to perform just simple regressions where SNPs are (numerically) recoded under genetic inheritance models (e.g., dominant, recessive) which nowadays is almost always the additive model (Balding, 2006)—though the sole use of additive models will miss certain effects (for more details, see the final section of this chapter and also Vormfelde & Brockmöller, 2007).

Quite recently, especially in the brain and behavioral sciences, several new approaches have been proposed to link genetic markers to various brain or behavioral measures. These techniques—which were largely spurred by the ADNI project—generally try to follow brain imaging conventions (e.g., general linear models, or statistical parametric mapping; see L. Shen et al., 2010; Stein et al., 2010) or a combination of SPM and principal components regression

(Hibar, Stein, Kohannim, Jahanshad, Saykin, et al., 2011). But these techniques have several drawbacks and, in most cases are not powerful enough (especially with small sample sizes) or do not provide suitable (population) inference estimates (as explained in the following sections).

3.1.1 Inference (and issues with current approaches) in association studies

Nearly since the inception of GWAS there has been one methodological constant: the dreaded "standard" p-value[7] for the field because a result is considered significant only if it reaches $p \leq 5 \times 10^{-8}$. This particular p-value is obtained as the value of $p = .05$ "Bonferroni corrected" for 1 million comparisons (Fadista, Manning, Florez, & Groop, 2016).

This "standard" value is used even for a study comprising fewer than one million gene (e.g., ~500k) or more than one million (e.g., ~15 million), or in some cases even for a single gene candidate gene. This particular threshold for a p-value has been challenged in recent years (X. Gao, Becker, Becker, Starmer, & Province, 2010)—in part because, often, studies cannot reach the standard significance threshold (e.g., Franke et al., 2016; Voineskos et al., 2015)—and criticized largely for one particular reason: even when *not* in linkage disequilibrium (relatively strong statistical association between two SNPs, which suggests non-independence between them), many SNPs are *not* independent but the Bonferroni correction assumes independence.

[7] This "standard" should be carefully reconsidered in light of particular "replication crises" and the ASA's statement about how p-values are typically used (Wasserman & Lazar, 2016).

Table 3.1

Inheritance models and analyses

Analysis or Model	Major Homozygote	Heterozygote	Minor Homozygote	Data Type
Genotypes and their general representations for a variety of analytical and inheritance models.				
HWE[a]	AA	Aa	aa	Categorical (3 levels)
Genotypic[a]	AA	Aa	aa	Categorical (3 levels)
Dominant (D)	Not D	D	D	Categorical (dichotomous)
Recessive (R)	Not R	Not R	R	Categorical (dichotomous)
Heterozygous (H)[b]	Not H	H	Not H	Categorical (dichotomous)
Linear Additive[c]	b	$b+r$	$b+(2r)$	Quantitative (interval or ratio scale)
Multiplicative[c]	b	br	br^2	Quantitative (interval or ratio scale)
Genotypes and their numeric representations for a variety of analytical and inheritance models.				
HWE	[1 0 0]	[0 1 0]	[0 0 1]	Categorical (3 levels)
Genotypic	[1 0 0]	[0 1 0]	[0 0 1]	Categorical (3 levels)
Dominant (D)	[0 1]	[1 0]	[1 0]	Categorical (dichotomous)
Recessive (R)	[0 1]	[0 1]	[1 0]	Categorical (dichotomous)
Heterozygous (H)[b]	[0 1]	[1 0]	[0 1]	Categorical (dichotomous)
Linear Additive[c]	b	$b+r$	$b+(2r)$	Quantitative (interval or ratio scale)
Multiplicative[c]	b	br	br^2	Quantitative (interval or ratio scale)

Note. HWE = Hardy-Weinberg Equilibrium. Note that, in general, many of these models are naturally categorical. [a]*Here, for HWE and the genotypic model, SNPs are presented generally where 'A' is the major allele and 'a' the minor allele. The major homozygote, heterozygote, and minor homozygote are denoted 'AA', 'Aa', and 'aa', respectively.* [b]The model codes for the heterozygote as different from either homozygote. [c]Where, *b* means "baseline" and *r* means "risk," assuming the risk is associated strictly with the minor homozygote (if the risk should be on the major homozygote, the scale can be reversed where *r* is associated with the major allele).

Though there have been a number of counter-suggestions to the threshold of $p = 5 \times 10^{-8}$, not many—such as multistage approaches (G. Kang et al., 2015)—have stuck. One of the earliest suggestions comes from Hirschhorn and Daly (2005) who point to permutation tests as an alternative to the traditionally conservative threshold. However, it must be noted that permutation tests are *exact* tests, and therefore converge to distributional solutions when the data fit distributional assumptions (Berry, 2011) and so permutation tests require corrections just like the other tests. So, to help provide better estimates, Gao et al., (2010) developed a method based on principal components analysis (PCA) to determine the number of independent tests (instead of 1 million) and correct for that threshold. In similar fashions, conservative methods such as bootstrap and jackknife resampling (Faye, Sun, Dimitromanolakis, & Bull, 2011; Manor & Segal, 2013; Y.-H. Zhou & Wright, 2015) as well as false discovery rate (FDR) procedures (Storey & Tibshirani, 2003) have been suggested as ways to avoid the traditionally strict thresholds. Finally, a relatively new technique called "jackstraw" (Chung & Storey, 2015) combines several of these ideas: specifically, jackstraw combines PCA (to determine the number of likely independent dimensions), assessment of component similarity, permutation resampling of variables, and FDR corrections. However, many of these techniques still approach association studies in the same way and find *the SNP* with the largest effect where *that SNP* is usually assumed to be a, or even *the "causal" SNP*.

The development and use of polygenic risk scores (Maher, 2015) have been on the rise since: (1) use in a major study (International Schizophrenia Consortium et al., 2009) and (2) the first (major) statistical assessment (Dudbridge, 2013) of such approaches. Polygenic risk scores—also sometimes referred to as "multi-locus (genetic) profile scores"—combine multiple markers into a single variable, often through adding up "risk markers" usually based on known or estimated (via meta-analyses) risk scores. By reducing many markers to a single variable it is believed that the same conservative penalties for testing no longer exist, and thus a much more relaxed threshold is often used for polygenic risk score studies. Risk score studies are most frequently found in the behavioral and brain sciences, even though problems can occur when risk is empirically defined (see final section in this chapter). Finally, in addition to finding significant SNPs, association studies have been particularly concerned with "post association" results (Hiersche, Rühle, & Stoll, 2013; Saccone et al., 2008) because, sometimes, significant SNPs are difficult to interpret (i.e., they are not involved in what a researcher would have expected). Even though researchers have a variety of tools that are sensitive to *p*-values (as well as effect sizes and confidence intervals), they have sought better ways to analyze genetic and genomic association studies, and so have turned to multivariate approaches.

3.2 Advanced Statistical Techniques in Association Studies

In association studies, the more advanced and sophisticated techniques (e.g., multivariate or regularization approaches) are designed to boost power (or prediction). In addition, multivariate methods analyze how measures work *together* (Chi, 2012; Allison et al., 1998). Regularization, shrinkage, sparsification, and penalization techniques all minimize the number of variables to evaluate. The upcoming section covers several multivariate, regularization, and multivariate + regularization approaches used in a variety of genetic, genomic, transcriptomic, and related domains. The remaining sections focus on: principal components analysis, partial least squares (with some mentions of canonical correlation analysis and how the two are related), correspondence analysis, ridge regression, and the LASSO technique (as well as discriminant analyses based on any of the aforementioned). However, while these are probably the most used techniques, these are not the only ones used in association studies. Some other examples include support vector techniques (Long, Gianola, Rosa, & Weigel, 2011; Mittag et al., 2012; Oliveira et al., 2014; Roshan, Chikkagoudar, Wei, Wang, & Hakonarson, 2011), self-organizing maps (Wellenreuther & Hansson, 2016), independent components analysis (Liu & Calhoun, 2014; Meda et al., 2010, 2012; Vergara et al., 2014), kernel methods (Ge et al., 2015; Yan et al., 2015), graph techniques (M. Kang et al., 2015), bivariate methods (Jiang, Li, & Zhang, 2014; Yarosh, Meda, Wit, Hart, & Pearlson, 2015; à la, generalized Kendall's tau Simon, 1977), and multivariate regression (Guo et al., 2015; Lippert et al., 2011; O'Reilly et al., 2012; Schifano et al., 2013; van der Sluis, Posthuma, & Dolan, 2013; Zapala & Schork, 2006; X. Zhou & Stephens, 2014), amongst several other techniques (see, e.g., Barrett, Taylor, & Iles, 2014; Frommlet et al., 2016).

3.2.1 Cleaning

Even in early GWAS studies, population stratification effects were known to create problems. Population stratification is some inherent confound in the genetic structure of individuals, often due to geographical or ancestral effects, largely linked to self-reported racial (e.g., American Indian or Alaskan Native, Black/African American, White/Caucasian) and ethnic (i.e., Hispanic or Latino vs. not) identity. There are a few ways to correct for population stratification, but only a subset will be discussed here. Both PCA and multidimensional scaling (MDS) are now used routinely to identify, and then adjust for, population stratification (Liu, Zhang, Liu, & Arendt, 2013; Miclaus, Wolfinger, & Czika, 2009; Tian, Gregersen, & Seldin, 2008; G. Tucker, Price, & Berger, 2014; D. Wang et al., 2009), which help remove confounding factors that could lead to incorrect conclusions about genetic contributions to the disease, disorder, or trait studied.

With respect to population structure identification and stratification, multivariate techniques are the simplest option (see, e.g., Figure 3.1, and Footnote 8) but the real utility of multivariate techniques was illustrated when a research group (Zuvich et al., 2011) learned that they had a stratification effect *due to a chip sample*: that is, someone in their research group had made a mistake with the output of genetic data (i.e., the strand orientation of some batches were in different formats); standard univariate techniques missed this problem while multivariate techniques revealed it.[8]

3.2.2 Some Mathematical Notation

Matrices are denoted with upper case bold letters (e.g., $\mathbf{X}$), vectors with lower case bold letters (e.g., $\mathbf{x}$); scalars are denoted by upper case italic letters (e.g., I), and indices by lower case italic letters (e.g., i). The identity matrix is denoted $\mathbf{I}$. The transpose operation is denoted by a superscript T (e.g., $\mathbf{X}^{\mathrm{T}}$) and the inverse of a matrix is denoted by the superscript $^{-1}$ (e.g., $\mathbf{X}^{-1}$). By default, vectors are column vectors, and so a transposed vector is a row vector (i.e., $\mathbf{x}$ is a column vector but $\mathbf{x}^{\mathrm{T}}$ is a row vector). The diag{} operator transforms a vector into a diagonal matrix when applied to a vector and extracts the vector of the diagonal elements of a matrix when applied to a matrix. Writing side-by-side matrices or vectors (e.g., $\mathbf{X}^{\mathrm{T}}\mathbf{Y}$) indicates ordinary matrix multiplication, when multiplication needs to be made explicit, we use the symbol "×".

[8] In fact, we detected this exact same type of mistake in our own data: there was a "strand flip" because the orientation option was different from one batch to the other batches (see Figure 3.1 with MEG3 vs. F48, OB, MRN).

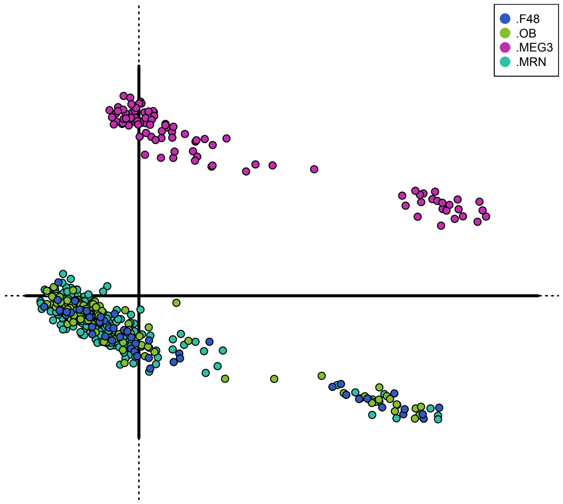


Figure 3.1 The above figure shows the results of a Multiple Correspondence Analysis (MCA) on a subset of SNPs from our combined data set, after the most recent batch was processed (MEG3). Because of this MCA, we immediately noticed something wrong with the MEG3 batch vs. all other batches. With some help from Illumina and the lab we work with at UTSW, it was revealed to be the same issue as Zuvich et al., (2011): one of the output options in the software was flipped (from top to forward strand). While this strand orientation issue is detectable through univariate methods it is extremely difficult and time consuming, as it requires testing each SNP stratified by their chip sample. Clearly – it is much easier and more direct to just visualize with a multivariate technique.

3.2.3 Principal Components Analysis

The oldest and most popular multivariate technique—not just in genetic and genomic association studies, but arguably any field—is principal components analysis (PCA; Abdi & Williams, 2010; Jolliffe, 2002). PCA is a multivariate technique designed to identify the largest possible—orthogonal—sources of variance in a matrix. PCA is performed through the singular value decomposition (SVD), where a data matrix $\mathbf{R}$ is preprocessed to have zero mean and unitary variance *per column*, is decomposed as:

$$\mathbf{R} = \mathbf{U}\Delta\mathbf{V}^{\mathrm{T}}$$

where $\mathbf{U}$ and $\mathbf{V}$ are (respectively) the left and right singular vectors—which represent rows (usually observations) and columns (usually variables), the singular vectors are orthonormal matrices and therefore $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I} = \mathbf{V}^{\mathrm{T}}\mathbf{V}$, where $\mathbf{I}$ is the identity matrix (1s on the diagonal, 0s off diagonal; i.e., all columns within each set of singular vectors are orthogonal to one another). The matrix Δ is diagonal with the diagonal elements—$\mathrm{diag}\{\Delta\}$—being the singular values (all off diagonal values are 0), and the squared singular values—$\Theta = \mathrm{diag}\{\Delta^2\}$—are the eigenvalues (which give the variance per component).

As noted before, PCA has been used to help correct (or adjust) for (overly conservative) multiple comparisons corrections (Chung & Storey, 2014; X. Gao et al., 2010; Tucker et al., 2014). PCA is frequently used to analyze haplotypes, ancestry, and population structures (which, as noted, can be used to correct for population stratification). Some of the earliest uses of PCA (or MDS) originated in the work of Cavalli-Sforza (Cavalli-Sforza & Edwards, 1967; Menozzi, Piazza, & Cavalli-Sforza, 1978). Following Cavalli-Sforza, MDS (as well as both PCA and correspondence analysis) was suggested for conducting these types of analyses (Lessa, 1990); though there have been many recent studies with almost an identical approach (Homburger et al., 2015; Z. Lin & Altman, 2004). However, PCA has also been a central tool for analyses (Chang & Keinan, 2014) because it is claimed: 1) to provide a boost in power (Duan et al., 2012), 2) to have the ability to detect gene-gene interactions (i.e., epistasis; Bhattacharjee et al., 2010) or, 3) to reduce the dimensionality of the data (Turgeon et al., 2016).

However, PCA has also been used to describe a number of techniques that are not really traditional PCA but are, rather, applications of or variations over the SVD. Among these applications of the SVD, principal component regression—which uses principal components as predictors in a multiple (univariate) regression—has been used with some success (Mei et al., 2010), including a study of the ADNI data (Hibar, Stein, Kohannim, Jahanshad, Jack, et al., 2011; and again, but, with a different name: Hibar, Stein, Kohannim, Jahanshad, Saykin, et al., 2011). PCA has also been used as a first step before a discriminant analysis ("discriminant analysis of principal components"; Jombart, Devillard, & Balloux, 2010), and again as the core analytic technique called *between gene sets* (Frost, Li, & Moore, 2014).

3.2.4 Canonical Methods: Canonical Correlation, Partial Least Squares, and Discriminant Techniques

The primary goal of canonical techniques is to discover some optimum relationship between two data sets measured on the same individuals. There exist two major families of techniques for two table analyses: canonical correlation analysis (CCA; Hotelling, 1936; Thompson, 2005) and partial least squares (PLS; Wold, 1975, 1984). Both CCA and PLS generally look for the maximum information common between two data sets. Though the two are closely related, they optimize different criteria (McIntosh & Mivsic, 2013; Sun, Ji, Yu, & Ye, 2009). Finally, it can be shown that CCA can be regarded as a particular case of PLS. CCA and (the core of) PLS are similar to PCA because they use the SVD to provide orthogonal slices of a matrix where the matrix is computed as the common information *between two data matrices* (e.g., a correlation or covariance matrix).

CCA, like many multivariate techniques, uses at its core the SVD and works as follows. If we have two data matrices, $\mathbf{X}$ and $\mathbf{Y}$, both measured on the same observations (rows) and assumed to be preprocessed in some way (usually column-wise mean-centered and unitary norm), we multiply them together such that $\mathbf{C} = \mathbf{X}^{\mathrm{T}}\mathbf{Y}$, where $\mathbf{C}$ is, generally, the *correlation matrix* between $\mathbf{X}$ and $\mathbf{Y}$. Next, we derive the two sets of constraints derived from $\mathbf{X}$ and $\mathbf{Y}$: $\mathbf{W}_J = \mathbf{X}^{\mathrm{T}}\mathbf{X}$ and $\mathbf{W}_K = \mathbf{Y}^{\mathrm{T}}\mathbf{Y}$, respectively. In the case of CCA, the *generalized* SVD (GSVD) would be performed on the matrix $\mathbf{R} = \mathbf{W}_J^{-1}\mathbf{C}\mathbf{W}_K^{-1}$, where the constraints of $\mathbf{W}_J$ and $\mathbf{W}_K$ are applied in the GSVD step. For more details on the GSVD see Chapter 4 for information on the GSVD, and see (Abdi, 2007) or the Appendix of (Greenacre, 1984). CCA is computed with the GSVD though for completeness, the SVD for CCA is performed on $\mathbf{R}$:

$$\mathbf{R} = \mathbf{U}\Delta\mathbf{V}^{\mathrm{T}},$$

with the constraints that

$$\mathbf{U}^{\mathrm{T}}\mathbf{W}_J\mathbf{U} = \mathbf{I} = \mathbf{V}^{\mathrm{T}}\mathbf{W}_K\mathbf{V}.$$

The matrices $\mathbf{U}$ and $\mathbf{V}$ are (respectively) the left and right singular vectors—which represent (respectively) rows (here: the variables of $\mathbf{X}$) and columns (here: the variables of $\mathbf{Y}$)—of $\mathbf{R}$, and Δ is a diagonal matrix, where the diagonal values—$\mathrm{diag}\{\Delta\}$—are the singular values. In canonical techniques, when one of the data tables is a simple design matrix (i.e., group coding), it is called a “discriminant” technique. In the CCA framework, if one table is a design matrix, this is the standard linear discriminant analysis (LDA). While both CCA and LDA have been used in association studies (e.g., Aebi et al., 2015; Biffani et al., 2015; Cahill & Levinton, 2015), they have not been employed as frequently as other methods but with one exception: when CCA or LDA are used in a *regularized* framework. The most common regularized framework is "sparsified CCA" (Witten, Tibshirani, & Hastie, 2009). Witten and colleagues propose that the basis of sparsified CCA is diagonal matrices in place of $\mathbf{W}_J$ and $\mathbf{W}_K$ (a.k.a., "diagonal penalized CCA"). In its simplest form, sparsified CCA would use identity matrices in place of $\mathbf{W}_J$ and $\mathbf{W}_K$ —an operation equivalent to "partial least squares correlation" (as in the following two paragraphs, and Chapter 4).

For canonical techniques, the other major "family" of techniques is the partial least squares (PLS) family. PLS techniques exist as regression (Tenenhaus, 1998; Abdi, 2010), path-modeling (Tenenhaus et al., 2005; Vinzi et al., 2010), and correlation (McIntosh & Lobaugh, 2004; Krishnan et al., 2011) types of methods with variations within each type. A variety of PLS approaches (mostly in the form of PLS regression) have been used to study genetics and genomics. The most common applications are in models of human diseases (Pérez-Enciso, Toro, Tenenhaus, & Gianola, 2003; Michaelson, Alberts, Schughart, & Beyer, 2010), traits in animals (e.g., cows, Bolormaa, Pryce, Hayes, & Goddard, 2010; Moser, Tier, Crump, Khatkar, & Raadsma, 2009), and traits in plants (e.g., crops, Y. Xu, Hu, Yang, & Xu, 2016). However, recent applications also appear in human populations for diseases (e.g., Crohn's disease: Chun, Ballard, Cho, & Zhao, 2011; schizophrenia: Tura, Turner, Fallon, Kennedy, & Potkin, 2008; gene-gene and gene-environment interactions in endometrial cancer: T. Wang, Ho, Ye, Strickler, & Elston, 2009). Finally, a hybrid form of PLSR and PLS correlation, which used sparsification to help identify important variables, was used to detect genetic correlates of brain networks (Le Floch et al., 2012).

While the most common form of PLS used in genetics studies is PLS regression (PLSR), this dissertation focuses on PLS correlation (PLSC) and an extension thereof to categorical and mixed data types (Beaton, Dunlop, et al., 2016; Beaton, Kriegsman, et al., 2016). PLSC works as follows: if we have two data matrices, $\mathbf{X}$ and $\mathbf{Y}$, both measured on the same observations (rows) and assumed to be preprocessed in some way (usually column-wise mean-centered and unitary norm), we multiply them together such that $\mathbf{R} = \mathbf{X}^{\mathrm{T}}\mathbf{Y}$, where $\mathbf{R}$ is, generally, the *correlation matrix* between $\mathbf{X}$ and $\mathbf{Y}$. Then, just as in PCA, the SVD is performed on $\mathbf{R}$ as:

$$\mathbf{R} = \mathbf{U}\Delta\mathbf{V}^{\mathrm{T}}$$

under the constraints that:

$$\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I} = \mathbf{V}^{\mathrm{T}}\mathbf{V}.$$

Like in PCA and CCA, the matrices $\mathbf{U}$ and $\mathbf{V}$ are the left and right singular vectors—which represent rows (here: the variables of $\mathbf{X}$) and columns (here: the variables of $\mathbf{Y}$)—of $\mathbf{R}$, respectively, where $\mathrm{diag}\{\Delta\}$ are the singular values and $\mathrm{diag}\{\Delta\}^2$ are the eigenvalues. Like CCA, the goal of PLSC is to identify *common information* between two data matrices, but the two techniques differ on what they optimize: CCA optimizes the relationship *between* two matrices ($\mathbf{X}$ and $\mathbf{Y}$) with respect to their *within matrix* variances ($\mathbf{W}_J = \mathbf{X}^{\mathrm{T}}\mathbf{X}$ and $\mathbf{W}_K = \mathbf{Y}^{\mathrm{T}}\mathbf{Y}$), where PLSC optimizes the relationship simply *between* two matrices ($\mathbf{X}$ and $\mathbf{Y}$), or, alternatively, CCA optimizes the correlation between two data sets where PLSC optimizes the covariance between two data sets. The relationship between PLSC and CCA is equivalent if for CCA we have $\mathbf{W}_J = \mathbf{I}$ and $\mathbf{W}_K = \mathbf{I}$: $\mathbf{U}^{\mathrm{T}}\mathbf{W}_J\mathbf{U} = \mathbf{U}^{\mathrm{T}}\mathbf{I}\mathbf{U} = \mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I} = \mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{V}^{\mathrm{T}}\mathbf{I}\mathbf{V} = \mathbf{V}^{\mathrm{T}}\mathbf{W}_K\mathbf{V}.$ Thus the simplest form of sparsified CCA is essentially PLSC.

Finally, a very recent paper by Mitteroecker et al., (2016) proposed a framework (called "multivariate analysis for genotype-phenotype association") to unify a number of techniques (e.g., CCA, redundancy analysis) via PLSC. Mitteroecker and colleagues note that PLSC ("maximizes covariance") is the superordinate method to a number of methods to elicit specific types of genotypic-phenotypic relationships: redundancy analysis ("maximizes genetic effect"), reduced-rank regression ("maximizes genetic variance"), and CCA ("maximizes heritability").

However, PLS-CA (Beaton, Dunlop, et al., 2016)—specifically the mixed-modality version of PLS-CA (Beaton, Kriegsman, et al., 2016)—generalizes PLSC to *any type of data* (categorical, continuous, ordinal). When both data tables are quantitative PLS-CA gives the same results as PLSC. Because PLS-CA generalizes PLSC—as a specific case of the GSVD—PLS-CA thus generalizes the framework outlined by Mitteroecker et al., (2016) but PLS-CA is more flexible with respect to inheritance models (see Chapter 4 and Appendix B of Beaton, Dunlop, et al., 2016).

3.2.5 Correspondence analysis

Up until this point, nearly all methods discussed in Chapter 3 are designed for the analyses of *quantitative* data (typically interval or ratio scale). This means that any study that used these techniques must have treated the genetic data—and likely some additional data, such as trait measures—as quantitative. To treat SNPs as quantitative is problematic because a SNP exists only as three genotypic categories: the major homozygote (*AA*), the heterozygote (*Aa*), and the minor homozygote (*aa*). Thus, if one were to use methods designed for analyses of interval, ratio, or even ordinal scale data, then certain assumptions must be applied to SNPs in order to transform them from categorical to numerical. Typically, SNP data are transformed under the assumption of the "additive" model: emphasis should be placed on the minor allele. So, the genotypes of *AA*, *Aa*, and *aa* are thus represented (usually) by the numbers [0, 1, 2], respectively. However, if the genetic effect is not additive, then this transformation could either miss or mischaracterize the effect of a SNP (a more detailed discussion of these problems are discussed in this chapter, section 3.3). In order to analyze SNP data in their categorical format we require a technique designed to analyze categorical data. Categorical (and contingency) data are analyzed with χ^2. For multivariate data, we require something more sophisticated than simple χ^2: CA is like to PCA but suited for matrices comprised of categorical data. For CA, we have a matrix $\mathbf{R}$ and we compute two matrices: an *observed* matrix ($\mathbf{O_R}$) and an *expected* matrix ($\mathbf{E_R}$) just as for a χ^2. First, we compute the observed matrix $\mathbf{O_R}$ as $\mathbf{R}$ divided by the total sum of the table:

$$\mathbf{O_R} = \mathbf{R} \times (\mathbf{1}^{\mathrm{T}}\mathbf{R1})^{-1},$$

where $\mathbf{1}$ is a conformable vector of 1s. Next, from $\mathbf{O_R}$ we compute (column) vectors of the marginal frequencies for the rows and columns, respectively called "masses" ($\mathbf{m}$) and "weights" ($\mathbf{w}$):

$$\mathbf{m} = \mathbf{O_R 1} \text{ and } \mathbf{w{=}1O_R},$$

where $\mathbf{1}$ is a conformable vector of 1s. We then compute the expected matrix from the masses and weights:

$$\mathbf{E_R} = \mathbf{mw}^{\mathrm{T}}.$$

Just as in χ^2, we compute the deviation of observed from expected:

$$\mathbf{Z_R} = \mathbf{O_R} - \mathbf{E_R}.$$

From here, we used the GSVD with the constraint matrices of $\mathbf{M} = \text{diag}\{\mathbf{m}^{-1}\}$ and $\mathbf{W} = \text{diag}\{\mathbf{w}^{-1}\}$:

$$\mathbf{Z_R} = \mathbf{U}\Delta\mathbf{V}^{\text{T}},$$

under the constraints that:

$$\mathbf{U}^{\text{T}}\mathbf{MU} = \mathbf{I} = \mathbf{V}^{\text{T}}\mathbf{WV}.$$

Just like in PCA $\mathbf{U}$, $\mathbf{V}$, and $\text{diag}\{\Delta\}$ are, respectively, the left singular vectors, right singular vectors, and singular values. When CA is applied to a contingency table (i.e., co-occurrences between two sets of categorical data) it as called "simple" CA or just CA. When CA is applied to a data table comprised of categorical variables—with observations on the rows and each level for all variables along the columns—the technique is called "multiple" CA (MCA).

While CA is a natural technique to analyze relationships between multiple genotypes, it is rarely used: Applications of CA to genetics studies exists—almost entirely—outside of North America and in most cases very strictly within biology and ecology (mostly to understand genetic diversity of animals and plants within the same geographical regions). For example, CA has been used to study the genetic underpinnings of ancestry, population structures, population divergence, and possible hybridization of a variety of animals: Eurasian otters (Geboes, Rosoux, Lemarchand, Hansen, & Libois, 2016), llamas and alpacas (Kadwell et al., 2001), manatees (Luna, 2013), sea turtles (Vilaça et al., 2012), platypuses (Furlan et al., 2012), mountain pygmy possums (Mitrovski, Heinze, Broome, Hoffmann, & Weeks, 2007), Little penguins (Burridge, Peucker, Valautham, Styan, & Dann, 2015), and monk parakeets (Edelaar et al., 2015).

Even though CA has been applied more in plants and animals, there have been some notable uses for human populations, the first of which was Greenacre and Degos (1977) to understand the distribution of human leukocyte antigen groups across many populations. CA has also been used—like MDS and PCA—to understand human "back migration" via ancestral genetic markers (Cruciani et al., 2002). More importantly, CA has been used to understand genetic aspects of psychiatric diseases such as major depressive disorder (Suchanek et al., 2011), schizophrenia (Paul-Samojedny et al., 2010), and alcohol use disorder (Onori et al., 2010).

3.2.6 Regularization techniques

Most multivariate methods can suffer from a variety of problems if particular conditions are not met, for example: when the number of observations is smaller than the number of variables (which is almost always the case in neuroimaging, genetics, and "imaging genetics"). One of these problems is often described by the umbrella term of "over-fitting," a name applied when a model appears to provide good results for a specific data set, but is more than likely fitting conditions specific to the data set (often times noise particular to the sample instead of the "true" underlying signal). To counteract over-fitting, a number of approaches are routinely used, such as cross-validation (e.g., jackknife or split half resampling), conservative resampling strategies (e.g., bootstrap), or approaches lumped under the generic term of "regularization" (a technique originally developed to handle the "multicollinearity" problem in multiple regression). Because several terms—such as regularization, sparsification, and shrinkage—have many uses, I will define several terms for use in this dissertation.

1. Shrinkage: the process of making values, such as β coefficients, go toward zero; usually effects small values the most.

2. Penalization (term): a particular value added to or subtracted from estimates (e.g., β coefficients) to shrink these values (towards zero) to counteract some assumed bias.
3. Regularization: any statistical approach meant to adjust for over-fitting through the use of penalization and shrinkage.
4. Sparsification: a particular form of regularization where shrinkage is used to minimize the number of non-zero elements in an estimate; often an iterative process. For example, shrinking β coefficients under the constraint that only 1 coefficient can be non-zero.

The most common and well-known regularization approaches are ridge regression (a.k.a., Tikhonov regularization (Hoerl & Kennard, 1970) and the LASSO ("least absolute shrinkage and selection operator"; Tibshirani, 1996). All these approaches are best understood from the perspective of ordinary least squares (OLS), as used, for example to estimate beta coefficients in multiple regression. OLS is defined as (Abdi, Edelman, Valentin, & Dowling, 2009; Chapter 24):

$$\mathbf{y} = \mathbf{X}\mathrm{B} + \varepsilon$$

where the estimation of β, assuming $\mathbf{X}$ is full rank, comes from:

$$\hat{\beta} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y},$$

where in $\hat{\mathbf{y}}$ is computed from

$$\hat{\mathbf{y}} = \mathbf{X}\beta = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y},$$

and $\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$ is the projection (or "hat") matrix. In OLS, we want to minimize the error sums of squares, so, the estimation equation can be written another way:

$$(\mathbf{y} - \mathbf{X}\beta)^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\beta) = \min.$$

For ridge regression we slightly modify the estimation equation to include a tuning parameter denoted λ:

$$\hat{\beta}^* = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y},$$

where $\mathbf{I}$ is the identity matrix (1s on the diagonal, 0s off-diagonal) and when λ is 0, we have the same results as OLS. Written another way:

$$(\mathbf{y} - \mathbf{X}\beta)^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\beta) + \lambda\beta^{\mathrm{T}}\beta = \min.$$

Where $\lambda\beta^{\mathrm{T}}\beta$ is called the *penalty term* for the standard OLS estimate for ridge regression. This penalty is the L_2-norm (a.k.a., Euclidean), multiplied by the tuning parameter λ, and often shown, more simply, as $\lambda\beta^{\mathrm{T}}\beta = \lambda\left\|\beta\right\|_2^2 = \lambda\sum_i \beta_i^2$. In a similar fashion to ridge regression, the LASSO also uses a penalty term with respect to this same minimization:

$$(\mathbf{y} - \mathbf{X}\beta)^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\beta) + \lambda\left\|\beta\right\|_1 = \min,$$

where $\lambda\left\|\beta\right\|_1 = \lambda\sum_i \left|\beta_i\right|$, which is also known as the L_1-norm (a.k.a., Manhattan, taxicab, or city-block distance), multiplied by the tuning parameter λ. By contrast with OLS and ridge regression, the LASSO requires an iterative process to come to a solution, and in addition, both ridge and LASSO procedures require an iterative search to find the optimal value of λ. When λ is a small positive value, the regularization process—to quote Takane and Hwang (2003)—"works almost magically" (pg. 3) to provide better (and possibly more stable) estimates of $\hat{\beta}^*$. Thus, in scenarios when the number of observations will almost always be lower than the number of variables (e.g., genomics, *f*MRI), we would likely benefit from techniques designed to handle (possibly) incorrect estimation, over-fitting, and collinearity.

3.2.6.1 Where it all began: practical and interpretable regularization with the SVD

When it comes to multivariate analyses of high dimensional data, PCA is the standard technique because it combines the original variables into new orthogonal variables that explain most of the variance of the data. But with large numbers of variables (e.g., millions), it can be difficult to understand what a component means (Jolliffe, 2002), and thus a number of solutions were developed to help interpretation. In the early days of PCA and factor analysis, rotations were developed (see, e.g., Tucker, 1944) in order to try to force as many values as possible—per component—towards zero, thus leaving only a few “interpretable” values for each component. However, sometimes it is not sufficient to just rotate, because there are still: (i) many non-zero values, (ii) many values still exist above some acceptable threshold, or, which is the more likely scenario today, (iii) unique solutions do not exist. Thus, Jolliffe et al., (2003) proposed an extension of rotation and regularization procedures for PCA that incorporated the LASSO regression technique (Tibshirani, 1996, 2011), and called this new form of PCA “SCoTLASS” (short for “simplified component technique LASSO”). SCoTLASS explicitly minimizes the number of non-zero loadings per component and thus facilitates the interpretation of components.

3.2.6.2 More modern approaches

SCoTLASS was followed by a number of proposed solutions (for PCA) to meet two goals simultaneously: (1) avoid over-fitting (or some other form of mis-estimation), and (2) minimize the number of interpretable variables to consider. Some of the most well-known and popular techniques are sparse PCA (Zou, Hastie, & Tibshirani, 2006), sparse PCA-regularized SVD (a.k.a., sPCA-rSVD; Shen & Huang, 2008), and—one of the “go-to” approaches—penalized matrix decomposition (PMD; Witten, Tibshirani, & Hastie, 2009), which was designed for both PCA and CCA. Though there are a number of other similar approaches to regularized PCA (e.g., shrinking singular values: Verbanck et al., 2013; rotations: Trendafilov & Adachi, 2014; or manifold optimizations: Genicot, Huang, & Trendafilov, 2015) most forms exist as ridge (Tikhonov) or sparsified (mostly via LASSO) PCA, multivariate regression, or related techniques (Allen & Maletić-Savatić, 2011; Gao et al., 2013; Hibar et al., 2015; Kohannim et al., 2011; Kohannim et al., 2012; Liu et al., 2013; Sill, Saadati, & Benner, 2015; Vounou et al., 2012; Vounou, Nichols, & Montana, 2010).

Like PLSC, CCA is a technique designed for the analysis of two data tables, where both techniques identify latent variables, but with a slightly different optimization criterion than PLSC (see Chapter 5). While both techniques are used for the analysis of genetic data, CCA is generally the more common technique used by behavioral and brain scientists (Chi et al., 2013; D. Lin et al., 2013; D. Lin, Calhoun, & Wang, 2014; Takane, Hwang, & Abdi, 2008; Jingwen Yan, Zhang, et al., 2014). There has also been recent development of sparse CCA approaches to include a third table that is used as a “guide” or “structure” (often applied via the observations), where in the “guide” or “structure” is some sort of targeted data or design matrix, or some preferred constraints (Du et al., 2016a; Yan, Du, Kim, Risacher, Huang, Moore, Saykin, & Shen, 2014; Jingwen Yan et al., 2013). While a sparsified PLS has been used in these domains (Le Floch et al., 2012), the majority of work exists in other domains (Chun et al., 2011; Chun & Keleş, 2009), such as genomic analysis of French dairy cattle (Colombani et al., 2012). However, one major issue still remains surrounding all of regularized approaches and nearly all of previously mentioned (regularized) multivariate methods: they are all designed under the assumptions that data—especially SNPs—are continuous.

3.3 The nature of SNP data and SNP models

Nearly all current methods and approaches to analyzing SNPs (alone or in conjunction with other data) treat SNPs as numerical data even though SNPs are inherently categorical variables. A SNP takes the general form of *AA* as the major homozygote, *aa* as the minor homozygote, and *Aa* as the heterozygote, where *A* is the major allele (most frequent of the two letters in a pair), and *a* the minor allele. The near universal standard of SNP coding, however, is in the form of [0,1,2], where it is most common to code 0 as the major homozygote—where there is no presumed risk—1 as the heterozygote, and 2 as the minor homozygote. This particular coding is referred to as the “additive model”, which assumes additive (and thus linear) effects on some trait. In some cases, however, an opposite coding is used—[2,1,0] for *AA*, *Aa*, and *aa*—though it must be noted this just produces a sign flip compared to [0,1,2]. While the additive model is the most commonly used, it is only one of several possible models to estimate genetic effects associated with some trait. The next most common models are the dominant and recessive, respectively. Both the dominant and recessive models dichotomize SNPs based on the minor or major allele. The dominant model would be *AA* vs. $\{Aa + aa\}$ where the recessive model would be $\{AA + Aa\}$ vs. *aa*. The multiplicative model exists as a compromise somewhere between the additive model and either a dominant or recessive model: emphasis is still placed on the minor allele, but the *difference* between each genotype is not uniform. The rarest type of effect is a heterozygous effect, and is estimated with the heterozygous model, which treats the heterozygote as different from both homozygotes: *Aa* vs. $\{AA + aa\}$. Finally, what is probably one of the rarest models in use—even though it is the most general—is the “genotypic” model, where each genotype’s effect is estimated. A summary of these models is shown in Table 3.1.

Though the additive model is almost universally used, there exists a simple counter example to show the inadequacy of additive assumptions: the risk of Alzheimer's Disease (AD) due to the *ApoE* gene is neither uniform nor linear (see e.g., Table 2 in Genin et al., 2011). Furthermore, "*ApoE* 4" gene alleles are considered risk factors for AD and the "*ApoE* 2" gene alleles are considered protective against AD (Corder et al., 1994). But, by contrast, "*ApoE* 2" is also a genetic risk factor for the inability to break down fats, and this inability could, in turn, predispose individuals to certain types of vascular diseases or obesity (Koopal et al., 2014). When risk factors are unknown, we—obviously—do not know (1) which direction, nor (2) how much risk is associated with which markers or diseases. So, applying these [0,1,2] values to SNPs is problematic because allelic counts do not represent *how much* of a SNP is present, but rather, *only which* allele pair is present (e.g*.,* AA, AT, or TT).

Thus far in the field, the [0,1,2] coding scheme is often used simply because most analytical methods are designed to handle numbers rather than qualitative data such as categorical or ordinal data. However, another part of the motivation is summarized by Balding (2006):

> "*For complex traits, it is widely thought that contributions to disease risk from individual SNPs will often be roughly additive—that is, the heterozygote risk will be intermediate between the two homozygote risks."* (pg. 784)

Another part of the motivation, also described by Balding (2006), is that "general tests"—which test all genotypes (*à la* one factor ANOVA), as opposed to [0,1,2] (*à la* regression, or as a contrast applied to the one factor ANOVA)—are less powerful than using an additive recoding scheme, and per the recommendation of Balding (2006):

> "*Using the Fisher test spreads the research investment over the full range of risk models, but this inevitably means investing less in the detection of additive risks."* (pg 785).

Though, just *before* this statement in the same paragraph, Balding makes a strong point about model selection for SNPs:

> *"There is no generally accepted answer to the question of which single-SNP test to use. We could design optimal analyses if we knew what proportion of undiscovered disease-predisposing variants function additively and what proportions are dominant, recessive or even over-dominant. Lacking this knowledge, researchers have to use their judgment to choose which 'horse' to back."* (pg 785),

and so implies that researchers *must choose only one* model to test—as opposed to using a general test. In their reply to Balding, Vormfelde and Brockmöller (2007) point out that if there are haplotypic effects, a [0,1,2] coding and subsequent test of *each SNP individually* will entirely miss the haplotype-phenotype relationship. Furthermore, in a separate study by Lettre, Lange, and Hirschhorn (2007)—conducted at almost the same time as these reviews and responses—showed that the additive model is not necessarily a ubiquitous, catch-all coding scheme that provides the most (statistically) powerful way to detect the effects of SNPs. In fact, this coding scheme can be detrimental when the effects are, for example, recessive: an effect shown by Lettre et al., (2007; see their Figures 1 and 2) that illustrates that when the additive model is used for a (true) recessive effect, power is very low. The same problem occurs when recessive models are applied to (true) inheritance patterns that are either dominant or additive. Furthermore, if an effect is detected (i.e., significant; through a statistical test) with, for example, the additive model but the (true) inheritance is—as some cases shown in Lettre et al. 2007)—dominant, or even heterozygous (not shown in Lettre et al.), then a researcher would conclude they have discovered an effect and would therefore commit two errors: (1) They would claim an effect was additive (a Type I error), while (2) missing the true effect (a Type II error). Thus, Lettre et al., (2007, p 361) point out that:

> "*We also note that testing the three models [additive, dominant, and recessive] together provided slightly more power than testing the two degrees of freedom co-dominant [general] model alone (compare 'Add+Dom+Rec' with 'Co-dominant' in Fig. 2), but these approaches were fairly similar."*

In fact, the similarity between the two options—three tests per SNP, vs. one test per SNP—had lead Lettre, et al. (2007, pg. 362) to conclude and recommend:

> *[…] testing the co-dominant statistical model alone, or alternatively testing the additive, dominant, and recessive models together but using empirically determined significance thresholds to correct for testing multiple correlated genetic models.*

In more simple terms: if you want to know which SNPs are significant *and* what their likely true effects are, the more general tests should be used or one must test *all* inheritance models (and then correct for the increase in number of tests). Note, though, that with respect to genome-wide (or large scale candidate) studies, SNP tests are already very conservative (i.e., $\alpha = 5 \times 10^{-8}$). If, on the genome-wide scale, we were to test all models, our already conservative threshold must become more conservative (i.e., a multiple comparisons correction applied to a multiple comparisons correction). Yet, it is extremely rare to find genome-wide (or even small and large scale candidate gene) association studies that use anything besides the additive model alone.

With respect to complex traits, conditions, diseases, and disorders, the almost exclusive use of the additive ([0, 1, 2]) model is, to some degree, perplexing given that many researchers believe that non-linear and non-additive polygenic effects are likely explanations for complex traits. For example, we turn to the ADNI, which is (relatively) open access data and has been studied by hundreds—or even thousands—of researchers at this point. The additive model is, without question, the only apparent one used on a large scale for genome-wide data (see, e.g., Potkin et al., 2009; Shen et al., 2010; Stein et al., 2011; Hibar et al., 2011; Shulman et al., 2013; Swaminathan et al., 2011; Meda et al., 2012; and Hohman, Koran, Thornton-Wells, & for the Alzheimer's Neuroimaging Initiative, 2013). However, in other studies with the ADNI data, it has been made clear that linear additive effects are not necessarily the appropriate assumption. Hibar et al., (2013) state:

> *"For many complex traits, the similarity of family members drops faster than would be expected as relatedness decreases [2]. This implies that there are non-additive (epistatic) interactions involved in the etiology of many complex traits."*

Yet, Hibar and colleagues (2013) only discuss the additive model in their analyses, and again, Hibar et al., (2015) note:

> *"Potential sources of the missing heritability might be caused by nonadditive effects like dominance and SNP-SNP interactions (Carlborg and Haley, 2004) and gene-by-environment interactions (Visscher et al., 2008), and rare genetic variants (Manolio et al., 2009)."*

However the authors only discuss the additive model in their analyses. And, in the ADNI data alone, a very high number of studies use only the [0, 1, 2] additive model, even though it is generally accepted that there are complex genetic contributions to Alzheimer’s disease (as well as many other diseases and disorders). The sole use of [0, 1, 2] can also be seen in many other domains, such as: cognitive ability (Arden, Harlaar, & Plomin, 2007), behavioral disinhibition (Derringer et al., 2015), intelligence (Loo et al., 2012), obesity and depression (Hung et al., 2014), personality traits (Kazantseva et al., 2015), onset of Alzheimer’s disease (Naj et al., 2014), CSF Tau levels in Alzheimer’s disease (Cruchaga et al., 2013), as well as various domains of psychology (Mõttus et al., 2015; Carr et al., 2013). This use of [0,1,2] can also be seen in fields outside of psychology and neuroscience (Kathiresan et al., 2008; Lango Allen et al., 2010; Takeuchi et al., 2009). The use of [0,1,2] makes little sense for such a myriad of complex traits. As Lehner (2011) states (in the “Concluding Remarks” section; *emphasis* mine):

> “It has been argued that phenotypic variation in a population could, in many cases, be accounted for by purely additive genetic models. *However, this is only a theoretical possibility [101], and it contradicts both the demonstrated importance of epistasis in particular human diseases [102, 103, 104 and 105] and the pervasive epistasis that has been detected in model organisms and highlighted here*. It is also somewhat inconsistent with patterns of sequence evolution [79, 81 and 106] *and inconsistent with our understanding of molecular biology and the abundance of non-linear regulatory interactions [86]*. Put simply, although they are very challenging to predict and detect in human populations because of a lack of statistical power [8 and 9], from what is currently understood about genetic architecture and biology, epistatic interactions between mutations are likely to be central to what makes us unique, both in health and disease.”

Thus, the additive model—while almost exclusively used—is not necessarily the best default approach, because it assumes that effects are uniform and linear, as well in the same direction, *across all SNPs*. Not only is this a problem for detecting genetic effects from individual SNPs, but it is also a problem when effects are polygenic *and* researchers use polygenic risk (a.k.a., multi-locus genetic profile) scores.

In recent years, polygenic risk (multi-locus profile) scores have become a popular (and simple) tool to assess how multiple markers (usually SNPs) may *together* contribute to a variety of traits or disorders (e.g., Davis et al., 2013; Mõttus et al., 2015; Nikolova, Ferrell, Manuck, & Hariri, 2011; Stice, Yokum, Burger, Epstein, & Smolen, 2012; Hart et al., 2014; Papiol et al., 2014). However, to truly estimate how a number of genetic markers should, essentially, be added up to create a single "risk" value, we require robust research on traits in advance, where we can estimate effect sizes (and directions) from either (1) meta-analytic strategies or (2) Bayesian estimates. However, if there is not enough information to apply particular effect sizes and directions to each marker, then polygenic scores must rely on *empirically defined* risk, which is usually the (empirical, from the data set) minor allele and SNPs are treated as [0,1,2]. In these cases if, for example, two SNPs with minor allele-defined (based on sample) "risks" contribute in an *opposite* way to some behavior or trait, then the effect is essentially nullified. Combining the same genotypes across two SNPs with the additive code would produce a "2" in three distinctly different configurations: 0 + 2 = 2, 1 + 1 = 2, and 2 + 0 = 2; there is no predictive power in this case. In fact, this exact pattern exists in one of the most well-known genes: *ApoE*.

The *ApoE* gene alleles are usually genotyped through two SNPs: rs429358 is C/T, where the minor allele is C (dbSNP MAF: 0.150), and rs7412 is C/T, where the minor allele is T (dbSNP MAF: 0.075). Table 3.2 shows a tabular format of *ApoE* gene alleles with respect to these two SNPs. Based on the unifying nomenclature in Zannis et al., (1982), and discussed in Nyholt, Yu, and Visscher (2008), to be *ApoE* E4/E4 requires a *minor* homozygote on rs429358 and a *major* homozygote on rs7412. However it is likely that, for Alzheimer's risk, the real genetic risk (due to *ApoE*) is likely only with respect to rs429358 (Bennet et al., 2010).

Furthermore, different SNPs are likely to express *different inheritance patterns*, depending on the trait studied (and possibly even *how* that trait is measured, i.e., choice of instrument). As pointed out by Genin et al., (2011) for disease risk, *ApoE* does indeed show a non-linear effect that roughly matches a quadratic (ordinal) risk (from E2/2 → E4/4). However, *ApoE*—with respect to Alzheimer's disease—has shown a variety of different patterns across a number of different Alzheimer's traits or phenotypes. Some of the clearest example of non-additivity can be found in Lebedeva et al., (2012): Figure 2a shows a recessive (e.g., [0 & 1] > 2) effect of a haplotype on beta-amyloid levels, where Figure 2b shows an over-dominant (e.g., [0 & 2] > 1) effect of a haplotype on beta-amyloid levels; Figure 3a and b show what appears to be an inverted U-shape effect contrary to what is typically considered an ordinal risk factor for Alzheimer's Disease: E2/3 → E2/4 → E3/3 → E3/4 → E4/4. We can also see (in Levedeva and colleagues' figures) that different levels of beta-amyloid—considered a likely pathological marker of AD—do not match the expected ordinal effect from least-risk associated *ApoE* allele (E2/3) to most (E4/4)—in fact, the first (E2/3) and last (E4/4) alleles have essentially the same relation to beta-amyloid. Next, they present the same data in the format of 0, 1, or 2 E4 alleles.

Here again is the inverted U-shape and most resembles an over-dominant (heterozygous) model. This suggests that *ApoE* genotypes are related to a variety of phenotypes, endophenotypes, and biomarkers under different inheritance models. Similar non-linear and non-additive contributions of *ApoE*—to some trait or phenotype—can be also be seen in (to name just a few) Lautner et al. (2014), Linnertz et al. (2014), Soares et al., (2012), and Trabzuni et al., (2012), Stranger et al., (2005). Likewise, the following papers also show a variety of non-linear expression effects with respect to genotype (Gibbs et al., 2010, Figure 4; Zhang et al., 2008, Figure 3; Myers et al., 2007, Figure 3; Okamoto et al., 2011, Figure 2). Finally, when it comes to how a gene is expressed, it is, in most cases, tissue dependent (that is, gene expression varies based on where a particular gene's expression is being measured). What will not vary, however, is the genotype. In fact, some of our own recent work shows how multiple genes are expressed quite differently all throughout the cortex (Cioli, Abdi, Beaton, Burnod, & Mesmoudi, 2014). Even though we did not analyze genotypes, the differential expression in various regions still aligns with the fact that a genotype does not necessarily reflect the same expression throughout the cortex. In this context, the same allele results in expression patterns corresponding to a variety of models simply depending on where expression was measured in the cortex.

Finally, while many—especially multivariate—methods are "data-driven" these methods cannot operate free of hypotheses if an additive model is applied to *all* SNPs *a priori*. The application of the additive code could hinder "data-driven" techniques. Thus, in association studies of complex diseases, disorders, traits, and behaviors the inheritance model for a given SNP is rarely, if ever, known, and as such, if the inheritance model is not known, applying the same inheritance model to every SNP could be problematic.

Table 3.2

ApoE Gene Allele, SNP, Additive Coding, and "Risk Factor"

Genotypes	**E2**		**E3**		**E4**	
	rs429358	rs7412	rs429358	rs7412	rs429358	rs7412
E2	TT	TT				
E3	TT	CT	TT	CC		
E4	CT	CT	CT	CC	CC	CC

Allelic	**E2**		**E3**		**E4**	
	rs429358	rs7412	rs429358	rs7412	rs429358	rs7412
E2	AA	aa				
E3	AA	Aa	AA	AA		
E4	Aa	Aa	Aa	AA	aa	AA

Additive	**E2**		**E3**		**E4**	
	rs429358	rs7412	rs429358	rs7412	rs429358	rs7412
E2	0	2				
E3	0	1	0	0		
E4	1	1	1	0	2	0

Note. The three representations of *ApoE*: as SNP genotypes, as general allelic format, and the additive model. With respect to polygenic risk scores, *ApoE2*/2, *ApoE4*/4, and *ApoE2*/4 (or *ApoE4*/2) all compute to the same (minor allele-based) risk value of "2".

CHAPTER 4

THE STATISTICS OF PLS, CA, AND PLS-CA

Parts of this chapter are adapted from my Beaton, Dunlop, ADNI, & Abdi (2016). Copyright © 2015 American Psychological Association. Reproduced with permission. The official citation that should be used in referencing this material is Beaton, D., Dunlop, J., Abdi, H., & Alzheimer's Disease Neuroimaging Initiative. (2016). Partial least squares correspondence analysis: A framework to simultaneously analyze behavioral and genetic data. Psychological Methods, 21(4), 621-651. http://dx.doi.org/10.1037/met0000053. This article may not exactly replicate the authoritative document published in the APA journal. It is not the copy of record. No further reproduction or distribution is permitted without written permission from the American Psychological Association.

4.1 The Generalized Singular Value Decomposition

The singular value decomposition (SVD)—and by extension the Generalized SVD (GSVD)—is the core tool for many statistical and analytical techniques such as PCA, correspondence analysis (CA), discriminant analysis (DA), canonical analysis (CCA) and partial least square methods (PLS) to name only but a few methods.

The SVD decomposes rectangular tables (Yanai et al., 2011). Typically, the rows of these rectangular matrices are observations, and the columns are variables (that describe the observations). The SVD produces orthogonal components (sometimes called dimensions, axes, principal axes, or factors) that are new variables computed as linear combinations of the original variables. Because components are orthogonal (i.e., two different components have zero correlation), they can also be obtained as a simple geometric rotation of axes with respect to the original variables (Jolliffe, 2002). The first component always explains the maximum variance in the data and each following component explains the next largest possible amount of remaining variance under the condition that components are mutually orthogonal.

Observations and measures are assigned values, called component (or factor) scores, for each component. Component scores reflect how much an observation contributes to the variance of a component. Additionally, component scores are often plotted to produce component maps (akin to scatter plots). These maps represent the relationship between observations, between measures, and some cases between observations and variables (Greenacre, 1984). In the maps, observations close to each other are similar and observations far apart differ.

Recall that the SVD decomposes a data matrix $\mathbf{R}$—with J rows and K columns—into three matrices:

$$\mathbf{R} = \mathbf{U\Delta V}^{\mathrm{T}}, \tag{4.1}$$

where $\mathbf{R}$ has rank L, $\mathbf{U}$ is a J by L matrix of left singular vectors, $\mathbf{V}$ is a K by L matrix of right singular vectors, and $\mathbf{\Delta}$ is an L by L diagonal matrix where $\mathrm{diag}\{\mathbf{\Delta}\}$ stores the singular values. Furthermore, $\mathbf{U}$ and $\mathbf{V}$ are orthonormal matrices such that

$$\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I} = \mathbf{V}^{\mathrm{T}}\mathbf{V}. \tag{4.2}$$

Component scores for the J rows and K columns are computed as

$$\mathbf{F}_J = \mathbf{U\Delta} \text{ and } \mathbf{F}_K = \mathbf{V\Delta}, \tag{4.3}$$

and can be plotted—often with two components at a time—to produce component maps. As described by Greenacre (1984), Lebart et al., (1984), and Abdi (2007c), the GSVD generalizes the standard SVD by imposing—on, respectively, the left and right singular vectors—constraints represented by positive definite matrices of sizes (respectively) J by J and K by K. These constraints matrices are often diagonal matrices and, when this is the case, they are usually called masses or weights. We denote the weights for the rows, $\mathbf{W}_{J,}$ and the weights for the columns, $\mathbf{W}_{K}$. Decomposition of a matrix is the same as in Eq. 4.1 with the following constraints:

$$\mathbf{U}^{\mathrm{T}}\mathbf{W}_{J}\mathbf{U} = \mathbf{I} = \mathbf{V}^{\mathrm{T}}\mathbf{W}_{K}\mathbf{V}, \tag{4.4}$$

where component scores for the J rows and K columns are computed as

$$\mathbf{F}_{J} = \mathbf{W}_{J}\mathbf{U}\boldsymbol{\Delta} \text{ and } \mathbf{F}_{K} = \mathbf{W}_{K}\mathbf{V}\boldsymbol{\Delta}. \tag{4.5}$$

The GSVD is a very powerful technique and, with the correct selection of weights, can implement or generalize many techniques (e.g., correspondence analysis, multi-dimensional scaling, Fisher's linear discriminant analysis, canonical correlation analysis). For a comprehensive list of techniques that the GSVD generalizes, see Appendix A of Greenacre (1984).

4.2 Partial Least Squares (Correlation)

In this section I present a summary of partial least squares correlation (PLSC)—sometimes also called Tucker's inter-battery analysis (Tucker, 1958), singular value decomposition of the covariance between two fields (Bretherton et al., 1992), or co-inertia analysis (Dray, 2014), or even, recently, "multivariate genotype-phenotype" (MGP) analysis (Mitteroecker, Cheverud, & Pavlicev, 2016)—in order to (1) provide required background and (2) establish the concepts and notations we need for a novel generalization of PLSC to varied data types (Beaton, Dunlop et al., 2016; Beaton, Kriegsman et al., 2016).

PLSC analyzes the relationship between two data matrices of sizes (respectively) I by J and I by K, denoted (respectively) $\mathbf{X}$ and $\mathbf{Y},$ that measure the same I observations (rows) described by (respectively) J and K quantitative variables (i.e., columns). The centered and normalized versions of $\mathbf{X}$ and $\mathbf{Y}$ are denoted $\mathbf{Z}_{\mathbf{X}}$ and $\mathbf{Z}_{\mathbf{Y}}$. The common information between these two data tables is represented by the matrices computed as:

$$\mathbf{R} = \mathbf{X}^{\mathrm{T}}\mathbf{Y} \text{ and } \mathbf{Z}_{\mathbf{R}} = \mathbf{Z}_{\mathbf{X}}^{\mathrm{T}}\mathbf{Z}_{\mathbf{Y}}. \tag{4.6}$$

This multiplication produces a J by K cross-product matrix ($\mathbf{R}$) or correlation matrix ($\mathbf{Z}_{\mathbf{R}}$). In PLSC (Krishnan et al., 2011; McIntosh et al., 1996; Bookstein, 1994; Abdi & Williams, 2013) the variables are, in general, centered and normalized (i.e., matrices $\mathbf{Z}_{\mathbf{X}}$ and $\mathbf{Z}_{\mathbf{Y}}$ are used) and therefore the matrix $\mathbf{Z}_{\mathbf{R}}$ is used for further analysis. This matrix $\mathbf{Z}_{\mathbf{R}}$ is decomposed with the singular value decomposition as:

$$\mathbf{Z}_{\mathbf{R}} = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}^{\mathrm{T}}, \tag{4.7}$$

where (1) L is the rank of $\mathbf{Z}_{\mathbf{R}}$, (2) $\mathbf{U}$ is the J by L orthonormal matrix of left singular vectors, (3) $\mathbf{V}$ is the K by L orthonormal matrix of right singular vectors, and (4) $\boldsymbol{\Delta}$ is an L by L diagonal matrix (i.e., the off-diagonal elements of $\boldsymbol{\Delta}$ are all 0) where the elements of the vector diag$\{\boldsymbol{\Delta}\}$ are the singular values (ordered from the largest to the smallest). The squared singular values—called eigenvalues—express the variance of the data extracted by the components. In the PLSC nomenclature, the matrices $\mathbf{U}$ and $\mathbf{V}$ are also called *saliences* (Bookstein, 1994; McIntosh & Lobaugh, 2004; Krishnan et al., 2011). The matrices $\mathbf{U}\boldsymbol{\Delta}$ and $\mathbf{V}\boldsymbol{\Delta}$ are akin to component scores for PCA (Abdi & Williams, 2010) and CA (Abdi & Béra, 2014).

In PLSC, the original variables of $\mathbf{Z_X}$ and $\mathbf{Z_Y}$ are linearly combined to create pairs of *latent variables* (each pair has one latent variable from $\mathbf{Z_X}$ and one from $\mathbf{Z_Y}$, see, Krishnan et al., 2011; Abdi & Williams, 2010b). The coefficients of these linear combinations are given by the singular vectors of $\mathbf{Z_R}$. The latent variables for $\mathbf{Z_X}$ and $\mathbf{Z_Y}$ are computed as

$$\mathbf{L_X} = \mathbf{Z_X}\mathbf{U} \quad \text{and} \quad \mathbf{L_Y} = \mathbf{Z_Y}\mathbf{V}. \tag{4.8}$$

4.2.1 What does PLSC maximize?

PLSC seeks two vectors of coefficients—denoted $\mathbf{u}$ (resp. $\mathbf{v}$)—that define a linear combination of the columns of $\mathbf{Z_X}$ (resp. $\mathbf{Z_Y}$) such that these two linear combinations—called latent variables, denoted $\mathbf{l_X}$ (resp. $\mathbf{l_Y}$), and computed as $\mathbf{l_X} = \mathbf{Z_X}\mathbf{u}$ (resp. $\mathbf{l_Y} = \mathbf{Z_Y}\mathbf{v}$)—have maximal covariance as stated by:

$$\delta = \arg\max(\mathbf{l}_\mathbf{X}^\mathrm{T}\mathbf{l_Y}) \; = \; \arg\max \operatorname{cov}(\mathbf{l}_\mathbf{X}^\mathrm{T}, \mathbf{l_Y}) \tag{4.9}$$

under the constraints that the set of coefficients of the linear transformation for $\mathbf{Z_X}$ (resp. $\mathbf{Z_Y}$) have unit norm:

$$\mathbf{u}_l^\mathrm{T}\mathbf{u}_l = 1 = \mathbf{v}_l^\mathrm{T}\mathbf{v}_l. \tag{4.10}$$

After the first pair of latent variables has been extracted, subsequent pairs are extracted under the additional condition that unpaired sets of latent variables are orthogonal:

$$\mathbf{l}_{\mathbf{X},l}^\mathrm{T}\mathbf{l}_{\mathbf{Y},l'} = 0 \text{ when } l \neq l'. \tag{4.11}$$

The coefficients of the successive linear transformations (stored in matrices $\mathbf{L}_{\mathbf{X}}$ and $\mathbf{L}_{\mathbf{Y}}$) are obtained from the SVD of $\mathbf{Z}_{\mathbf{R}}$ (see Eq. 4.7) as shown by

$$\mathbf{L}_{\mathbf{X}}^{\mathrm{T}}\mathbf{L}_{\mathbf{Y}} = \mathbf{U}^{\mathrm{T}}\mathbf{Z}_{\mathbf{X}}^{\mathrm{T}}\mathbf{Z}_{\mathbf{Y}}\mathbf{V} \; = \; \mathbf{U}^{\mathrm{T}}\mathbf{Z}_{\mathbf{R}}\mathbf{V} = \; \mathbf{U}^{\mathrm{T}}\mathbf{U}\boldsymbol{\Delta}\mathbf{V}^{\mathrm{T}}\mathbf{V} = \; \boldsymbol{\Delta}. \tag{4.12}$$

When $l = 1$, the covariance between $\mathbf{L}_{\mathbf{X}}$ and $\mathbf{L}_{\mathbf{Y}}$ has the largest possible value, when $l = 2$, the covariance between $\mathbf{L}_{\mathbf{X}}$ and $\mathbf{L}_{\mathbf{Y}}$ has the largest possible value under the constraints that the second pair of latent variables are orthogonal (as defined by Eq. 4.11) to the first pair of latent variables. This property holds for each subsequent value of l; for proofs, see, Tucker (1958) and Bookstein (1994), in addition to Section 3.1.2.3, Lebart et al., (1984) and Greenacre (1984).

4.3 Partial Least Squares-Correspondence Analysis

The properties of PLSC hold when matrices **X** and **Y** contain quantitative variables (and therefore $\mathbf{Z}_{\mathbf{R}}$ is a correlation matrix). However, SNPs and many types of behavioral data (e.g., surveys, clinical assessments, diagnostic groups) are inherently categorical. Here, I present and formalize a PLS method recently developed hat was designed to handle categorical data (Beaton, Filbey, & Abdi, 2013; Beaton, Dunlop et al., 2016) or "heterogeneous" (aka mixed) data (Beaton, Dunop, et al., 2016; Beaton, Kriegsman, et al., 2016), called "Partial Least Squares-Correspondence Analysis" (PLSCA). The following section is adapted or directly taken from Beaton, Dunlop, et al., (2016).

4.3.1 Formalization of PLS-CA

PLSCA analyzes the relationships between two tables of categorical data (denoted **X** and **Y**) that describe the same set of *I* observations (i.e., rows). Both **X** and **Y** store categorical variables that are expressed with group coding (a.k.a. “disjunctive coding” or “indicator matrix coding,” see, e.g., Lebart et al., 1984; Greenacre, 1984), as illustrated in Table 2 of Beaton, Dunop, et al., (2016). With this coding scheme, the *N* levels of a categorical variable are coded with *N* binary vectors. The level describing an observation has a value of 1 and the other levels have a value of 0, and so the product $\mathbf{X}^{\mathrm{T}}\mathbf{Y}$ creates a contingency table. Contingency tables are routinely analyzed with χ^2 statistical approaches and thus we developed PLSCA in such a χ^2 framework.

First, compute the vectors of the proportional column sums for **X** and **Y**, and call these vectors *masses*:

$$\mathbf{m}_{\mathbf{X}} = (\mathbf{1}^{\mathbf{T}}\mathbf{X}\mathbf{1})^{-1} \times (\mathbf{1}^{\mathbf{T}}\mathbf{X}) \text{ and } \mathbf{m}_{\mathbf{Y}} = (\mathbf{1}^{\mathbf{T}}\mathbf{Y}\mathbf{1})^{-1} \times (\mathbf{1}^{\mathbf{T}}\mathbf{Y}) \tag{4.13}$$

(with **1** being a conformable vector of ones). In PLSCA, each level of a variable is weighted according to the information it provides. Assuming that rare occurrences are more informative than frequent occurrences, these weights are computed as the inverse of the relative frequencies (masses) and stored in diagonal matrices computed as:

$$\mathbf{W}_{\mathbf{X}} = \mathrm{diag}\{\mathbf{m}_{\mathbf{X}}\}^{-1} \text{ and } \mathbf{W}_{\mathbf{Y}} = \mathrm{diag}\{\mathbf{m}_{\mathbf{Y}}\}^{-1}\ . \tag{4.14}$$

As in PLSC, the disjunctive data matrices **X** and **Y** are, in general, pre-processed to have zero mean and unitary norm. Here, centered and normalized matrices are denoted $\mathbf{Z}_{\mathbf{X}}$ and $\mathbf{Z}_{\mathbf{Y}}$, and with $N_{\mathbf{X}}$ and $N_{\mathbf{Y}}$ denoting the number of (original) variables (i.e., before disjunctive coding) for **X** and **Y** respectively, matrices $\mathbf{Z}_{\mathbf{X}}$ and $\mathbf{Z}_{\mathbf{Y}}$ are computed as

$$\mathbf{Z}_{\mathbf{X}} = (\mathbf{X} - (\mathbf{1}(\mathbf{1}^{\mathbf{T}}\mathbf{X} \times I^{-1}))) \times (N_{\mathbf{X}}\, I^{\frac{1}{2}})^{-1} \tag{4.15}$$

and

$$\mathbf{Z_Y} = (\mathbf{Y} - (\mathbf{1}(\mathbf{1^T Y} \times I^{-1}))) \times (N_\mathbf{Y}\, I^{\frac{1}{2}})^{-1}. \tag{4.16}$$

From here we compute $\mathbf{Z_R}$ as:

$$\mathbf{Z_R} = \mathbf{Z_X^T Z_Y}\,. \tag{4.17}$$

Then we decompose $\mathbf{Z_R}$ with the GSVD as

$$\mathbf{Z_R} = \mathbf{U \Delta V}^\mathrm{T} \text{ with } \mathbf{U}^\mathrm{T}\mathbf{W_X U} = \mathbf{I} = \mathbf{V}^\mathrm{T}\mathbf{W_Y V}. \tag{4.18}$$

Similarly to PLSC, the latent variables are computed as weighted projections on the left and right singular vectors:

$$\mathbf{L_X} = \mathbf{Z_X W_X U} \text{ and } \quad \mathbf{L_Y} = \mathbf{Z_Y W_Y V}, \tag{4.19}$$

where—by analogy with PLSC—$\mathbf{W_X U}$ and $\mathbf{W_Y V}$ are called saliences. PLSCA performs a maximization similar PLSC, namely that the first pair of latent variables have maximum covariance evaluated just as in Eq. 4.12, except under the constraints that $\mathbf{u}$ and $\mathbf{v}$ each have unit $\mathbf{W_X}$ -norm and $\mathbf{W_Y}$-norm, respectively:

$$\mathbf{u}_l^\mathrm{T}\mathbf{W_X u}_l = 1 = \mathbf{v}_l^\mathrm{T}\mathbf{W_Y v}_l. \tag{4.20}$$

Just like with PLSC, after the first pair of latent variables has been extracted, subsequent pairs are extracted under the additional condition that unpaired sets of latent variables are orthogonal. The coefficients of the successive linear transformations (stored in matrices $\mathbf{L_X}$ and $\mathbf{L_Y}$) are obtained from the GSVD of $\mathbf{Z_R}$:

$$\mathbf{L_X^T L_Y} = \mathbf{U}^\mathrm{T}\mathbf{W_X Z_X^T Z_Y W_Y V} = \mathbf{U}^\mathrm{T}\mathbf{W_X Z_R W_Y V} = \mathbf{U}^\mathrm{T}\mathbf{W_X U \Delta V}^\mathrm{T}\mathbf{W_Y V} = \mathbf{\Delta}. \tag{4.21}$$

When $l = 1$, the covariance between $\mathbf{L_X}$ and $\mathbf{L_Y}$ has the largest possible value, when $l = 2$, the covariance between $\mathbf{L_X}$ and $\mathbf{L_Y}$ has the largest possible value under the constraints that the second pair of latent variables is orthogonal to the first pair of latent variables, and therefore:

$$\text{diag}\{\mathbf{L}_\mathbf{X}^\text{T}\mathbf{L_Y}\} = \text{diag}\{\mathbf{\Delta}\}. \tag{4.22}$$

4.3.2 Links to Correspondence Analysis

PLSCA can be seen as a generalization of PLSC for two categorical data tables, but also as an extension of Correspondence Analysis (CA; Abdi & Williams, 2010a; Greenacre, 1984; Lebart et al., 1984). Correspondence analysis, in turn, is often presented as a generalization of PCA to be used for qualitative data. PCA decomposes the total variance of a quantitative data table, whereas CA—as a generalized PCA—decomposes the χ^2 of a data table because this statistic is analogous to the variance of a contingency table. First, CA computes **R** (a contingency table) as:

$$\mathbf{R} = \mathbf{X}^\text{T}\mathbf{Y}. \tag{4.23}$$

Next, CA computes two matrices related to **R**, referred to in the χ^2 framework as *observed* ($\mathbf{O_R}$) and *expected* ($\mathbf{E_R}$). The observed matrix is computed as

$$\mathbf{O_R} = \mathbf{R} \times (\mathbf{1}^\mathbf{T}\mathbf{R}\mathbf{1})^{-1}, \tag{4.24}$$

and the computation of expected values of **R** (under independence) comes from the marginal frequencies of **R** (which are also the masses—and relative frequencies of the columns—of **X** and **Y**, see Eq. 4.13):

$$\mathbf{E_R} = \mathbf{m_X}\mathbf{m}_{\mathbf{Y}}^{\mathrm{T}}. \tag{4.25}$$

Next, just as when computing the χ^2, we compute the deviations:

$$\mathbf{Z_R} = \mathbf{O_R} - \mathbf{E_R}, \tag{4.26}$$

a formula which is equivalent to Eq. 4.17 and, so, $\mathbf{Z_R}$ can be decomposed according to Eq. 4.18. In CA, the component scores for the rows and the columns of a matrix (the *J* and *K* elements of **R**) are computed as:

$$\mathbf{F}_J = \mathbf{W_X U \Delta} \text{ and } \mathbf{F}_K = \mathbf{W_Y V \Delta}. \tag{4.27}$$

Like in CA and PCA, several additional indices can be computed from the component scores. These indices are called contributions, cosines, and squared distances. Each of the indices provide additional information on how variables, from each variable set (*J* and *K* variables) contribute to the structure of the components. For more information, see Lebart et al. (1984), Greenacre (1984), Abdi & Williams (2010a), and Beaton et al., 2014.

Component scores for the *I* observations, of both **X** and **Y**, can be computed via supplementary projections. The component scores for observations of **X** and **Y**, are projected as supplementary elements by projecting them onto their respective singular vectors. Specifically, the first step computes **X** observed and **Y** observed, (cf. Eq. 4.24):

$$\mathbf{O_X} = \mathbf{X} \times (\mathbf{1}^{\mathbf{T}}\mathbf{X}\mathbf{1})^{-1} \text{ and } \mathbf{O_Y} = \mathbf{Y} \times (\mathbf{1}^{\mathbf{T}}\mathbf{Y}\mathbf{1})^{-1}, \tag{4.28}$$

then $\mathbf{O_X}$ and $\mathbf{O_Y}$ are projected as supplementary elements:

$$\mathbf{F_X} = \mathbf{O_X}\mathbf{F}_J\mathbf{\Delta}^{-1} = \mathbf{O_X W_X U \Delta \Delta}^{-1} = \mathbf{O_X W_X U}, \tag{4.29}$$

$$\mathbf{F_Y} = \mathbf{O_Y}\mathbf{F}_K\mathbf{\Delta}^{-1} = \mathbf{O_Y W_Y V \Delta \Delta}^{-1} = \mathbf{O_Y W_Y V}. \tag{4.30}$$

Finally, we compute the latent variables—which are proportional to the supplementary projections obtained by re-scaling the component scores (in Eqs. 4.29 and 4.30):

$$\mathbf{L_X} = \mathbf{F_X} \times I^{\frac{1}{2}} \text{ and } \mathbf{L_Y} = \mathbf{F_Y} \times I^{\frac{1}{2}}. \tag{4.31}$$

Equivalently, the latent variables could be directly computed as:

$$\mathbf{L_X} = \mathbf{Z_X}\mathbf{F}_J\mathbf{\Delta}^{-1} = \mathbf{Z_X}\mathbf{W_X}\mathbf{U} \text{ and } \mathbf{L_Y} = \mathbf{Z_Y}\mathbf{F}_K\mathbf{\Delta}^{-1} = \mathbf{Z_Y}\mathbf{W_Y}\mathbf{V}. \tag{4.32}$$

So, in conclusion—as the name Partial Least Squares-Correspondence Analysis indicates—the computations and rationale of the analysis can be interpreted either as a generalization of PLSC or an extension of CA. Both perspectives provide a basis of how to extend PLS-CA into a regularized version of PLS-CA.

CHAPTER 5

SMOOTHED PARTIAL LEAST SQUARES-CORRESPONDENCE ANALYSIS

Regularization is a somewhat broad term in statistics, used as an umbrella term for almost any techniques that addresses ill-posed problems. Generally, regularization techniques aim to improve stability of results, prevent overfitting, or aid interpretation. Parts of this chapter are adapted from Beaton, Dunlop, ADNI, & Abdi (2016). Copyright © 2015 American Psychological Association. Reproduced with permission. The official citation that should be used in referencing this material is Beaton, D., Dunlop, J., Abdi, H., & Alzheimer's Disease Neuroimaging Initiative. (2016). Partial least squares correspondence analysis: A framework to simultaneously analyze behavioral and genetic data. Psychological Methods, 21(4), 621-651. http://dx.doi.org/10.1037/met0000053. This article may not exactly replicate the authoritative document published in the APA journal. It is not the copy of record. No further reproduction or distribution is permitted without written permission from the American Psychological Association.

In practice, many regularization approaches shrink values (e.g., β estimates, component scores or loadings) toward zero. These approaches are popular because shrinking adjusts for biases (e.g., overfitting). Because values shrink toward or become zero, regularization can also make interpretation of the results easier (by reducing the number of variables to consider). The two most well known and commonly used forms of regularization are ridge (a.k.a. Tikhonov) and the LASSO, which are different forms of regularization (see definitions and equations in Chapter 3 for more details). While ridge and LASSO share a number of features and theoretical principles they are different techniques: Ridge regularization shrinks values towards zero—but does not guarantee a minimal set of null values—whereas the LASSO guarantees a minimal and unique set of non-null values via L_1-norm (but the LASSO is iterative which is a substantial drawback for large data sets).

Because PLS-CA is designed for large data sets, a parsimonious regularization procedure is more suited for this technique. Therefore, I use ridge (i.e., Tikhonov) regularization as the basis to extend PLS-CA into a regularized method. Three approaches can be used to incorporate ridge-like regularization in PLS-CA (especially for large genetics data sets): (1) regularized χ^2 for SNPs (Li et al., 2014), (2) regularized multiple correspondence analysis (Takane & Hwang, 2006), and (3) sparse and functional (a.k.a. "two-way") PCA (Allen, 2013).

5.1 Ridge OLS, Correspondence Analysis, and reguarlized SVD approaches

Recall (see Chapters 3 and 4) that CA *generalizes* PCA to nominal data and decomposes a data matrix under the assumptions of independence: essentially, CA is a χ^2 version of PCA. Therefore we can use some basic principles of χ^2 and ridge regularization to understand how to regularize CA and thus PLS-CA. For convenience, the OLS and ridge OLS equations (from Chapter 3) are recalled below:

$$\hat{\beta} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \text{ and} \tag{5.1}$$

$$\hat{\beta}^{*} = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}\,. \tag{5.2}$$

Before moving on we should also further define a projection matrix with respect to OLS (beyond what is provided in Chapter 3) as the concept is revisited for SmooPLS-CA. In OLS we estimate $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}.$ If we expand this with respect to Eq. 5.1 we have $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}.$ For OLS, we project $\mathbf{y}$ onto the orthogonal subspace defined by $\mathbf{P} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}.$ We refer to $\mathbf{P}$ as a *projection matrix*, and thus $\hat{\mathbf{y}}$ is the estimate of $\mathbf{y}$ with respect to $\mathbf{P}$: $\hat{\mathbf{y}} = \mathbf{P}\mathbf{y}.$ For ridge OLS we define $\mathbf{P}^{*} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^{\mathrm{T}},$ as a ridge projection matrix (*cf* Eq. 5.2).

Equations 5.1 and 5.2 include an inverse term [e.g., the term $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$] and so can interpreted as having a "numerator" and a "denominator" (i.e., the inverse). In Eqs. 5.1 and 5.2, the denominators would be, respectively: $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ and $(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I})^{-1}$, wherein the term $\lambda\mathbf{I}$ in Eq. 5.2 *inflates* the diagonal of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ (compared to that in Eq. 5.1). The inflation of the diagonal boosts the sample size and forces the numerator (predicted)—$\mathbf{X}^{\mathrm{T}}\mathbf{y}$—to exist *within* a larger subspace of the denominator (predictors)—$\mathbf{X}^{\mathrm{T}}\mathbf{X}$. The inflation is because of the (diagonalized) tuning parameter—$\lambda\mathbf{I}$. This procedure reduces bias and improves the estimate of $\mathbf{b}$. Ridge regularization forces the "numerator" to fit into a smaller subspace (with respect to the "denominator"). When it comes to PCA-like methods (based on the SVD), the generalized SVD (GSVD) is particularly suited for ridge-like approaches.

When CA is interpreted in the standard "observed minus expected" χ^2 framework, it starts with a contingency table $\mathbf{R}$ and requires two matrices related to $\mathbf{R}$ that are called the *observed* ($\mathbf{O}_{\mathbf{R}}$) and *expected* ($\mathbf{E}_{\mathbf{R}}$) matrices. The observed matrix is computed as:

$$\mathbf{O}_{\mathbf{R}} = \mathbf{R} \times (\mathbf{1}^{\mathbf{T}}\mathbf{R}\mathbf{1})^{-1}, \tag{5.3}$$

which is just $\mathbf{R}$ divided by the total of all its entries. The computation of expected values of $\mathbf{R}$ (under independence) comes from the relative marginal frequencies (row sums divided by total sum; column sums divided by total sum) of $\mathbf{R}$:

$$\mathbf{E}_{\mathbf{R}} = \mathbf{m}_J \mathbf{m}_K^{\mathrm{T}}. \tag{5.4}$$

From the marginal probabilities, we also compute weight matrices: the *inverse* of these frequencies (masses) are stored in diagonal matrices:

$$\mathbf{W}_J = \mathrm{diag}\{\mathbf{m}_J\}^{-1} \text{ and } \mathbf{W}_K = \mathrm{diag}\{\mathbf{m}_K\}^{-1} . \tag{5.5}$$

Next, just as when computing χ^2, we compute the deviation of *observed* from *expected*:

$$\mathbf{Z}_{\mathbf{R}} = \mathbf{O}_{\mathbf{R}} - \mathbf{E}_{\mathbf{R}}, \tag{5.6}$$

For CA, we then decompose $\mathbf{Z}_{\mathbf{R}}$ with the GSVD (SVD; see also Chapter 4) as

$$\mathbf{Z}_{\mathbf{R}} = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}^{\mathrm{T}} \text{ with } \mathbf{U}^{\mathrm{T}}\mathbf{W}_J\mathbf{U} = \mathbf{I} = \mathbf{V}^{\mathrm{T}}\mathbf{W}_K\mathbf{V}. \tag{5.7}$$

The GSVD can be presented with the convenient “triplet notation” that integrates (1) data to be analyzed, (2) column constraints, and (3) row constraints as (for CA): $\mathrm{GSVD}(\mathbf{Z}_{\mathbf{R}}, \mathbf{W}_K, \mathbf{W}_J)$.

Here CA (through the GSVD) is closely linked to both χ^2 and the OLS through the “numerator” and “denominator” analogy: Eq. 5.6 is equivalent to the numerator of χ^2 whereas the weight matrices in Eq. 5.5 are equivalent to the denominator (see Chapters 3 and 4). CA is a multivariate χ^2 in that the deviations from independence matrix (Eq. 5.6) is the numerator and the observed value matrix is the denominator (Eq. 5.5). Furthermore, the numerator-denominator analogy suggests a regularization strategy for CA: regularize the weights in Eq. 5.5 because this forces the deviations (Eq. 5.6) into a larger subspace as if the sample size was larger.

This chapter is outlined as follows. First, I present an alternate formulation of MCA, followed by an established form of regularized MCA (Takane & Hwang, 2006). Next, I present the relationship between sparse and functional (“two-way”) PCA (SFPCA; Allen, 2013) and the GSVD. With both RMCA and SFPCA, I propose an approach to regularize PLS-CA called “smoothed” (terminology used in Allen, 2013) PLS-CA (SmooPLS-CA, pronounced “Smooples-

C-A”). Finally, I discuss the regularized χ^2 approach of Li et al., (2014) and suggest a slightly different—and more computationally efficient version of SmooPLS-CA.

5.1.1 Regularized Multiple Correspondence Analysis

Currently, there exists a particular form of regularized CA that respects the group-structured nature of the variables (i.e., disjunctive tables): regularized multiple correspondence analysis (MCA; Takane & Hwang, 2006). In this section, I recreate the formulation of Takane and Hwang (2006), but within the notation and framework of PLS-CA, as established in Chapter 4 (and in Beaton, Dunlop et al., 2016, and Beaton, Kriegsman, et al., 2016) in order to develop the RPLSCA framework. In multiple correspondence analysis (MCA) the data are first coded in a disjunctive format (see Table 5.1) and then analyzed with CA. The disjunctive form of a categorical table creates a block structure for the variables (columns) so that $\mathbf{X}$ can written as $\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{X}_2 & \dots & \mathbf{X}_d & \mathbf{X}_D \end{bmatrix}$ with each $\mathbf{X}_d$ being a disjunctive matrix, that represents the presence (i.e., “1”) or absence (i.e., “0”) of each level for each item (variable). MCA can also be reformulated differently from CA (as in Eqs. 5.3-5.7) as a version of centered, non-scaled, and weighted PCA—via the GSVD (see Secton 4.2)—of the matrix $\mathbf{X}$ as follows.

Call $\mathbf{Z}$ the centered version of $\mathbf{X}$. The MCA of $\mathbf{X}$ would be:

$$\mathbf{Z}_{\mathbf{X}}\mathbf{W}_J^{-1} = \mathbf{U}\Delta\mathbf{V}^{\mathrm{T}} \text{ with } \mathbf{U}^{\mathrm{T}}\mathbf{I}\mathbf{U} = \mathbf{I} = \mathbf{V}^{\mathrm{T}}\mathbf{W}_J\mathbf{V} \tag{5.8}$$

where $\mathbf{W}_J$ is a block diagonal matrix wherein $\mathbf{W}_{J,d} = \mathbf{X}_d^{\mathrm{T}}\mathbf{X}_d$; or more simply, a diagonal matrix of the column sums of $\mathbf{X}$. This particular formulation of MCA in triplet notation is: $\mathrm{GSVD}(\mathbf{Z}_{\mathbf{X}}\mathbf{W}_J^{-1}, \mathbf{W}_J, \mathbf{I})$.

Table 5.1

Nominal and disjunctive formats of SNP data.

	SNP1	SNP2
Nominal		
Subject 1	Aa	Aa
Subject 2	aa	Aa
Subject i	Aa	aa
Subject I	AA	AA

	SNP1			SNP2		
	AA	Aa	aa	AA	Aa	aa
Disjunctive						
Subject 1	0	1	0	0	1	0
Subject 2	0	0	1	0	1	0
Subject i	0	1	0	0	0	1
Subject I	1	0	0	1	0	0

Note. Example of nominal, and disjunctive coding of illustrative SNPs referred to as SNP 1 and SNP 2. *Here, both illustrative SNPs are presented generally where 'A' is the major allele and 'a' the minor allele. The major homozygote, heterozygote, and minor homozygote are denoted 'AA', 'Aa', and 'aa', respectively.*

Though the equations here are different from the previously established CA approaches in this dissertation, this formulation provides an identical result within a constant scaling factor. With this particular formulation of MCA, the component scores are simpler to compute:

$$\mathbf{F}_J = \mathbf{V}\Delta \tag{5.9}$$

and

$$\mathbf{F}_I = \mathbf{U}\Delta \,. \tag{5.10}$$

MCA maximizes the variance (called inertia) of component scores—see Eqs. 5.9-10:

$$\underset{\mathbf{f}_J}{\arg\max}\,\mathbf{f}_J^{\mathrm{T}}\mathbf{W}_J\mathbf{f}_J = \delta^2 = \underset{\mathbf{f}_I}{\arg\max}\,\mathbf{f}_I^{\mathrm{T}}\mathbf{I}\mathbf{f}_I, \tag{5.11}$$

under the following constraints applied to the left and right singular vectors:

$$\mathbf{v}_l^{\mathrm{T}}\mathbf{W}_J\mathbf{v}_l = 1 = \mathbf{u}_l^{\mathrm{T}}\mathbf{I}\mathbf{u}_l,\ \text{and}\ \mathbf{v}_l^{\mathrm{T}}\mathbf{W}_J\mathbf{v}_{l'} = 0 = \mathbf{u}_l^{\mathrm{T}}\mathbf{I}\mathbf{u}_{l'}, \tag{5.12}$$

where $l \neq l'$. Rewriting Eq. 5.11 and 5.12 indicated that MCA optimizes

$$\mathbf{F}_J^{\mathrm{T}}\mathbf{W}_J\mathbf{F}_J = \Delta\mathbf{V}^{\mathrm{T}}\mathbf{W}_J\mathbf{V}\Delta = \Delta\mathbf{I}\Delta = \Delta^2 = \Delta\mathbf{I}\Delta = \Delta\mathbf{U}^{\mathrm{T}}\mathbf{I}\mathbf{U}\Delta = \mathbf{F}_I^{\mathrm{T}}\mathbf{I}\mathbf{F}_I, \tag{5.13}$$

where Δ^2 is the diagonal matrix of the eigenvalues (variance; see also the SVD procedure (*cf.* Eq. 5.8).

How do we regularize MCA? Recall that $\lambda\mathbf{I}$ *inflates* the diagonal of a predictor subspace—a procedure equivalent to increasing the sample size. In MCA, the sample size is directly reflected in $\mathbf{W}_J$ (the sum of each $\mathbf{W}_{J,d}$ is the sample size).

In 2006, Takane and Hwang formally defined a regularization procedure for MCA ("regularized MCA"; RMCA). RMCA is defined in a fashion similar to regularized OLS, where, initially, a regularization of $\mathbf{W}_J$ is defined as:

$$\mathbf{W}(\lambda)_J = \mathbf{W}_J + \lambda\mathbf{P}_J \tag{5.14}$$

where λ is the regularization parameter, and $\mathbf{P}_J$ is a block-diagonal projection matrix. The block-diagonal matrix is used here because $\mathbf{X}$ has a block structure where within each $\mathbf{X}_d$, the presence of a variable for an individual is exclusive *within* the block, but not exclusive *across* the blocks. The matrix $\mathbf{P}_J$ is defined per block as $\mathbf{P}_{J,d} = \mathbf{X}_d^{\mathrm{T}}(\mathbf{X}_d\mathbf{X}_d^{\mathrm{T}})^{-1}\mathbf{X}_d$. However, this procedure can be simplified by using the centered matrix, $\mathbf{Z}$, where $\mathbf{P}_J = \mathbf{Z}_\mathbf{X}^\mathrm{T}(\mathbf{Z}_\mathbf{X}\mathbf{Z}_\mathbf{X}^\mathrm{T})^{-1}\mathbf{Z}_\mathbf{X}$, which is a standard projection matrix.

Because there is a defined regularization on the columns, there is also a required regularization on the rows. In this particular formulation of MCA (see Eq. 5.8 and its respective triplet), the identity matrix $\mathbf{I}$ tis the constraint applied to the rows. The regularized row constraints is formulated as:

$$\Xi(\lambda) = \mathbf{I} + \lambda(\mathbf{Z}_\mathbf{X}\mathbf{Z}_\mathbf{X}^\mathrm{T})^{+} \tag{5.15}$$

where $(\mathbf{X}\mathbf{X}^\mathrm{T})^{+}$ is the Moore-Penrose inverse of $\mathbf{X}\mathbf{X}^\mathrm{T}$. This regularization scheme corresponds to the analysis of the triplet: $\mathrm{GSVD}(\mathbf{Z}\mathbf{W}_J^{-1}, \mathbf{W}(\lambda)_J, \Xi(\lambda))$. The optimization of RMCA is similar to MCA; namely, RMCA maximizes

$$\arg\max_{\mathbf{f}_J} \mathbf{f}_J^\mathrm{T}\mathbf{W}(\lambda)_J\mathbf{f}_J = \delta^2 = \arg\max_{\mathbf{f}_I} \mathbf{f}_I^\mathrm{T}\Xi(\lambda)\mathbf{f}_I, \tag{5.16}$$

under the constraints that the component are normalized and orthogonal to each other:

$$\mathbf{v}_l^\mathrm{T}\mathbf{W}(\lambda)_J\mathbf{v}_l = 1 = \mathbf{u}_l^\mathrm{T}\Xi(\lambda)\mathbf{u}_l, \text{ and } \mathbf{v}_l^\mathrm{T}\mathbf{W}(\lambda)_J\mathbf{v}_{l'} = 0 = \mathbf{u}_l^\mathrm{T}\Xi(\lambda)\mathbf{u}_{l'}, \tag{5.17}$$

where $l \neq l'$. Equivalently RMCA optimizes

$$\mathbf{F}_J^\mathrm{T}\mathbf{W}(\lambda)_J\mathbf{F}_J = \Delta\mathbf{V}^\mathrm{T}\mathbf{W}(\lambda)_J\mathbf{V}\Delta = \Delta\mathbf{I}\Delta = \Delta^2 = \Delta\mathbf{I}\Delta = \Delta\mathbf{U}^\mathrm{T}\Xi(\lambda)\mathbf{U}\Delta = \mathbf{F}_I^\mathrm{T}\Xi(\lambda)\mathbf{F}_I. \tag{5.18}$$

Finally, Takane and Hwang (2006) propose that the regularization procedure can be further generalized (to incorporate, e.g., “degrees smoothness”) to allow for any block-diagonal matrix $\mathbf{B}$, with the same block structure as $\mathbf{X}$, where Eqs. 5.14 and 5.15 would be rewritten as:

$$\mathbf{W}(\lambda)_J = \mathbf{W}_J + \lambda\mathbf{B} \tag{5.19}$$

and

$$\Xi(\lambda) = \mathbf{I} + \lambda(\mathbf{Z}_{\mathbf{X}}\mathbf{B}^{+}\mathbf{Z}_{\mathbf{X}}^{\mathrm{T}})^{+} . \tag{5.20}$$

When λ is 0, we have standard MCA. The effect of an increased λ can be seen in Figure 5.1; the first effects of note are that, as λ increases, the component scores for column and row scores smooth and begin to shrink toward 0 (Fig. 5.1a-b). Furthermore, the inertia (total variance) shrinks (Fig. 5.1c), and in most (but not all) cases the singular values shrink as well (Fig. 5.1d), where eventually results approach sphericity (i.e., approximately equal singular values).

5.1.2 Sparse Functional Principal Components Analysis

Currently, there exists a technique for PCA that allows for regularization—in the form of sparsification—and "smoothing" applied to both the rows and columns of a matrix. This technique is called sparse functional (a.k.a., "two-way") PCA (Allen, 2013). SFPCA is a broadly defined method that encompasses a variety of regularization and smoothness approaches. Let us assume a Rank 1 problem (i.e., a single source of variance via e.g., PCA), where we have only one pair of (left and right) singular vectors and one singular value. Given a matrix, $\mathbf{X}$ (whose columns are centered and normalized), SFPCA maximizes:

$$\underset{\mathbf{u},\mathbf{v}}{\arg\max}\ \mathbf{u}^{\mathrm{T}}\mathbf{X}\mathbf{v} - \alpha_{\mathbf{u}}P_{\mathbf{u}}(\mathbf{u}) - \alpha_{\mathbf{v}}P_{\mathbf{v}}(\mathbf{v}), \tag{5.21}$$

under the constraints that:

$$\mathbf{v}^{\mathrm{T}}(\mathbf{I} + \lambda_{J}\mathbf{W}_{J})\mathbf{v} \leq 1 \text{ and } \mathbf{u}^{\mathrm{T}}(\mathbf{I} + \lambda_{I}\mathbf{W}_{I})\mathbf{u} \leq 1. \tag{5.22}$$

The sparsity, regularization, and smoothness parameters of SFPCA are as follows. First, penalties (sparsity) applied to the singular vectors are defined by their respective $P_{*}(*)$, and the regularization of the penalization is controlled by α_* . Next, smoothness of the structure is defined by respective their $\mathbf{W}_*$, and regularization of smoothness is controlled by λ_* . Allen (2013) suggests that $\mathbf{W}_*$ could be "second or fourth differences" matrices to control for the *L*2 smoothness, but other smoothness constraints could be used. The inequality in the smoothness constraints (Eq. 5.22) does not require orthogonality (as per usual with the SVD) in Rank > 1 problems because the penalization procedure corresponds to a non-linear problem.

For the purposes of my proposed method—which is a regularization scheme *a là* Takane and Hwang (2006), for PLSCA—we can ignore part of Allen (2013)'s proposed approach; specifically, we are not concerned with the part of Eq. 5.21 that applies convex penalties and regularization of those penalties. Therefore, Eq. 5.21 can be rewritten (for our purposes) simply as $\underset{\mathbf{u},\mathbf{v}}{\arg\max}\ \mathbf{u}^{\mathrm{T}}\mathbf{X}\mathbf{v}$; a maximization that we can achieve through the plain singular value decompositions. Let us also define $\mathbf{W}_*(\lambda_*) = \mathbf{I} + \lambda_* \mathbf{W}_*$, just as in Eq. 5.14 and 5.15 and reminiscent of the ridge OLS in Eq. 5.2. Furthermore, if we require orthogonality and no longer assume a Rank 1 problem, Eq. 5.22 changes specifically to:

$$\mathbf{v}_l^{\mathrm{T}}\mathbf{W}(\lambda_J)_J\mathbf{v}_l = 1 \text{ and } \mathbf{u}_l^{\mathrm{T}}\mathbf{W}(\lambda_I)_I\mathbf{u}_l = 1 \tag{5.23}$$

where, no longer under the assumption of a Rank 1 problem where $l \neq l'$,

$$\mathbf{v}_l^{\mathrm{T}}\mathbf{W}(\lambda_J)_J\mathbf{v}_{l'} = 0 \text{ and } \mathbf{u}_l^{\mathrm{T}}\mathbf{W}(\lambda_I)_I\mathbf{u}_{l'} = 0. \tag{5.24}$$

Therefore, we can present a regularized "*L*2-ball smoothed two-way" (Eqs. 5.23–24) PCA with the following triplet notation: $\mathrm{GSVD}(\mathbf{X}, \mathbf{W}(\lambda_J)_J, \mathbf{W}(\lambda_I)_I)$. Thus, we can think of RMCA as an "L2-ball smoothed MCA", where the constraints are defined under a χ^2 metric. There is a final feature to present that, while not discussed by Allen (2013), can be incorporated into "L_2-ball smoothed two-way" by way of RMCA. Takane and Hwang (2006)'s RMCA provides an approach to handle group or structured regularization; that is when distinct items (columns or rows) of a matrix should be grouped together during the regularization process. This is a requirement in MCA and RMCA by Takane and Hwang (2006), as multiple columns belong to a single variable (see Table 5.1). However, the same principle could be applied to any type of data—for PCA, MCA, or any related technique—where some *a priori* structure exists for the rows and/or columns. This broader group-regularization procedure can be performed with a block-diagonal matrix, as in Eq. 5.19 and incorporated into the smoothing matrices.

5.2 How to regularize PLS-CA

Now that we have established two regularized approaches to single data tables—Takane and Hwang (2006)'s RMCA and Allen (2013)'s SFPCA—we can now extend these ideas to the analysis of two tables (e.g., PLS) with a particular focus on categorical and mixed data types via PLS-CA (see Chapter 4).

5.2.1 PLSC and PLS-CA

Let us revisit the formulations of PLSC and PLSCA. Recall that PLSC maximizes the common information between two data tables. The maximization criteria equations are presented in Chapter 4, section 3.1, though they are copied here for convenience. Maximization is the covariance between two (normalized) data tables, $\mathbf{Z_X}$ and $\mathbf{Z_Y}$:

$$\mathbf{Z_R} = \mathbf{Z_X^T Z_Y} = \mathbf{U \Delta V^T}, \qquad (5.25)$$

where the latent variables are computed as:

$$\mathbf{L}_{\mathbf{X}} = \mathbf{Z}_{\mathbf{X}}\mathbf{U} \text{ and } \mathbf{L}_{\mathbf{Y}} = \mathbf{Z}_{\mathbf{Y}}\mathbf{V}. \tag{5.26}$$

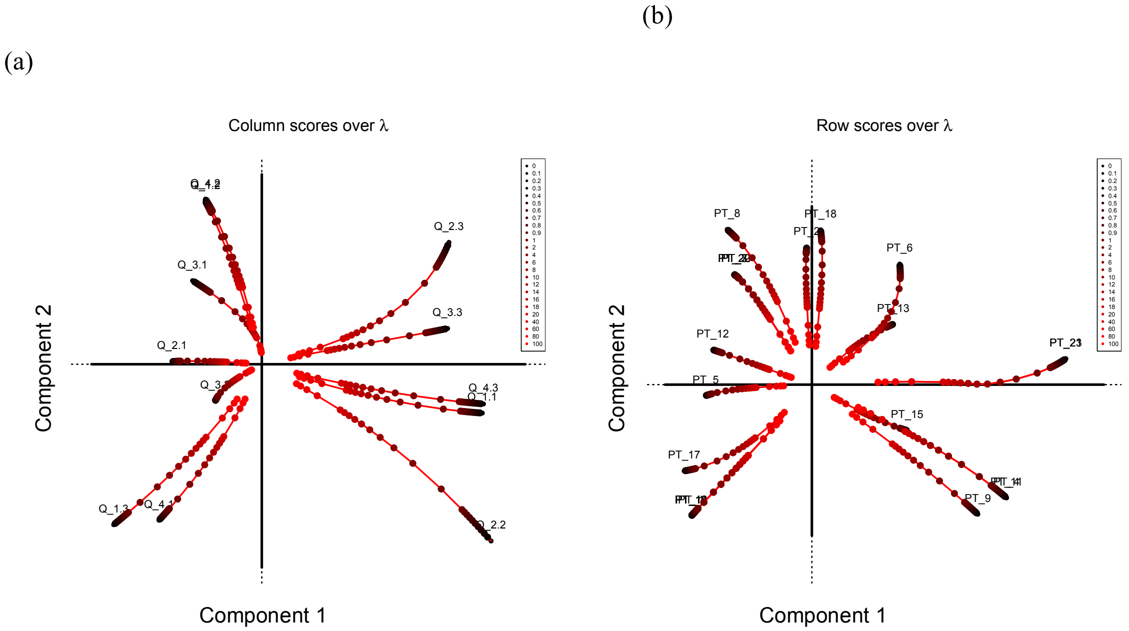


Figure 5.1 Takane and Hwang (2006)'s Regularized Multiple Correspondence Analysis, with the Nishisato data as provided in their paper, with regularization effects from λ = 0 (black) to λ = 100 (red). (a) and (b) show the effects of increased λ on the row and column component scores for Components 1 (horizontal) and 2 (vertical), respectively. The component scores are shown within the constraints of λ = 0 (black dots). Relatively, as λ increases, component scores smooth and approach zero. This effect can be further seen in (c) and (d).

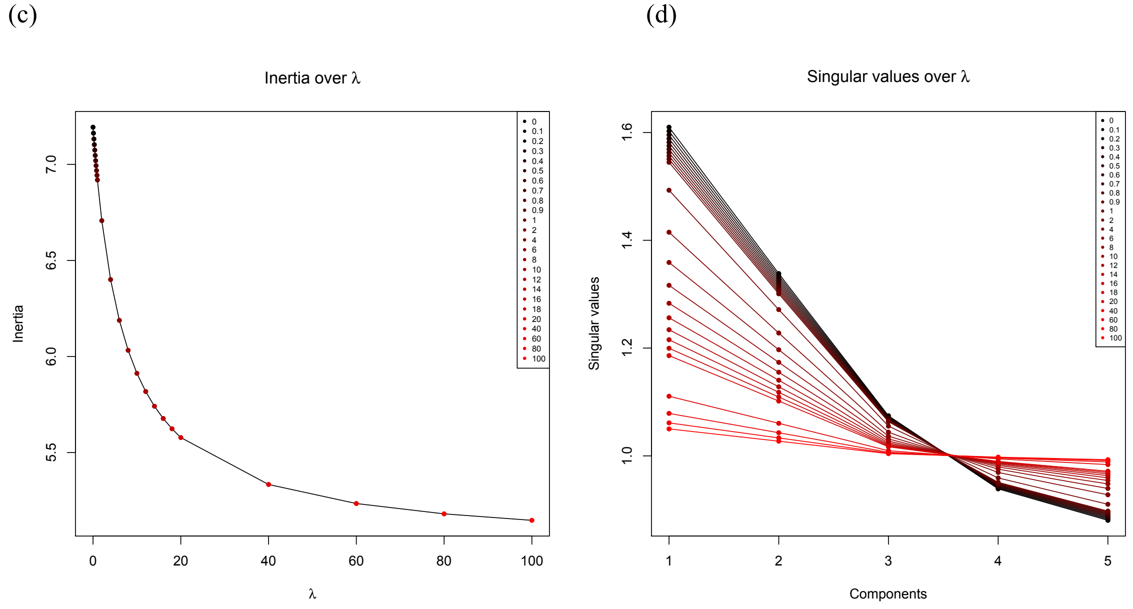


Figure 5.1 cont'd Takane and Hwang (2006)'s Regularized Multiple Correspondence Analysis, with the Nishisato data as provided in their paper, with regularization effects from λ = 0 (black) to λ = 100 (red). In (c), the effects of λ on the total variance (a.k.a., "inertia") of the analyses are shown. As λ increases, total variance decreases. Finally, (d) the singular values (for the first 5 components) show that, generally, the first sources of variance (e.g., the first and second component) decrease rapidly in explained variance. As λ becomes large, the space becomes spherical (i.e., components explain approximately equal variance, as seen by the flattening of the scree plot).

The goal of PLSC is to maximize the relationship between these latent variables,

$$\arg\max \operatorname{cov}(\mathbf{l}_{\mathbf{X}}^{\mathrm{T}}\mathbf{l}_{\mathbf{Y}}), \tag{5.27}$$

under the constraints of unit norm per singular vector,

$$\mathbf{u}_{l}^{\mathrm{T}}\mathbf{u}_{l} = 1 = \mathbf{v}_{l}^{\mathrm{T}}\mathbf{v}_{l}, \tag{5.28}$$

and the additional orthogonality constraint between pairs of latent variables (and singular vectors)

$$\mathbf{l}_{\mathbf{X},l}^{\mathrm{T}}\mathbf{l}_{\mathbf{Y},l'} = 0 \text{ when } l \neq l', \tag{5.29}$$

so that the pairs of latent variables provide maximal variance, conditional to the orthogonality to the subsequent pairs (i.e., via the SVD):

$$\mathbf{L}_{\mathbf{X}}^{\mathrm{T}}\mathbf{L}_{\mathbf{Y}} = \mathbf{U}^{\mathrm{T}}\mathbf{Z}_{\mathbf{X}}^{\mathrm{T}}\mathbf{Z}_{\mathbf{Y}}\mathbf{V} = \mathbf{U}^{\mathrm{T}}\mathbf{Z}_{\mathbf{R}}\mathbf{V} = \mathbf{U}^{\mathrm{T}}\mathbf{U}\boldsymbol{\Delta}\mathbf{V}^{\mathrm{T}}\mathbf{V} = \boldsymbol{\Delta}. \tag{5.30}$$

When $l = 1$, the covariance between $\mathbf{L}_{\mathbf{X}}$ and $\mathbf{L}_{\mathbf{Y}}$ has the largest possible value, when $l = 2$, the covariance between $\mathbf{L}_{\mathbf{X}}$ and $\mathbf{L}_{\mathbf{Y}}$ has the largest possible value under the constraints that the second pair of latent variables are orthogonal (as defined by Eq. 5.29) to the first pair of latent variables and so on for higher order latent variables.

While standard, mixed-data, and the numerous variants of PLSCA have been formally defined in Chapter 4 (see also, Beaton et al., 2013, Beaton, Dunlop, et al., 2016, and Beaton, Kriegsman, et al., 2016), we need to define an alternative formulation of PLSCA that closely resembles RMCA, in order to extend PLSCA to Smoothed PLS-CA ("SmooPLS-CA"). We start in a fashion similar to Takane and Hwang (2006), with disjunctive matrices, $\mathbf{X}$ and $\mathbf{Y}$, that have been centered $\mathbf{Z}_{\mathbf{X}}$ and $\mathbf{Z}_{\mathbf{Y}}$. Next, we compute the column sums of $\mathbf{X}$ and $\mathbf{Y}$, and then put these column sums into two diagonal matrices denoted (respectively) $\mathbf{W}_{\mathbf{X}}$ and $\mathbf{W}_{\mathbf{Y}}$. Next, we compute $\mathbf{Z}_{\mathbf{R}} = \mathbf{Z}_{\mathbf{X}}^{\mathrm{T}}\mathbf{Z}_{\mathbf{Y}}$, where we have PLS-CA as the analysis of the triplet: $\mathrm{GSVD}(\mathbf{W}_{\mathbf{X}}^{-1}\mathbf{Z}_{\mathbf{R}}\mathbf{W}_{\mathbf{Y}}^{-1}, \mathbf{W}_{\mathbf{Y}}, \mathbf{W}_{\mathbf{X}})$. Like RMCA when compared to MCA, this formulation of PLSCA differs from PLSCA as presented in Chapter 4 by constant scaling factors. Here, the latent variables are computed as:

$$\mathbf{L}_{\mathbf{X}} = \mathbf{Z}_{\mathbf{X}}\mathbf{U} \text{ and } \mathbf{L}_{\mathbf{Y}} = \mathbf{Z}_{\mathbf{Y}}\mathbf{V}. \tag{5.31}$$

The PLSC-style maximization here is the same as in both PLSC and PLSCA as presented in this chapter, and in Chapter 4, but aligns more closely with the reformulation based on RMCA (to match the more generalized maximization of SFPCA). First, we want to maximize the covariance between the latent variables, a problem stated as $\arg\max \operatorname{cov}(\mathbf{l}_{\mathbf{X}}^{\mathrm{T}}\mathbf{l}_{\mathbf{Y}})$, with the constraints that $\mathbf{l}_{\mathbf{X}}^{\mathrm{T}}\mathbf{l}_{\mathbf{Y}} = \mathbf{u}^{\mathrm{T}}\mathbf{Z}_{\mathbf{X}}^{\mathrm{T}}\mathbf{Z}_{\mathbf{Y}}\mathbf{v}$, and thus this problem is equivalent to (as SFPCA) solving: $\underset{\mathbf{u},\mathbf{v}}{\arg\max}\ \mathbf{u}^{\mathrm{T}}\mathbf{Z}_{\mathbf{R}}\mathbf{v}$. For PLS-CA, we add the orthogonality constraints expressed by, $\mathbf{u}_{l}^{\mathrm{T}}\mathbf{W}_{\mathbf{X}}\mathbf{u}_{l} = 1 = \mathbf{v}_{l}^{\mathrm{T}}\mathbf{W}_{\mathbf{Y}}\mathbf{v}_{l}$ and $\mathbf{u}_{l}^{\mathrm{T}}\mathbf{W}_{\mathbf{X}}\mathbf{u}_{l'} = 0 = \mathbf{v}_{l}^{\mathrm{T}}\mathbf{W}_{\mathbf{Y}}\mathbf{v}_{l'}$, where $l \neq l'$. Thus, this last problem is equivalent to

$$\mathbf{L}_{\mathbf{X}}^{\mathrm{T}}\mathbf{L}_{\mathbf{Y}} = \mathbf{U}^{\mathrm{T}}\mathbf{Z}_{\mathbf{X}}^{\mathrm{T}}\mathbf{Z}_{\mathbf{Y}}\mathbf{V} = \mathbf{U}^{\mathrm{T}}\mathbf{Z}_{\mathbf{R}}\mathbf{V} = \Delta, \tag{5.32}$$

where, per usual with the GSVD, the orthogonality are

$$\mathbf{U}^{\mathrm{T}}\mathbf{W}_{\mathbf{X}}\mathbf{U} = \mathbf{I} = \mathbf{V}^{\mathrm{T}}\mathbf{W}_{\mathbf{Y}}\mathbf{V}. \tag{5.33}$$

An alternative maximization can be presented as a CA or an MCA variance maximization problem, similar to Eqs. 5.11-5.13:

$$\mathbf{F}_{J}^{\mathrm{T}}\mathbf{W}_{\mathbf{X}}\mathbf{F}_{J} = \Delta^{2} = \mathbf{F}_{K}^{\mathrm{T}}\mathbf{W}_{\mathbf{Y}}\mathbf{F}_{K}. \tag{5.34}$$

5.2.2 Smoothed PLS-CA ("SmooPLS-CA")

The goal of SmooPLS-CA is to extend PLS-CA into a two-way regularization via smoothness—like Allen (2013)—but for categorical data (though, PLS-CA can easily accommodate mixed data types)—like Takane and Hwang (2006). Therefore, to adhere to the principles of PLS-CA and RMCA, the maximization problem is $\arg\max \operatorname{cov}(\mathbf{l}_{\mathbf{X}}^{\mathrm{T}}\mathbf{l}_{\mathbf{Y}})$, under the (orthogonality) constraints that $\mathbf{L}_{\mathbf{X}}^{\mathrm{T}}\mathbf{L}_{\mathbf{Y}} = \Delta$, and also with the constraints imposed by the structured (i.e., disjunctive) nature of the data: $\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{X}_2 & \dots & \mathbf{X}_d & \mathbf{X}_D \end{bmatrix}$. First define the regularized (column-wise) weight matrices for (respectively) $\mathbf{Z}_{\mathbf{X}}$ and $\mathbf{Z}_{\mathbf{Y}}$:

$$\mathbf{W}(\lambda_{\mathbf{X}})_{\mathbf{X}} = \mathbf{W}_{\mathbf{X}} + \lambda_{\mathbf{X}}\mathbf{P}_{\mathbf{X}} \text{ and } \mathbf{W}(\lambda_{\mathbf{Y}})_{\mathbf{Y}} = \mathbf{W}_{\mathbf{Y}} + \lambda_{\mathbf{Y}}\mathbf{P}_{\mathbf{Y}}, \tag{5.35}$$

where $\mathbf{P}_{\mathbf{X}} = \mathbf{Z}_{\mathbf{X}}^{\mathrm{T}}(\mathbf{Z}_{\mathbf{X}}\mathbf{Z}_{\mathbf{X}}^{\mathrm{T}})^{-1}\mathbf{Z}_{\mathbf{X}}$, $\mathbf{P}_{\mathbf{Y}} = \mathbf{Z}_{\mathbf{Y}}^{\mathrm{T}}(\mathbf{Z}_{\mathbf{Y}}\mathbf{Z}_{\mathbf{Y}}^{\mathrm{T}})^{-1}\mathbf{Z}_{\mathbf{Y}}$, and where $\mathbf{W}_{*}$ are diagonal matrices with the column sums of their respective disjunctive matrix. Next, the row regularization matrices for $\mathbf{Z}_{\mathbf{X}}$ and $\mathbf{Z}_{\mathbf{Y}}$, are (respectively) defined as:

$$\Xi(\lambda_{\mathbf{X}})_{\mathbf{X}} = \mathbf{I} + \lambda_{\mathbf{X}}(\mathbf{Z}_{\mathbf{X}}\mathbf{Z}_{\mathbf{X}}^{\mathrm{T}})^{+} \text{ and } \Xi(\lambda_{\mathbf{Y}})_{\mathbf{Y}} = \mathbf{I} + \lambda_{\mathbf{Y}}(\mathbf{Z}_{\mathbf{Y}}\mathbf{Z}_{\mathbf{Y}}^{\mathrm{T}})^{+}. \tag{5.36}$$

Recall that the goal of regularization—and in the context of "smoothing"—we want to bring values closer to 0, while minimizing the noise, and thus the variance (see the Figure 5.1 for RMCA as a baseline of how regularization should work). The same goals with PLS-CA is obtained following Eq. 5.36, that describes row (i.e., observation) regularized versions of $\mathbf{Z}_{\mathbf{X}}$ and $\mathbf{Z}_{\mathbf{Y}}$:

$$\Omega_{\mathbf{X}} = \Xi(\lambda_{\mathbf{X}})_{\mathbf{X}}^{1/2}\mathbf{Z}_{\mathbf{X}} \text{ and } \Omega_{\mathbf{Y}} = \Xi(\lambda_{\mathbf{Y}})_{\mathbf{Y}}^{1/2}\mathbf{Z}_{\mathbf{Y}}. \tag{5.37}$$

When $\lambda_{*} = 0$, $\mathbf{W}(\lambda_{*})_{*} = \mathbf{W}_{*}$, $\Xi(\lambda_{*})_{*} = \mathbf{I}$, and $\Omega_{*} = \mathbf{Z}_{*}$, which would be the usual form of PLS-CA. Next, the cross-product matrix is computed as:

$$\Omega_{\mathbf{R}} = \Omega_{\mathbf{X}}^{\mathrm{T}}\Omega_{\mathbf{Y}} \tag{5.38}$$

and so, SmooPLS-CA is equivalent to the analysis of the triplet

$\mathrm{GSVD}(\mathbf{W}(\lambda_{\mathbf{X}})_{\mathbf{X}}^{-1}\Omega_{\mathbf{R}}\mathbf{W}(\lambda_{\mathbf{Y}})_{\mathbf{Y}}^{-1},\mathbf{W}(\lambda_{\mathbf{Y}})_{\mathbf{Y}},\mathbf{W}(\lambda_{\mathbf{X}})_{\mathbf{X}})$.

What does SmooPLS-CA maximize?

Like plain PLS-CA, SmooPLS-CA maximizes the covariance between latent variables. In SmooPLS-CA, the latent variables are computed as in PLS-CA:

$$\mathbf{L}_{\mathbf{X}} = \Omega_{\mathbf{X}}\mathbf{U} \text{ and } \mathbf{L}_{\mathbf{Y}} = \Omega_{\mathbf{Y}}\mathbf{V}. \tag{5.39}$$

and so the maximization problem corresponds to

$\arg\max\operatorname{cov}(\mathbf{l}_{\mathbf{X}}^{\mathrm{T}}\mathbf{l}_{\mathbf{Y}}) = \arg\max\limits_{\mathbf{u},\mathbf{v}}\ \mathbf{u}^{\mathrm{T}}\Omega_{\mathbf{X}}^{\mathrm{T}}\Omega_{\mathbf{Y}}\mathbf{v} = \arg\max\limits_{\mathbf{u},\mathbf{v}}\ \mathbf{u}^{\mathrm{T}}\Omega_{\mathbf{R}}\mathbf{v}$.

under the (orthogonality) constrains imposed on the columns: $\mathbf{u}_{l}^{\mathrm{T}}\mathbf{W}(\lambda_{\mathbf{X}})_{\mathbf{X}}\mathbf{u}_{l} = 1 = \mathbf{v}_{l}^{\mathrm{T}}\mathbf{W}(\lambda_{\mathbf{Y}})_{\mathbf{Y}}\mathbf{v}_{l}$

and $\mathbf{u}_{l}^{\mathrm{T}}\mathbf{W}(\lambda_{\mathbf{X}})_{\mathbf{X}}\mathbf{u}_{l'} = 0 = \mathbf{v}_{l}^{\mathrm{T}}\mathbf{W}(\lambda_{\mathbf{Y}})_{\mathbf{Y}}\mathbf{v}_{l'}$, where $l \neq l'$.

Thus

$$\mathbf{L}_{\mathbf{X}}^{\mathrm{T}}\mathbf{L}_{\mathbf{Y}} = \mathbf{U}^{\mathrm{T}}\Omega_{\mathbf{X}}^{\mathrm{T}}\Omega_{\mathbf{Y}}\mathbf{V} = \mathbf{U}^{\mathrm{T}}\Omega_{\mathbf{R}}\mathbf{V} = \Delta, \tag{5.40}$$

with

$$\mathbf{U}^{\mathrm{T}}\mathbf{W}(\lambda_{\mathbf{X}})_{\mathbf{X}}\mathbf{U} = \mathbf{I} = \mathbf{V}^{\mathrm{T}}\mathbf{W}(\lambda_{\mathbf{Y}})_{\mathbf{Y}}\mathbf{V}. \tag{5.41}$$

Finally, just as in PLS-CA, an alternative maximization can be presented as either a smoothed (regularized) CA or MCA problem, where the eigenvalues give the inertia of the component scores:

$$\mathbf{F}_{J}^{\mathrm{T}}\mathbf{W}(\lambda_{\mathbf{X}})_{\mathbf{X}}\mathbf{F}_{J} = \Delta^{2} = \mathbf{F}_{K}^{\mathrm{T}}\mathbf{W}(\lambda_{\mathbf{Y}})_{\mathbf{Y}}\mathbf{F}_{K}. \tag{5.42}$$

5.3 Examples of SmooPLS-CA

In this section, I illustrate SmooPLS-CA with data from the Alzheimer's Disease Neuroimaging Initiative (ADNI). This provides a basis for comparing SmooPLS-CA and PLS-CA (defined in Beaton, Dunlop, et al. 2016), which used the same ADNI data. Following the illustration of how SmooPLS-CA works, I discuss the choices of λ_* as applied to both sides of the analysis (selection of optimal λ_* is in the following section).

5.3.1 SmooPLS-CA in action

This example illustrates the behavior(s) of SmooPLS-CA as λ changes, with reference to $\lambda = 0$ (which is the plain form of PLS-CA). This base example includes only the form of $\lambda_{\mathbf{X}} = \lambda_{\mathbf{Y}} = \lambda$; the single λ -value approach is the recommended baseline form of SmooPLS-CA in part because the ability to find two optimal λ 's is not only computationally expensive, but may not be tractable. Therefore, the recommended baseline form of SmooPLS-CA makes the following changes to Eqs. 5.35-5.37:

$$\mathbf{W}(\lambda)_{\mathbf{X}} = \mathbf{W}_{\mathbf{X}} + \lambda\mathbf{P}_{\mathbf{X}} \text{ and } \mathbf{W}(\lambda)_{\mathbf{Y}} = \mathbf{W}_{\mathbf{Y}} + \lambda\mathbf{P}_{\mathbf{Y}}, \tag{5.43}$$

$$\Xi(\lambda)_{\mathbf{X}} = \mathbf{I} + \lambda(\mathbf{Z}_{\mathbf{X}}\mathbf{Z}_{\mathbf{X}}^{\mathrm{T}})^{+} \text{ and } \Xi(\lambda)_{\mathbf{Y}} = \mathbf{I} + \lambda(\mathbf{Z}_{\mathbf{Y}}\mathbf{Z}_{\mathbf{Y}}^{\mathrm{T}})^{+}, \text{ and} \tag{5.44}$$

$$\Omega_{\mathbf{X}} = \Xi(\lambda)_{\mathbf{X}}^{1/2}\mathbf{Z}_{\mathbf{X}} \text{ and } \Omega_{\mathbf{Y}} = \Xi(\lambda)_{\mathbf{Y}}^{1/2}\mathbf{Z}_{\mathbf{Y}}. \tag{5.45}$$

Here, I present two examples of SmooPLS-CA with the ADNI data used previously in Beaton, Dunlop, et al., (2016): 1) standard SmooPLS-CA (behavioral data + genetic data) and 2) discriminant SmooPLS-CA (group data + genetic data). Following these two illustrations, I provide an example of what happens when we only want to regularize *one* data set (e.g., genetics), when the non-regularized dataset has some sort of *a priori* structure that we do not want to shrink or alter. This subsection and its illustrations serve as a mechanistic description of SmooPLS-CA; and illustrate how to decide on the values of λ, which items are significant, which effects are reliable (how to interpret the results will be discussed in section 5.4 and onward).

5.3.1.1 Standard SmooPLS-CA

This example uses PLS-CA to analyze the relationships between behavioral data (the mini-mental state exam, clinical dementia rating, and geriatric depression scales) and genetic data (SNPs recoded as their genotypes) in order to identify which genotypes are most associated with particular types of behaviors. In PLS-CA, very rare items can influence the structure of the data by stretching the components (i.e., increasing the variance). While rare items can be useful (especially in genetics), they may also bias the overall structure and to regularization can palliate this problem. Figure 5.2 illustrates the behavior of SmooPLS-CA when the regularization parameter varies as: $\lambda_{\mathbf{X}} = \lambda_{\mathbf{Y}} = \lambda = [0, 1, 2, 3, 4, 5, 10, 15, 20, 25, 50, 75, 100]$, where $\lambda = 0$ is equivalent to plain PLS-CA, and as λ increases, (1) variance in the data decreases, and (2) smoothing becomes more evident as items move toward zero. While an increase of λ will smooth the structure, and brings far away (i.e., rare) items closer to zero, *stable* items are less susceptible to this effect.

The effects of λ on the component scores can be seen in Figures 5.2a (rows, behavior) and 5.2b (columns, genotypes). As λ increases, the component scores of the items approach zero. In the behavioral data, there are some rare and *unstable* items that rapidly approach zero (top left in Figure 5.2a), whereas frequent, but stable, items barely move (upper right and lower left quadrants, near origin in Figure 5.2a). The same can be also be seen in the genetics data (Figure 5.2b)—the data of interest in this example. We want to smooth out instability in the genetics data in order to find the most reliable markers associated with behaviors (reliability is discussed in more detail in Section 5.4). The effects of regularization can also be observed with respect to the reduction of the overall variance as well the variance per component and both decrease as λ increases. Though in this example, as seen in Figure 5.2d, the proportion of variance per component retains its shape (unlike in RMCA, see Figure 5.1). Finally, the effect of increased λ can be seen in the latent variables, where the fit lines (Figures 5.2e and f), and correlation between latent variables (Table 5.2) increases as λ increases, thus showing more estimated similarity between latent variables.

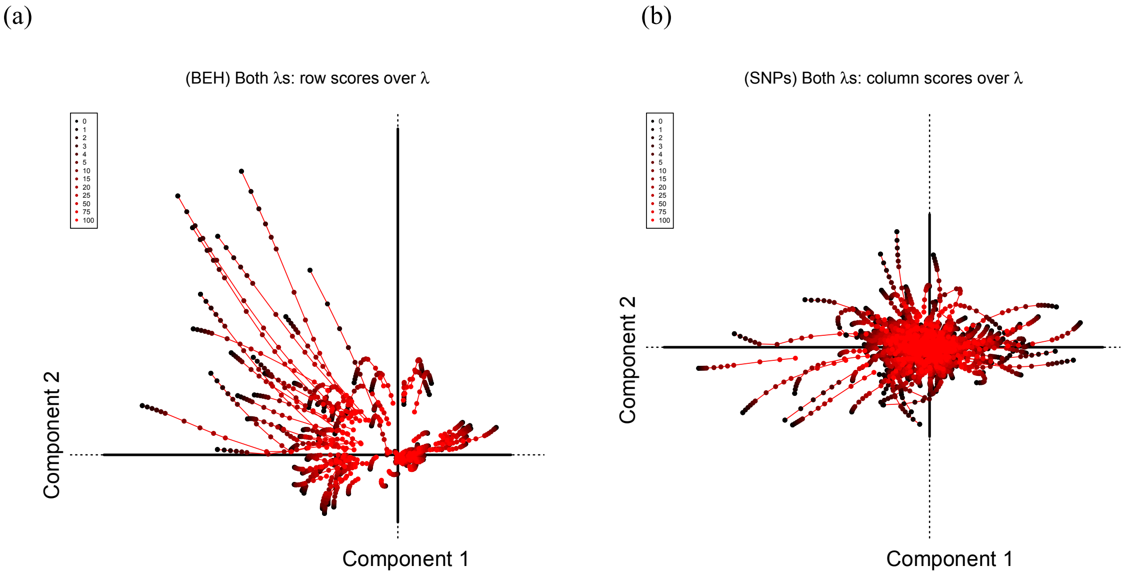


Figure 5.2 SmooPLS-CA data from Beaton et al., (2016) with λ = 0-100 (black to red). (a; b) show the effects of increased λ on the behavioral and genetic component scores for Components 1 (horizontal) and 2 (vertical). Component scores are shown as λ = 0 (black dots). As λ increases, component scores smooth and approach zero. This effect can be further seen in (c) and (d).

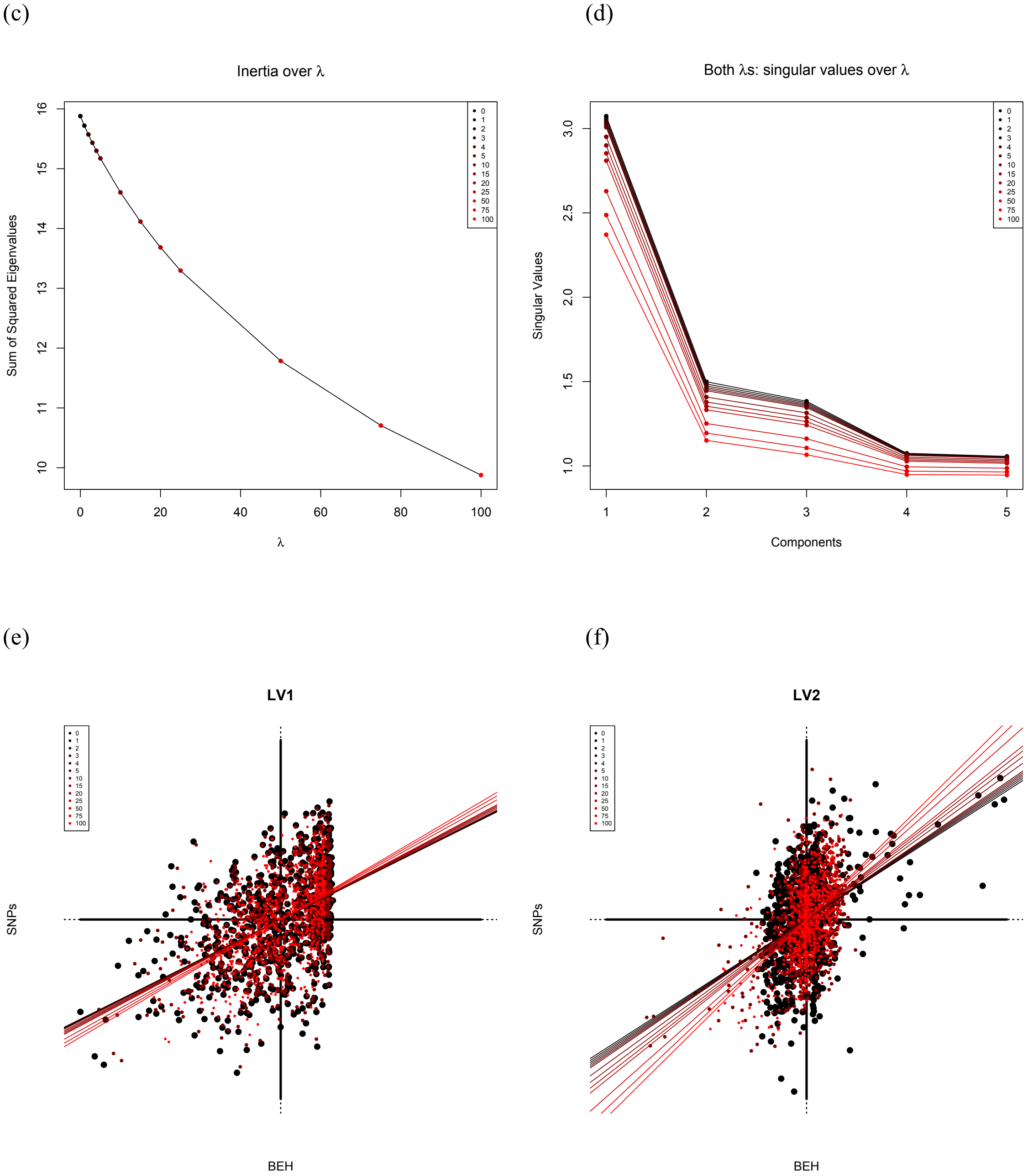


Figure 5.2 cont'd Smoothed Partial Least Squares-Correspondence Analysis (SmooPLS-CA) with the Behavior (BEH) + Genetic (SNPs) ADNI subset data from Beaton et al., (2016), with regularization effects from λ = 0 (black) to λ = 100 (red). In (c), the effects of λ on the total variance (a.k.a., "inertia") of the analyses are shown. As λ increases, total variance decreases. Finally, (d) the singular values (for the first 5 components) show a rapid decrease. Though,

showing a different pattern than in RMCA, as λ becomes large, the components retain their shape, as opposed to becoming spherical. In both (e) and (f), the latent variables are shown (respectively) for LV1 and LV2. The lines show fit lines and as λ increases, so does the slope, a configuration which illustrates that increased λ makes for stronger relationships between the latent variables. Correlation values between LVs are shown in Table 5.2.

Table 5.2

Latent variable correlations for increased λ.

	BEH+GENETICS		**GRP+GENETICS**	
λ	***LV1***	***LV2***	***LV1***	***LV2***
0	0.4644	0.3974	0.4688	0.3828
1	0.4660	0.3999	0.4700	0.3854
2	0.4675	0.4024	0.4711	0.3878
3	0.4688	0.4049	0.4722	0.3901
4	0.4701	0.4073	0.4733	0.3923
5	0.4714	0.4096	0.4743	0.3944
10	0.4771	0.4200	0.4788	0.4040
15	0.4821	0.4287	0.4829	0.4126
20	0.4866	0.4362	0.4866	0.4203
25	0.4908	0.4427	0.4901	0.4275
50	0.5085	0.4673	0.5053	0.4569
75	0.5229	0.4855	0.5180	0.4795
100	0.5354	0.5005	0.5291	0.4977

Note. Correlation values between latent variables (from each data set) for both the behavioral + genetic data analysis (see Fig. 5.2) and group + genetic data analysis (see Fig. 5.3). As λ increases, the correlations between the LVs become stronger.

5.3.1.2 Discriminant SmooPLS-CA

Here PLS-CA is used to analyze the relationship between group association and genetic data (SNPs recoded as their genotypes) in order to identify the genotypes most associated with each group (i.e., a discriminant analysis; classification). In this example, there are three groups (control, mild cognitive impairment, and Alzheimer's disease); because there are three groups there are only two components. Again, we want to smooth out possible instability in the data, this time with a particular emphasis on the genetic data, because group association is consider a stable—and often static—variable.

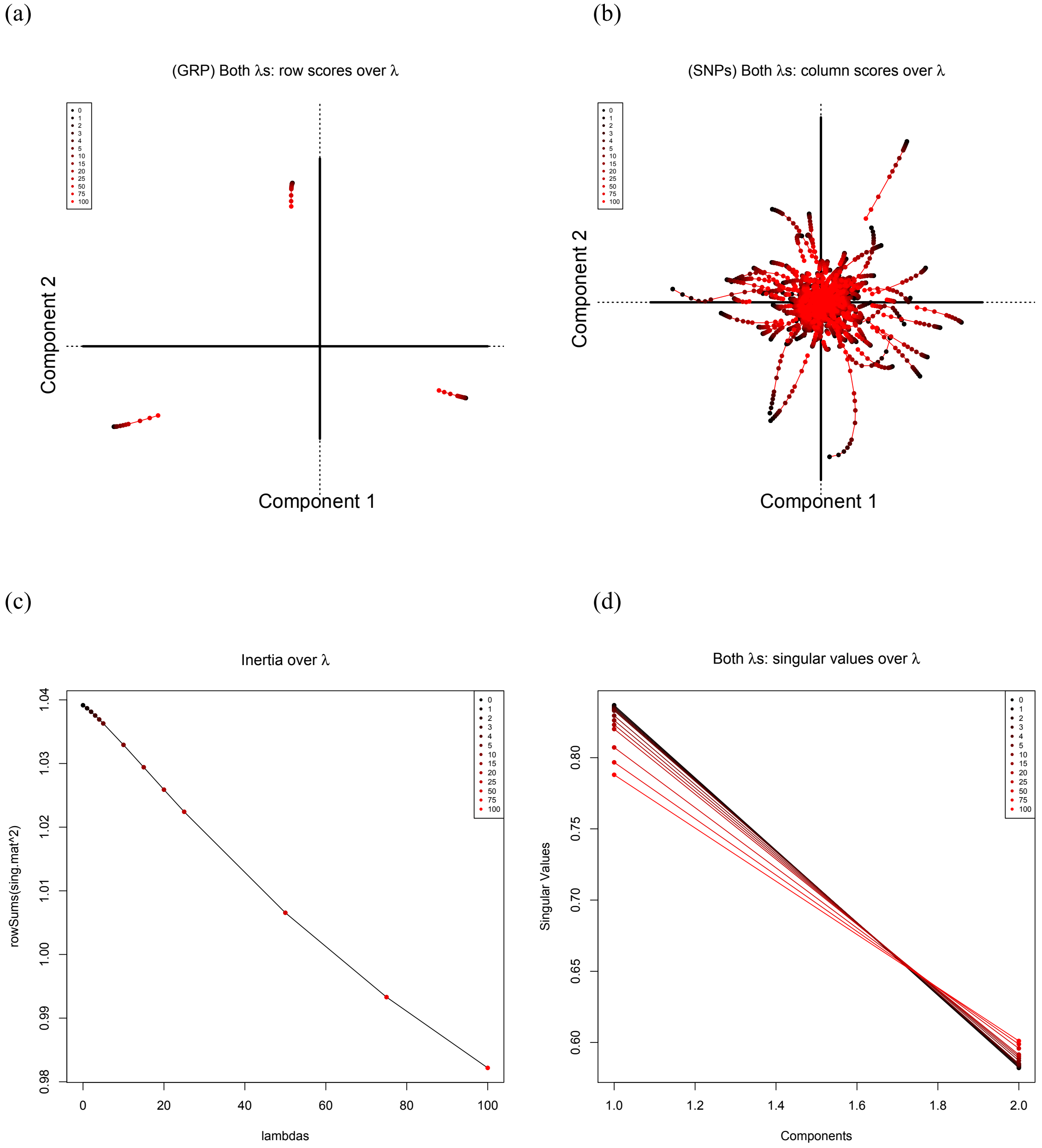
The effects of λ on the component scores are shown in Figures 5.3a (rows, groups) and 5.3b (columns, genotypes). As λ increases *column* items approach zero, a pattern that suggests that the group variables—which move very little from their initial positions—are already quite stable. Like with the prior example (behavior + genetics), the inertia decreases as λ increases (Figure 5.3c), but that only the variance for the *first component* decreases as λ increases; the variance of the second component actually increases. This is because, like in the RMCA example, increased λ causes the space to become more spherical. Finally, it seems that the regularization procedure does not make much of a difference from baseline, with respect to the latent variables (Figures 5.3e and f); though there is a reduction in noise, the fit does not improve (Figures 5.3e and f), but the similarity between latent variables does increase (Table 5.2).

5.3.1.3 Asymmetric regularization for SmooPLS-CA

The setup and results of the prior analyses invite a particular question: What if we do not want to shrink values of an already strong, and well-defined structure (e.g., behavioral data, group association)? Can we use the inherently strong structure in one set, while regularizing the other set (i.e., genetics)?

Both prior examples illustrate this type of problem. Let **Y** denote the genetics data and **X** denote the other set (i.e., behavior or group), if we re-do these analyses we have a fixed λ for **X** (λ = 0), but allow λ to increase for **Y**. So, what are the consequences of this asymmetric regularization? Figure 5.4 shows the effects of asymmetric regularization for the two prior analyses. The top of Figure 5.4 (a and b) shows the results when λ = 0 (for data sets a: behavior and b: groups). Both results show little-to-no movement of most items, in fact some items actually move *further away* from zero as λ increases for the genetics data set. Figures 5.4c and d show the effects of increased λ on the regularized set (i.e., genetics); the effects are similar to, but less dramatic than, those in the symmetric regularization analyses (Figures 5.2 and 5.3). Finally, the usual behavior of decreasing overall variance can be observed in Figures 5.4e and f, but again, the changes are less dramatic than the symmetric regularization analyses.

Thus, in asymmetric SmooPLS-CA we see the non-regularized data stay relatively in the same position as standard PLS-CA, while the regularized data continue to shrink towards zero. This particular approach can be extremely useful when we do not want to shrink the variance in a known, well-defined data set, but we do want to minimize noise in a data set with little-to-no known structure.


(a)
(GRP) Both λs: row scores over λ
0
1
2
3
4
5
10
15
20
25
50
75
100
Component 2
Component 1
(b)
(SNPs) Both λs: column scores over λ
0
1
2
3
4
5
10
15
20
25
50
75
100
Component 2
Component 1
(c)
Inertia over λ
rowSums(sing.mat^2)
0.98
0.99
1.00
1.01
1.02
1.03
1.04
0
20
40
60
80
100
lambdas
0
1
2
3
4
5
10
15
20
25
50
75
100
(d)
Both λs: singular values over λ
Singular Values
0.60
0.65
0.70
0.75
0.80
1.0
1.2
1.4
1.6
1.8
2.0
Components
0
1
2
3
4
5
10
15
20
25
50
75
100

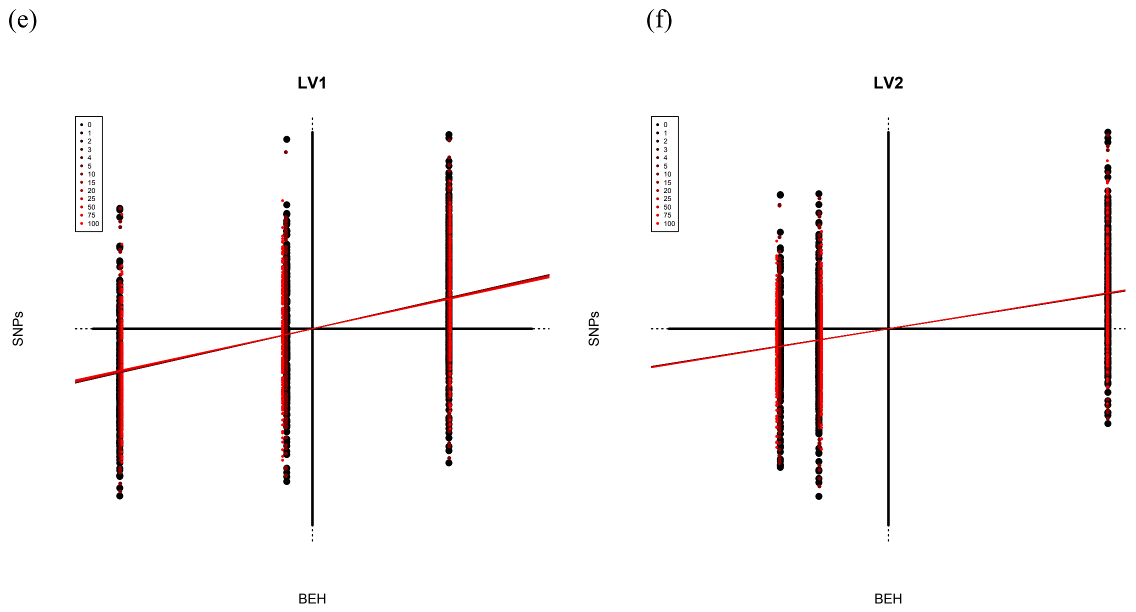
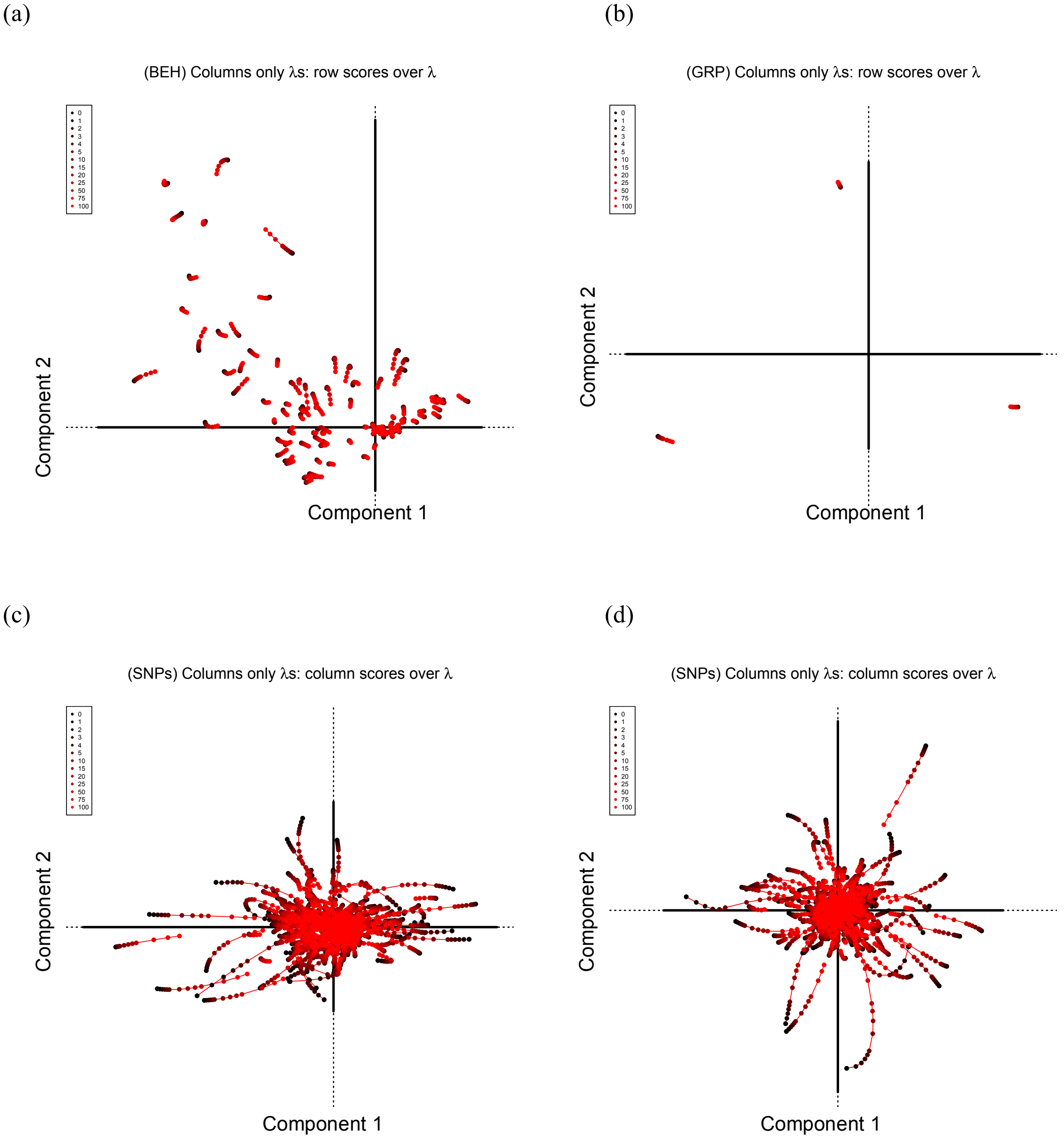


Figure 5.3 Smoothed Partial Least Squares-Correspondence Analysis (SmooPLS-CA) with the Group (GRP) + Genetic (SNPs) ADNI subset data from Beaton et al., (2016), with regularization effects from λ = 0 (black) to λ = 100 (red). (a) and (b) show the effects of increased λ on the behavioral data (row) and genetic data (column) component scores for Components 1 (horiztontal) and 2 (vertical), respectively; though (a) does show very little change by comparison to 5.2(a). The component scores are shown within the constraints that λ = 0 (black dots). Relatively, as λ increases, component scores smooth and approach zero. This effect can be further seen in (c) and (d). In (c), the effects of λ on the total variance (a.k.a., "inertia") of the analyses are shown. As λ increases, the total variance decreases. Finally, (d) the singular values (there exists only 2 components) shows that only the first singular value always decreases, whereas the second eventually increases by comparison to λ = 0. In both (e) and (f), the latent variables are shown for LV1 and LV2, respectively. The lines show fit lines and as λ increases, the slope changes very little; the lack of change shows that there is little-to-no effect on the overall structure. However, as λ increases, the correlation between latent variables does indeed increase (see Table 5.2).


(a)
(BEH) Columns only λs: row scores over λ
0
1
2
3
4
5
10
15
20
25
50
75
100
Component 2
Component 1
(b)
(GRP) Columns only λs: row scores over λ
0
1
2
3
4
5
10
15
20
25
50
75
100
Component 2
Component 1
(c)
(SNPs) Columns only λs: column scores over λ
0
1
2
3
4
5
10
15
20
25
50
75
100
Component 2
Component 1
(d)
(SNPs) Columns only λs: column scores over λ
0
1
2
3
4
5
10
15
20
25
50
75
100
Component 2
Component 1

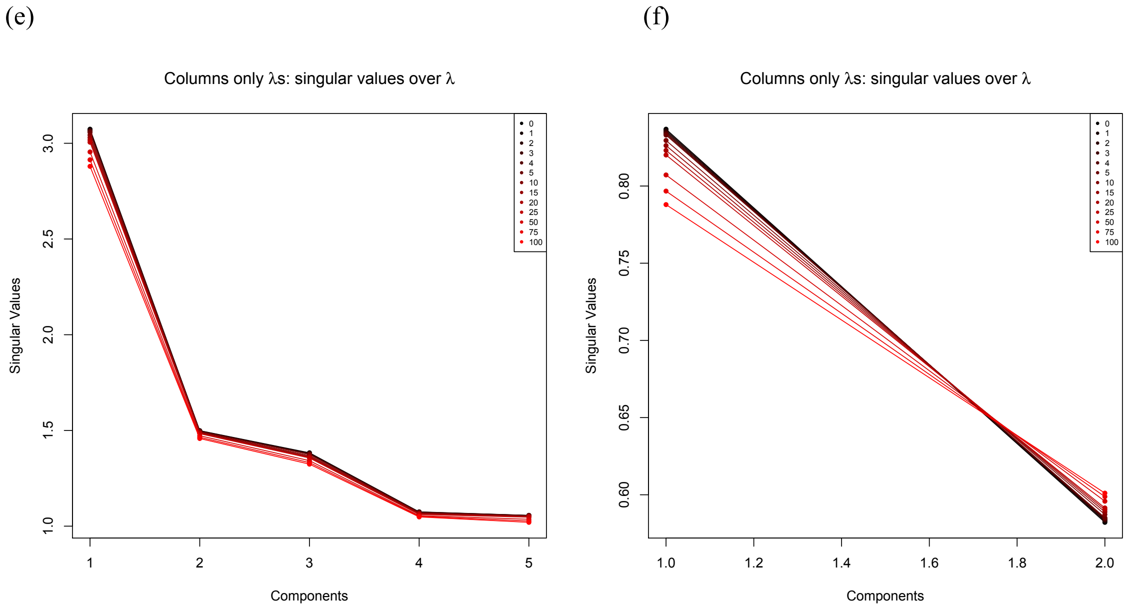


Figure 5.4 Smoothed Partial Least Squares-Correspondence Analysis (SmooPLS-CA) with the row items (left: behavioral data, right: group data) set at λ = 0 (top row) for the ADNI subset data from Beaton et al., (2016). Regularization effects from λ = 0 (black) to λ = 100 (red) were applied to the column items (genotypes). Panels a-d show Components 1 (horizontal) and 2 (vertical). Compare (a) and (b) against Figures 5.2a and 5.3a, respectively. Generally, row items (top) do not move when λ = 0, but some items do move closer to 0, while others actually move further away. Compare (c) and (d) to Figures 5.2b and 5.3b. Generally, items still move toward 0, but as λ increases, the move towards 0 is less dramatic than in Figures 5.2b and 5.3b. Finally, compare Figures (e) and (f) to Figures 5.2(d) and 5.3(d). The same general trend of decreased variance (overall and per component) still exists with a single fixed λ, but like the column items (c and d), the decrease is less dramatic than with equal λs for both sides. Not shown: latent variables, though the effects are like those observed in (c)-(f).

5.4 An Alternative form of SmooPLS-CA

SmooPLS-CA is essentially a combination of Takane and Hwang (2006)'s RMCA and (part of) Allen (2013)'s SFPCA; this makes SmooPLS-CA a smoothed PCA with orthogonal components (like Allen) under the χ^2 assumptions of CA, with specific λ-regularized constraint matrices for the data matrices (like Takane & Hwang). However, as pointed out by Allen (2013) smoothing can be an expensive procedure, especially as data matrices become very large. Additionally, the λ-regularized constraint matrices—which are square, block-diagonal matrices; see for example Eq. 5.43—could pose a problem for both multiplication (computational time) and storage (extremely large memory footprint). In fact, with very large data sets—such as large-scale genetic and genomic data sets—this approach can be extremely difficult and expensive, or even impossible to implement. Therefore, we need a regularization procedure (like the established one) that is practical and does not require excessive resources. An ideal alternative for a regularized PLS-CA should require no more memory or computational power than standard PLS-CA. Thus, in this section, I provide an alternative form of SmooPLS-CA that has the same memory and computational requirements as standard PLS-CA while adhering to the general purpose of ridge regularization: to inflate the space from which we estimate. Consequently, this alternative form of SmooPLS-CA can be seen as a generalization of many techniques, and thus provides the basis of a family of regularized procedures (i.e., Multidimensional Scaling, PCA, CA, MCA, standard PLSC, and PLS-CA).

5.4.1 SmooPLS-CA as a form of regularized CA

PLS-CA can either be viewed as generalization of PLSC to categorical and mixed-data types, or as a special case of CA (cf. Chapter 4an also Eqs. 5.3-5.7). If we consider PLS-CA a special case of CA then we make less assumptions than in PLS-CA: (1) we may not know about the structure of individual observations, and (2) we may not know of or have any particular structure of the variables. Therefore, in these cases, we assume a simple contingency table and the problem becomes: “How do we regularize CA as applied to simple contingency tables?” This approach is referred to as “truncated” SmooPLS-CA

To address this question we have to return to the basis of ridge regularization in OLS: $\hat{\beta} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$ vs. $\hat{\beta}^{*} = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y},$ where $\lambda\mathbf{I}$ *inflates* the diagonal of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ so that the estimates are made within a larger subspace. As previously noted, this inflates the effect of the observations (an effect sometimes called the “phoney data” principle; Basilevsky, 2009; Draper, Smith, Draper, & Smith, 1998).

Currently there is only one form of regularized (standard and simple) χ^2 (by Li et al., 2014). Li and colleagues simply added a small inflation parameter to the denominator of the χ^2 equation. Thus, with the purpose of ridge regularization, and the proposed regularized χ^2 of Li et al., (2014), we can define a simpler approach to SmooPLS-CA that still adheres to the framework of Allen (2013), while still being reminiscent of Takane and Hwang (2006). Let us begin from the perspective of CA wherein we have the original data matrices to form a contingency table: $\mathbf{Z_R} = \mathbf{Z_X^{\mathrm{T}}}\mathbf{Z_Y}$ where $\mathbf{Z_R}$ is a double centered contingency table and $\mathbf{Z_X}$ and $\mathbf{Z_Y}$ are column-wise centered 0/1 matrices (typically these would be disjunctive matrices like in Table 5.1). Next, define inflated weight matrices:

$$\mathbf{W}(\lambda)_{\mathbf{X}} = \mathbf{W}_{\mathbf{X}} + \lambda\mathbf{I} \text{ and } \mathbf{W}(\lambda)_{\mathbf{Y}} = \mathbf{W}_{\mathbf{Y}} + \lambda\mathbf{I}, \tag{5.46}$$

where there is just a single λ for both sets of weights, as in Eq. 5.43. The GSVD triplet for this alternative form is $\text{GSVD}(\mathbf{W}(\lambda)_{\mathbf{X}}^{-1}\mathbf{Z}_{\mathbf{R}}\mathbf{W}(\lambda)_{\mathbf{Y}}^{-1}, \mathbf{W}(\lambda)_{\mathbf{Y}}, \mathbf{W}(\lambda)_{\mathbf{X}})$. This alternative form requires less memory and computational requirements than the original proposed SmooPLS-CA (see Section XX), but ignores the *a priori* structure of variables in the weight matrices. This particular reformulation is more in line with simple correspondence analysis (as applied to a contingency table) than MCA, RMCA, PLS-CA, or SmooPLS-CA and thus the optimization should be considered more in line with that of Eq. 5.42. However, through the magic of the GSVD, the PLSC-style optimization of $\mathbf{L}_{\mathbf{X}} = \mathbf{Z}_{\mathbf{X}}\mathbf{U}$ and $\mathbf{L}_{\mathbf{Y}} = \mathbf{Z}_{\mathbf{Y}}\mathbf{V}$, where $\mathbf{L}_{\mathbf{X}}^{\mathrm{T}}\mathbf{L}_{\mathbf{Y}} = \Delta$ still holds.

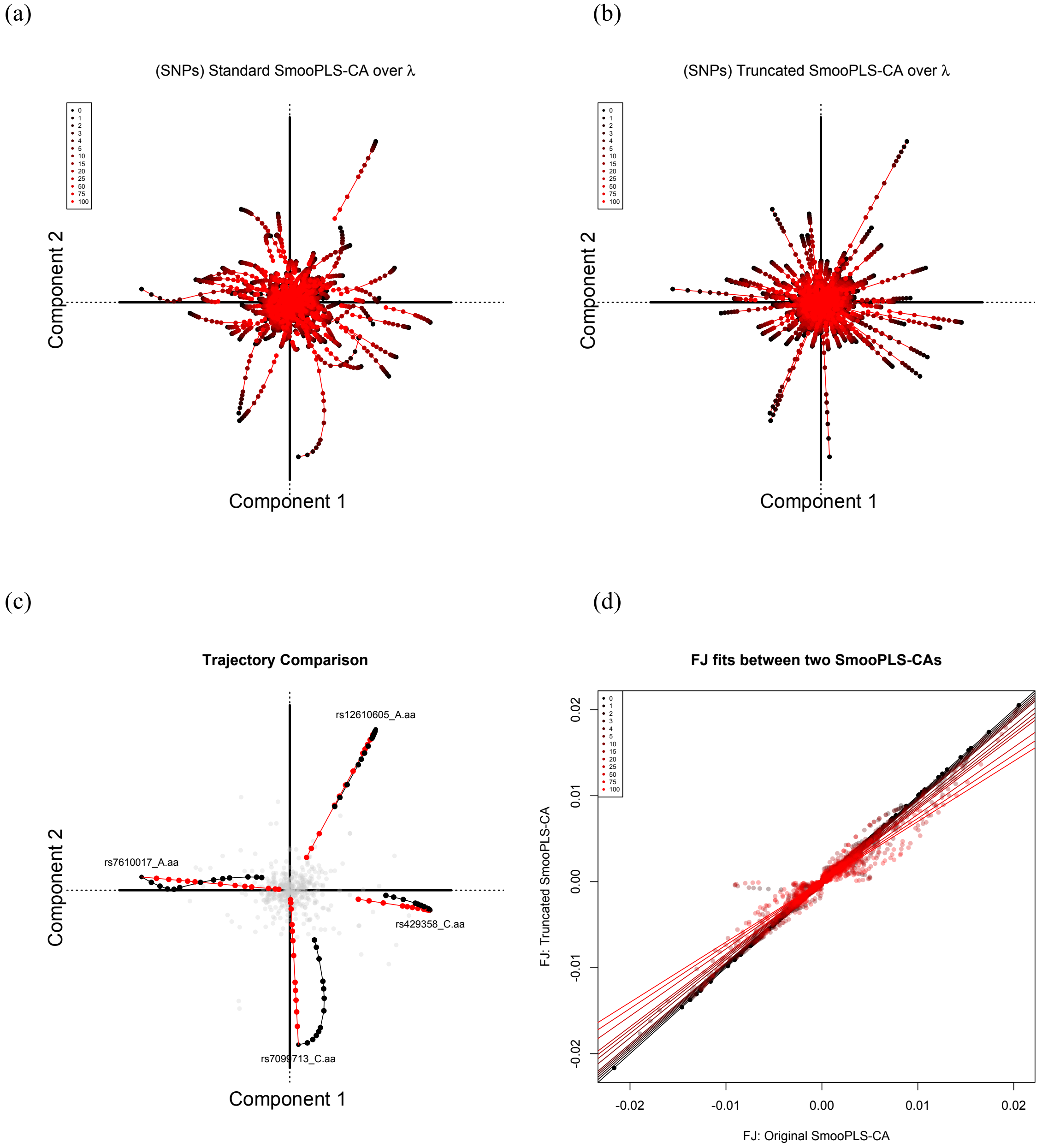
5.4.2 Similarities and Differences Between Formulations

Having two formulations of SmooPLS-CA (i.e., one based on RMCA and one based on simple CA) invite the questions: How do these two approaches differ? What is lost or gained by using either approach? In the following section, I show the differences and similarities between standard SmooPLS-CA and truncated SmooPLS-CA. First, the discriminant version is presented followed by the behavioral + genetics version.

5.4.2.1 Standard vs. Truncated Discriminant SmooPLS-CA

Figure 5.5 illustrates symmetric regularization approach (i.e., $\lambda_{\mathbf{X}} = \lambda_{\mathbf{Y}} = \lambda = [0, 1, 2, 3, 4, 5, 10, 15, 20, 25, 50, 75, 100]$) for the standard vs. truncated forms of Discriminant SmooPLS-CA. Figure 5.5a shows that the genotype component scores for standard and Fig. 5.5b shows the truncated version of SmooPLS-CA. The two versions differ in their trajectories toward zero, where the truncated version is approaches zero faster than the stnadard version as λ increases (see Figures 5.5a, b, c, and f). Furthermore the standard, as opposed to the truncated, can exhibit erratic movements (see rs7099713_C.aa in Fig. 5.5c) before convergence towards zero begins, where this effect results in part from the constraints imposed on regularization (i.e., rs7099713_C.aa must move conditionally on rs7099713_C.Aa and rs7099713_C.AA, which are near 0 anyways). However, the overall difference between standard and truncated component scores are not that drastic, as seen in Fig. 5.5d each fit line is for each level of λ-regularization. In general, the component scores differ very little between the two approaches. In fact the correlations between the genotype component scores, as well as the latent variables (individuals projected onto the genotypes) are highly correlated ($r > .88$), and relatively low correlation values appear only in late components with high λ. Finally, the truncated version actually approaches zero much faster than the standard version (Fig. 5.5e); an effect that could be inferred from Figures 5.5a and b.

Though the truncated version makes fewer assumptions and thus applies fewer constraints on the analysis, for the discriminant version there are negligible qualitative differences in how the method behaves.


(a)
(SNPs) Standard SmooPLS-CA over λ
0
1
2
3
4
5
10
15
20
25
50
75
100
Component 2
Component 1
(b)
(SNPs) Truncated SmooPLS-CA over λ
0
1
2
3
4
5
10
15
20
25
50
75
100
Component 2
Component 1
(c)
Trajectory Comparison
rs12610605_A.aa
rs7610017_A.aa
rs429358_C.aa
rs7099713_C.aa
Component 2
Component 1
(d)
FJ fits between two SmooPLS-CAs
0
1
2
3
4
5
10
15
20
25
50
75
100
FJ: Truncated SmooPLS-CA
FJ: Original SmooPLS-CA
0.02
0.01
0.00
-0.01
-0.02
-0.02
-0.01
0.00
0.01
0.02

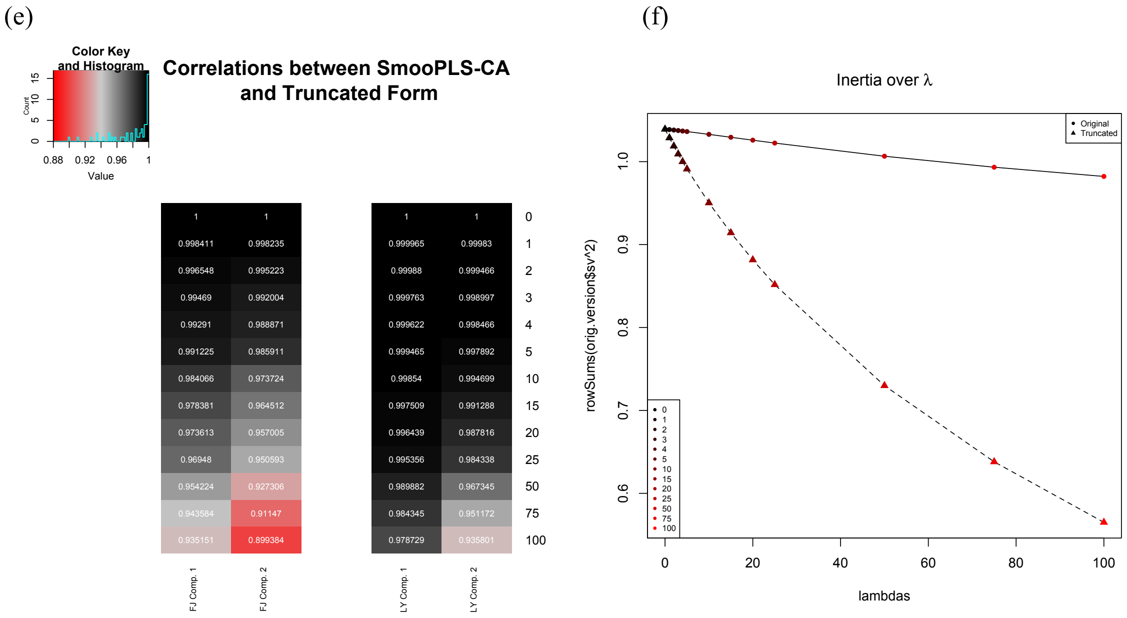
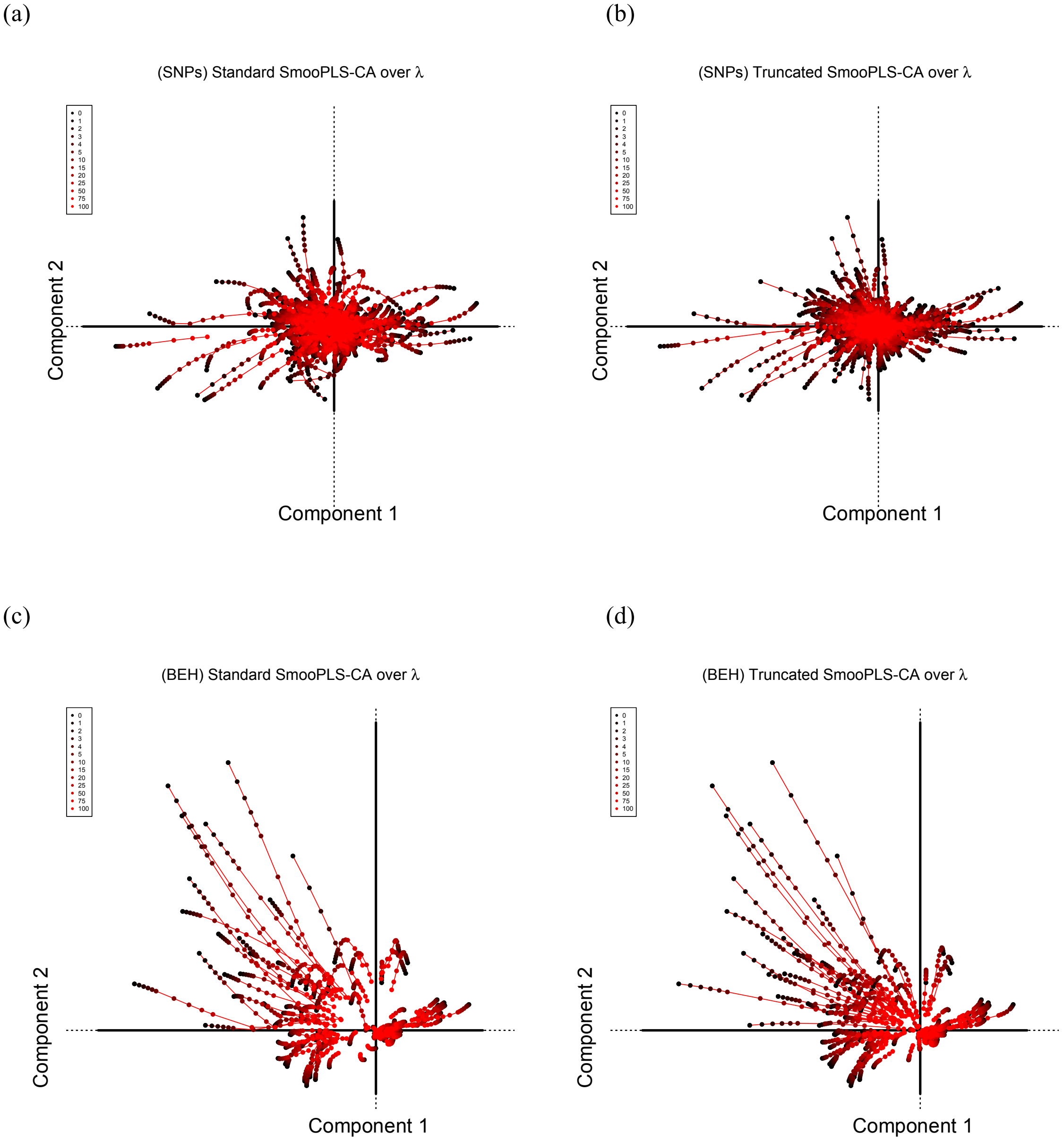


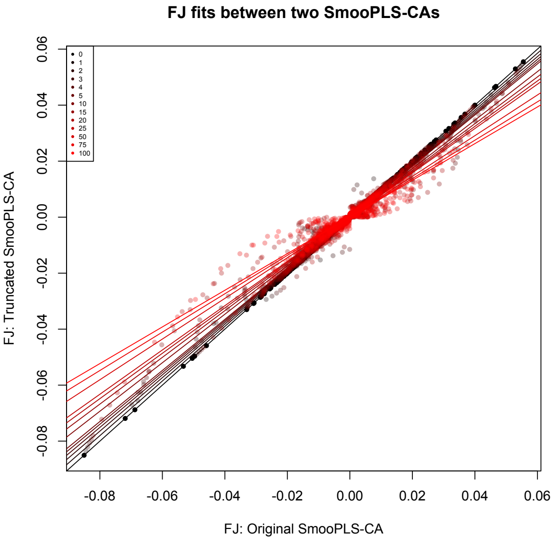
Figure 5.5 Genotype component scores for standard (a) vs. truncated (b) SmooPLS-CA. Panels a-c show Components 1 (horizontal) and 2 (vertical). Standard shows a less drastic descent of component scores toward zero. Occassional erractic movements are highlighted in (c), where 4 (extreme) genotypes were selected to show this effect; in particular, 'rs7099713_C.aa' shows that small values in standard regularization do not immediately converge towards zero, where as they do in the truncated version. Relationship between standard and truncated versions of λ levels, via fit lines, are shown in (d), where (e) shows the correlation values between the standard and trucnated genotype component scores, as well as the latent variables (i.e., scores for the individuals). Finally, (e) shows how the standard vs. truncated approaches impact the inertia (i.e., total variance).

5.4.2.2 Standard vs. Truncated SmooPLS-CA

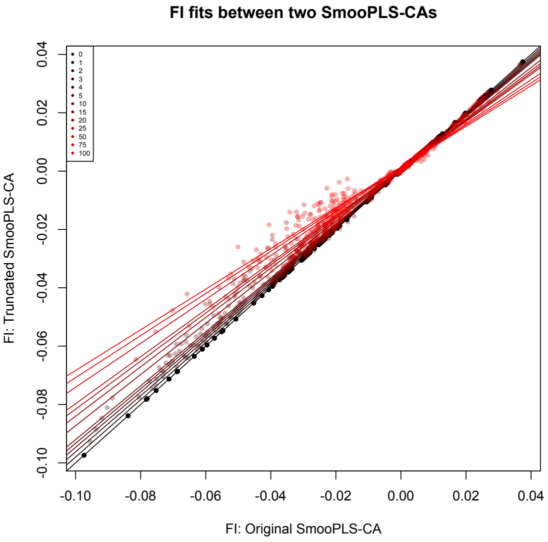
Figure 5.6 illustrates symmetric regularization approach (i.e., $\lambda_{\mathbf{X}} = \lambda_{\mathbf{Y}} = \lambda = [0, 1, 2, 3, 4, 5, 10, 15, 20, 25, 50, 75, 100]$) for the standard vs. truncated forms of SmooPLS-CA. Like the discriminant forms (see previous section), the truncated version approaches zero faster than the standard for genotype component scores (Fig 5.6a vs. b for standard vs. truncated), behavioral scores (Fig 5.6c vs. d for standard vs. truncated), and the inertia (Fig 5.6h). The fit lines between the the standard and truncated approaches show high similarity (Fig 5.6e and f). Trajectories can be erratic at first (Fig 5.6g), but as λ increase, and values approach zero there is a stabilization of this effect. Finally, the latent variables between the standard and truncated SmooPLS-CA—much like the discriminant form—show a high similarity (Fig 5.6i and j), where the only correlations to reflect deviance between the models are late components with high λ; in general, correlations remain very high ($r > .8$).


(a)
(SNPs) Standard SmooPLS-CA over λ
0
1
2
3
4
5
10
15
20
25
50
75
100
Component 2
Component 1
(b)
(SNPs) Truncated SmooPLS-CA over λ
0
1
2
3
4
5
10
15
20
25
50
75
100
Component 2
Component 1
(c)
(BEH) Standard SmooPLS-CA over λ
0
1
2
3
4
5
10
15
20
25
50
75
100
Component 2
Component 1
(d)
(BEH) Truncated SmooPLS-CA over λ
0
1
2
3
4
5
10
15
20
25
50
75
100
Component 2
Component 1

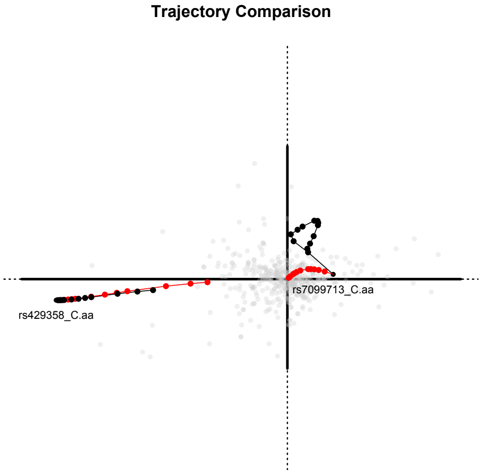
(e)


FJ fits between two SmooPLS-CAs
FJ: Truncated SmooPLS-CA
FJ: Original SmooPLS-CA
0
1
2
3
4
5
10
15
20
25
50
75
100
-0.08
-0.06
-0.04
-0.02
0.00
0.02
0.04
0.06

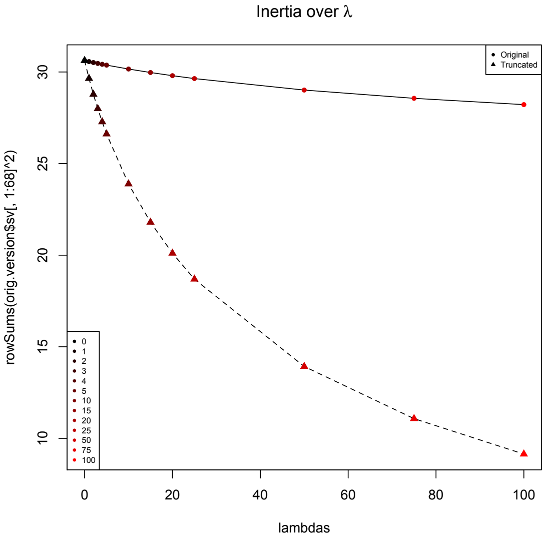
(f)


FI fits between two SmooPLS-CAs
FI: Truncated SmooPLS-CA
FI: Original SmooPLS-CA
0
1
2
3
4
5
10
15
20
25
50
75
100
-0.10
-0.08
-0.06
-0.04
-0.02
0.00
0.02
0.04

(g)


Trajectory Comparison
rs7099713_C.aa
rs429358_C.aa

(h)


Inertia over λ
rowSums(orig.version$sv[, 1:68]^2)
lambdas
Original
Truncated
0
1
2
3
4
5
10
15
20
25
50
75
100
10
15
20
25
30
0
20
40
60
80
100

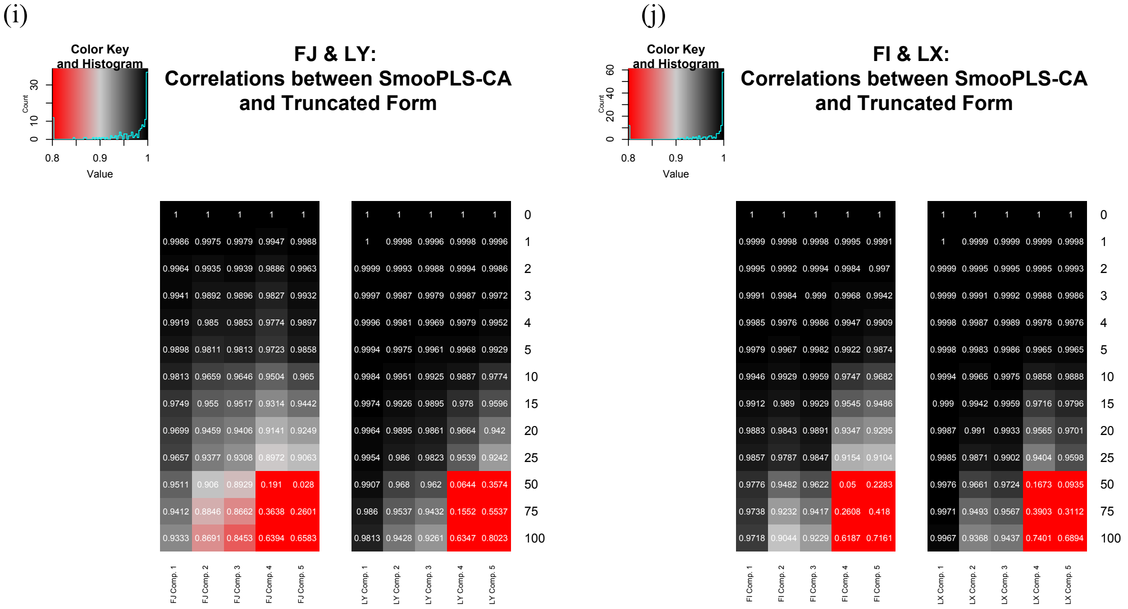


Figure 5.6 Genotype component scores for standard vs. truncated SmooPLS-CA. Panels a-d, and g show Components 1 (horizontal) and 2 (vertical). Standard shows a less drastic descent of component scores toward zero. Similar to Figure 5.5, except this example is the behavioral + genetic data. In panels (a-d) we can see the standard version of SmooPLS-CA (a, c) vs. the truncated version of SmooPLS-CA (b, d). In both casees—like in Figure 5.5—the standard version is slower to approach zero with some erratic movements. The truncated version more directly approaches zero. Panels (e) and (f) show the latent variable fits between the two; standard and truncated are highly similar. Panels (g) shows the similarity and differences in particular genotypes ('rs429358_C.aa' and 'rs7099713_C.aa') between standard and truncated; as λ increases, most values rapidly approach zero and converge across techniques. Panel (h) shows the inertia decreases as λ increases. Finally, the component scores (genetics, behaviors) and latent variables (individuals for genetics and behaviors) are shown in panels (i) and (j). Similar to Figure 5.5, the smoothed and truncated versions of SmooPLS-CA are virtually identical and only deviate as: (1) λ becomes very large and (2) later (likely noise) components are used.

5.5 Selection of λ and inference in SmooPLS-CA

PLSCA is a multivariate descriptive (i.e., fixed-effect) technique, but, it can be complemented by a variety of inference tests. These inference tests are usually computed from non-parametric resampling methods (a.k.a., cross-validation) such as permutation and bootstrap (see Beaton, Dunlop, et al., 2016). Resampling methods generate a large number (e.g., thousands) of new data sets which are then used to derive the distributions of various statistics. The observed statistics are then compared against their resampling based distributions to determine if the observed effects are "significant."

However, while SmooPLS-CA can also use the same inference approaches as PLS-CA, there is an additional problem: how to select the best λ. Generally, as seen in Figures 5.2-5.4, and Table 5.2, as λ increases, the results are easier to interpret. If this were tru, then we could simply use the largest possible value of λ (without compromising numerical precision). However, very high values of λ suffer from the same drawback as very low values of λ or when λ = 0: overfitting.

Overfitting happens when a particular method models idiosyncrasies or peculiarities in a specific data set, likely capturing particular forms of noise specific to these data, and (artificially) increasing performance. Thus if a model is overfit, the conclusions obtained from the analysis of a sample do not generalize to the population. To palliate this problem, data-driven approaches are used to identify—for the data set of interest—the "best" λ (i.e., the best compromise between good results for the data set and good generalization to the population).

The following sections first present two of resampling-based inference approaches we can use to assess the stability, predictability, and thus replicability of effects. Next, I describe how Takane and Hwang (2006) and Allen (2013) approach selection to regularization parameters, and finally I propose an alternative criteria to Allen (2013)'s and Takane and Hwang (2006)'s approaches that is better suited for this dissertation.

5.5.1 Split-half

Split-half resampling (SHR) is a compromise between leave-one-out (LOO) and bootstrap resampling. In SHR a data set is randomly split into two, and the same analyses are performed on each split (Strother et al., 2002). Each split is then compared with one another, either by predicting one set from the other, or estimating which effects are reproducible between the splits. Split-half resampling has been used primarily in neuroimaging research to provide: (1) an optimization of parameters, and (2) an estimate of both the quality of prediction (e.g., of individuals to their respective groups) and reproducibility (e.g., of components) for SVD-based and other multivariate techniques (Churchill, Spring, Afshin-Pour, Dong, & Strother, 2015). SHR can be used in PLS-CA to estimate the reproducibility of effects from one split to the other, as well as—if there are *a priori* groups—prediction estimation from one split to the other.

5.5.2 Bootstrap

The bootstrap is a resampling *with replacement* technique (Efron & Tibshirani, 1993; Chernick, 2008). In PLSCA, observations are assumed to represent a population of interest. New samples are generated by resampling (in general the observations) with replacement from the original sample (i.e., the rows of both **X** and **Y** are resampled with the same resampling scheme).

The distribution of the statistics computed from bootstrap resampling is a maximum likelihood estimation of the distribution of the statistic of interest (for the population of the observations). In addition, the bootstrap can be stratified (a.k.a. “constrained”) to resample within *a priori* groups (e.g., marijuana, nicotine, control). The bootstrap is used to derive two different types of inferential statistics: bootstrap ratios and confidence intervals.

5.5.2.1 Bootstrap Ratio Tests

Bootstrap ratios (BSR) originated in the neuroimaging literature (McIntosh & Lobaugh, 2004) but are related to other tests based on the bootstrap (see Hesterberg, 2011, for “interval-*t*”) or on asymptotic theory (see Lebart et al., 1984, for “test-value”). The BSR test is a *t*-like statistic computed by dividing the bootstrap computed mean of a measure by its bootstrap derived standard deviation. Just as for the usual *t*-statistic, a value of 2 would (roughly) correspond to significance level of $\alpha = .05$ [i.e., $P\ (|t| > 2) \approx .05$] and can be considered as a critical value for a single null-hypothesis test. Corrections for multiple comparisons (e.g., Bonferroni) can be implemented when performing a large number of tests simply by increasing the BSR threshold to correspond to a particular α value (e.g., $P\ (|t| > 3) \approx .0013$ or $P(|t| > 4) \approx 3.17 \times 10^{-5}$).

5.5.2.2 Confidence Intervals

Confidence intervals are created from percentile cut-offs of the bootstrap distributions. Confidence intervals are generated for anything with component scores except observations (because the observations are the units for resampling). Confidence intervals can be created for each measure (just like the BSR) or around groups of participants (e.g., marijuana, nicotine, control). Confidence intervals can be displayed on component maps as peeled convex hulls (Greenacre, 2007) or as ellipsoids (Abdi et al., 2009). When the confidence intervals of two measures or groups do not overlap, these measures or groups are significantly different at the chosen level (Abdi et al., 2009).

5.5.3 Selection of λ

Both Takane and Hwang (2006) and Allen (2013) note that components-based regularization procedures should "greedily extract" (Allen, 2013) or "presuppose[s] we already know" (Takane & Hwang, 2006) the number of components *before* we perform regularization. This assumption is made because we must, for each λ, run the analysis again—a procedure likely to be time consuming especially with large data sets. When the number of components is very large, the number of relevant components should be estimated from the baseline (non-regularized) analysis. However, the selection of λ is a procedure agnostic to the number of components; it simply helps reduce computational cost if the number of components are preselected.

Both Allen (2013) and Takane and Hwang (2006) provide recommendations on how to select λ. Allen (2013) describes a combined approach to estimate (all four) regularization parameters using both (1) a minimal estimate of degrees of freedom for sparse penalization (through computing the trace of **u** and **v** from a model), and (2) a BIC criterion for the estimate of both **u** and **v** (separately). Takane and Hwang (2006) suggest several possible approaches, but ultimately present a simple approach based on *K*-folds cross-validation, where the dataset is split into *K* sets and in turn the *k*-th set is left out and predicted from the larger set. Once all *K* sets have been estimated, they are aggregated into a single matrix; the λ that produces the minimal difference (i.e., trace) between RMCA on the full set and all RMCAs of the predicted *K*-sets, is regarded as the "optimal" λ. Selection procedures in both approaches aim to find the best λ, while minimizing the risk of overfitting (by minimizing error variance). While both approaches aim very broadly to capture a best estimate of the population—instead of the sample—neither actually address how predictive (i.e., classification) or reproducible the results are.

5.5.3.1 Split half

With the various types of data sets and study designs we could analyze, we actually have many types of questions we would like to answer: (1) How predictive are the results? (2) How stable are the results? (3) How reproducible are the results? In fact, not all of these questions apply to all problems: (1) we only care about classification accuracy if we have some *a priori* groups, (2) we (may) only care about stability to find the best items, and (3) we almost always want to ensure that our results generalize to the population. Different resampling techniques help us answer these questions in different ways. For the purposes of this dissertation, I used split-half resampling to estimate the selection of λ. The motivation to use split-half resampling was because: (1) the number of components in all analyses are guaranteed to be very few (see next Chapter for study outlines), and (2) I want to maximize predictive utility of genotypes between a discovery and a validation sample.

The SHR approach can be used in two ways: (1) estimates of predictability (i.e., classification between splits; *K*-fold CV where *K* = 2), and (2) reproducibility of model configuration (i.e., similarity between splits). However, for this dissertation only the first approach (prediction) is used. With SHR, we want to predict one split-half data set (e.g., SH2) from the solution of the other split-half (e.g., SH1). By doing so, we can get random effects estimates of classification. Because SHR is repeated many times, we can estimate intervals of the prediction accuracy.

Prediction estimates via SHR can be used in two ways with PLS techniques: (1) on a whole model (i.e., component scores across all the components), or (2) on a component-by-component basis (i.e., the latent variables scores). Both approaches are feasible when there are only a few components. However, the second approach is only feasible if a relatively small number of components are extracted and tested; this second case fits the same criteria of "greedily extract[ing]" or "presuppose[ing]" the number of components. Though as components increase, the prediction accuracy should decrease as later components are typically—especially through regularization—comprised of noise.

CHAPTER 6

APPLICATION OF SMOOPLS-CA TO SUBSTANCE USE DISORDERS

In this chapter, I present and discuss the results from the application of standard and SmooPLS-CA to genetics data of substance use disorders (SUDs). Two data sets were used: a discovery set and an independent validation set. The data sets comprised a panel of single nucleotide polymorphisms (SNPs) based on a large set of “knowledge-informed” or “literature-based” candidate-genes. Parts of this chapter are adapted Beaton, Filbey, & Abdi (2014). Beaton, D., Abdi, H., & Filbey, F. M. (2014) Unique aspects of impulsive traits in substance use and overeating: specific contributions of common assessments of impulsivity. The American journal of drug and alcohol abuse, 40(6), 463-475. http://dx.doi.org/10.3109/00952990.2014.937490 reprinted by permission of Taylor & Francis LLC (http://www.tandfonline.com).

The broad goal of this dissertation was to identify (discovery) and then verify (validation) possible genetic markers of SUDs. Yet, there exists methodological challenges behind this goal. One challege is knowing the best study configuration to identify genetics of SUDs: case-control, group-based, or trait-based analyses. Another challenge is that large scale genetics studies rarely reach a sufficient sample size. Therefore, we require analytical approaches to help boost power. The expecation in this dissertation was that SmooPLS-CA would outperform standard PLS-CA and find “better” genetic markers of SUDs.

Because of these challenges (i.e., study configuration, sample size), there were three sets of results from the discovery sample: (1) a case-control (control vs. any SUD) analysis, (2) a multi-group (control vs. marijuana vs. marijuana+nicotine vs. nicotine groups) analysis, and (3) a trait-based (self-reported impulsivity) analysis. Each study configuration was repeated for standard PLS-CA and each iteration of SmooPLS-CA. Each set of results was then tested for predictive accuracy in an external data set (SAGE), in the hopes to find a “winning” configuration that outperformed all others.

The various study configurations, standard vs. regularized, and discovery to prediction pipeline was designed to specifically answer three questions: (1) *How does regularization help find better markers?* (2) *Which configuration is best (e.g., case v. control or impulsivity)?* and, if those worked out, (3) *Which specific SNPs, and which associated genes contribute to the identified effects?*

The results were not as expected. It appeared as though no set of genetic markers identified in the archived data were predictive in the external data set. Prediction was also low when the analysis pipelines were flipped: SAGE was used for discovery, and the archived data then used for validation. The overall lack of prediction suggested a failure to find genetic markers of SUDs. However, upon closer inspection of the archived-as-discovery and the SAGE-as-discovery analyses, I noticed that both discovery analyses identified many common effects. In other words, common markers of SUDs were identified across archived and external data sets through independent analyses. In the end, all of the analyses in this dissertation—when understood together—answered the three primary questions as: (1) *Regularization did not help: standard PLS-CA either outperformed SmooPLS-CA or the differences were negligible*, (2) *case-control (i.e., CON vs. SUD) analyses provided the most generalizable results*, and (3) *several expected and several unexpected genes contributed to the CON vs. SUD effect*.

This chapter is outlined as follows. First, I provide a description of the participants (e.g., demographics, usage), measures, and a brief description of preprocessing and quality control for these studies. Next, I provide an overview of the proposed, and also the "flipped," analysis pipelnes. Then, I provide a description of the key results from the studies in this dissertation. Finally, I discuss the results, conclusions, and limitiations.

6.1 Data Sets

This study was approved by the University of New Mexico and The University of Texas at Dallas Institutional Review Boards. This dissertation includes two data sets: an archived (local) set and an external set (Study of Addiction: Genetics and Environment, a.k.a., SAGE: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000092.v1.p1). Both sets are described in the following sections. However, because some analyses require preprocessing and quality control independently of others, some details of preprocessing and quality control are described whenever relevant.

6.1.1 SNPs and genes of interest

The archived data set used the Illumina Human OmniExpress chip with 709,362 SNPs. The SAGE set used the Illumina Human1Mv1_C BeadChips with 1,040,107 SNPs. However, in these studies do not use the entire set of available genome-wide SNPs but only a "knowledge informed" or "literature based" candidate approach to a large-scale genetic association study. I focused on a particular set of genes for two reasons: (1) the selected genes are the largely agreed upon genes where some contribution to SUDs *should* exist, and (2) one of the latest trends in large-scale genetic and genomic-association studies of complex traits or disorders either extract genes of interest from a general chip (e.g., as in Ashen et al., 2016) or use a specialized chipset to target specific diseases and disorders or *types* of diseases and disorders, for examples: (1) both the NeuroX chip (Nalls et al., 2015), and the ONDRIseq chip (Farhan et al., 2016) have been designed to specifically target genes of neurodegenerative disorders, (2) Illumina offers a "PsychArray" (http://www.illumina.com/products/by-type/microarray-kits/infinium-psycharray.html) that specifically targets genes associated with psychiatric disorders (e.g., schizophrenia, bipolar disorder), and (3) the chip that inspires this study: the "Addictions Array" (Hodgkinson et al., 2008; see also Section 2.3.4) was designed with a broad range of SUDs in mind.

Therefore, like these studies, I used only SNPs that are associated with a large-scale, *a priori* defined set of genes. These genes are almost entirely those listed in NIAAA et al., (2008), but I have also included genes from other suggested, much smaller, panels (i.e., substance-specific, reward, and stress-based panels; see Section 2.3). In total, there were 149 genes selected from this panel, though for all analyses, sex-based genes were excluded (e.g., *MAO-A* on Chr. X; see Section 6.3 on preprocessing).

Because different chips have different SNPs, the selection procedure must be done based on gene names and Ensembl IDs. The selection of SNPs was done as followed. Given a set of human gene names (a.k.a., symbols) such as "*COMT*", "*GABRB3*", "*CNR1*", all Ensembl IDs that correspond with those names were retrieved, such as "ENSG00000118432", "ENSG00000093010", and "ENSG00000166206". From these Ensembl IDs, all possible associated SNPs (a.k.a., "variants") were then retrieved. This procedure was done with the `biomaRt` package in R, using both the "gene mart" and "SNP mart" of the human genome build from September 2015 (when this and related work began). From the list of all possible SNPs, only those SNPs that existed on a given chip (i.e., the archived or SAGE chips) were extracted for analyses. Additional details on the total number of SNPs are found in the subsequent sections and supplemental material.

6.1.2 Archived ("discovery") data

The archived data set originally contained 475 participants across control and a number of SUD groups. These individuals were recruited to participate in studies on marijuana use, nicotine use, or binge-eating. However, for the focus of this study is psychoactive substance use disorders binge-eating participants were excluded. Furthermore, some of the participants did not have (suitably) complete genetic data and were therefore excluded. There were 431 participants included in these analyses. Participants were recruited from the general community in Albuquerque (NM) or Dallas (TX) and took part in a larger set of studies to determine markers of substance use disorders and addiction. Participants were excluded from the studies if they had (1) past or present diagnosis of a neurological disorder, (2) psychosis or other substance use disorder besides their primary substance use disorder (assessed via the Psychotic Symptoms and Substance Use Disorders modules of the SCID), or (3) currently taking prescribed psychoactive medication. Non-using controls did not report any current regular use of illicit substances (including marijuana) within the past 6 months of recruitment. The control group is referred to as "CON" (*n* = 122). For the current study, the focus here is on psychoactive SUDs: nicotine participants (a.k.a., "NIC"; *n* = 74) self-reported current nicotine use of at least 10 cigarettes per day (positive use verified via breath CO monitor), where marijuana participants (a.k.a., "MJ"; *n* = 173) self-reported current marijuana use of at least 4 occasions per week over the previous 6 months (positive use verified via urinalysis). Within the marijuana recruitment group, we identified individuals that also used nicotine at least daily; these users are referred to as marijuana+nicotine co-use (a.k.a., “MJ+NIC”; *n* = 62).

Participants within the groups were also assessed on several usage measures. The NIC group had a mean (SD) score of 4.247 (2.332) on the Fagerstrom’s Test for Nicotine Dependence (FTND). Marijuana dependence—for MJ and MJ+NIC groups—was assessed with the structured clinical interview for DSM-IV-TR (Research Version). Within the MJ group 61 individuals qualified as lifetime dependent, whereas 66 are qualified as currently dependent (35.26 and 38.15% of the whole MJ group, respectively). Within the MJ+NIC group, 27 individuals qualified as lifetime dependent, whereas 25 qualified as currently dependent (43.548% and 40.323% of the whole MJ+NIC group, respectively). To note, the SAGE data (see Section 6.1.3) recruitment was based primarily on alcohol use disorders. For comparable reference, lifetime and current alcohol dependence are also provided for the archived data. Very few individuals qualify for current alcohol dependence (less than 17% per group). Demographics and usage characteristics for these groups are provided in Table 6.1a.

Table 6.1

Archived and SAGE Demographics and Depdendency.

	Groups			
(a) Archived data demographics.	**CON**	**MJ**	**MJ+NIC**	**NIC**
Sex				
Female	72	57	13	27
Male	50	116	49	47
Ethnicity				
Hispanic or Latino	28	65	23	27
Not Hispanic or Latino	91	108	39	47
Race				
American Indian or Alaska Native	2	4	1	3
Asian	25	2	1	1
Black or African American	16	20	2	4
White or Caucasian	55	88	38	39
Multiracial/Other/No Response	13	34	6	10
Alcohol Dependence				
Lifetime	14	54	29	26
Current	0	24	10	4

	Groups				
(b) SAGE data demographics.	**CON**	**MJ**	**MJ+NIC**	**NIC**	**OTHER**
Sex					
Female	985	24	173	659	166
Male	514	112	337	481	288
Ethnicity					
Not Hispanic or Latino	1448	125	495	1104	432
Hispanic or Latino	51	11	15	36	22
Race					
Black or African American	380	56	178	339	146
White or Caucasian	1113	80	331	800	308
Multiracial/Unknown/No Response	6	0	1	1	0

	Groups				
(c) SAGE dependency characteristics.	**CON**	**MJ**	**MJ+NIC**	**NIC**	**OTHER**
MJ. Dependence					
NO	1499	0	0	1140	454
YES	0	136	510	0	0
Nic. Dependence					
NO	1499	136	0	0	454
YES	0	0	510	1140	0
Alc. Dependence					
NO	1499	0	0	389	0
YES	0	136	510	751	454
Coc. Dependence					
NO	1499	41	131	867	319
YES	0	95	379	273	135
Op. Dependence					
NO	1499	115	385	1063	422
YES	0	21	125	77	32
Oth. Dependence					
NO	1499	79	288	1012	405
YES	0	57	222	128	49
FTND	0.71 (1.454)	3.267 (2.466)	5.925 (2.4)	3.857 (3.238)	2.896 (2.644)

Note. Demographics of archived and SAGE data sets. (a) Provides the demographics of the archived data, where (b) provides the demographics of the SAGE data set. (c) Provides the usage characteristics by group configuration in the SAGE data (Bierut et al., 2010 and Agrawal et al., 2011). For SAGE, only the CON group has dependency status of "no" across all dependence criteria.

Participants also provided self-assessments of impulsivity on two scales: the Impulsive and Sensation Seeking (ImpSS; Zuckerman, Kuhlman, Joireman, Teta, & Kraft, 1993) scale, and the Barratt's Impulsivity Scale (BIS; Patton, Stanford, & Barrat, 1995). Both measures have been extensively studied in SUD populations (Beaton et al., 2014), and are typically regarded as a robust phenotype of SUD. For analyses, both impulsivity scales were recoded in a categorical format (i.e., disjunctive coding; see Table 5.1 for the SNPs example) for every question-response item. Very low frequency responses were combined with the next most similar response. In this particular data set some questions (1, 2, 5, 8, 12, 14, 16, 17, 18, 20, 21, 24) on the BIS had < 5% response rates on "Almost Always/Always" and were thus combined with "Often".

6.1.2.1 Missing data and simple imputation

In the case of missing data—either missing SNPs or responses— per participant, the missing cells were replaced by the disjunctive mean. However, there was a subset of participants (*n* = 72; with *n* = 42 from CON and *n* = 30 from MJ) without impulsivity data. For these participants, the impulsivity data were entirely replaced by *grand* mean (not group mean) responses for the following reasons: (1) the grand mean does not add variance to the structure, but still allows for these individuals to be analyzed, especially with respect to the fact that (2) they had relatively complete—and thus useful—genetic data. Additional details of preprocessing and quality control, with a particular emphasis on genetics, can be found in Section 6.3.

6.1.3 External ("validation") data

The external data set is an archived data set available through the NIH (`https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000092.v1.p1`). The SAGE data is comprised of many individual studies—and is an early exemplar of multi-site, collaborative, and eventually open access studies—each with a particular focus on alcohol use disorders for individuals and within families (Bierut et al., 1998). However, many individuals also have dependence on other substances, where most individuals qualify as co-dependent (two) or multi-dependent (many) classification. The SAGE data set has been used to investigate a variety of SUDs, and in particular the genetic associations of alcohol use (Bierut et al., 2010), nicotine use (Hancock et al., 2015), cannabis use (Agrawal et al., 2011, Agrawal et al., 2014), as well as broader and individual aspects of various substance use (Weatherhill et al., 2015). The original SAGE data set contains 4,121 individuals, wherein nearly 25% of individuals are related (i.e., family-based recruitment). Because family members tend to be more genetically similar to one another than non-related individuals (and thus would inflate frequency of genetic markers), all family-based recruited individuals within SAGE were excluded for this study.

Because of the complex phenotypic (i.e., dependency status) data in the SAGE set, very few individuals qualify as a direct match to the individuals in the archived data set. As previously noted, the SAGE study was initially designed as a study of alcohol use disorders (AUD), thus the recruitment primarily focused on AUD individuals (and families) and control individuals. However, there was one exception to the recruitment of control individuals in SAGE: they could be nicotine dependent. Thus the control group in SAGE exists as two groups: non-using controls, and exclusively nicotine using "controls". Additionally, during the recruitment stages of SAGE, many AUD individuals were also co- or multi-use or dependent in other ways (alcohol plus cannabis, nicotine, cocaine, opioids, and/or "other").

But even with the complex (and often multi-) dependency statuses of SAGE individuals, it was still possible to create "roughly matched" (to the archived data) groups. Control individuals—for this study—had to be recruited as a control individual *and* had no dependency status otherwise. The nicotine individuals were required to have nicotine dependence but not marijuana dependence status. The marijuana individuals were required to have marijuana dependence but not nicotine dependence status. The marijuana+nicotine co-use individuals were required to have dependence for both. At this point, there are 5 mutually exclusive groups: control (CON), nicotine (NIC), marijuana (MJ), marijuana+nicotine co-use (MJ+NIC), and "OTHER" which were comprised of individuals with no {marijuana or nicotine} dependence, and {alcohol or cocaine or opioids or "other"} dependence. For this configuration, only some individuals in the NIC group were not alcohol dependent, though all individuals in the marijuana and marijuana+nicotine groups were also alcohol dependent.

In the end, the groups were as such: CON (*n* = 1,499), MJ (*n* = 136), MJ+NIC (*n* = 510), NIC (*n* = 1,140), and OTHER (*n* = 454; comprised of individuals with alcohol or cocaine or opioid or "other" dependency, but not marijuana or nicotine). For a detailed list of demographics and dependency in the SAGE sample (for this study), see Table 6.1b and 6.1c. Because of the study configurations, either a set of "matched" individuals (CON, MJ, MJ+NIC, NIC; *N* = 3,285) or all individuals (CON, MJ, MJ+NIC, NIC, OTHER; *N* = 3,739) qualified for analyses, however these numbers fluctuate slightly based on which analysis was performed (due to preprocessing and quality control steps; for more details on these analyses, see Section 6.3).

The SAGE data available for download is a relatively complete data set, with only a small number of missing genotypic markers. Like in the archived data, missing data in SAGE were imputed to the grand mean.

6.2 Proposed and “flipped” analysis pipelines

The discovery and validation data sets are completely unique from one another and are henceforth referred to as “archived” (discovery set) and “SAGE” (validation set) respectively. Furthermore, the validation set (SAGE) was completely sequestered until all analyses were *completed and finalized* within the discovery set; this procedure precludes the inclusion of bias and ensures that the validation set is independent of the archived set. Therefore, both sets were preprocessed and analyzed separately, to ensure the integrity and independence of a predictive set. However, the sequestration did create some issues with respect to preprocessing and confound correction (which happens in the cases of true, independent replications). These issues are described in later sections.

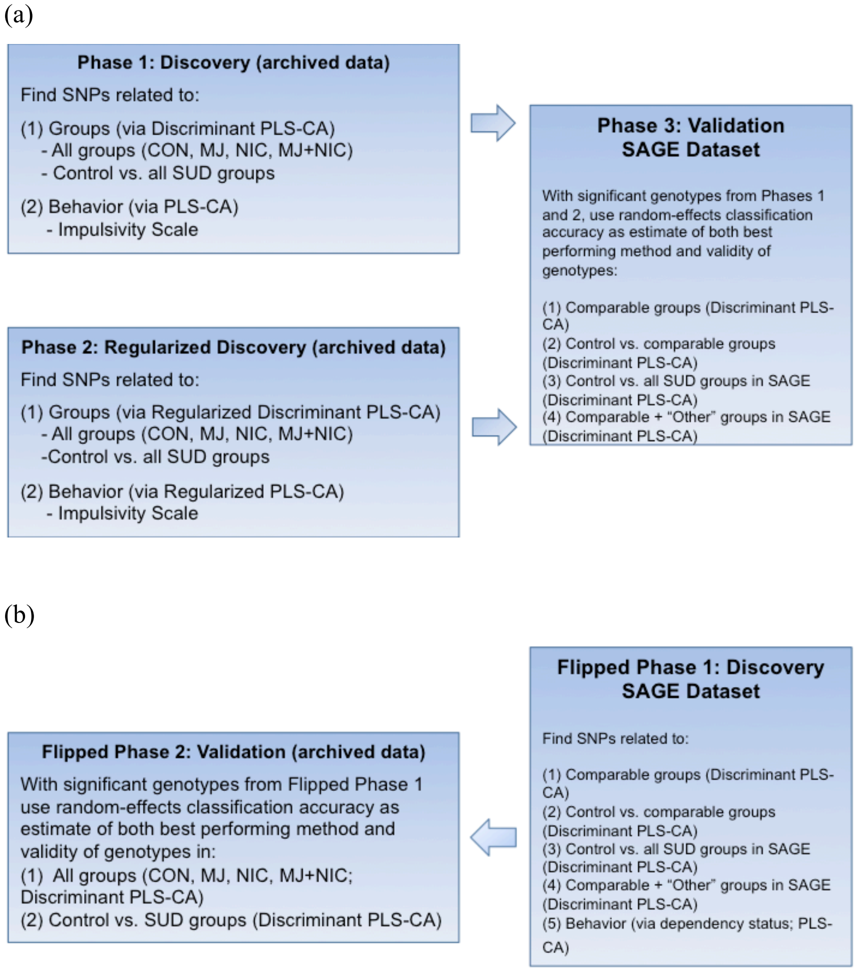
Figure 6.1 provides a schematic of the analysis pipelines. Figure 6.1a shows the original proposed pipeline with the archived data as a discovery set, and the SAGE data as a validation set. The motivation for small → big discovery validation was to see if PLS-CA—in standard or regularized forms—could utilize small data effectively. The motivation behind the small data set as the discovery set is because a major issue within many fields right now is that sample sizes are regarded as “too small” to analyze large scale genetic data. Therefore, if standard or SmooPLS-CA provides generalizable results (validated in SAGE), then we have a framework to analyze high dimensional, low sample size genetic association data.

However, as previously noted, the analyses in Phase 3 yielded very low prediction accuracies, thus implying that the genotypes discovered in Phases 1 and 2 were not sufficient contributors to SUDs. Because the low prediction accuracy in Phase 3 implied that either (1) there are no genetic contributions to SUDs or (2) something could be wrong in the analysis pipeline, I wanted to test a more typical pipeline: discovery with the large set (SAGE; $N > 3,285$) and validation with the small set (archived; $N = 431$). Figure 6.1b shows this alternative ("flipped") pipeline. If the models in Phases 1 and 2 were bad *because of the small sample*, then prediction accuracy in the archived set should be high (or at least higher) based on models built in the SAGE set.

But prediction accuracies in this flipped pipeline were also very low. The details of the results—and how the apparent failure to predict did not indicate a failure to identify replicable genotypes—are discussed in later sections.

6.2.1 Proposed Phases 1 and 2: Discovery with standard and SmooPLS-CA

Phases 1 and 2 included three base analyses on an archived data set with standard (Phase 1) and smoothed (Phase 2) PLS-CA. Phase 1 acts as a baseline for analyses in Phase 2. Standard and SmooPLS-CA were performed on three configurations of the data between a set of candidate SNPs and: (1) a case-control analysis: CON vs. SUDs, (2) a multi-group analysis: CON vs. MJ vs. MJ+NIC vs. NIC, and (3) a trait-based analysis: the ImpSS & BIS (with no explicit group coding). These three configurations help answer the question: what is the best way to find genetic markers of SUDs? Phase 2 repeated these analyses, but with a regularization parameter applied to PLS-CA: $\lambda = [1, 2, 3, 4, 5, 10, 15, 20, 25, 100]$.



Phase 1: Discovery (archived data)

Find SNPs related to:

(1) Groups (via Discriminant PLS-CA)
- All groups (CON, MJ, NIC, MJ+NIC)
- Control vs. all SUD groups

(2) Behavior (via PLS-CA)
- Impulsivity Scale

Phase 2: Regularized Discovery (archived data)

Find SNPs related to:

(1) Groups (via Regularized Discriminant PLS-CA)
- All groups (CON, MJ, NIC, MJ+NIC)
-Control vs. all SUD groups

(2) Behavior (via Regularized PLS-CA)
- Impulsivity Scale

Phase 3: Validation SAGE Dataset

With significant genotypes from Phases 1 and 2, use random-effects classification accuracy as estimate of both best performing method and validity of genotypes:

(1) Comparable groups (Discriminant PLS-CA)
(2) Control vs. comparable groups (Discriminant PLS-CA)
(3) Control vs. all SUD groups in SAGE (Discriminant PLS-CA)
(4) Comparable + "Other" groups in SAGE (Discriminant PLS-CA)

Flipped Phase 1: Discovery SAGE Dataset

Find SNPs related to:

(1) Comparable groups (Discriminant PLS-CA)
(2) Control vs. comparable groups (Discriminant PLS-CA)
(3) Control vs. all SUD groups in SAGE (Discriminant PLS-CA)
(4) Comparable + "Other" groups in SAGE (Discriminant PLS-CA)
(5) Behavior (via dependency status; PLS-CA)

Flipped Phase 2: Validation (archived data)

With significant genotypes from Flipped Phase 1 use random-effects classification accuracy as estimate of both best performing method and validity of genotypes in:
(1) All groups (CON, MJ, NIC, MJ+NIC; Discriminant PLS-CA)
(2) Control vs. SUD groups (Discriminant PLS-CA)

Figure 6.1 (a) Proposed Studies Pipeline Schematic, and (b) "Flipped" Studies Pipeline Schematic. Panel (a) shows the schematic of the (intended) three analysis Phases. Phases 1 and 2 are identical, except that Phase 2 uses regularized PLS-CA methods. The results from Phases 1 and 2 are then tested in a validation set, where the external set will be used to (1) if the results identified in each analyses can be repeated in an external sample, and (2) identify which method in Phases 1 and 2 was the best method (via highest classification accuracy in Phase 3, based on results provided from Phases 1 and 2). Panel (b) shows an alternative to the proposed pipeline, where the larger data set (SAGE) was used as the discovery set, and the smaller set (archived) was used as validation. The "flipped" pipeline is essentially the same as the proposed, except there are no regularization analyses.

Recall that λ = 0 is standard PLS-CA. SmooPLS-CA answer the questions: Does regularization help identify better genetic markers of SUDs compared to standard PLS-CA? If so, to what degree does the model need to be regularized (i.e., which λ), and how does this regularized model differ from standard PLS-CA?

Bootstrap resampling was used to identify which genotypes were significant contributors to each model, and thus those SNPs were selected for follow up analyses in Phase 3. In all analyses for Phase 1 and Phase 2, a “likely best” analysis was to be identified by highest (relative) random-effects prediction accuracy *within the archived data*. Prediction accuracy in Phases 1 and 2 was assessed with split-half resampling in part because it has a distinct advantage over other prediction frameworks for small samples: interval estimates of predication accuracy through the course of resampling. However, the actual tests of Phases 1 and 2 analyses are the prediction accuracies in Phase 3 (a validation set).

6.2.2 Proposed Phase 3: Validation on external data (with standard PLS-CA)

Phase 3 used standard PLS-CA in its discriminant analysis form (Beaton, Dunlop, et al., 2016) on an external data set (SAGE) to both validate which genotypes are likely contributors to SUDs, and also to identify which configuration from Phases 1 and 2 produced the best set of predictive genetic markers. Only SNPs that were identified in Phases 1 and 2 were used in Phase 3. To note, the genotypes in the validation phase were coded as they exist in the SAGE set, and were not coded to match the genotypes as they existed (via preprocessing) in the archived set. This was in part because the SAGE set was completely sequestered during Phases 1 and 2, thus making it a completely independent set from the archived set. Because of this sequestration, an exact prediction of one data set from the coefficients of the other (SAGE from archived or vice versa) was not possible because of a particular step in preprocessing, which had to be applied to each set independently: confound correction (for race and ethnicity). Additionally, it is best at this stage of research inquiry to leave the data in a genotypic model for both discovery and prediction sets in order to identify—instead of assume—the most likely genetic effects (e.g., additive, dominant, recessive) in SUDs.

Phase 3 was performed on the SAGE data set with group configurations comparable to the archived data set, and then expanded to include other SUDs. This is because SAGE data were initially collected as a way to identify genetic markers of alcohol use disorder (AUD; Beirut et al., 2010). However, as previously noted, SAGE has been used for a variety of other disorders.

Phase 3 used leave-one-out (LOO) cross validation—instead of split-half—to assess random effects prediction accuracy. LOO works by predicting each individual—in turn—from an analysis that they were left out of. In this case, LOO provides an overestimate of prediction error due in part because it maximizes the size of the "training set" (Hastie, Tibshirani, & Friedman, 2013). Therefore, LOO in this case provides us with a likely "worst case scenario" of a real world application: how likely could we predict SUD from a given panel of genetic markers? Finally, bootstrap resampling was also used in Phase 3 to identify which, if any, genetic markers remain significant in Phase 3, and are thus more likely to generalize to the broader population of SUDs (both specific to the groups studied here, and in general).

6.2.3 Flipped pipeline

As previously noted, the prediction accuracies in Phase 3 were relatively low (see Section 6.4 for details, and Table 6.6). Typically, studies that use independent discovery and validation samples often use the larger set for discovery and the smaller set for validation. In my dissertation, I made the choice to use the smaller specifically to test if PLS methods are sufficient to detect effects in small samples. However, as previously noted, the failure of prediction could have been because of the small sample in the archived data or an overall issue with the analysis pipeline. Thus, to test if this were the case, I "flipped" the pipeline. Furthermore, I also excluded regularization analyses in this "flipped" pipeline, because the SAGE data set should not require the "boost" provided by regularization. The procedure to select SNPs—described in Section 6.1.1—was also applied to the SAGE data set, and (because the chipsets were different) yielded a different set of SNPs than the archived set. For additional details on the SNPs selected for all analyses, see Section 6.3.

A schematic of the flipped version of the pipeline can be found in Figure 6.1b. This flipped pipeline included slightly different analysis configurations to ask the same questions as in the proposed pipeline. "Flipped" Phase 1 was standard PLS-CA run on five configurations of the SAGE data: (1) a "matched" case-control (CON vs. matched SUDs), (2) case-control (CON vs. all SUDs), (3) a "matched" multi-group analysis (CON vs. MJ vs. MJ+NIC vs. NIC), (4) a "matched plus" multi-group analysis (CON vs. MJ vs. MJ+NIC vs. NIC vs. "other"), and (5) a dependency-based analysis where, instead of exclusive group classification, the actual dependency status was used; dependency status for all individuals was a "yes" or "no" across 6 categories: alcohol, marijuana, nicotine, cocaine, opioids, and "other." In the "flipped" Phase 1, only bootstrap resampling was performed in order to identify which genotypes were significant. The SNPs that corresponded with these genotypes were then extracted from the archived data, and—similar to proposed Phase 3—prediction accuracy of these SNPs were tested with LOO in two configurations: (1) archived CON vs. SUD and (2) archived multi-group.

Not only were the prediction accuracies in “flipped” Phase 2 low, but they were generally worse than in the proposed pipeline (archived → SAGE). Details from the analysis pipelines can be found in Section 6.4 and onward, as well as the supplemental/appendix.

6.3 Quality control, preprocessing, and data preparation

Quality control and preprocessing were performed with `plink` (Purcell, 2007) and custom made code in `R` (R Core Team). All sex-linked SNPs (i.e., Chromosomes X and Y) were excluded. Standard qualities control parameters were the same for, and performed strictly within, their respective data set (archived or SAGE), for only the extracted subsets of SNPs: both observations and SNPs had to have ≥ 90% call rates (i.e., complete data). SNPs had to have a minor allele frequency ≥ 5%. Because SNPs were recoded in a disjunctive format, there was an additional step for preprocessing. If a genotype (i.e., either the heterozygote or the minor homozygote) had less than 5% frequency it was combined with another genotype. If the minor homozygote was rare (≤ 5%), it was simply combined with the heterozygote (i.e., the dominant model). However, if the heterozygote was rare (≤ 5%), it was essentially treated as variant of “missing” data, and individuals with the heterozygote were recoded [.5 .5], thus giving equal weight to the major and minor alleles.

Because individuals within each data set are from diverse racial and/or ethnic backgrounds, the data required stratification correction based on race and ethnicity *for each individual set of data*. Not only did this occur on the large sets (proposed Phases 1 and 2, and flipped Phase 1), but was also applied within the validation sets because different SNPs can be confounded with race or ethnicity in different ways. Typically, race and ethnicity confound correction is done on the additive-model coded data (i.e., [0, 1, 2]) and an "eigen stratification" is performed. This "eigen stratification" is simply a PCA or MDS applied to the (presumed) quantitative data, and then components are individually inspected for confounding effects. Components that show race or ethnicity effects are then regressed out of the data (or the data are reconstructed from a lower rank representation of the data excluding those components). Because the data here are coded in a disjunctive form, correspondence analysis (CA) was used in place of PCA or MDS, and the same procedure (i.e., removal of components showing a confound) was performed. At their largest (i.e., the "discovery" versions of) the archived and SAGE data sets had 3,381 and 4,781 SNPs, respectively. The number of SNPs for the validation phases varied (so that each initial discovery set could be tested) but never exceeded 255 SNPs (recall the validation phases, see Figure 6.1).

6.3.1 Archived "discovery" set

Of the 709,362 original SNPs in the archived data there were a total of 4,472 associated with the 149 gene candidate panel. After quality control and removal of sex-linked SNPs, there were a total of 3,381 SNPs. The first 5 components were removed for race and/or ethnicity effects. The final matrix analyzed was 431 (participants) × 8771 (disjunctive genotypes).

6.3.2 SAGE "discovery" set

Of the 1,040,107 original SNPs in the SAGE data there were a total of 5,053 associated with the 149 gene candidate panel. After quality control and removal of sex-linked SNPs, the total number of SNPs were between 4,777 (for matched individuals) and 4,781 (for matched + "other" individuals), with 12,392 or 12,406 disjunctive genotypes (columns), respectively. The first component was removed for race effects; no other components showed a confounding effect.

6.4 Overview of results

Because the majority of analyses (i.e., almost all of the regularized) did not differ much from one another and were similar to the baseline, and because the prediction accuracies were quite low for both the proposed and flipped pipelines, only a truncated version of the results are presented here. This section provides a high-level overview of all the analyses performed, and segues into a discussion section.

6.4.1 Proposed Pipeline

A truncated version of the results from Phases 1 and 2—because they were the same techniques and configurations, but with different λ parameters—is presented here.

6.4.1.1 Phases 1 and 2

Figure 6.2 presents the prediction accuracies (Fig. 6.2a) and number of significant bootstrap ratios (Fig. 6.2b) for the case-control (i.e., CON vs. SUD) configuration. Figure 6.2a shows that there was virutally no change in prediction accuracy as λ increased; in fact, the "best" model (via random-effects split-half resampling) was λ=2 and, on average, performed close to baseline, even though baseline had a higher lower bound estimate. All random-effects prediction accuracies were above chance. Figure 6.2b shows the total number of significant bootstrap ratios, which remained virtually unchanged at the more stringent threshold of |BSR| > 3. In sum, regularization had virtually no effect on the prediction accuracy. Regularization also had virtually no impact on the number of, as well as which genotypes were identified as significant.

The positive side of Component 1 (which was the only component in this analysis) was associated with the CON group, where the negative side was associated with the SUD group. Table 6.2 shows just the top (|BSR| > 3.5 in the baseline λ=0 analysis) genotypes for the case-control configuration, where a negative value means a genotype was more associated with the SUD group than the CON group. Note that the BSR values are also virtually identical for λ = 0, and the average from λ > 0. A larger table with all significant genotypes at |BSR| > 3 can be found in the online supplemental material.

Table 6.2

Significant Genotypes from Archived Two Group Analyses.

SNP	**Genotype**	**BSR λ=0**	**BSR Mean λ > 0**	**Chr.**	**Gene**
rs7188171	TT	-4.079	-3.924	16	GRIN2A
rs1396860	AA	-3.84	-4.028	11	CCKBR
rs1396860	GA+GG	3.84	4.028	11	CCKBR
rs173766	AA	-3.826	-3.787	5	GABRB2
rs12441474	GG	-3.791	-3.81	15	*GABRB3*
rs2941023	GG	-3.775	-3.734	11	CCKBR
rs2929183	GG	-3.616	-3.585	11	CCKBR
rs3846448	CC	-3.653	-3.445	4	*ADH1C*
rs3846448	TC+TT	3.653	3.445	4	*ADH1C*

Note. Top significant genotypes from the two-group configuration (ranked by bootstrap ratio value). Even though the selection criteria for Phase 3 was |BSR| > 3, the threshold for this table was |BSR| > 3.5 in the baseline (i.e., λ=0) analysis. The full list of significant genotypes are available in the electronic supplemental material. Genes identified by NCBI2R package. Those in *italics* are closest genes through other resources (e.g., ALFRED). BSR λ=0 is the bootstrap ratio value for the baseline analysis, where BSR Mean λ > 0 is the average BSR value across all λ>0, because no single λ appeared to outperform one another or baseline (see Figure 6.3).

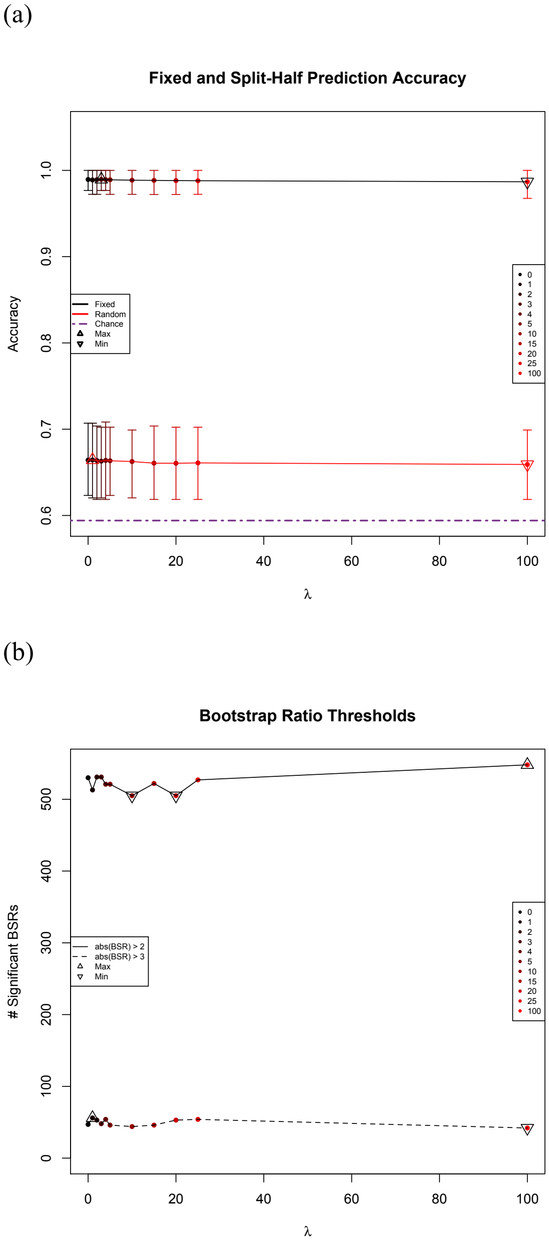


Figure 6.2 *Two group analysis.* BSR = Bootstrap ratio. Inference effects in the two-group configuration in archived data set. Figures show (a; top) split-half prediction (fixed and random) accuracies with error bars, and (b; bottom) shows the number of significant (via BSR) genotypes as λ increases for a threshold of |BSR| > 2 and |BSR| > 3.

Figure 6.3 presents the prediction accuracies (Fig. 6.3a) and number of significant bootstrap ratios (Fig. 6.3b) for the four group (i.e., CON vs. MJ vs. MJ+NIC vs. NIC) configuration. Random-effects prediction accuracies were above chance. This configuration provides the most striking example of the failure of regularization: Not only did prediction accuracy decrease as λ increased, but the value of λ = 0 gave the highest prediction accuracy for *both* fixed- and random-effects (Fig. 6.3a). The results from the significant number of BSRs shows that regularization more than likely just "pushed" the noise from high variance components to lower variance components, rather than eliminate noise. Baseline and regularized analyses were still virtually identical on prediction accuracy as well as which genotypes were identified as significant. From the four-group configuration, Component 1 was essentially the CON (positive side) vs. other groups (negative side). As λ increased, Component 1 become more strictly CON (positive side) vs. MJ (negative side). Component 2 was more associated with the NIC group (negative side), and Component 3 was more associated with the MJ+NIC group (negative side). However, these same associations can be found via the original component scores, and the BSRs for each group in just the baseline analysis. Table 6.3 shows just the top genotypes (|BSR| > 3.5, in the baseline λ=0 analysis) for the four-group configuration. Note that the BSR values are also virtually identical for λ=0, and the average from λ>0. However, for Component 1, the λ=0 BSRs have a greater magnitude than λ>0, where the opposite is generally true for Components 2 and 3. A larger table with all significant genotypes at |BSR| > 3 can be found in the online supplemental material.

Table 6.3

Significant Genotypes from Archived Four Group Analyses.

		BSR	**BSR Mean**	**BSR**	**BSR Mean**	**BSR**	**BSR Mean**		
SNP	**Geno.**	**λ=0 C1**	**λ>0 C1**	**λ=0 C2**	**λ>0 C2**	**λ=0 C3**	**λ>0 C3**	**Chr**	**Gene**
rs7188171	TT	-4.752	-4.047					16	GRIN2A
rs10217351	TT			4.236	3.857			9	MPDZ
rs2036108	AA					3.833	4.179	8	ADRA1A
rs2941023	GG	-3.983	-3.8					11	CCKBR
rs17683096	AA	-4.264	-3.47					16	GRIN2A
rs35586628	TT			-3.736	-3.959			15	GABRA5
rs220597	GG					3.523	3.957	12	GRIN2B
rs2929183	GG	-3.794	-3.641					11	CCKBR
rs2350786	AG			-3.685				7	CHRM2/ LOC349160
rs28426996	GG			-3.523	-3.757			15	GABRA5
rs1872688	AA	3.672	3.607					16	ADCY7
rs2113545	AG			-3.587				7	CHRM2/ LOC349160
rs10500373	CC	-3.762	-3.407					16	LOC105371076/ GRIN2A
rs9934026	TT	3.503	3.558					16	ADCY7
rs17087959	TT					-3.605	-3.445	9	NTRK2
rs17087959	CT+CC					3.605	3.445	9	NTRK2
rs2070673	AA					3.52	3.514	10	CYP2E1
rs6535594	GG			-3.502				4	NR3C2
rs9926046	TT	-3.888	-3.075					16	GRIN2A
rs17750208	AA	-3.518	-3.232					16	GRIN2A
rs362817	CC	3.508	3.133					6	GRM1

Note. C1 = Component 1, C2 = Component 2, C3 = Component 3.Top significant genotypes from the four-group configuration. Even though the selection criteria for Phase 3 was |BSR| > 3, the threshold for this table was |BSR| > 3.5 in the baseline (i.e., λ=0) analysis. The full list of significant genotypes are available in the electronic supplemental material. Genes identified by NCBI2R package. BSR λ=0 is the bootstrap ratio value for the baseline analysis, where BSR Mean λ > 0 is the average BSR value across all λ>0, because no single λ appeared to outperform one another or baseline (see Figure 6.3).

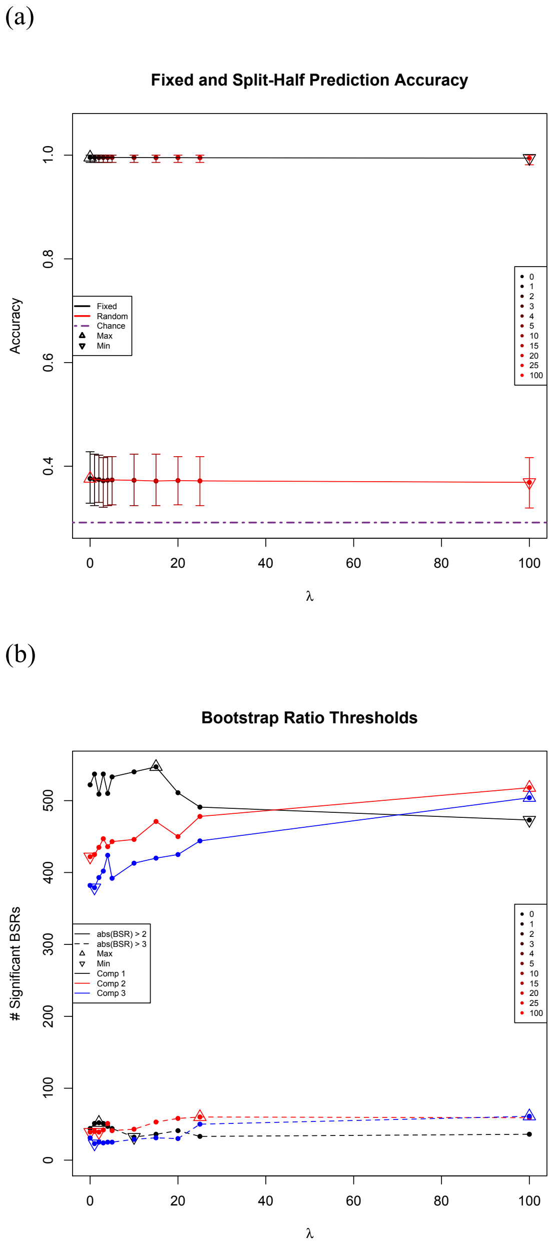


Figure 6.3 *Four group analysis.* BSR = Bootstrap ratio. Inference effects in the four-group configuration in archived data set. Figures show (a; top) split-half prediction (fixed and random) accuracies with error bars, and (b; bottom) shows the number of significant (via BSR) genotypes (or all components) as λ increases for a threshold of |BSR| > 2 and |BSR| > 3.

Figure 6.4 presents the prediction accuracies (Fig. 6.4a, b, c) and number of significant bootstrap ratios (Fig. 6.4d) for the impulsivity-based configuration. All fixed-effects prediction accuracies were above chance. For prediction—either with two groups, seen in Fig. 6.4a, or with four groups, seen in Fig. 6.4b—within the genetic latent variables alone, all (average) random-effects prediction accuracies were below chance. For the joint (bi-dimensional) latent variables (one from impulsivity, one from genetics; see Figure 6.4c), prediction accuracy was above chance for fixed- and random-effects but only for the first component. Interestingly, when prediction was performed just within the genetic latent variables (Fig. 6.4a and b), as λ increased, performance generally decreased. The opposite is true for prediction within the joint latent variables: A higher λ increased prediction, but only for the first component. This effect maybe due to group association being highly collinear and predictive within impulsivity (see Beaton et al., 2014). Figure 6.4d shows the number of significant bootstrap ratios per component. Again, it is clear that regularization had little to no effect on the analysis. Table 6.4 shows just the top (|BSR| > 3.5, in the baseline λ = 0 analysis) genotypes for the impulsivity configuration. Because an increased λ showed some effect on the first component, Table 6.4 shows BSRs for λ = 0, λ = 100, and the average where λ > 0 and λ < 100. Like previous analyses, the BSRs are virtually identical for changes in λ. The impulsivity analysis clearly performed much worse than the group-level analyses, and only identified four genotypes with |BSR| > 3.5; these four genotypes are from 2 SNPs associated with the same gene: GABRB2. A larger table with all significant genotypes at |BSR| > 3 can be found in the online supplemental material.

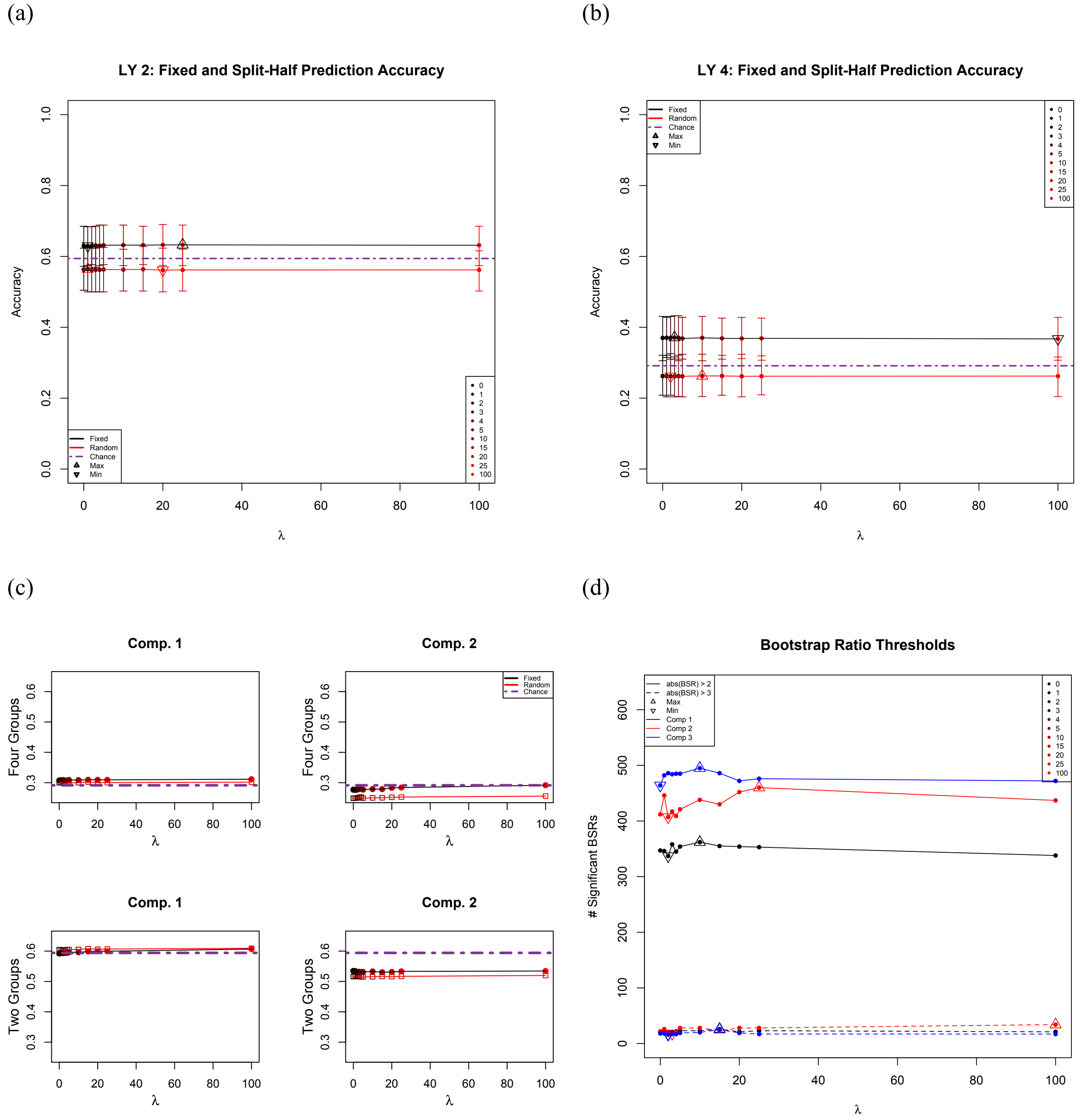


Figure 6.4 *Impulsivity Analyses.* BSR = bootstrap ratio. CON = control group. SUD = Substance use disorder (group). Inference effects in the impulsivity configuration in archived data set. Panels (a) and (b) show split half prediction accuracies (with error bars): (a; top left) shows prediction estimates for CON vs. SUD in the impulsivity genotype subspace, where (b; top right) shows prediction estimates for the four groups in the impulsivity genotype subspace. Panel (c; bottom left) shows prediction accuracy per latent variable (combination of impulsivity

and genotype scores) for the first two components in each configuration (four groups on top, two groups on the bottom). Note that only the first component appears to show an effect across λ, though there appears to be a slight effect of regularization (λ=100) for CON v. SUD. Panel (d) shows the number of significant (via BSR) genotypes (or all components) as λ increases for a threshold of |BSR| > 2 and |BSR| > 3.

Table 6.4

Significant Genotypes from Archived Impulsivity Analyses.

		BSR	**BSR**	**BSR Mean**		
SNP	**Genotype**	**λ=0 C1**	**λ=100 C1**	**λ>0 & λ<100 C1**	**Chr.**	**Gene**
rs13362062	CT+CC	3.907	3.862	3.871	5	GABRB2
rs13362062	TT	-3.907	-3.862	-3.871	5	GABRB2
rs4921390	CT+CC	3.794	3.742	3.751	5	LOC105377694,GABRB2
rs4921390	TT	-3.794	-3.742	-3.751	5	LOC105377694,GABRB2

Note. C1 = Component 1. Top significant genotypes from the impulsivity configuration (ranked by bootstrap ratio value). Even though the selection criteria for Phase 3 was |BSR| > 3, the threshold for this table was |BSR| > 3.5 in the baseline (i.e., λ=0) analysis. The full list of significant genotypes are available in the electronic supplemental material. Genes identified by NCBI2R package. All significant SNPs are associated with GABRB2. Only BSRs of the first component are presented, but in this particular analysis, they are presented for BSR λ=0, BSR λ=100, and the mean of λ that was not 0 or 100. Only λ=100 appeared to show a small advantage (via prediction) over other λs, though, according to the bootstrap results presented here, provides no advantage to detect stable and significant genotypes (see also Figure 6.4).

6.4.1.2 Common SNPs in Phases 1 and 2

Even though some analyses identified specific genotypes (i.e., different genotypes than the other analyses), there were some common genotypes across analyses. Table 6.5 presents all genotypes that were significant at |BSR| > 3—in either standard or any regularized condition—in *at least two* configurations (case-control, multi-group, or impulsivity-based) from Phases 1 and 2. While the CON vs. SUD and multi-group analyses produced only 1 and 3 components respectively, the impulsivity based analyses produced many more. However, as noted in the previous section, only the first component of the impulsivity analysis was interpreted (i.e., a higher λ increased prediction, but only for the first component).

Nearly all of the common genotypes were identified from the opposition between the CON and SUD groups. Furthermore, only 1 genotype was identified in the impulsivity-based analysis that was common with any another analysis. While there are a number of different genotypes across genes, in the archived data there appear to be only a few core genes that reflect a dissociation of CON vs. SUD, namely: GABRB2, GABRG2, GABRB3, CCKBR, GRIN2B, and GRIN2A. Finally, one gene in particular—but via different genotypes—came up in all analyses: PRKCE.

Table 6.5

Significant Genotypes Across Archived Analyses

				Two Groups		*Four Groups*		*ImpSS*
SNP	**Geno.**	**Chr.**	**Gene**	**Comp. 1**	**Comp. 1**	**Comp. 2**	**Comp. 3**	**Comp. 1**
rs1562653	AA	2	PRKCE	1	1			
rs751237	GG	2	PRKCE	1	1			
rs642200	CC	2	PRKCE				1	1
rs3846448	CC	4	*ADH1C*	1			1	
rs3846448	TC+TT	4	*ADH1C*	1			1	
rs1229979	CC	4	*ADH1C*	1			1	
rs1662037	GG	4	*ADH1C*	1			1	
rs1662033	TT	4	ADH1C	1			1	
rs17024437	AG+AA	4	NR3C2	1		1		
rs17024437	GG	4	NR3C2	1		1		
rs173766	AA	5	GABRB2	1	1			
rs365054	CC	5	GABRG2	1	1			
rs1396860	AA	11	CCKBR	1	1			
rs1396860	GA+GG	11	CCKBR	1	1			
rs2941023	GG	11	CCKBR	1	1			
rs2929183	GG	11	CCKBR	1	1			
rs7297761	GG	12	GRIN2B	1	1			
rs12441474	GG	15	*GABRB3*	1	1			
rs7188171	TT	16	GRIN2A	1	1			
rs9927871	TT	16	GRIN2A	1	1			
rs17683096	AA	16	GRIN2A	1	1			
rs17670276	CC	16	GRIN2A	1	1			
rs17670276	AC+AA	16	GRIN2A	1	1			
rs9926046	TT	16	GRIN2A	1	1			
rs4552023	CC	16	GRIN2A	1	1			
rs10500373	CC	16	GRIN2A	1	1			

Note. Common (significant) SNPs where at least 2 of the 3 configurations (two group, four group, impulsivity) identified a genotype as significant where either λ=0, or average λ where |BSR| > 3. This threshold choice was because regularization appeared to provide no substantial advantage; bootstrap estimates were relatively robust across all λ.

6.4.1.3 Predictions in Phase 3

There were two goals in Phase 3, both considered a form of validation: (1) identify which analysis from Phases 1 and 2 had the highest prediction (a "winning model") in Phase 3, and (2) which genotypes are also significant from the "winning model".

Because the standard and regularized results were identical in Phases 1 and 2 for all configurations, many SNP panels were made from the results. First, panels were derived from each configuration, where SNPs were selected if they had a genotype (in Phases 1 and 2) with |BSR| > 3. Next, two panels were derived from each configuration: a standard set and a regularized set. Because no individual regularized iteration clearly outperformed any other, the regularized panels were derived from the average |BSR| value across λ = [1, 2, 3, 4, 5, 10, 15, 20, 25, 100]; the threshold was still also |BSR| > 3. At this stage, there would have been 6 separate panels to test. However, because all results were virtually identical, I created two more panels *within* each study configuration (i.e., within case-control, within multi-group, within impulsivity-based): the intersection of standard and regularized SNPs with genotypes of |BSR| > 3, and the union of standard and regularized SNPs with genotypes of |BSR| > 3. One final panel was created: a "full" panel, where any SNP that was included in the previously mentioned panels was included in one large panel. Furthermore, each panel was then tested on 4 possible configurations of the external data: (1) CON vs. matched SUDs, (2) CON vs. matched MJ vs. matched MJ+NIC vs. matched NIC, (3) con vs. SUDs, and (4) CON vs. matched MJ vs. matched MJ+NIC vs. matched NIC vs. "other." Thus, there were a total of 52 possible validation analyses. The prediction accuracy of these analyses is shown in Table 6.6.

As previously noted, prediction accuracies—especially the LOO accuracies—for Phase 3 were very low. In general, the few LOO accuracies that were above chance were panels identified in the case-control, standard PLS-CA analyses in Phases 1 and 2, and tested best on case-control analyses in Phase 3. Furthermore, the best Phase 3 configuration appeared to be the "matched case"-control analyses, as this SAGE configuration had above chance LOO accuracies for all archived case-control panels, as well as the final "full" panel. However, these accuracies were barely above chance (chance = 50.37%, best LOO was 51.25%). Because each of these panels was small, and because the prediction accuracies were so low, it was not worthwhile to identify significant genotypes in the external set.

6.4.2 Flipped Pipeline

Because the results in the proposed pipeline yielded such poor predictive accuracy, an alternative "flipped" pipeline was performed. This flipped pipeline did not include any regularization (see Figure 6.1b). Furthermore, the "flipped" pipeline was, essentially, a test of whether building models on small data was not sufficient for prediction in a larger data set.

Table 6.6

Prediction accuracy from archived to SAGE.

		Archived Configuration					
		Two Group		Four Group		ImpSS	
SAGE Configuration	Panel	FIXED	LOO	FIXED	LOO	FIXED	LOO
CON v. {MJ, NIC, MJ+NIC}	$\lambda=0$	53.811*	50.9756*	55.5793*	51.25*	53.7805*	50.5793*
Chance: 50.3696	ave($\lambda>0$)	53.2317*	50.7927*	54.6037*	49.939	51.2805*	46.3415
	Union	53.6585*	50.6402*	55.8841*	51.0366*	53.3841*	50.3659
	Intersection	53.4451*	50.7622*	53.628*	50.1524	50.5488*	45.8232
Full :							
FIXED: 56.4939*							
LOO: 50.4878*							
CON v. MJ vs. NIC vs. MJ+NIC	$\lambda=0$	27.9268	24.2683	29.878	24.1159	25.061	21.7988
Chance: 35.4683	ave($\lambda>0$)	26.5854	23.3232	27.378	22.4695	21.9512	19.4207
	Union	27.5	23.7195	30.6402	24.2378	24.3293	21.372
	Intersection	26.9817	23.811	26.189	22.4085	21.8598	19.7561
Full :							
FIXED: 32.439							
LOO: 24.4817							
CON v. SUD	$\lambda=0$	53.2815*	51.4332	54.7013*	51.1117	53.2012*	50.4152
Chance: 51.9383	ave($\lambda>0$)	53.3351*	51.1117	54.3263*	50.6295	51.1385	47.6025
	Union	53.5494*	51.0046	56.0139*	50.442	52.9869*	49.8795
	Intersection	53.4155*	51.4064	53.3619*	49.9866	50.4956	47.0131
Full :							
FIXED: 55.88*							
LOO: 50.817							
Five Groups	$\lambda=0$	21.9395	18.7249	24.6451	18.0284	20.4661	17.6266
Chance: 28.855	ave($\lambda>0$)	21.2965	18.1088	22.2073	16.3943	18.6981	15.7514
	Union	21.9127	18.2695	25.0737	18.3231	20.1179	16.8765
	Intersection	21.1894	18.0552	21.0019	16.7961	19.1267	16.0729
Full :							
FIXED: 26.2523							
LOO: 18.082							

Note. Phase 3 predictions (based on panels from Phases 1 and 2 in the archived data). Because SAGE had several configurations and there were three panels, there were many possible analyses to perform. Prediction accuracy is presented here for those possible analyses, where $\lambda=0$ indicates baseline, ave($\lambda>0$) indicates averaged regularization results, union indicates the union of $\lambda=0$ and ave($\lambda>0$), intersection indicates the intersection of $\lambda=0$ and ave($\lambda>0$) for each possible analysis. Full indicates the entire set of SNPs identified across all analyses. In general, only two group configurations presented results that were above chance (indicated with *).

6.4.2.1 Discovery with the SAGE data

The SAGE-as-discovery set included six analyses (see proposed pipeline, and Table 6.1): (1) "matched" case-control, (2) full case-control, (3) "matched" multi-group, (4) multi-group, (5) a dependency-based analysis (not yet discussed), and (6) the combined ("all") SNPs. Because each individual could—and in nearly all cases does—have multiple dependencies, a dependency-based analysis was also run. The dependency-based analysis exists somewhere between the impulsivity-based analysis in terms of structure (in Phases 1 and 2), and discriminant analyses; in the dependency case, each individual is allowed to belong to multiple groups.

In nearly all SAGE-as-discovery analyses, the same general effect as in the archived-as-discovery was observed: the first component generally dissociates control from SUD groups. For that reason, only the significant genotypes—with a threshold of |BSR| > 3.5—on the first component, across all analyses here, are presented in Table 6.7. Genotypes with |BSR| > 3—on the first component only, for each analysis—can be found in the online supplemental material.

Table 6.7

Significant Genotypes from the "SAGE-as-discovery" analysis

SNP	Geno.	Chr.	Gene	2M	4M	2	5	Dep
rs10865212	GT	2	PRKCE	-3.6335				
rs1008400	TC	16	FTO	4.386	3.8281			
rs11076017	TC	16	FTO	4.9859	4.5966	4.0686		
rs363538	CC	21	GRIK1	3.5502				
rs363510	TT	21	GRIK1	3.5678				
rs2832431	CC	21	GRIK1	3.6982				
rs10802779	TC	1	LOC105373225,CHRM3		-3.7362			
rs610529	CT	9	ALDH1A1		3.5609			
rs10772703	TC	12	GRIN2B		-3.7598			
rs220563	AA	12	GRIN2B		3.6608			
rs541098	AG	15	AVEN/CHRM5		3.5319	3.5742	4.5997	
rs6800622	CA	3	GSK3B			3.6912		
rs1732170	GA	3	GSK3B			3.7254		
rs1719894	GA	3	GSK3B			3.6468		
rs9813864	TC	3	GSK3B			3.6912		
rs6795653	TC	3	GSK3B			3.7889		
rs2319398	GT	3	GSK3B			3.7659		
rs13321783	TC	3	GSK3B			3.543		
rs6438552	AG	3	GSK3B			3.7718		
rs6792572	AC	3	GSK3B			3.6218		
rs9873477	AG	3	GSK3B			3.7307		
rs9878473	TC	3	GSK3B			3.8229		
rs7644234	GT	3	GSK3B			3.7643		
rs12630592	GT	3	GSK3B			3.7451		
rs17204878	GT	3	GSK3B			3.866		
rs334563	GT	3	GSK3B			3.7799		
rs334533	GA	3	GSK3B			3.8817		
rs7026417	CT+CC	9	NTRK2			-3.648		
rs7026417	TT	9	NTRK2			3.648		
rs980365	GG	12	GRIN2B			-3.8258		
rs3121819	GG	1	GABRD				3.6015	
rs3121819	AG+AA	1	GABRD				-3.6015	
rs6545976	TG	2					-3.5102	
rs4586906	GT	4	GABRB1				-3.6008	
rs1519480	CT	11	BDNF-AS				-3.6205	
rs1519480	TT	11	BDNF-AS				3.6417	
rs1984490	CT	1	FAAH					4.3623
rs6703669	TC	1	FAAH					4.4378

rs17361936	TC	1	FAAH	4.6422
rs17361950	TC	1	FAAH	4.1166
rs12624279	AA	2	FOSL2	3.5295
rs6719779	CT	2	PRKCE	-3.7158
rs6578750	GG	11	*CCKBR*	4.1651
rs7175581	AA	15	CHRNA7	3.7155

Note. Bootstrap ratios (BSRs) and corresponding information for genotypes on the first component across each SAGE analysis configuration (2M = Two matched groups, 4M = Four matched groups, 2 = Two groups, 5 = Five groups, Dep = Dependency analysis). Interestingly, some of the SNPs and genes found here were also found in the previous (archived) analyses, but had poor predictive accuracy in the SAGE data (see Table 6.6).

6.4.2.2 Validation with the archived data

Like the proposed pipeline, the flipped pipeline included a "validation" phase. For the flipped pipeline, panels were built with the SAGE data, and then tested on the archived data. Prediction accuracy—from discovery to validation—in the flipped pipeline was generally worse than the proposed pipeline (Table 6.8). This effect is striking because it is unexpected: usually models are built on large test sets—mostly because this increases power—and then tested on smaller sets, often yielding suitable predictions. Yet, the flipped pipeline (big set → small set) had worse prediction accuracy than the proposed pipline (small set → big set). Only two analyses—on two different configurations—showed prediction accuracy above chance: (1) the SAGE dependence model SNPs could predict (above chance) CON v. SUD in the archived data, and (2) the SAGE four matched group model could predict (above chance) the four groups in the archived data. See Table 6.8 for prediction accuracies.

Table 6.8

Prediction accuracies on archived (from SAGE).

	CON v. SUD			*Local Data*	*CON v. MJ v. MJ+NIC v. NIC*		
SAGE Configurations	**CHANCE**	**FIXED**	**LOO**		**CHANCE**	**FIXED**	**LOO**
CON v. {MJ, MJ+NIC, NIC}	59.4124	59.3968*	54.0603		29.1412	37.587*	27.1462
CON v. MJ v. MJ+NIC v. NIC	59.4124	68.2135*	55.6845		29.1412	47.7958*	29.6984*
CON v. SUD	59.3553	64.6512*	56.0465		29.2104	37.2093*	26.7442
CON v. MJ v. MJ+NIC v. NIC v. OTHER	59.3553	64.6512*	56.0465		29.2104	37.2093*	26.7442
Dependence	59.4124	73.7819*	61.7169*		29.1412	57.5406*	28.0742
Full	59.4124	81.2065*	58.4687		29.1412	67.0534*	27.8422

Note. Prediction accuracies on archived data from the SAGE models (use significant genotypes identified in SAGE data to predict group configurations in archived data). Because SAGE and the archived data sets each have several possible configurations, each are tested and shown here based on SAGE model (rows) and group configuration in the archived data (columns). Full indicates the entire set of SNPs identified across all analyses.

6.5 Are the panels really this bad?

The low prediction accuracies in both pipelines suggest the absence of predictive genetic markers of SUDs (within the proposed, large-scale panel; see Section 6.2-6.4) and therefore perhaps no genetic contributions to SUDs. Moreover, the proposed panel is generally comprised of genes that are (i) strongly suspected, or (ii) have already been shown in the literature, to contribute to SUDs (see Chapter 2); yet it appears that there was no prediction from the panels devised during the (proposed and flipped) discovery phases. Thus, the question then becomes: are the panels created in the discovery analyses really this deficient at predicting SUDs *compared to other panels*?

To answer this question, a final panel—based on independent (of this dissertation) analyses of SAGE—was created and then tested. For this panel, we turn to the work of Agrawal et al., (2011) and Agrawal et al., (2014). The SAGE data were used in both studies, but unlike my dissertation, Agrawal and colleagues used genome-wide data. Both studies emphasized cannabis use: via dependence (Agrawal et al., 2011) and via phenotypic characteristics (Agrawal et al., 2014). While in their work there was no traditionally significant GWAS SNP ($p < .05^{-8}$), Agrawal and colleagues did report top SNPs in both papers; though some of those SNPs do show relatively strong—albeit "not significant"—effects.

6.5.1 An external SNP panel

A final panel was created based on the top listed SNPs in Agrawal et al., (2011) and Agrawal et al., (2014). The SNPs in both studies by Agrawal and colleagues had no overlap with the proposed panel (see Chapter 2, and Sections 6.2-6.4), thus making this panel an excellent candidate for a final test of prediction in SUDs.

The goal was to validate the SNPs identified in the studies by Agrawal and colleagues in the archived data set. The primary validation criterion, like in all other validation steps, was based on predictive accuracy. Again, random-effects prediction accuracies were very low and in general, below chance (Table 6.9). However, with some creative restructuring of groups to match the intended studies of Agrawal and colleagues (i.e., cannabis use), two configurations were just above chance: (1) CON v. {MJ, MJ+NIC}, which excluded the NIC group entirely, and (2) CON v. MJ v. {NIC, MJ+NIC}, where a single "nicotine" group was made from the NIC and MJ+NIC groups. Neither configuration was in the original analyses, nor intended as part of study recruitments, and only tested as a set of possible configurations. Regardless, prediction accuracies were still barely above chance.

Table 6.9

Prediction accuracy in "archived" from an external panel.

Agrawal et al., Panels	CHANCE	FIXED	LOO
Original configurations			
CON v. {MJ, NIC, MJ+NIC}	59.4124	*64.6512**	58.8372
CON v. MJ v. NIC v. MJ+NIC	29.1412	*37.6744**	27.907
Possible configurations			
{MJ, MJ+NIC} v. {CON, NIC}	50.4094	*58.8372**	49.5349
CON v. {MJ, MJ+NIC}	54.9489	*64.8876**	*56.7416**
CON v. MJ v. {NIC, MJ+NIC}	34.0809	*47.2093**	*38.8372**
CON v. NIC v. {MJ, MJ+NIC}	40.6894	*45.3488**	33.7209

Note. A final test of prediction accuracy with a set of (small effect) SNPs associated with cannabis and/or nicotine use independently discovered in Agrawal et al., (2011) and Agrawal et al., (2014). These tests were performed to assess whether the previous predictions were poor because of the SNPs or because of other factors. The prediction accuracy for these SNPs was also very poor, with only random-effects model with accuracy above chance. To note, neither of the original group configurations showed adequate prediction, only "possible" configurations, where groups were restructure to test all the possible configurations of cannabis & nicotine use to best match the expected results of Agrawal et al., (2011) and Agrawal et al., (2014).

6.5.2 Common effects in discovery analyses

In both the proposed and flipped discovery analyses, the first component dissociated control from SUDs. However, upon close inspection of significant genotypes and their associated genes, it became clear that the first component is not the only commonality.

Table 6.10 presents the significant SNPs with |BSR| > 3: (i) in at least one configuration, (ii) on any component, and (iii) for both the archived-as-discovery and SAGE-as-discovery analyses. Both sets "as-discovery" were essentially the same analyses applied to two independent data sets. In total, there were 7 common SNPs identified in each discovery phase. These SNPs were associated with the following genes: CHRM3, PRKCE, DRD3, GRM1, NTRK2, and GRIN2A. One SNP— rs1562653—was associated with PRKCE and identified in the archived data set for the case-control, and multi-group analyses, then again in the SAGE set for the dependency-based analysis.

Furthermore, one of the most apparent effects was the dissociation of control from SUDs on the first component of nearly all analyses (besides the obvious case-control analyses). Thus, I decided to identify which genes, through their associated genotypes, were significant for *both* the archived-as-discovery (regularized or non-regularized) and SAGE-as-discovery analyses but only on the first component. Table 6.11 shows which genes were identified at |BSR| thresholds of 3.5, 3.25, and 3 for both sets-as-discovery phases. Of the genes listed in Table 6.11, the strongest CON v. SUD effects were associated with the following genes: GRIK1 and GRIN2B, which both have |BSR| > 3.5. However, some genes at a lower threshold—i.e., PRKCE at 3.25—also appear to be major contributors to the effect (see Section 6.6 for the discussion).

Table 6.10

Common SNPs across analyses.

			LOCAL			***SAGE***				
SNP	**Chr**	**Gene**	**Two Group**	**Four Group**	**Imp.**	**Two Matched**	**Four Matched**	**Two Group**	**Five Group**	**Dep.**
rs6703930	1	CHRM3		1			1			
rs1562653	2	PRKCE	1	1						1
rs167771	3	DRD3		1			1			
rs362817	6	GRM1		1						1
rs1307279	9	*NTRK2*			1		1			
rs17087959	9	NTRK2		1						1
rs7195732	16	GRIN2A		1		1				

Note. Common SNPs across *any* discovery (not predictive) analysis for both the archived and SAGE sets, with a bootstrap ratio (BSR) threshold of |BSR| > 3. Though the chipsets differ quite a bit, there were some common SNPs, and of those common SNPs, these 7 were identified as significant. The SNP rs1562653 (and its corresponding gene: PRKCE) is of particular interest here, because it was identified in both group configurations (two, and four) in the archived data, and the dependency-based analysis in the SAGE data. Nearly all other SNPs here show a significant effect in the archived four group configuration, and either the SAGE four (matched) group, or dependency analysis.

6.6 Discussion

The goals of this dissertation were to (1) extend partial least squares-correspondence analysis with regularization in order to increase power to detect effects in relatively small sample sizes, and possibly with that extension, (2) find a suitably predictive set of genetic markers for SUDs. This dissertation did not appear to achieve that goal. Through two separate discovery-validation pipelines (proposed, then flipped), and even one post-hoc panel (Agrawal and colleagues), there were no panels of genetic markers that reached a suitable level of prediction. Additionally, Phases 1 and 2 show that regularization made no difference for prediction or identification of significant genotypes compared to baseline.

Embedded within the broad goal of identifying sutiably predictive genetic markers of SUDs, was a methodological challenge: Rarely can a data set—and especially the data sets used in this dissertation—reach a sample size required for large-scale genetics analyses of complex phenotypes, traits, or disorders. Therefore, we need some methodological approach that can handle (relatively) small sample sizes, that also avoids false-positives. The methodological goal of this dissertation was to extend an already powerful method—partial least squares-correspondence analysis—with regularization, in order to make it more powerful. As results in Phases 1 and 2 show, regularization made virtually no difference—for prediction or identification of significant genotypes—compared to baseline.

6.6.1 Common effects

Nearly every analysis revealed the same type of effect: presence vs. absence of SUDs. In the cases of group-based analyses the effect was presence of substance use (SUDs) vs. absence of substance use (control). Presence vs. absence of substance use was either the explicit goal of the analysis (i.e., case-control) or just so happened to appear (i.e., first components in multi-group analyses). This effect also existed, obviously, in the dependence-based (SAGE) analyses. And finally, the presence vs. absence effect also appeared in the impulsivity analysis (archived data); as research has shown, individuals with SUDs typically display moderate to high levels of impulsivity, while non-using controls display low levels (Beaton et al., 2014). Thus, one of the strongest conclusions to draw from these studies is that it is highly likely that genetic differences exist between control and SUDs. The control v. SUD effect suggests possible protective and risk factors. Furthermore, the control vs. SUD effect was not just a simple common factor: there were specific genes that contributed to this effect (see Section 6.6.2).

The next strongest conclusion to draw is that a small number of genotypic markers may suffice to explain such complex phenomena as the genetic bases of: (1) control vs. SUD or (2) individual SUDs. In every validation analysis—from archived to SAGE included somewhere between 12 and 107 SNPs each, and from SAGE to archived included somewhere between 20 and 255 each—prediction accuracy was either below or barely above chance. Such a low prediction is odd, given that many other neurological or psychiatric disorders appear to have a strong genetic basis (e.g., schizophrenia, Alzheimer's disease, and alcohol use disorder) and are considered polygenic. Therefore, complex disorders—such as SUDs in general, or specific SUDs—could be the product of extremely complex interactions between a very large number of genes, and/or between multiple systems (e.g., genes and environment). Furthermore, like other polygenic disorders, the genetic effects may not be large effects, but rather reflect the combination of a high number of small (with a few large) effects. However, because this could lead to false positives, simply identifying a large number of small, or even moderate, effects from a *single analysis* is not a suitable approach to find robust, and likely replicable, genetic markers of SUDs, as that could lead to false positives. Some form of replication is *required* if small-to-moderate effects are to be believed.

Thus, through independent analyses—namely, the archived-as-discovery and SAGE-as-discovery—a number of small-to-moderate genetic contributions became apparent.

6.6.2 Likely genes of the control vs. SUDs effect

A number of genes identified in Table 6.11 are already very well known to SUDs, such as the GABR* family of genes, DRD2, SLC6A4, OPRM1, and HTR2A (see Chapter 2, as these are already reviewed). Furthermore some genes—such as TH and MPDZ—are known, albeit to a lesser extent than, for example DRD2, in the SUD literature. For earlier reviews, see (Palmer & de Wit, 2012; Paus, Keshavan, & Giedd, 2008), though more recent work exists (Aurelian, Warnock, Balan, Puche, & June, 2016; Kruse, Walter, & Buck, 2014). Therefore, instead of a discussion on genes that are already well-known in the genetics of SUDs research, I will focus my discussion on five particular genes that showed (1) at least moderate effects, and more importantly, (2) came up in multiple analyses: GRIK1, CCKBR, GRIN2A, GRIN2B, and PRKCE. The observed effects of these genes also further support the conclusion that there is a general control vs. SUD effect.

Table 6.11

Genes with genotypes above BSR thresholds in both "as-discovery" analyses.

	BSR Threshold		
Gene	**> 3.5**	**> 3.25**	**> 3.0**
GRIN2B	1	1	1
CCKBR	1	1	1
GRIK1		1	1
PRKCE		1	1
GRIN2A		1	1
HTR2A		1	1
GABRB1		1	1
MPDZ		1	1
CHRM3			1
NTRK2			1
TH			1
SLC6A3			1
DRD2			1
OPRM1			1
GABRB3			1
GABRG3			1

Note. Common genes (identified via SNPs), at a variety of thresholds, that appeared in *both* the archived-as-discovery (regularized or non-regularized) and SAGE-as-discovery analyses. Only the first component was selected here because, when taken together, all results point towards a broad effect of CON v. SUD genetic effects, with only small (possible) genetic effects for each particular group. The genes here likely best reflect the CON v. SUD effects.

6.6.2.1 GRIK1

While GRIK1 does not appear in any of this chapter's tables for the archived analysis, it does appear in supplemental material. GRIK1 has only one genotype in the archived multi-group analysis, but several in the archived case-control (see online supplmental material). However, in the "matched" case-control SAGE analysis, GRIK1 shows several moderate effects (Table 6.7).

GRIK1's name is derived from "**G**lutamate **I**onotropic **R**eceptor **K**ainate Type Subunit **1**". GRIK1 encodes for glutamate receptors (glutamate is, in general, excitatory see here: http://www.genecards.org/cgi-bin/carddisp.pl?gene=GRIK1). There are few studies that show an effect of GRIK1 in SUDs, and most show effects of alcohol use disorder (Kranzler et al., 2016). See also (Jones, Comer, & Kranzler, 2015) for a review.

However, a relatively early work in the genetics of SUDs research showed that GRIK1 was one of the contributing genetic factors separating control from SUDs (Johnson et al., 2008). What made the work of Johnson and colleagues so unique is that they had a volunteer sample that was ethnically and racially matched with an epidemiological sample across a number of SUDs (e.g., nicotine, methamphetamine, and polysubstance). Furthermore, both Johnson et al., (2008) and this dissertation are unique in that samples were not restricted—as they usually are—to European American and occasionally African American individuals.

Finally, GRIK1 has been linked to both suicidal behaviors (Sokolowski, Wasserman, & Wasserman, 2015) and mood disorders (Deo et al., 2013). The Deo et al., (2013) study included individuals with diagnoses of "major depressive disorder, dysthymia, or bipolar disorder". Furthermore, Deo et al., note that, while most cases were major depressive, a large proportion of cases were also co-morbid for alcohol use disorder or other SUDs.

This dissertation shares two properties with the Deo et al., (2013) study. First both studies were a “large-scale candidate gene analysis”, where Deo and colleagues explicitly used the NIAAA “addictions array” (Hodgkinson et al., 2008), I created a panel based on that same array. Second, studies used genetic models besides the typical additive code (specifically, Deo et al., use dominant and recessive). More interestingly, Deo et al., (2013) also found effects of genes common with this dissertation besides GRIK1: CHRM3, NTRK2, GRM1, DRD2, GABRG2, and—two that are discussed here—GRIN2B and PRKCE.

6.6.2.2 CCKBR

Of the five genes to discuss, CCKBR was frequently a top contributor in the archived data (Tables 6.2 and 6.3), but not as frequently in the SAGE data (Table 6.7). However, the effects of this gene are particularly intriguing. Not only were the only effects of CCKBR in the archived data on the first component, but only contributed to the dependency-based analysis in SAGE. Furthermore, the BSR values of the CCKBR genotypes are amongst the highest throughout all analyses.

CCKBR’s (an initialism of “**C**hole**c**ysto**k**inin **B** **r**eceptor”) regulates peptides in the brain (http://www.genecards.org/cgi-bin/carddisp.pl?gene=CCKBR). CCKBR has also been shown—in mice—to regulate dopamine release (Altar & Boyar, 1989).

CCKBR is rarely identified in SUDs literature. Tyndale (2003) suggested that there were weak effects of CCKBR for alcohol and nicotine use. CCKBR could regulate cocaine use—albeit in rat models (Lull, Freeman, Vrana, & Mash, 2008). However, CCKBR (like GRIK1) is a risk factor for comorbid depressive and alcohol use disorders (Kertes et al., 2011), suicide in bipolar disorder (Costa et al., 2015), and panic disorders (Wilson, Markie, & Fitches, 2012).

6.6.2.3 GRIN2A and GRIN2B

Both GRIN2A and GRIN2B were amongst the top significant genes (via many genotypes) in the archived multi-group analysis, especially on Component 1 (see Table 6.3), and survived the threshold of |BSR| > 3.5 for both the archived multi-group analysis and archived case-control analysis (see Table 6.5). Furthermore, GRIN2A was significant in the archived case-control analysis. Subsequently, we see both genes as top contributors (to the first components) in the matched multi-group and full case-control in the SAGE-as-discovery analysis (Table 6.7). Finally, a specific SNP— rs7195732—was significant in the archived multi-group, and SAGE "matched" case-control analyses.

GRIN2A and GRIN2B (acronyms of "**G**lutamate **I**onotropic **R**eceptor **N**MDA Type Subunit **2**: **A** and **B**) are both, like the GRIK1 gene previously mentioned, involved with glutamate receptors. Both GRIN2A and GRIN2B have been of interest, or shown significant results in: alcohol disorders (Chen et al., 2015), disorders comorbid with alcohol or substance use disorders (Dalvie, Fabbri, Ramesar, Serretti, & Stein, 2016; Edwards et al., 2012), and psychiatric disorders more broadly (Abdolmaleky, Thiagalingam, & Wilcox, 2005). GRIN2A has also been associated with heroin and cocaine use (Levran et al., 2016). GRIN2B has been associated with opioid use (Xie et al., 2014). However, GRIN2B has shown effects in a number of interesting areas. One area in particular is addictions in Parkinson's Disease (Ceravolo, Frosini, Rossi, & Bonuccelli, 2010). Finally, GRIN2B has been linked to a variety of physiological and SUD traits (Nikpay et al., 2012).

6.6.2.4 PRKCE

While there were no top contributor genotypes (i.e., |BSR| > 3.5) to any archived analysis, genotypes associated with PRKCE showed effects in every archived analysis (see online supplemental material). Furthermore, PRKCE genotypes did appear as top contributors (i.e., |BSR| > 3.5) to both the "matched" case-control analysis and the dependency-based analysis in the SAGE-as-discovery set. Furthermore, a particular SNP— rs1562653—also came up in three analyses: archived case-control, archived multi-group, and SAGE dependency-based.

The effects of PRKCE in all analyses were the most interesting. Table 6.12 shows all genotypes from PRKCE significant at |BSR| > 3 and only for the respective first components; every "as-discovery" analysis had significant PRKCE genotypes for at least the first component. The effects of PRKCE are, generally, lower than most other genotypes and below the "top contributor" threshold of |BSR| > 3.5. However, common and unique genotypes contribute to the control vs. SUD effect in every discovery analysis, and so in all analyses, this gene always had at least one significant genotype for a control vs. SUD effect.

PRKCE's (an initialism of "**Pr**otein **K**inase **C** **E**psilon") is part of a larger PKC* gene family. While PRKCE's role is unclear—PRKCE and the PKC family appear to play many roles (Dekker & Parker, 1994)—it does appear to play a role in many cardiac-related functions including ischemia and anxiety (http://www.genecards.org/cgi-bin/carddisp.pl?gene=PRKCE).

PRKCE has been studied in animal models of SUDs (Lesscher et al., 2009; Newton et al., 2007; Newton & Messing, 2006). Effects of PRKCE have been observed in opioid dependence (Levran et al., 2015), alcohol use disorders (Han et al., 2013; Rodd et al., 2006), alcohol consumption in younger populations (Adkins et al., 2015), broader SUD phenotypes (Uhl et al., 2008), and even pathological gambling (Lang et al., 2016).

Additionally, several studies already mentioned in this discussion section also presented effects of PRKCE: physiological, obesity-related, and SUDs traits (Nikpay et al., 2012), mood disorders (Deo et al., 2013) a possible contributor to suicidal behaviors (Sokolowski et al., 2015), alcohol use disorders (Chen et al., 2015), and—like GRIK1—effects across: (i) research and epidemiological samples, (ii) race and ethnicity, and (iii) many SUDs (Johnson et al., 2008). Though PRKCE was identified in both samples in the Johnson et al., (2008) study, it did not meet the final criteria for significance: an effect that shows perhaps—like in this dissertation—there are many robust, albeit small, effects as opposed to only a few large effects.

6.6.2.5 Genetic commonalities across many neuro-psychiatric disorders

Some of the strongest and most replicable genotypes found in the studies within this dissertation are not amongst the “usual” or expected genotypes in SUDs. The majority of genetic studies on SUDs focus either on particular systems—dopaminergic via, e.g., DRD2—or target endogenous genes—such as cannabinoid for cannabis, nicotinic for nicotine, and opioid receptor genes for opioids (see Chapter 2). However, the results of this dissertation—with respect to which genotypes—are very clear: there are other, unexpected, genes involved too. Nearly all of these unexpected genes (i.e., CCKBR, GRIK1, GRIN2A, GRIN2B, and PRKCE) have shown some effects in the SUD literature before, albeit far fewer than more popular (to study) genes. Furthermore, as already noted, these unexpected genes also appear in a variety of other psychiatric disorders.

In sum, the genetics of SUDs—and other neurological or psychiatric disorders—appear to be very complex. This complexity is likely rooted in concepts of polygenic—many genes contribute to one trait—or even pleiotropic—many genes contribute to many traits—in nature (Kendler et al., 2012; Latvala, Kuja-Halkola, D'Onofrio, Larsson, & Lichtenstein, 2016; O'Donovan & Owen, 2016). Thus, we have to move away from supposing simpler genetic explanations of these disorders, and with that move away, we must also move towards more sophisticated analytical approaches—like those in this dissertation—designed specifically to detect complex genetic contributions.

6.6.3 Methodological discussion

Recall that one of the primary goals of this dissertation was to address methodological issues of small sample sizes—and thus potential false positives—for large-scale genetics analyses. Typically, when sample sizes are too small, robust estimation techniques should be used in order to avoid "ill-posed problems" and of the issues that come with those "ill-posed problems" (e.g., rank deficiency). Therefore, PLS-CA (Beaton, Dunlop, et al., 2016), was extended with a "ridge-like" approach (à la Takane & Hwang, 2006) that combines and generalizes the concepts of regularized multiple correspondence analysis (Takane and Hwang, 2006) and the smoothed approach to two-way functional PCA (Allen, 2011). In all analyses, resampling procedures were performed to (1) determine prediction accuracy, and (2) identify likely robust effects; these resampling procedures provided inferential, as opposed to descriptive, estimations of the observed effects.

Table 6.12

Significant PRKCE Genotypes.

			Archived				*SAGE*		
SNP	***genotype***	**2G**	**4G**	**Imp**	**M2G**	**M4G**	**2G**	**5G**	**DEP**
rs10169469	*TG*								-3.033
rs10865212	*GT*				-3.633	-3.198	-3.284		
rs11674329	*CA*								-3.305
rs11677077	*CC*					3.206			
rs2245633	*CC*							-3.059	
rs608139	*CT+CC*					-3.07		-3.123	
rs608139	*TT*					3.07		3.123	
rs6719779	*CT*								-3.716
rs7557421	*AG+AA*							3.05	
rs7557421	*GG*							-3.05	
rs935653	*CC*								-3.344
rs935653	*TC+TT*								3.344
rs12622193	*AC+AA*	3.083							
rs12622193	*CC*	-3.083							
rs1562653	*AA*	-3.334	-3.139						
rs4557033	*GG*		-3.007						
rs642200	*CC*			-3.005					
rs666334	*CC*	3.306							
rs751237	*GG*	-3.237	-3.098						
rs7581914	*GG*		3.122						

Note. All SNPs and genotypes from the PRKCE gene—for Component 1—across all data-as-discovery analyses. While there were no common genotypes across the archived and SAGE ("as-discovery") sets for Component 1, there were a high number of significant PRKCE genotypes for each and across both analyses. However, common significant genotypes exist *within* each analysis.

As previously noted, regularized PLS-CA—as a ridge regularized approach— decreased variance, while forcing the component scores closer to zero (i.e., “shrinkage”). However, it did not provide any of the expected advantages (i.e., better estimation) a regularized method should. In the end, SmooPLS-CA provided no advantage over standard PLS-CA for this dissertation.

6.6.3.1 Failure of regularization

Phases 1 and 2 (standard PLS-CA and SmooPLS-CA, respectively) suggested that there were slight but inconsequential differences in the bootstrap identified genotypes between phases. Furthermore, prediction accuracies were either comparable, or in some cases *better* in standard PLS-CA. Taken together, it would appear as though regularization failed. Phase 3 confirmed this failure: no advantage was afforded by regularization from discovery to validation.

The failure of regularization is not an expected behavior, as regularization is a technique designed specifically to boost power for situations exactly like those in this dissertation: small sample sizes, high-dimensional, and often noisy data. Therefore, instead of the perspective that regularization has failed, perhaps we could consider that PLS is already a powerful technique and does not require regularization.

6.6.3.2 Robustness of PLS & Resampling

There have been two recent studies with partial least squares that help explain why standard PLS-CA appeared to be as good as, and in some cases better than SmooPLS-CA.

The first study comes from imaging genetics (Grellmann et al., 2015). In their study, Grellmann and colleagues used simulated data—to control the “ground truth” of SNP-voxel associations—and tested three techniques: partial least squares (correlation), sparse canonical correlation analysis (à la Witten et al., 2009), and “Bayesian inter-battery factor analysis” (a Bayes version of IBFA à la Tucker, 1958). As noted in Chapter 3, Section 3.2.4, Grellmann et al., (2015) showed that standard PLS outperformed the other two techniques, under the condition that the number of variables greatly exceeded the sample size, that is: PLSC did a better job at identifying SNP-voxel associations under *very high dimensionality*, where sparse CCA performed well when the number of variables was small. Furthermore, Grellmann et al., (2015) also showed that even when sparse CCA was the superior technique, that sparse CCA and PLSC performed at comparable levels. Finally, Grellmann and colleagues also showed that PLSC, regardless of dimensionality, usually detected effects within a few, and sometimes only one, component. This dissertation and Grellmann et al., (2015) together suggest that a "standard" PLS technique is better than or equal to comparable regularized methods *for these particular domains*.

The next study is by Churchill et al., (2014), with further support from an unpublished study (Churchill, personal communication, January 2016). Churchill and colleagues were studying the effects of pipeline choices on functional neuroimaging, with an emphasis on maximizing prediction and reproducibility. In both works the support vector machine—a standard predictive technique—maintained high prediction, but low reproducibility. In their 2014 work, Churchill et al., showed that PCA, as opposed to regularized techniques, maximized reproducibility. Furthermore, in their unpublished work, they also included discriminant (a.k.a., "mean-centered") PLS that is (a between-class covariance technique). In their unpublished work, a slightly modified (estimable) linear discriminant analysis and discriminant PLS produced the most reproducible effects, well beyond other techniques. Taken together, the work of Churchill and colleagues, as well as this dissertation suggest that PLS is not a suitable technique to build highly predictive models. However, the results from the PLS models are highly reproducible.

Reproducible effects are precisely what we observed when both the archived and SAGE data sets were treated as “discovery” (i.e., the same analysis applied to two independent sets) and produced highly similar results, yet could not adequately predict group relationships.

Finally, SmooPLS-CA worked the way it was supposed to: component scores, especially of distant outliers, shrunk towards zero and thus eliminated potentially spurious effects, and only the strongest signals per component remained far from zero. However, when used in conjunction with resampling-based inference techniques, there was no substantial difference between standard and SmooPLS-CA to detect significant genotypes. At least in these studies, it seems that regularization approximates and eventually converges to the same information provided by BSRs: unstable items shrink towards zero (have non-significant BSRs), whereas stable items stay in place (have significant BSRs).

6.7 Limitations and conclusions

At first it appeared that there were null effects from both the proposed and flipped pipelines. However, upon closer inspection it was clear that there were some moderate-to-strong reproducible effects. These effects were genotypes from the same genes that contributed to a separation of control from SUDs. Though these reproducible effects were discovered with independent analyses via PLS-CA, there are some limitations that should be discussed.

First, the archived and SAGE sets are not entirely comparable. The SAGE study focuses on alcohol use disorders, and so, the non-control participants were alcohol dependent. In addition, many individuals in the SAGE set are co- or polysubstance users—an artifact of an early, multi-site data set—and this makes the study of any one SUD difficult. In an opposite sense, the archived data had relatively clean study groups—an “artifact” of recruitment—with extremely low incidence of alcohol dependence. In the end, the closest comparable groups across the archived and SAGE were only the control groups. And while the primary effect observed in this dissertation appears to be control vs. SUD, I do not believe it is entirely due to the group configurations. Some of the multi-group analyses performed relatively well—within discovery stages—and even some of the genotypes of later (i.e., not the first) components showed common genotypes (see online supplemental material for all genotypes with |BSR| > 3).

Another possible limitation was with respect to the proposed Phase 3: perhaps a more robust approach for prediction should have been used in this phase. While this is reasonable, there is no other available statistical method for strictly categorical data. To use another method would result in using only the additive model, and thus limit all genetic assumptions to be strictly linear and additive.

Another limitation is that the two data sets had different chipsets for SNPs. This actually made the transition from discovery phases to validation phases difficult. Not all SNPs discovered could be validated. However, it would appear as though the parallel analyses (i.e., same analysis applied to two independent sets) are a very reasonable work around to different chip sets. Furthermore, the two parallel analyses—where one data set is completely sequestered from the other—actually matches how replication analyses would be performed.

Finally, there is a major issue that ties together all the limitations: confound correction. In genomics, the usual approach to confound correction is that after standard quality control and preprocessing have been completed, a multivariate assessment of the individuals is made. This multivariate assessment is almost exclusively PCA (a.k.a., "eigen stratification"). Next, either sets of individuals are excluded (which now changes minor allele frequencies and missingness) or more commonly, genomic components are removed from the data set. There are two issues with confound correction. First, in order to have a unified confound correction, the data sets would have to be combined, cleaned, preprocessed, and then separated into a "discovery" and "validation" sample. This is akin to a training/testing paradigm. But doing so in genetics will create another problem: When the data sets are cleaned through the same pipelines, they are no longer independent. The next issue is that when data sets are left as independent data sets, confound correction will be different in each set. As noted in the supplemental methods, the archived-as-discovery data had five components removed, where as the SAGE-as-discovery had only removed one component. Furthermore, confound correction was applied to *every* panel in Phase 3 individually: Every unique set showed a slightly different confounding factor. This, therefore, made Phase 3 a very time consuming process. In the end, it was not an easy choice, but I believe that a complete sequestration—like I did—is the best way forward. While it precluded me from actually predicting SAGE data from the archived models, this is also the same scenario independent researchers would face in replication attempts.

While there were limitations, I do not believe them to hinder the results of these studies. In fact, some limitations—such as keeping data completely independent—made for a stronger set of results: the same effects and genotypes were observed in separate analyses of separate data.

REFERENCES

Abdi, H. (2007). Singular Value Decomposition (SVD) and Generalized Singular Value Decomposition (GSVD). In Encyclopedia of Measurement and Statistics (pp. 907–912).

Abdi, H. (2007c). Eigen-decomposition: eigenvalues and eigenvecteurs. In N.J. Salkind (Ed.): Encyclopedia of measurement and statistics. (pp. 304–308) Thousand Oaks, CA: Sage.

Abdi, H., Dunlop, J. P., & Williams, L. J. (2009). How to compute reliability estimates and display confidence and tolerance intervals for pattern classifiers using the bootstrap and 3-way multidimensional scaling (DISTATIS). NeuroImage, 45, 89–95.

Abdi, H., Edelman, B., Valentin, D., & Dowling, W.J. (2009). Experimental design and analysis for psychology. Oxford University Press.

Abdi, H., & Williams, L. J. (2010a). Principal Component Analysis. Wiley Interdisciplinary Reviews: Computational Statistics, 2.

Abdi, H. & Williams, L.J. (2010b). Correspondence analysis. In N.J. Salkind, D.M., Dougherty, & B. Frey (Eds.): Encyclopedia of Research Design. Thousand Oaks (CA): Sage. pp. 267–278

Abdi, H., & Williams, L.J. (2013). Partial least squares methods: Partial least squares correlation and partial least square regression. In: B. Reisfeld & A. Mayeno (Eds.), Methods in molecular biology: computational toxicology. New York: Springer Verlag. pp. 549-579.

Abdi, H., & Béra, M. (2014). Correspondence analysis. In R. Alhajj and J. Rokne (Eds.), Encyclopedia of Social Networks and Mining. New York: Springer Verlag. pp 275–284.

Abdolmaleky, H. M., Thiagalingam, S., & Wilcox, M. (2005). Genetics and Epigenetics in Major Psychiatric Disorders. *American Journal of Pharmacogenomics*, *5*(3), 149–160. https://doi.org/10.2165/00129785-200505030-00002

Adkins, D. E., Clark, S. L., Copeland, W. E., Kennedy, M., Conway, K., Angold, A., … Costello, E. J. (2015). Genome-Wide Meta-Analysis of Longitudinal Alcohol Consumption Across Youth and Early Adulthood. *Twin Research and Human Genetics*, *18*(4), 335–347. https://doi.org/10.1017/thg.2015.36

Adinoff, B., Devous, M. D., Williams, M. J., Best, S. E., Harris, T. S., Minhajuddin, A., ... & Cullum, M. (2010). Altered neural cholinergic receptor systems in cocaine-addicted subjects. Neuropsychopharmacology, 35(7), 1485-1499.

Aebi, M., van Donkelaar, M. M., Poelmans, G., Buitelaar, J. K., Sonuga-Barke, E. J., Stringaris, A., ... & van Hulzen, K. J. (2015). Gene-set and multivariate genome-wide association analysis of oppositional defiant behavior subtypes in attention-deficit/hyperactivity disorder. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics.

Agrawal, A., & Lynskey, M. T. (2009). Candidate genes for cannabis use disorders: findings, challenges and directions. Addiction, 104(4), 518–532. http://doi.org/10.1111/j.1360-0443.2009.02504.x

Agrawal, A., Lynskey, M. T., Hinrichs, A., Grucza, R., Saccone, S. F., Krueger, R., … Bierut, L. J. (2011). A genome-wide association study of DSM-IV cannabis dependence. Addiction Biology, 16(3), 514–518. http://doi.org/10.1111/j.1369-1600.2010.00255.x

Agrawal, A., Lynskey, M. T., Kapoor, M., Bucholz, K. K., Edenberg, H. J., Schuckit, M., … Bierut, L. J. (2015). Are genetic variants for tobacco smoking associated with cannabis involvement? Drug and Alcohol Dependence, 150, 183–187. http://doi.org/10.1016/j.drugalcdep.2015.02.029

Agrawal A, Pergadia ML, Saccone SF, & et al. (2008). AN autosomal linkage scan for cannabis use disorders in the nicotine addiction genetics project. Archives of General Psychiatry, 65(6), 713–721. http://doi.org/10.1001/archpsyc.65.6.713

Agrawal, A., Wetherill, L., Dick, D. M., Xuei, X., Hinrichs, A., Hesselbrock, V., … Foroud, T. (2009). Evidence for association between polymorphisms in the cannabinoid receptor 1 (CNR1) gene and cannabis dependence. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics, 150B(5), 736–740. http://doi.org/10.1002/ajmg.b.30881

Ahrens, S., Markett, S., Breckel, T. P. K., Behler, O., Reuter, M., & Thiel, C. M. (2015). Modulation of nicotine effects on selective attention by DRD2 and CHRNA4 gene polymorphisms. Psychopharmacology, 232(13), 2323–2331. http://doi.org/10.1007/s00213-015-3869-2

Ali, M. A., Way, M. J., Marks, M., Guerrini, I., Thomson, A. D., Strang, J., … Morgan, M. Y. (2015). Phenotypic heterogeneity in study populations may significantly confound the results of genetic association studies on alcohol dependence. Psychiatric Genetics, 25(6), 234–240.

Allen, G. I. (2013). Sparse and Functional Principal Components Analysis. arXiv:1309.2895 [stat]. Retrieved from http://arxiv.org/abs/1309.2895

Allen, G. I., & Maletić-Savatić, M. (2011). Sparse non-negative generalized PCA with applications to metabolomics. Bioinformatics, 27(21), 3029–3035. http://doi.org/10.1093/bioinformatics/btr522

Allison, D. B., Thiel, B., St. Jean, P., Elston, R. C., Infante, M. C., & Schork, N. J. (1998). Multiple Phenotype Modeling in Gene-Mapping Studies of Quantitative Traits: Power

Advantages. The American Journal of Human Genetics, 63(4), 1190–1201. http://doi.org/10.1086/302038

Altar, C. A., & Boyar, W. C. (1989). Brain CCK-B receptors mediate the suppression of dopamine release by cholecystokinin. *Brain Research*, *483*(2), 321–326.

Anokhin, A. P., Grant, J. D., Mulligan, R. C., & Heath, A. C. (2015). The Genetics of Impulsivity: Evidence for the Heritability of Delay Discounting. Biological Psychiatry, 77(10), 887–894. http://doi.org/10.1016/j.biopsych.2014.10.022

Ardizzone, S., Maconi, G., Bianchi, V., Russo, A., Colombo, E., Cassinotti, A., … Bianchi Porro, G. (2007). Multidrug resistance 1 gene polymorphism and susceptibility to inflammatory bowel disease. Inflammatory Bowel Diseases, 13(5), 516–523. http://doi.org/10.1002/ibd.20108

Arnedo, J., Svrakic, D. M., del Val, C., Romero-Zaliz, R., Hernández-Cuervo, H., Fanous, A. H., … Zwir, I. (2014). Uncovering the Hidden Risk Architecture of the Schizophrenias: Confirmation in Three Independent Genome-Wide Association Studies. American Journal of Psychiatry, 172(2), 139–153. http://doi.org/10.1176/appi.ajp.2014.14040435

Ashare, R. L., Valdez, J. N., Ruparel, K., Albelda, B., Hopson, R. D., Keefe, J. R., … Lerman, C. (2013). Association of abstinence-induced alterations in working memory function and COMT genotype in smokers. Psychopharmacology, 230(4), 653–662. http://doi.org/10.1007/s00213-013-3197-3

Aurelian, L., Warnock, K. T., Balan, I., Puche, A., & June, H. (2016). TLR4 signaling in VTA dopaminergic neurons regulates impulsivity through tyrosine hydroxylase modulation. *Translational Psychiatry*, *6*(5), e815. https://doi.org/10.1038/tp.2016.72

Aydin, S., Shahi, A., Ozbayram, E. G., Ince, B., & Ince, O. (2015). Use of PCR-DGGE based molecular methods to assessment of microbial diversity during anaerobic treatment of antibiotic combinations. Bioresource Technology, 192, 735–740. http://doi.org/10.1016/j.biortech.2015.05.086

Azadeh, S., Hobbs, B. P., Ma, L., Nielsen, D. A., Gerard Moeller, F., & Baladandayuthapani, V. (in press, 2016). Integrative Bayesian analysis of neuroimaging-genetic data with application to cocaine dependence. NeuroImage. http://doi.org/10.1016/j.neuroimage.2015.10.033

Bael, S. V., & Pruett-Jones, S. (1996). Exponential Population Growth of Monk Parakeets in the United States. The Wilson Bulletin, 108(3), 584–588.

Ballard, M. E., Dean, A. C., Mandelkern, M. A., & London, E. D. (2015). Striatal Dopamine D2/D3 Receptor Availability Is Associated with Executive Function in Healthy Controls but Not Methamphetamine Users. PLoS ONE, 10(12), e0143510. http://doi.org/10.1371/journal.pone.0143510

Ballon, N., Leroy, S., Roy, C., Bourdel, M. C., Charles-Nicolas, A., Krebs, M. O., & Poirier, M. F. (2005). (AAT)n repeat in the cannabinoid receptor gene (CNR1): association with cocaine addiction in an African-Caribbean population. The Pharmacogenomics Journal, 6(2), 126–130. http://doi.org/10.1038/sj.tpj.6500352

Bankston, S. M., Carroll, D. D., Cron, S. G., Granmayeh, L. K., Marcus, M. T., Moeller, F. G., … Liehr, P. R. (2009). Substance abuser impulsivity decreases with a nine-month stay in a therapeutic community. The American Journal of Drug and Alcohol Abuse, 35(6), 417–420. http://doi.org/10.3109/00952990903410707

Barrett, J. H., Taylor, J. C., & Iles, M. M. (2014). Statistical Perspectives for Genome-Wide Association Studies (GWAS). In R. Trent (Ed.), Clinical Bioinformatics (pp. 47–61). Springer New York.

Bart, G., Kreek, M. J., Ott, J., LaForge, K. S., Proudnikov, D., Pollak, L., & Heilig, M. (2004). Increased Attributable Risk Related to a Functional |[mu]|-Opioid Receptor Gene Polymorphism in Association with Alcohol Dependence in Central Sweden. Neuropsychopharmacology, 30(2), 417–422. http://doi.org/10.1038/sj.npp.1300598

Baumeister, R. F., & Heatherton, T. F. (1996). Self-Regulation Failure: An Overview. Psychological Inquiry, 7(1), 1–15. http://doi.org/10.1207/s15327965pli0701_1

Beaton, D., Abdi, H., & Filbey, F. M. (2014). Unique aspects of impulsive traits in substance use and overeating: specific contributions of common assessments of impulsivity. The American Journal of Drug and Alcohol Abuse, 40(6), 463–475. http://doi.org/10.3109/00952990.2014.937490

Beaton, D., Abdi, H., & Filbey, F. M. (in prep). Identifying the best candidate polygenic panels for chronic marijuana, nicotine, and marijuana + nicotine co-use.

Beaton, D., Chin Fatt, C. R., & Abdi, H. (2014). An ExPosition of multivariate analysis with the singular value decomposition in R. Computational Statistics & Data Analysis, 72, 176–189.

Beaton, D., Dunlop, J., Abdi, H., & Alzheimer's Disease Neuroimaging Initiative. (2016). Partial Least Squares Correspondence Analysis: A Framework to Simultaneously Analyze Behavioral and Genetic Data. Psychological Methods. http://doi.org/10.1037/met0000053

Beaton, D., Kriegsman, M., ADNI, Dunlop, J. P., Filbey, F. M., & Abdi, H. (2016). Partial Least Squares for mixed-data types: An application for imaging genetics. In Abdi, H., Vinzi, V.E., Russolillo, G., Saporta, G., Trinchera, L. (Eds.), The Multiple Facets of Partial Least Squares Methods.. New York, NY, USA: Springe-Verlag.

Beaton, D., Filbey, F., & Abdi, H. (2013). Integrating partial least squares correlation and correspondence analysis for nominal data. In H. Abdi, W. Chin, V. Esposito Vinzi, G. Russolillo, & L. & Trinchera (Eds.), New perspectives in partial least squares and related methods (pp. 81–94). New York: Springer Verlag.

Beaton, D., Rieck, J.R., Alhazmi, F., ADNI, & Abdi, H. (in prep). Genome-wide Pattern Analysis in Alzheimer's Disease: Multivariate genotypic approaches to identify multiple, but specific, genotypes to characterize disease and control groups.

Beck, A., Schlagenhauf, F., Wüstenberg, T., Hein, J., Kienast, T., Kahnt, T., … Wrase, J. (2009). Ventral Striatal Activation During Reward Anticipation Correlates with Impulsivity in Alcoholics. Biological Psychiatry, 66(8), 734–742. http://doi.org/10.1016/j.biopsych.2009.04.035

Belcher, A. M., Volkow, N. D., Moeller, F. G., & Ferré, S. (2014). Personality traits and vulnerability or resilience to substance use disorders. Trends in Cognitive Sciences, 18(4), 211–217. http://doi.org/10.1016/j.tics.2014.01.010

Belin, D., Mar, A. C., Dalley, J. W., Robbins, T. W., & Everitt, B. J. (2008). High Impulsivity Predicts the Switch to Compulsive Cocaine-Taking. Science, 320(5881), 1352–1355. http://doi.org/10.1126/science.1158136

Belsky, D. W., Moffitt, T. E., Baker, T. B., Biddle, A. K., Evans, J. P., Harrington, H., … others. (2013). Polygenic risk and the developmental progression to heavy, persistent smoking and nicotine dependence: evidence from a 4-decade longitudinal study. JAMA Psychiatry, 70(5), 534–542.

Benyamina, A., Bonhomme-Faivre, L., Picard, V., Sabbagh, A., Richard, D., Blecha, L., … Reynaud, M. (2009). Association between ABCB1 C3435T polymorphism and increased risk of cannabis dependence. Progress in Neuro-Psychopharmacology and Biological Psychiatry, 33(7), 1270–1274. http://doi.org/10.1016/j.pnpbp.2009.07.016

Bertario, L., Russo, A., Sala, P., Varesco, L., Giarola, M., Mondini, P., … Radice, P. (2003). Multiple Approach to the Exploration of Genotype-Phenotype Correlations in Familial Adenomatous Polyposis. Journal of Clinical Oncology, 21(9), 1698–1707. http://doi.org/10.1200/JCO.2003.09.118

Berry, K. J., Johnston, J. E., & Mielke, P. W. (2011). Permutation methods. Wiley Interdisciplinary Reviews: Computational Statistics, 3, 527–542.

Bevilacqua, L., & Goldman, D. (2011). Genetics of emotion. Trends in Cognitive Sciences. http://doi.org/10.1016/j.tics.2011.07.009

Bhattacharjee, S., Wang, Z., Ciampa, J., Kraft, P., Chanock, S., Yu, K., & Chatterjee, N. (2010). Using Principal Components of Genetic Variation for Robust and Powerful Detection of

Gene-Gene Interactions in Case-Control and Case-Only Studies. The American Journal of Human Genetics, 86(3), 331–342. http://doi.org/10.1016/j.ajhg.2010.01.026

Bickel, W. K., Koffarnus, M. N., Moody, L., & Wilson, A. G. (2014). The behavioral- and neuro-economic process of temporal discounting: A candidate behavioral marker of addiction. Neuropharmacology, 76, Part B, 518–527. http://doi.org/10.1016/j.neuropharm.2013.06.013

Bidwell, L. C., Metrik, J., McGeary, J., Palmer, R. H. C., Francazio, S., & Knopik, V. S. (2013). Impulsivity, variation in the cannabinoid receptor (CNR1) and fatty acid amide hydrolase (FAAH) genes, and marijuana-related problems. Journal of Studies on Alcohol and Drugs, 74(6), 867–878.

Bierut L, Dinwiddie SH, Begleiter H, & et al. (1998). Familial transmission of substance dependence: Alcohol, marijuana, cocaine, and habitual smoking: a report from the collaborative study on the genetics of alcoholism. Archives of General Psychiatry, 55(11), 982–988. http://doi.org/10.1001/archpsyc.55.11.982

Bierut, L. J., Agrawal, A., Bucholz, K. K., Doheny, K. F., Laurie, C., Pugh, E., … Consortium, E. A. S. (GENEVA). (2010). A genome-wide association study of alcohol dependence. Proceedings of the National Academy of Sciences, 107(11), 5082–5087. http://doi.org/10.1073/pnas.0911109107

Biffani, S., Dimauro, C., Macciotta, N., Rossoni, A., Stella, A., & Biscarini, F. (2015). Predicting haplotype carriers from SNP genotypes in Bos taurus through linear discriminant analysis. Genetics, Selection, Evolution : GSE, 47(1). http://doi.org/10.1186/s12711-015-0094-8

Bi, J., Gelernter, J., Sun, J., & Kranzler, H. R. (2014). Comparing the Utility of Homogeneous Subtypes of Cocaine Use and Related Behaviors With DSM-IV Cocaine Dependence as Traits for Genetic Association Analysis. American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics : The Official Publication of the International Society of Psychiatric Genetics, 0(2), 148–156. http://doi.org/10.1002/ajmg.b.32216

Binder, D. K., & Scharfman, H. E. (2004). Brain-derived Neurotrophic Factor. Growth Factors (Chur, Switzerland), 22(3), 123–131. http://doi.org/10.1080/08977190410001723308

Binder, E. B., Bradley, R. G., Liu, W., Epstein, M. P., Deveau, T. C., Mercer, K. B., … others. (2008). Association of FKBP5 polymorphisms and childhood abuse with risk of posttraumatic stress disorder symptoms in adults. Jama, 299(11), 1291–1305.

Bloss, C. S., Schiabor, K. M., & Schork, N. J. (2010). Human behavioral informatics in genetic studies of neuropsychiatric disease: Multivariate profile-based analysis. Brain Research Bulletin, 83(3-4), 177–188. http://doi.org/doi: DOI: 10.1016/j.brainresbull.2010.04.012

Blum, K., Braverman, E. R., Holder, J. M., Lubar, J. F., Monastra, V. J., Miller, D., … Comings, D. E. (2000). Reward deficiency syndrome: a biogenetic model for the diagnosis and treatment of impulsive, addictive, and compulsive behaviors. Journal of Psychoactive Drugs, 32 Suppl, i–iv, 1–112.

Blum, K., Oscar-Berman, M., Demetrovics, Z., Barh, D., & Gold, M. S. (2014). Genetic Addiction Risk Score (GARS): Molecular Neurogenetic Evidence for Predisposition to Reward Deficiency Syndrome (RDS). Molecular Neurobiology, 50(3), 765–796. http://doi.org/10.1007/s12035-014-8726-5

Bobadilla, L., Vaske, J., & Asberg, K. (2013). Dopamine receptor (D4) polymorphism is related to comorbidity between marijuana abuse and depression. Addictive Behaviors, 38(10), 2555–2562. http://doi.org/10.1016/j.addbeh.2013.05.014

Bolormaa, S., Pryce, J. E., Hayes, B. J., & Goddard, M. E. (2010). Multivariate analysis of a genome-wide association study in dairy cattle. Journal of Dairy Science, 93(8), 3818–3833. http://doi.org/10.3168/jds.2009-2980

Bookstein, F. (1994). Partial least squares: a dose–response model for measurement in the behavioral and brain sciences. Psycoloquy, 5(23).

Bousman, C. A., Glatt, S. J., Cherner, M., Atkinson, J. H., Grant, I., Tsuang, M. T., & Everall, I. P. (2010). Preliminary evidence of ethnic divergence in associations of putative genetic variants for methamphetamine dependence. Psychiatry Research, 178(2), 295–298. http://doi.org/10.1016/j.psychres.2009.07.019

Braak, C. J. F., & Verdonschot, P. F. M. (1995). Canonical correspondence analysis and related multivariate methods in aquatic ecology. Aquatic Sciences - Research Across Boundaries, 57(3), 255–289. http://doi.org/10.1007/BF00877430

Breckel, T. P. K., Giessing, C., Gieseler, A., Querbach, S., Reuter, M., & Thiel, C. M. (2015). Nicotinergic Modulation of Attention-Related Neural Activity Differentiates Polymorphisms of DRD2 and CHRNA4 Receptor Genes. PLoS ONE, 10(6), e0126460. http://doi.org/10.1371/journal.pone.0126460

Bretherton, C. S., Smith, C., & Wallace, J. M. (1992). An intercomparison of methods for finding coupled patterns in climate data. Journal of Climate, 5, 541–560.

Buchmann, A. F., Hohm, E., Witt, S. H., Blomeyer, D., Jennen-Steinmetz, C., Schmidt, M. H., … Laucht, M. (2014). Role of CNR1 polymorphisms in moderating the effects of psychosocial adversity on impulsivity in adolescents. Journal of Neural Transmission, 1–9. http://doi.org/10.1007/s00702-014-1266-3

Bühler, M., Vollstädt-Klein, S., Kobiella, A., Budde, H., Reed, L. J., Braus, D. F., … Smolka, M. N. (2010). Nicotine Dependence Is Characterized by Disordered Reward Processing in a

Network Driving Motivation. Biological Psychiatry, 67(8), 745–752. http://doi.org/10.1016/j.biopsych.2009.10.029

Buhrman-Deever, S. C., Rappaport, A. R., & Bradbury, J. W. (2007). Geographic variation in contact calls of feral north american populations of the monk parakeet. The Condor, 109(2), 389–398. http://doi.org/10.1650/0010-5422(2007)109[389:GVICCO]2.0.CO;2

Buka, S. L., Shenassa, E. D., & Niaura, R. (2003). Elevated Risk of Tobacco Dependence Among Offspring of Mothers Who Smoked During Pregnancy: A 30-Year Prospective Study. American Journal of Psychiatry, 160(11), 1978–1984. http://doi.org/10.1176/appi.ajp.160.11.1978

Burridge, C. P., Peucker, A. J., Valautham, S. K., Styan, C. A., & Dann, P. (2015). Nonequilibrium Conditions Explain Spatial Variability in Genetic Structuring of Little Penguin (Eudyptula minor). Journal of Heredity, 106(3), 228–237. http://doi.org/10.1093/jhered/esv009

Cahill, A. E., & Levinton, J. S. (2016). Genetic differentiation and reduced genetic diversity at the northern range edge of two species with different dispersal modes. Molecular ecology, 25(2), 515-526.

Callicott, J. H., Egan, M. F., Mattay, V. S., Bertolino, A., Bone, A. D., Verchinksi, B., & Weinberger, D. R. (2003). Abnormal fMRI response of the dorsolateral prefrontal cortex in cognitively intact siblings of patients with schizophrenia. The American Journal of Psychiatry, 160(4), 709–719. http://doi.org/10.1176/appi.ajp.160.4.709

Camps, F. E. (1972). Genetics and Alcoholism. Annals of the New York Academy of Sciences, 197(1), 134–137. http://doi.org/10.1111/j.1749-6632.1972.tb28134.x

Cantor, R. M., Lange, K., & Sinsheimer, J. S. (2010). Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application. The American Journal of Human Genetics, 86(1), 6–22. http://doi.org/10.1016/j.ajhg.2009.11.017

Carlson, J. M., Cha, J., Harmon-Jones, E., Mujica-Parodi, L. R., & Hajcak, G. (2014). Influence of the BDNF Genotype on Amygdalo-Prefrontal White Matter Microstructure is Linked to Nonconscious Attention Bias to Threat. Cerebral Cortex, 24(9), 2249–2257. http://doi.org/10.1093/cercor/bht089

Carver, C. S., LeMoult, J., Johnson, S. L., & Joormann, J. (2014). Gene Effects and G × E Interactions in the Differential Prediction of Three Aspects of Impulsiveness. Social Psychological and Personality Science, 5(6), 730–739. http://doi.org/10.1177/1948550614527116

Cavalli-Sforza, L. L., & Edwards, A. W. F. (1967). Phylogenetic analysis. Models and estimation procedures. American Journal of Human Genetics, 19(3 Pt 1), 233–257.

Ceravolo, R., Frosini, D., Rossi, C., & Bonuccelli, U. (2010). Spectrum of addictions in Parkinson's disease: from dopamine dysregulation syndrome to impulse control disorders. *Journal of Neurology*, *257*(2), 276–283. https://doi.org/10.1007/s00415-010-5715-0

Chakrabarti, B., & Baron-Cohen, S. (2011). Variation in the human Cannabinoid Receptor (CNR1) gene modulates gaze duration for happy faces. Molecular Autism, 2(1), 10.

Chang, D., & Keinan, A. (2014). Principal Component Analysis Characterizes Shared Pathogenetics from Genome-Wide Association Studies. PLoS Comput Biol, 10(9), e1003820. http://doi.org/10.1371/journal.pcbi.1003820

Chang, L., Wang, Y., Ji, H., Dai, D., Xu, X., Jiang, D., … Wang, Q. (2014). Elevation of Peripheral BDNF Promoter Methylation Links to the Risk of Alzheimer's Disease. PLoS ONE, 9(11), e110773. http://doi.org/10.1371/journal.pone.0110773

Cheah, S.-Y., Lawford, B. R., Young, R. M., Connor, J. P., Morris, C. P., & Voisey, J. (2014). BDNF SNPs Are Implicated in Comorbid Alcohol Dependence in Schizophrenia But Not in Alcohol-Dependent Patients Without Schizophrenia. Alcohol and Alcoholism, agu040. http://doi.org/10.1093/alcalc/agu040

Chen, C.-K., Hu, X., Lin, S.-K., Sham, P. C., Loh, E.-W., Li, T., … Ball, D. M. (2004). Association analysis of dopamine D2-like receptor genes and methamphetamine abuse. Psychiatric Genetics, 14(4), 223–226.

Chen, C.-K., Lin, S.-K., Chiang, S.-C., Su, L.-W., & Wang, L.-J. (2014). Polymorphisms of COMT Val158Met and DAT1 3′-UTR VNTR in Illicit Drug Use and Drug-Related Psychiatric Disorders. Substance Use & Misuse, 140407111357004. http://doi.org/10.3109/10826084.2014.901391

Chen, J., Hutchison, K. E., Calhoun, V. D., Claus, E. D., Turner, J. A., Sui, J., & Liu, J. (2015). CREB-BDNF pathway influences alcohol cue-elicited activation in drinkers. Human brain mapping, 36(8), 3007-3019.

Chen, L.-S., Baker, T. B., Grucza, R., Wang, J. C., Johnson, E. O., Breslau, N., … Bierut, L. J. (2012). Dissection of the Phenotypic and Genotypic Associations With Nicotinic Dependence. Nicotine & Tobacco Research, 14(4), 425–433. http://doi.org/10.1093/ntr/ntr231

Chen, X., Liu, H., & Carbonell, J. G. (2012). Structured sparse canonical correlation analysis. In International Conference on Artificial Intelligence and Statistics (pp. 199–207).

Chernick, M. (2008). Bootstrap methods: A guide for practitioners and researchers (2nd ed., Vol. 619). New York, New York: Wiley.

Chi, E. C., Allen, G. I., Zhou, H., Kohannim, O., Lange, K., & Thompson, P. M. (2013). Imaging genetics via sparse canonical correlation analysis. In 2013 IEEE 10th International Symposium on Biomedical Imaging (ISBI) (pp. 740–743). http://doi.org/10.1109/ISBI.2013.6556581

Chikova, A., Bernard, H.-U., Shchepotin, I. B., & Grando, S. A. (2012). New associations of the genetic polymorphisms in nicotinic receptor genes with the risk of lung cancer. Life Sciences, 91(21-22), 1103–1108. http://doi.org/10.1016/j.lfs.2011.12.023

Chi, Y. (2012). Multivariate methods. Wiley Interdisciplinary Reviews: Computational Statistics, 4(1), 35–47. http://doi.org/10.1002/wics.185

Chung, N. C., & Storey, J. D. (2014). Statistical Significance of Variables Driving Systematic Variation in High-Dimensional Data. Bioinformatics, btu674. http://doi.org/10.1093/bioinformatics/btu674

Chun, H., Ballard, D. H., Cho, J., & Zhao, H. (2011). Identification of association between disease and multiple markers via sparse partial least-squares regression. Genetic Epidemiology, 35(6), 479–486. http://doi.org/10.1002/gepi.20596

Chun, H., & Keleş, S. (2009). Expression Quantitative Trait Loci Mapping With Multivariate Sparse Partial Least Squares Regression. Genetics, 182(1), 79–90. http://doi.org/10.1534/genetics.109.100362

Churchill, N. W., Spring, R., Afshin-Pour, B., Dong, F., & Strother, S. C. (2015). An Automated, Adaptive Framework for Optimizing Preprocessing Pipelines in Task-Based Functional MRI. PLoS ONE, 10(7), e0131520. http://doi.org/10.1371/journal.pone.0131520

Cioli, C., Abdi, H., Beaton, D., Burnod, Y., & Mesmoudi, S. (2014). Differences in Human Cortical Gene Expression Match the Temporal Properties of Large-Scale Functional Networks. PLoS ONE, 9(12), e115913. http://doi.org/10.1371/journal.pone.0115913

Cippitelli, A., Astarita, G., Duranti, A., Caprioli, G., Ubaldi, M., Stopponi, S., … Ciccocioppo, R. (2011). Endocannabinoid regulation of acute and protracted nicotine withdrawal: effect of FAAH inhibition. PloS One, 6(11), e28142. http://doi.org/10.1371/journal.pone.0028142

Clarke, T.-K., Weiss, A. R. D., Ferarro, T. N., Kampman, K. M., Dackis, C. A., Pettinati, H. M., … Berrettini, W. H. (2014). The Dopamine Receptor D2 (DRD2) SNP rs1076560 is Associated with Opioid Addiction. Annals of Human Genetics, 78(1), 33–39. http://doi.org/10.1111/ahg.12046

Clarke, T.-K., Bloch, P. J., Ambrose-Lanci, L. M., Doyle, G. A., Ferraro, T. N., Berrettini, W. H., … Lohoff, F. W. (2013). Further evidence for association between genetic variants in the cannabinoid receptor 1 (CNR1) gene and cocaine dependence: Confirmation in an independent sample and meta-analysis. Addiction Biology, 18(4), 702–708. http://doi.org/10.1111/j.1369-1600.2011.00346.x

Cole, J., Logan, T. K., & Walker, R. (2011). Social exclusion, personal control, self-regulation, and stress among substance abuse treatment clients. Drug and Alcohol Dependence, 113(1), 13–20. http://doi.org/10.1016/j.drugalcdep.2010.06.018

Colizzi, M., Fazio, L., Ferranti, L., Porcelli, A., Masellis, R., Marvulli, D., … Bertolino, A. (2015). Functional Genetic Variation of the Cannabinoid Receptor 1 and Cannabis Use Interact on Prefrontal Connectivity and Related Working Memory Behavior. Neuropsychopharmacology, 40(3), 640–649. http://doi.org/10.1038/npp.2014.213

Colombani, C., Croiseau, P., Fritz, S., Guillaume, F., Legarra, A., Ducrocq, V., & Robert-Granié, C. (2012). A comparison of partial least squares (PLS) and sparse PLS regressions in genomic selection in French dairy cattle. Journal of Dairy Science, 95(4), 2120–2131. http://doi.org/10.3168/jds.2011-4647

Congdon, E., & Canli, T. (2005). The endophenotype of impulsivity: reaching consilience through behavioral, genetic, and neuroimaging approaches. Behavioral and Cognitive Neuroscience Reviews, 4(4), 262–281. http://doi.org/10.1177/1534582305285980

Coon, H., Piasecki, T. M., Cook, E. H., Dunn, D., Mermelstein, R. J., Weiss, R. B., & Cannon, D. S. (2014). Association of the CHRNA4 Neuronal Nicotinic Receptor Subunit Gene with Frequency of Binge Drinking in Young Adults. Alcoholism, Clinical and Experimental Research, 38(4), 930–937. http://doi.org/10.1111/acer.12319

Corominas-Roso, M., Roncero, C., Daigre, C., Grau-Lopez, L., Ros-Cucurull, E., Rodríguez-Cintas, L., … Casas, M. (2015). Changes in brain-derived neurotrophic factor (BDNF) during abstinence could be associated with relapse in cocaine-dependent patients. Psychiatry Research, 225(3), 309–314. http://doi.org/10.1016/j.psychres.2014.12.019

Costa, L. da S., Alencar, Á. P., Neto, P. J. N., Santos, M. do S. V. dos, da Silva, C. G. L., Pinheiro, S. de F. L., … Neto, M. L. R. (2015). Risk factors for suicide in bipolar disorder: A systematic review. *Journal of Affective Disorders*, *170*, 237–254. https://doi.org/10.1016/j.jad.2014.09.003

Crawley, J. N., Heyer, W.-D., & LaSalle, J. M. (2016). Autism and Cancer Share Risk Genes, Pathways, and Drug Targets. Trends in Genetics, 32(3), 139–146. http://doi.org/10.1016/j.tig.2016.01.001

Cruciani, F., Santolamazza, P., Shen, P., Macaulay, V., Moral, P., Olckers, A., … Underhill, P. A. (2002). A Back Migration from Asia to Sub-Saharan Africa Is Supported by High-Resolution Analysis of Human Y-Chromosome Haplotypes. The American Journal of Human Genetics, 70(5), 1197–1214. http://doi.org/10.1086/340257

Cruz-Coke, R. (1982). Genetics and alcoholism. Neurobehavioral Toxicology and Teratology, 5(2), 179–180.

Culverhouse, R. C., Johnson, E. O., Breslau, N., Hatsukami, D. K., Sadler, B., Brooks, A. I., ... & Saccone, N. L. (2014). Multiple distinct CHRNB3–CHRNA6 variants are genetic risk factors for nicotine dependence in African Americans and European Americans. Addiction, 109(5), 814-822.

Dalley, J. W., Everitt, B. J., & Robbins, T. W. (2011). Impulsivity, Compulsivity, and Top-Down Cognitive Control. Neuron, 69(4), 680–694. http://doi.org/10.1016/j.neuron.2011.01.020

Dalvie, S., Fabbri, C., Ramesar, R., Serretti, A., & Stein, D. J. (2016). Glutamatergic and HPA-axis pathway genes in bipolar disorder comorbid with alcohol- and substance use disorders. *Metabolic Brain Disease*, *31*(1), 183–189. https://doi.org/10.1007/s11011-015-9762-1

Dawe, S., & Loxton, N. J. (2004). The role of impulsivity in the development of substance use and eating disorders. Neuroscience & Biobehavioral Reviews, 28(3), 343–351. http://doi.org/10.1016/j.neubiorev.2004.03.007

Davis, C., Zai, C., Levitan, R. D., Kaplan, A. S., Carter, J. C., Reid-Westoby, C., … Kennedy, J. L. (2011). Opiates, overeating and obesity: a psychogenetic analysis. International Journal of Obesity, 35(10), 1347–1354. http://doi.org/10.1038/ijo.2010.276

Dekker, L. V., & Parker, P. J. (1994). Protein kinase C - a question of specificity. *Trends in Biochemical Sciences*, *19*(2), 73–77. https://doi.org/10.1016/0968-0004(94)90038-8

Desrivières, S., Lourdusamy, A., Müller, C., Ducci, F., Wong, C. P., Kaakinen, M., … Schumann, G. (2011). Glucocorticoid receptor (NR3C1) gene polymorphisms and onset of alcohol abuse in adolescents. Addiction Biology, 16(3), 510–513. http://doi.org/10.1111/j.1369-1600.2010.00239.x

Deo, A. J., Huang, Y., Hodgkinson, C. A., Xin, Y., Oquendo, M. A., Dwork, A. J., … Haghighi, F. (2013). A large-scale candidate gene analysis of mood disorders: evidence of neurotrophic tyrosine kinase receptor and opioid receptor signaling dysfunction. *Psychiatric Genetics*, *23*(2). https://doi.org/10.1097/YPG.0b013e32835d7028

Devor, E. J., & Cloninger, C. R. (1989). Genetics of Alcoholism. Annual Review of Genetics, 23(1), 19–36. http://doi.org/10.1146/annurev.ge.23.120189.000315

DeWitt, S. J., Aslan, S., & Filbey, F. M. (2014). Adolescent risk-taking and resting state functional connectivity. Psychiatry Research: Neuroimaging, 222(3), 157–164. http://doi.org/10.1016/j.pscychresns.2014.03.009

DeWitt, S. J., Ketcherside, A., McQueeny, T. M., Dunlop, J. P., & Filbey, F. M. (2015). The hyper-sentient addict: an exteroception model of addiction. The American Journal of Drug and Alcohol Abuse, 41(5), 374–381. http://doi.org/10.3109/00952990.2015.1049701

Di Nicola, M., Tedeschi, D., De Risio, L., Pettorruso, M., Martinotti, G., Ruggeri, F., … Janiri, L. (2015). Co-occurrence of Alcohol Use Disorder and behavioral addictions: relevance of impulsivity and craving. Drug and Alcohol Dependence. http://doi.org/10.1016/j.drugalcdep.2014.12.028

Domschke, K., Dannlowski, U., Ohrmann, P., Lawford, B., Bauer, J., Kugel, H., … Baune, B. T. (2008). Cannabinoid receptor 1 (CNR1) gene: Impact on antidepressant treatment response and emotion processing in Major Depression. European Neuropsychopharmacology, 18(10), 751–759. http://doi.org/10.1016/j.euroneuro.2008.05.003

Drakenberg, K., Nikoshkov, A., Horváth, M. C., Fagergren, P., Gharibyan, A., Saarelainen, K., … Hurd, Y. L. (2006). Mu opioid receptor A118G polymorphism in association with striatal opioid neuropeptide gene expression in heroin abusers. Proceedings of the National Academy of Sciences of the United States of America, 103(20), 7883–7888. http://doi.org/10.1073/pnas.0600871103

Duan, F., Ogden, D., Xu, L., Liu, K., Lust, G., Sandler, J., … Zhang, Z. (2012). Principal component analysis of canine hip dysplasia phenotypes and their statistical power for genome-wide association mapping. Journal of Applied Statistics, 1–17. http://doi.org/10.1080/02664763.2012.740617

Dudbridge, F. (2013). Power and Predictive Accuracy of Polygenic Risk Scores. PLoS Genet, 9(3), e1003348. http://doi.org/10.1371/journal.pgen.1003348

Du, L., Huang, H., Yan, J., Kim, S., Risacher, S. L., Inlow, M., … Shen, L. (2016). Structured Sparse Canonical Correlation Analysis for Brain Imaging Genetics: An Improved GraphNet Method. Bioinformatics, btw033. http://doi.org/10.1093/bioinformatics/btw033

Du, L., Yan, J., Kim, S., Risacher, S. L., Huang, H., Inlow, M., … ADNI (2015). GN-SCCA: GraphNet Based Sparse Canonical Correlation Analysis for Brain Imaging Genetics. In Y. Guo, K. Friston, F. Aldo, S. Hill, & H. Peng (Eds.), Brain Informatics and Health (pp. 275–284). Springer International Publishing.

Dray, S. (2014). Analyzing a pair of tables: co-inertia analysis and duality diagrams. In J. Blasius & M. Greenacre, (Eds.), Visualization of verbalization of data (p. 289–300). Boca Raton, FL: CRC Press.

Edelaar, P., Roques, S., Hobson, E. A., Gonçalves da Silva, A., Avery, M. L., Russello, M. A., ... & Tella, J. L. (2015). Shared genetic diversity across the global invasive range of the monk parakeet suggests a common restricted geographic origin and the possibility of convergent selection. Molecular ecology, 24(9), 2164-2176.

Edenberg, H. J., & Foroud, T. (2006). The genetics of alcoholism: identifying specific genes through family studies. Addiction Biology, 11(3-4), 386–396. http://doi.org/10.1111/j.1369-1600.2006.00035.x

Edwards, A. C., Aliev, F., Bierut, L. J., Bucholz, K. K., Edenberg, H., Hesselbrock, V., … Dick, D. M. (2012). Genome-wide association study of comorbid depressive syndrome and alcohol dependence. *Psychiatric Genetics*, *22*(1), 31–41. https://doi.org/10.1097/YPG.0b013e32834acd07

Efron, B., & Tibshirani, R. (1993). An introduction to the bootstrap (Vol. 57). Boca Raton FL: Chapman & Hall/CRC.

Enoch, M.-A., Hodgkinson, C. A., Shen, P.-H., Gorodetsky, E., Marietta, C. A., Roy, A., & Goldman, D. (2016). GABBR1 and SLC6A1, Two Genes Involved in Modulation of GABA Synaptic Transmission, Influence Risk for Alcoholism: Results from Three Ethnically Diverse Populations. Alcoholism: Clinical and Experimental Research, 40(1), 93–101. http://doi.org/10.1111/acer.12929

Ernst, M., Grant, S. J., London, E. D., Contoreggi, C. S., Kimes, A. S., & Spurgeon, L. (2003). Decision making in adolescents with behavior disorders and adults with substance abuse. The American Journal of Psychiatry, 160(1), 33–40. http://doi.org/10.1176/appi.ajp.160.1.33

Ersche, K. D., Jones, P. S., Williams, G. B., Turton, A. J., Robbins, T. W., & Bullmore, E. T. (2012). Abnormal Brain Structure Implicated in Stimulant Drug Addiction. Science, 335(6068), 601–604. http://doi.org/10.1126/science.1214463

Ersche, K. D., Turton, A. J., Pradhan, S., Bullmore, E. T., & Robbins, T. W. (2010). Drug Addiction Endophenotypes: Impulsive Versus Sensation-Seeking Personality Traits. Biological Psychiatry, 68(8), 770–773. http://doi.org/10.1016/j.biopsych.2010.06.015

Estrada, G., Fatjó-Vilas, M., Muñoz, M. J., Pulido, G., Miñano, M. J., Toledo, E., … Fañanás, L. (2011). Cannabis use and age at onset of psychosis: further evidence of interaction with COMT Val158Met polymorphism. Acta Psychiatrica Scandinavica, 123(6), 485–492. http://doi.org/10.1111/j.1600-0447.2010.01665.x

Evenden, J. L. (1999). Varieties of impulsivity. Psychopharmacology, 146(4), 348–361. http://doi.org/10.1007/PL00005481

Everitt, B. J., & Robbins, T. W. (2016). Drug Addiction: Updating Actions to Habits to Compulsions Ten Years On. Annual Review of Psychology, 67(1), 23–50. http://doi.org/10.1146/annurev-psych-122414-033457

Fadista, J., Manning, A. K., Florez, J. C., & Groop, L. (2016). The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. European Journal of Human Genetics. http://doi.org/10.1038/ejhg.2015.269

Falcón, E., & McClung, C. A. (2009). A role for the circadian genes in drug addiction. Neuropharmacology, 56, Supplement 1, 91–96. http://doi.org/10.1016/j.neuropharm.2008.06.054

Faye, L. L., Sun, L., Dimitromanolakis, A., & Bull, S. B. (2011). A flexible genome-wide bootstrap method that accounts for ranking and threshold-selection bias in GWAS interpretation and replication study design. Statistics in Medicine, 30(15), 1898–1912. http://doi.org/10.1002/sim.4228

Feder, A., Nestler, E. J., & Charney, D. S. (2009). Psychobiology and molecular genetics of resilience. Nature Reviews Neuroscience, 10(6), 446–457. http://doi.org/10.1038/nrn2649

Feldstein Ewing, S. W., LaChance, H. A., Bryan, A., & Hutchison, K. E. (2009). Do genetic and individual risk factors moderate the efficacy of motivational enhancement therapy? Drinking outcomes with an emerging adult sample. Addiction Biology, 14(3), 356–365. http://doi.org/10.1111/j.1369-1600.2009.00149.x

Fellenberg, K., Hauser, N. C., Brors, B., Neutzner, A., Hoheisel, J. D., & Vingron, M. (2001). Correspondence analysis applied to microarray data. Proceedings of the National Academy of Sciences, 98(19), 10781–10786. http://doi.org/10.1073/pnas.181597298

Fernandez-Castillo, N., Ribases, M., Roncero, C., Casas, M., Gonzalvo, B., & Cormand, B. (2010). Association study between the DAT1, DBH and DRD2 genes and cocaine dependence in a Spanish sample. Psychiatric Genetics, 20(6), 317–320.

Fernie, G., Cole, J. C., Goudie, A. J., & Field, M. (2010). Risk-taking but not response inhibition or delay discounting predict alcohol consumption in social drinkers. Drug and Alcohol Dependence, 112(1–2), 54–61. http://doi.org/10.1016/j.drugalcdep.2010.05.011

Fernie, G., Peeters, M., Gullo, M. J., Christiansen, P., Cole, J. C., Sumnall, H., & Field, M. (2013). Multiple behavioural impulsivity tasks predict prospective alcohol involvement in adolescents. Addiction, 108(11), 1916–1923. http://doi.org/10.1111/add.12283

Filbey, F. M., & DeWitt, S. J. (2012). Cannabis Cue-elicited Craving and the Reward Neurocircuitry. Progress in Neuro-Psychopharmacology & Biological Psychiatry, 38(1), 30–35. http://doi.org/10.1016/j.pnpbp.2011.11.001

Filbey, F. M., Ray, L., Smolen, A., Claus, E. D., Audette, A., & Hutchison, K. E. (2008). Differential Neural Response to Alcohol Priming and Alcohol Taste Cues Is Associated With DRD4 VNTR and OPRM1 Genotypes. Alcoholism: Clinical and Experimental Research, 32(7), 1113–1123. http://doi.org/10.1111/j.1530-0277.2008.00692.x

Filbey, F. M., Schacht, J. P., Myers, U. S., Chavez, R. S., & Hutchison, K. E. (2009a). Individual and Additive Effects of the CNR1 and FAAH Genes on Brain Response to Marijuana Cues. Neuropsychopharmacology, 35(4), 967–975.

Filbey, F. M., Schacht, J. P., Myers, U. S., Chavez, R. S., & Hutchison, K. E. (2009b). Marijuana craving in the brain. Proceedings of the National Academy of Sciences, 106(31), 13016 – 13021. http://doi.org/10.1073/pnas.0903863106

First M, Spitzer R, Gibbon M, al. e. User's guide for the Structured Clinical Interview for DSM-IV Axis I Disorders - SCID. Washington, D. C.: American Psychiatric Press; 1997.

Franke, B., Stein, J. L., Ripke, S., Anttila, V., Hibar, D. P., van Hulzen, K. J. E., … Sullivan, P. F. (2016). Genetic influences on schizophrenia and subcortical brain volumes: large-scale proof of concept. Nature Neuroscience, advance online publication. http://doi.org/10.1038/nn.4228

Frommlet, F., Bogdan, M., & Ramsey, D. (2016). Statistical Analysis of GWAS. In Phenotypes and Genotypes (pp. 105–161). Springer London.

Frost, H. R., Li, Z., & Moore, J. H. (2015). Principal component gene set enrichment (PCGSE). BioData mining, 8(1), 25.

Furlan, E., Stoklosa, J., Griffiths, J., Gust, N., Ellis, R., Huggins, R. M., & Weeks, A. R. (2012). Small population size and extremely low levels of genetic diversity in island populations of the platypus, Ornithorhynchus anatinus. Ecology and Evolution, 2(4), 844–857. http://doi.org/10.1002/ece3.195

Gancarz, A., Jouroukhin, Y., Saito, A., Shevelkin, A., Mueller, L. E., Kamiya, A., … Pletnikov, M. V. (in press, 2016). DISC1 signaling in cocaine addiction: Towards molecular mechanisms of co-morbidity. Neuroscience Research. http://doi.org/10.1016/j.neures.2015.09.001

Garcia-Ratés, S., Camarasa, J., Escubedo, E., & Pubill, D. (2007). Methamphetamine and 3,4-methylenedioxymethamphetamine interact with central nicotinic receptors and induce their up-regulation. Toxicology and Applied Pharmacology, 223(3), 195–205. http://doi.org/10.1016/j.taap.2007.05.015

Gao, H., Wu, Y., Li, J., Li, H., Li, J., & Yang, R. (2013). Forward LASSO analysis for high-order interactions in genome-wide association study. Briefings in Bioinformatics. http://doi.org/10.1093/bib/bbt037

Gao, J., Xu, H., Weinberg, C., Huang, X., Park, Y., Hollenbeck, A., … Chen, H. (2011). An Exploratory Study on the CHRNA3-CHRNA5-CHRNB4 Cluster, Smoking, and Parkinson’s Disease. Neurodegenerative Diseases, 8(5), 296–299. http://doi.org/10.1159/000323190

Gao, X., Becker, L. C., Becker, D. M., Starmer, J. D., & Province, M. A. (2010). Avoiding the high Bonferroni penalty in genome-wide association studies. Genetic Epidemiology, 34(1), 100–105. http://doi.org/10.1002/gepi.20430

Gearhardt, A. N., Yokum, S., Orr, P. T., Stice, E., Corbin, W. R., & Brownell, K. D. (2011). Neural Correlates of Food Addiction. Arch Gen Psychiatry, archgenpsychiatry.2011.32. http://doi.org/10.1001/archgenpsychiatry.2011.32

Geboes, A.-L., Rosoux, R., Lemarchand, C., Hansen, E., & Libois, R. (2016). Genetic diversity and population structure of the Eurasian otter (Lutra lutra) in France. Mammal Research, 1–9.

Genicot, M., Huang, W., & Trendafilov, N. T. (2015). Weakly Correlated Sparse Components with Nearly Orthonormal Loadings. In Geometric Science of Information (pp. 484–490). Springer.

Ge, T., Nichols, T. E., Ghosh, D., Mormino, E. C., Smoller, J. W., Sabuncu, M. R., & Alzheimer's Disease Neuroimaging Initiative. (2015). A kernel machine method for detecting effects of interaction between multidimensional variable sets: an imaging genetics application. NeuroImage, 109, 505–514. http://doi.org/10.1016/j.neuroimage.2015.01.029

Ghosh, J., Pradhan, S., & Mittal, B. (2012). Identification of a Novel ANKK1 and Other Dopaminergic (DRD2 and DBH) Gene Variants in Migraine Susceptibility. NeuroMolecular Medicine, 15(1), 61–73. http://doi.org/10.1007/s12017-012-8195-9

Goenjian, A. K., Noble, E. P., Steinberg, A. M., Walling, D. P., Stepanyan, S. T., Dandekar, S., & Bailey, J. N. (2015). Association of COMT and TPH-2 genes with DSM-5 based PTSD symptoms. Journal of Affective Disorders, 172, 472–478. http://doi.org/10.1016/j.jad.2014.10.034

Greenacre, M. J. (1984). Theory and Applications of Correspondence Analysis. New York: Academic Press.

Graham, D. L., Edwards, S., Bachtell, R. K., DiLeone, R. J., Rios, M., & Self, D. W. (2007). Dynamic BDNF activity in nucleus accumbens with cocaine use increases self-administration and relapse. Nature Neuroscience, 10(8), 1029–1037. http://doi.org/10.1038/nn1929

Greenacre, M. J., & Degos, L. (1977). Correspondence analysis of HLA gene frequency data from 124 population samples. American Journal of Human Genetics, 29(1), 60–75.

Greenbaum, L., Rigbi, A., Teltsh, O., & Lerer, B. (2009). Role of genetic variants in the CHRNA5–CHRNA3–CHRNB4 cluster in nicotine dependence risk: importance of gene–

environment interplay. Molecular Psychiatry, 14(9), 828–830. http://doi.org/10.1038/mp.2009.25

Greenwald, M. K., Steinmiller, C. L., Śliwerska, E., Lundahl, L., & Burmeister, M. (2013). BDNF Val66Met genotype is associated with drug-seeking phenotypes in heroin-dependent individuals: a pilot study. Addiction Biology, 18(5), 836–845. http://doi.org/10.1111/j.1369-1600.2011.00431.x

Grellmann, C., Bitzer, S., Neumann, J., Westlye, L. T., Andreassen, O. A., Villringer, A., & Horstmann, A. (2015). Comparison of variants of canonical correlation analysis and partial least squares for combined analysis of MRI and genetic data. NeuroImage, 107, 289–310. http://doi.org/10.1016/j.neuroimage.2014.12.025

Guo, X., Li, Y., Ding, X., He, M., Wang, X., & Zhang, H. (2015). Association Tests of Multiple Phenotypes: ATeMP. PLoS ONE, 10(10), e0140348. http://doi.org/10.1371/journal.pone.0140348

Haller, G., Kapoor, M., Budde, J., Xuei, X., Edenberg, H., Nurnberger, J., … Goate, A. (2014). Rare missense variants in CHRNB3 and CHRNA3 are associated with risk of alcohol and cocaine dependence. Human Molecular Genetics, 23(3), 810–819. http://doi.org/10.1093/hmg/ddt463

Hancock, D. B., Levy, J. L., Gaddis, N. C., Glasheen, C., Saccone, N. L., Page, G. P., ... & Johnson, E. O. (2015). Replication of ZNF804A gene variant associations with risk of heroin addiction. Genes, Brain and Behavior, 14(8), 635-640.

Hancock, D. B., Reginsson, G. W., Gaddis, N. C., Chen, X., Saccone, N. L., Lutz, S. M., … Stefansson, K. (2015). Genome-wide meta-analysis reveals common splice site acceptor variant in CHRNA4 associated with nicotine dependence. Translational Psychiatry, 5(10), e651. http://doi.org/10.1038/tp.2015.149

Handley, E. D., Rogosch, F. A., & Cicchetti, D. (2015). Developmental pathways from child maltreatment to adolescent marijuana dependence: Examining moderation by FK506 binding protein 5 gene (FKBP5). Development and Psychopathology, 27(Special Issue 4pt2), 1489–1502. http://doi.org/10.1017/S0954579415000899

Han, S., Yang, B.-Z., Kranzler, H. R., Oslin, D., Anton, R., Farrer, L. A., & Gelernter, J. (2012). Linkage Analysis Followed by Association Show NRG1 Associated with Cannabis Dependence in African Americans. Biological Psychiatry, 72(8), 637–644. http://doi.org/10.1016/j.biopsych.2012.02.038

Han, S., Yang, B.-Z., Kranzler, H. R., Liu, X., Zhao, H., Farrer, L. A., … Gelernter, J. (2013). Integrating GWASs and Human Protein Interaction Networks Identifies a Gene Subnetwork Underlying Alcohol Dependence. *The American Journal of Human Genetics*, *93*(6), 1027–1034. https://doi.org/10.1016/j.ajhg.2013.10.021

Harrell, P. T., Lin, H.-Y., Park, J. Y., Blank, M. D., Drobes, D. J., & Evans, D. E. (2015). Dopaminergic genetic variation moderates the effect of nicotine on cigarette reward. Psychopharmacology, 1–10. http://doi.org/10.1007/s00213-015-4116-6

Hart, A. B., de Wit, H., & Palmer, A. A. (2013). Candidate Gene Studies of a Promising Intermediate Phenotype: Failure to Replicate. Neuropsychopharmacology, 38(5), 802–816. http://doi.org/10.1038/npp.2012.245

Hartz, S. M., Lin, P., Edenberg, H. J., Xuei, X., Rochberg, N., Saccone, S., … Rice, J. P. (2011). Genetic association of bipolar disorder with the β3 nicotinic receptor subunit gene. Psychiatric Genetics, 21(2), 77–84. http://doi.org/10.1097/YPG.0b013e32834135eb

Hasler, B. P., Smith, L. J., Cousins, J. C., & Bootzin, R. R. (2012). Circadian rhythms, sleep, and substance abuse. Sleep Medicine Reviews, 16(1), 67–81. http://doi.org/10.1016/j.smrv.2011.03.004

Hasler, R., Salzmann, A., Bolzan, T., Zimmermann, J., Baud, P., Giannakopoulos, P., & Perroud, N. (2015). DAT1 and DRD4 genes involved in key dimensions of adult ADHD. Neurological Sciences, 36(6), 861–869. http://doi.org/10.1007/s10072-014-2051-7

Hassan, A., Heckman, M. G., Ahlskog, J. E., Wszolek, Z. K., Serie, D. J., Uitti, R. J., … Ross, O. A. (2016). Association of Parkinson disease age of onset with DRD2, DRD3 and GRIN2B polymorphisms. Parkinsonism & Related Disorders, 22, 102–105. http://doi.org/10.1016/j.parkreldis.2015.11.016

Hastie, T., Tibshirani, R., & Friedman, J. (2013). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Tenth Edition*. Springer Science & Business Media.

Haughey, H. M., Marshall, E., Schacht, J. P., Louis, A., & Hutchison, K. E. (2008). Marijuana withdrawal and craving: influence of the cannabinoid receptor 1 (CNR1) and fatty acid amide hydrolase (FAAH) genes. Addiction, 103(10), 1678–1686. http://doi.org/10.1111/j.1360-0443.2008.02292.x

Heatherton, T. F., & Wagner, D. D. (2011). Cognitive neuroscience of self-regulation failure. Trends in Cognitive Sciences, 15(3), 132–139. http://doi.org/10.1016/j.tics.2010.12.005

Heinz, A., Mann, K., Weinberger, D. R., & Goldman, D. (2001). Serotonergic Dysfunction, Negative Mood States, and Response to Alcohol. Alcoholism: Clinical and Experimental Research, 25(4), 487–495. http://doi.org/10.1111/j.1530-0277.2001.tb02240.x

Heinzerling, K. G., Demirdjian, L., Wu, Y., & Shoptaw, S. (2016). Single nucleotide polymorphism near CREB1, rs7591784, is associated with pretreatment methamphetamine use frequency and outcome of outpatient treatment for methamphetamine use disorder. Journal of Psychiatric Research, 74, 22–29. http://doi.org/10.1016/j.jpsychires.2015.12.008

Hesterberg, T. (2011). Bootstrap. Wiley Interdisciplinary Reviews: Computational Statistics, 3, 497–526.

Hibar, D. P., Stein, J. L., Jahanshad, N., Kohannim, O., Hua, X., Toga, A. W., … Thompson, P. M. (2015). Genome-wide interaction analysis reveals replicated epistatic effects on brain structure. Neurobiology of Aging, 36, Supplement 1, S151–S158. http://doi.org/10.1016/j.neurobiolaging.2014.02.033

Hibar, D. P., Stein, J. L., Kohannim, O., Jahanshad, N., Jack, C. R., Weiner, M. W., … Thompson, P. M. (2011). Principal components regression: Multivariate, gene-based tests in imaging genomics. In 2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro (pp. 289–293). http://doi.org/10.1109/ISBI.2011.5872408

Hibar, D. P., Stein, J. L., Kohannim, O., Jahanshad, N., Saykin, A. J., Shen, L., … Thompson, P. M. (2011). Voxelwise gene-wide association study (vGeneWAS): Multivariate gene-based association testing in 731 elderly subjects. NeuroImage, 56(4), 1875 – 1891. http://doi.org/DOI: 10.1016/j.neuroimage.2011.03.077

Hiersche, M., Rühle, F., & Stoll, M. (2013). Postgwas: Advanced GWAS Interpretation in R. PLoS ONE, 8(8), e71775. http://doi.org/10.1371/journal.pone.0071775

Hirschhorn, J. N., & Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. Nature Reviews Genetics, 6(2), 95–108. http://doi.org/10.1038/nrg1521

Hodgkinson, C. A., Yuan, Q., Xu, K., Shen, P.-H., Heinz, E., Lobos, E. A., … Goldman, D. (2008). Addictions biology: haplotype-based analysis for 130 candidate genes on a single array. Alcohol and Alcoholism (Oxford, Oxfordshire), 43(5), 505–515. http://doi.org/10.1093/alcalc/agn032

Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics, 12(1), 55–67. http://doi.org/10.1080/00401706.1970.10488634

Homburger, J. R., Moreno-Estrada, A., Gignoux, C. R., Nelson, D., Sanchez, E., Ortiz-Tello, P., … Bustamante, C. D. (2015). Genomic Insights into the Ancestry and Demographic History of South America. PLoS Genet, 11(12), e1005602. http://doi.org/10.1371/journal.pgen.1005602

Honea, R., Verchinski, B. A., Pezawas, L., Kolachana, B. S., Callicott, J. H., Mattay, V. S., … Meyer-Lindenberg, A. (2009). Impact of interacting functional variants in COMT on regional gray matter volume in human brain. NeuroImage, 45(1), 44–51. http://doi.org/10.1016/j.neuroimage.2008.10.064

Hopfer, C. J., Stallings, M. C., Hewitt, J. K., & Crowley, T. J. (2003). Family Transmission of Marijuana Use, Abuse, and Dependence. Journal of the American Academy of Child &

Adolescent Psychiatry, 42(7), 834–841. http://doi.org/10.1097/01.CHI.0000046874.56865.85

Horstmann, A., Busse, F., Mathar, D., Mueller, K., Lepsien, J., Schloegl, H., … Pleger, B. (2011). Obesity-related differences between women and men in brain structure and goal-directed behavior. Frontiers in Human Neuroscience, 5. http://doi.org/10.3389/fnhum.2011.00058

Hotelling, H. (1936). Relations Between Two Sets of Variates. Biometrika, 28(3/4), 321–377. http://doi.org/10.2307/2333955

Hung, R. J., McKay, J. D., Gaborieau, V., Boffetta, P., Hashibe, M., Zaridze, D., … Brennan, P. (2008). A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. Nature, 452(7187), 633–637. http://doi.org/10.1038/nature06885

Hutchison, K. E. (2010). Substance Use Disorders: Realizing the Promise of Pharmacogenomics and Personalized Medicine. Annual Review of Clinical Psychology, 6(1), 577–589. http://doi.org/10.1146/annurev.clinpsy.121208.131441

Hutchison, K. E., Haughey, H., Niculescu, M., Schacht, J., Kaiser, A., Stitzel, J., … Filbey, F. (2008). The incentive salience of alcohol: translating the effects of genetic variant in CNR1. Archives of General Psychiatry, 65(7), 841–850.

Hutchison, K. E., Haughey, H., Niculescu, M., Schacht, J., Kaiser, A., Stitzel, J., … Filbey, F. (2008). The incentive salience of alcohol: translating the effects of genetic variant in CNR1. Archives of General Psychiatry, 65(7), 841–850.

Iacono, W. G., Vaidyanathan, U., Vrieze, S. I., & Malone, S. M. (2014). Knowns and unknowns for psychophysiological endophenotypes: Integration and response to commentaries. Psychophysiology, 51(12), 1339–1347. http://doi.org/10.1111/psyp.12358

Inoue, A., Akiyoshi, J., Muronaga, M., Masuda, K., Aizawa, S., Hirakawa, H., … Kawano, Y. (2015). Association of TMEM132D, COMT, and GABRA6 genotypes with cingulate, frontal cortex and hippocampal emotional processing in panic and major depressive disorder. International Journal of Psychiatry in Clinical Practice, 0(0), 1–9. http://doi.org/10.3109/13651501.2015.1043133

International Schizophrenia Consortium, Purcell, S. M., Wray, N. R., Stone, J. L., Visscher, P. M., O'Donovan, M. C., … Sklar, P. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature, 460(7256), 748–752. http://doi.org/10.1038/nature08185

Jensen, K. P., & Sofuoglu, M. (2015). Stress response genes and the severity of nicotine withdrawal. Pharmacogenomics, 17(1), 1–3. http://doi.org/10.2217/pgs.15.149

Jentsch, J. D., Ashenhurst, J. R., Cervantes, M. C., Groman, S. M., James, A. S., & Pennington, Z. T. (2014). Dissecting impulsivity and its relationships to drug addictions. Annals of the New York Academy of Sciences, 1327(1), 1-26.

Jiang, Y., Li, N., & Zhang, H. (2014). Identifying Genetic Variants for Addiction via Propensity Score Adjusted Generalized Kendall's Tau. Journal of the American Statistical Association, 0(ja), 00–00. http://doi.org/10.1080/01621459.2014.901223

John, B., & Lewis, K. R. (1966). Chromosome variability and geographic distribution in insects. Science (New York, N.Y.), 152(3723), 711–721. http://doi.org/10.1126/science.152.3723.711

Johnson, C., Drgon, T., Liu, Q.-R., Zhang, P.-W., Walther, D., Li, C.-Y., … Uhl, G. R. (2008). Genome wide association for substance dependence: convergent results from epidemiologic and research volunteer samples. *BMC Medical Genetics*, *9*, 113. https://doi.org/10.1186/1471-2350-9-113

Johnston, J. H., Linden, D., & van den Bree, M. B. M. (2015). Combining Stress and Dopamine Based Models of Addiction. Current Drug Abuse Reviews.

Johnstone, E. C., Elliot, K. M., David, S. P., Murphy, M. F. G., Walton, R. T., & Munafò, M. R. (2007). Association of COMT Val108/158Met Genotype with Smoking Cessation in a Nicotine Replacement Therapy Randomized Trial. Cancer Epidemiology Biomarkers & Prevention, 16(6), 1065–1069. http://doi.org/10.1158/1055-9965.EPI-06-0936

Jolliffe, I. T. (2002). Principal Components Analysis. Springer New York.

Jolliffe, I. T., Trendafilov, N. T., & Uddin, M. (2003). A modified principal component technique based on the LASSO. Journal of Computational and Graphical Statistics, 12(3), 531–547.

Jombart, T., Devillard, S., & Balloux, F. (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genetics, 11(1), 94. http://doi.org/10.1186/1471-2156-11-94

Jones, J. D., Comer, S. D., & Kranzler, H. R. (2015). The Pharmacogenetics of Alcohol Use Disorder. *Alcoholism: Clinical and Experimental Research*, *39*(3), 391–402. https://doi.org/10.1111/acer.12643

Jones, J. D., Luba, R. R., Vogelman, J. L., & Comer, S. D. (2016). Searching for evidence of genetic mediation of opioid withdrawal by opioid receptor gene polymorphisms. The American Journal on Addictions, 25(1), 41–48. http://doi.org/10.1111/ajad.12316

Josse, J., Chavent, M., Liquet, B., & Husson, F. (2012). Handling Missing Values with Regularized Iterative Multiple Correspondence Analysis. Journal of Classification, 29(1), 91–116. http://doi.org/10.1007/s00357-012-9097-0

Josse, J., & Sardy, S. (2013). Selecting thresholding and shrinking parameters with generalized SURE for low rank matrix estimation. arXiv:1310.6602 [stat]. Retrieved from http://arxiv.org/abs/1310.6602

Juhasz, G., Chase, D., Pegg, E., Downey, D., Toth, Z. G., Stones, K., … Deakin, J. W. (2009). CNR1 Gene is Associated with High Neuroticism and Low Agreeableness and Interacts with Recent Negative Life Events to Predict Current Depressive Symptoms. Neuropsychopharmacology, 34(8), 2019–2027. http://doi.org/10.1038/npp.2009.19

Jutras-Aswad, D., Jacobs, M. M., Yiannoulos, G., Roussos, P., Bitsios, P., Nomura, Y., … Hurd, Y. L. (2012). Cannabis-Dependence Risk Relates to Synergism between Neuroticism and Proenkephalin SNPs Associated with Amygdala Gene Expression: Case-Control Study. PLoS ONE, 7(6), e39243. http://doi.org/10.1371/journal.pone.0039243

Kadwell, M., Fernandez, M., Stanley, H. F., Baldi, R., Wheeler, J. C., Rosadio, R., & Bruford, M. W. (2001). Genetic analysis reveals the wild ancestors of the llama and the alpaca. Proceedings of the Royal Society of London. Series B: Biological Sciences, 268(1485), 2575–2584. http://doi.org/10.1098/rspb.2001.1774

Kalapatapu, R. K., & Delucchi, K. L. (2013). APOE e4 genotype and cigarette smoking in adults with normal cognition and mild cognitive impairment: a retrospective baseline analysis of a national dataset. The American Journal of Drug and Alcohol Abuse, 39(4), 219–226. http://doi.org/10.3109/00952990.2013.800084

Kang, G., Liu, W., Cheng, C., Wilson, C. L., Neale, G., Yang, J. J., … Srivastava, D. K. (2015). Evaluation of a two-step iterative resampling procedure for internal validation of genome-wide association studies. Journal of Human Genetics, 60(12), 729–738. http://doi.org/10.1038/jhg.2015.110

Kang, M., Park, J., Kim, D.-C., Biswas, A. K., Liu, C., & Gao, J. (2015). An integrative genomic study for multimodal genomic data using multi-block bipartite graph. In 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 563–568). http://doi.org/10.1109/BIBM.2015.7359744

Kang, S.-G., Lee, B.-H., Lee, J.-S., Chai, Y. G., Ko, K.-P., Lee, H.-J., … Shin, H. E. (2014). DRD3 Gene rs6280 Polymorphism May Be Associated with Alcohol Dependence Overall and with Lesch Type I Alcohol Dependence in Koreans. Neuropsychobiology, 69(3), 140–146. http://doi.org/10.1159/000358062

Kapur, S., Sharad, S., Singh, R. A., & Gupta, A. K. (2007). A118g polymorphism in mu opioid receptor gene (oprm1): association with opiate addiction in subjects of indian origin.

Journal of Integrative Neuroscience, 06(04), 511–522. http://doi.org/10.1142/S0219635207001635

Karch, C. M., & Goate, A. M. (2015). Alzheimer's Disease Risk Genes and Mechanisms of Disease Pathogenesis. Biological Psychiatry, 77(1), 43–51. http://doi.org/10.1016/j.biopsych.2014.05.006

Kauppi, K., Nilsson, L.-G., Persson, J., & Nyberg, L. (2014). Additive genetic effect of APOE and BDNF on hippocampus activity. NeuroImage, 89, 306–313. http://doi.org/10.1016/j.neuroimage.2013.11.049

Kazantseva, A., Gaysina, D., Kutlumbetova, Y., Kanzafarova, R., Malykh, S., Lobaskova, M., & Khusnutdinova, E. (2015). Brain derived neurotrophic factor gene (BDNF) and personality traits: The modifying effect of season of birth and sex. Progress in Neuro-Psychopharmacology and Biological Psychiatry, 56, 58–65. http://doi.org/10.1016/j.pnpbp.2014.08.001

Keith, T. (2016). GOP Candidates Address Forum On Addiction In New Hampshire. Retrieved February 11, 2016, from http://www.npr.org/2016/01/06/462114373/gop-candidates-address-forum-on-addiction-in-new-hampshire

Kemper, K. E., Daetwyler, H. D., Visscher, P. M., & Goddard, M. E. (2012). Comparing linkage and association analyses in sheep points to a better way of doing GWAS. Genetics Research, 94(04), 191–203. http://doi.org/10.1017/S0016672312000365

Kendler, K. S., Chen, X., Dick, D., Maes, H., Gillespie, N., Neale, M. C., & Riley, B. (2012). Recent advances in the genetic epidemiology and molecular genetics of substance use disorders. *Nature Neuroscience*, *15*(2), 181–189. https://doi.org/10.1038/nn.3018

Kendler, K. S., Jacobson, K. C., Prescott, C. A., & Neale, M. C. (2003). Specificity of Genetic and Environmental Risk Factors for Use and Abuse/Dependence of Cannabis, Cocaine, Hallucinogens, Sedatives, Stimulants, and Opiates in Male Twins. American Journal of Psychiatry, 160(4), 687–695. http://doi.org/10.1176/appi.ajp.160.4.687

Kendler KS, Prescott CA, Myers J, & Neale MC. (2003). THe structure of genetic and environmental risk factors for common psychiatric and substance use disorders in men and women. Archives of General Psychiatry, 60(9), 929–937. http://doi.org/10.1001/archpsyc.60.9.929

Kendler, K. S., Schmitt, E., Aggen, S. H., & Prescott, C. A. (2008). Genetic and Environmental Influences on Alcohol, Caffeine, Cannabis, and Nicotine Use From Early Adolescence to Middle Adulthood. Arch Gen Psychiatry, 65(6), 674–682. http://doi.org/10.1001/archpsyc.65.6.674

Kendler K. S., Sundquist K., Ohlsson H., & et al. (2012). Genetic and familial environmental influences on the risk for drug abuse: A national swedish adoption study. Archives of General Psychiatry, 69(7), 690–697. http://doi.org/10.1001/archgenpsychiatry.2011.2112

Kenna, G. A., Zywiak, W. H., Swift, R. M., McGeary, J. E., Clifford, J. S., Shoaff, J. R., … Leggio, L. (2014). Ondansetron Reduces Naturalistic Drinking in Nontreatment-Seeking Alcohol-Dependent Individuals with the LL 5′-HTTLPR Genotype: A Laboratory Study. Alcoholism: Clinical and Experimental Research, 38(6), 1567–1574. http://doi.org/10.1111/acer.12410

Kertes, D. A., Kalsi, G., Prescott, C. A., Kuo, P.-H., Patterson, D. G., Walsh, D., … Riley, B. P. (2011). Neurotransmitter and Neuromodulator Genes Associated With a History of Depressive Symptoms in Individuals With Alcohol Dependence. *Alcoholism: Clinical and Experimental Research*, *35*(3), 496–505. https://doi.org/10.1111/j.1530-0277.2010.01366.x

Khantzian, E. J. (2013). Addiction as a self-regulation disorder and the role of self-medication. Addiction, 108(4), 668–669. http://doi.org/10.1111/add.12004

Kishino, H., & Waddell, P. J. (2000). Correspondence analysis of genes and tissue types and finding genetic links from microarray data. Genome Informatics. Workshop on Genome Informatics, 11, 83–95.

Klengel, T., Mehta, D., Anacker, C., Rex-Haffner, M., Pruessner, J. C., Pariante, C. M., … Binder, E. B. (2013). Allele-specific FKBP5 DNA demethylation mediates gene-childhood trauma interactions. Nature Neuroscience, 16(1), 33–41. http://doi.org/10.1038/nn.3275

Kohannim, O., Hibar, D. P., Stein, J. L., Jahanshad, N., Hua, X., Rajagopalan, P., … Thompson, P. M. (2012). Discovery and Replication of Gene Influences on Brain Structure Using LASSO Regression. Frontiers in Neuroscience, 6. http://doi.org/10.3389/fnins.2012.00115

Kohannim, O., Hibar, D. P., Stein, J. L., Jahanshad, N., Jack, C. R., Weiner, M. W., … Thompson, P. M. (2011). Boosting power to detect genetic associations in imaging using multi-locus, genome-wide scans and ridge regression. In 2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro (pp. 1855–1859). http://doi.org/10.1109/ISBI.2011.5872769

Kolla, N. J., Dunlop, K., Downar, J., Links, P., Bagby, R. M., Wilson, A. A., ... & Meyer, J. H. (2016). Association of ventral striatum monoamine oxidase-A binding and functional connectivity in antisocial personality disorder with high impulsivity: A positron emission tomography and functional magnetic resonance imaging study. European Neuropsychopharmacology, 26(4), 777-786.

Koller, G., Zill, P., Rujescu, D., Ridinger, M., Pogarell, O., Fehr, C., … Preuss, U. W. (2012). Possible Association Between OPRM1 Genetic Variance at the 118 Locus and Alcohol Dependence in a Large Treatment Sample: Relationship to Alcohol Dependence Symptoms. Alcoholism: Clinical and Experimental Research, 36(7), 1230–1236. http://doi.org/10.1111/j.1530-0277.2011.01714.x

Koob, G. F., & Le Moal, M. (2001). Drug Addiction, Dysregulation of Reward, and Allostasis. Neuropsychopharmacology, 24(2), 97–129. http://doi.org/10.1016/S0893-133X(00)00195-0

Koob, G., & Kreek, M. J. (2007). Stress, Dysregulation of Drug Reward Pathways, and the Transition to Drug Dependence. The American Journal of Psychiatry, 164(8), 1149–1159. http://doi.org/10.1176/appi.ajp.2007.05030503

Kranzler, H. R., Armeli, S., Wetherill, R., Feinn, R., Tennen, H., Gelernter, J., … Pond, T. (2016). Self-efficacy mediates the effects of topiramate and GRIK1 genotype on drinking. *Addiction Biology*, *21*(2), 450–459. https://doi.org/10.1111/adb.12207

Kranzler, H. R., Hernandez-Avila, C. A., & Gelernter, J. (2002). Polymorphism of the 5-HT1B Receptor Gene (HTR1B): Strong Within-Locus Linkage Disequilibrium without Association to Antisocial Substance Dependence. Neuropsychopharmacology, 26(1), 115–122. http://doi.org/10.1016/S0893-133X(01)00283-4

Kreek, M. J., Nielsen, D. A., Butelman, E. R., & LaForge, K. S. (2005). Genetic influences on impulsivity, risk taking, stress responsivity and vulnerability to drug abuse and addiction. Nature Neuroscience, 8(11), 1450–1457. http://doi.org/10.1038/nn1583

Krishnan, A., Williams, L. J., McIntosh, A., & Abdi, H. (2011). Partial least squares (PLS) methods for neuroimaging: A tutorial and review. NeuroImage, 56, 455 – 475.

Kruse, L. C., Walter, N. a. R., & Buck, K. J. (2014). Mpdz expression in the caudolateral substantia nigra pars reticulata is crucially involved in alcohol withdrawal. *Genes, Brain and Behavior*, *13*(8), 769–776. https://doi.org/10.1111/gbb.12171

Kühn, A. B., Feis, D.-L., Schilbach, L., Kracht, L., Hess, M. E., Mauer, J., … Tittgemeyer, M. (2016). FTO gene variant modulates the neural correlates of visual food perception. NeuroImage, 128, 21–31. http://doi.org/10.1016/j.neuroimage.2015.12.049

Kupiainen, H., Kuokkanen, M., Kontto, J., Virtamo, J., Salomaa, V., Lindqvist, A., … Laitinen, T. (2016). CHRNA5/CHRNA3 Locus Associates with Increased Mortality among Smokers. COPD: Journal of Chronic Obstructive Pulmonary Disease, 0(0), 1–7. http://doi.org/10.3109/15412555.2015.1049260

Labate, A., Mumoli, L., Fratto, A., Quattrone, A., & Gambardella, A. (2013). Hippocampal sclerosis worsens autosomal dominant nocturnal frontal lobe epilepsy (ADNFLE) phenotype related to CHRNB2 mutation. European Journal of Neurology, 20(3), 591–593. http://doi.org/10.1111/j.1468-1331.2012.03839.x

Lamb, Y. N., Thompson, J. M. D., Murphy, R., Wall, C., Kirk, I. J., Morgan, A. R., … Waldie, K. E. (2014). Perceived stress during pregnancy and the catechol-O-methyltransferase (COMT) rs165599 polymorphism impacts on childhood IQ. Cognition, 132(3), 461–470. http://doi.org/10.1016/j.cognition.2014.05.009

Landgren, S., Engel, J. A., Andersson, M. E., Gonzalez-Quintela, A., Campos, J., Nilsson, S., … Jerlhag, E. (2009). Association of nAChR gene haplotypes with heavy alcohol use and body mass. Brain Research, 1305, Supplement, S72–S79. http://doi.org/10.1016/j.brainres.2009.08.026

Lang, M., Leménager, T., Streit, F., Fauth-Bühler, M., Frank, J., Juraeva, D., … Mann, K. F. (2016). Genome-wide association study of pathological gambling. *European Psychiatry*, *36*, 38–46. https://doi.org/10.1016/j.eurpsy.2016.04.001

Lang, U. E., Sander, T., Lohoff, F. W., Hellweg, R., Bajbouj, M., Winterer, G., & Gallinat, J. (2006). Association of the met66 allele of brain-derived neurotrophic factor (BDNF) with smoking. Psychopharmacology, 190(4), 433–439. http://doi.org/10.1007/s00213-006-0647-1

Latvala, A., Kuja-Halkola, R., D’Onofrio, B. M., Larsson, H., & Lichtenstein, P. (2016). Cognitive ability and risk for substance misuse in men: genetic and environmental correlations in a longitudinal nation-wide family study. *Addiction*, *111*(10), 1814–1822. https://doi.org/10.1111/add.13440

Lebart, L., Morineau, A., & Warwick, K. M. (1984). Multivariate descriptive statistical analysis: correspondence analysis and related techniques for large matrices. London: Wiley.

Lee, W., Ray, R., Bergen, A. W., Swan, G. E., Thomas, P., Tyndale, R. F., … Conti, D. V. (2012). DRD1 Associations with Smoking Abstinence Across Slow and Normal Nicotine Metabolizers. Pharmacogenetics and Genomics, 22(7), 551–554. http://doi.org/10.1097/FPC.0b013e3283539062

Le Floch, É., Guillemot, V., Frouin, V., Pinel, P., Lalanne, C., Trinchera, L., … Duchesnay, É. (2012). Significant correlation between a set of genetic polymorphisms and a functional brain network revealed by feature selection and sparse Partial Least Squares. NeuroImage, 63(1), 11–24. http://doi.org/10.1016/j.neuroimage.2012.06.061

Lessa, E. P. (1990). Multidimensional Analysis of Geographic Genetic Structure. Systematic Zoology, 39(3), 242–252. http://doi.org/10.2307/2992184

Lesscher, H. M. B., Wallace, M. J., Zeng, L., Wang, V., Deitchman, J. K., McMahon, T., … Newton, P. M. (2009). Amygdala Protein Kinase C Epsilon Controls Alcohol

Consumption. *Genes, Brain, and Behavior*, *8*(5), 493–499. https://doi.org/10.1111/j.1601-183X.2009.00485.x

Lettre, G., Lange, C., & Hirschhorn, J. N. (2007). Genetic model testing and statistical power in population-based association studies of quantitative traits. Genetic Epidemiology, 31(4), 358–362. http://doi.org/10.1002/gepi.20217

Levran, O., Peles, E., Randesi, M., Correa da Rosa, J., Ott, J., Rotrosen, J., … Kreek, M. J. (2015). Synaptic Plasticity and Signal Transduction Gene Polymorphisms and Vulnerability to Drug Addictions in Populations of European or African Ancestry. *CNS Neuroscience & Therapeutics*, *21*(11), 898–904. https://doi.org/10.1111/cns.12450

Levran, O., Peles, E., Randesi, M., Correa da Rosa, J., Ott, J., Rotrosen, J., … Kreek, M. J. (2016). Glutamatergic and GABAergic susceptibility loci for heroin and cocaine addiction in subjects of African and European ancestry. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *64*, 118–123. https://doi.org/10.1016/j.pnpbp.2015.08.003

Li, M. D., Lou, X.-Y., Chen, G., Ma, J. Z., & Elston, R. C. (2008). Gene-Gene Interactions Among CHRNA4, CHRNB2, BDNF, and NTRK2 in Nicotine Dependence. Biological Psychiatry, 64(11), 951–957. http://doi.org/10.1016/j.biopsych.2008.04.026

Li, T., Chen, C., Hu, X., Ball, D., Lin, S.-K., Chen, W., … Collier, D. A. (2004). Association analysis of the DRD4 and COMT genes in methamphetamine abuse. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics, 129B(1), 120–124. http://doi.org/10.1002/ajmg.b.30024

Li, T., Du, J., Yu, S., Jiang, H., Fu, Y., Wang, D., … Zhao, M. (2012). Pathways to Age of Onset of Heroin Use: A Structural Model Approach Exploring the Relationship of the COMT Gene, Impulsivity and Childhood Trauma. PLoS ONE, 7(11), e48735. http://doi.org/10.1371/journal.pone.0048735

Lin, D., Calhoun, V. D., & Wang, Y.-P. (2014). Correspondence between fMRI and SNP data by group sparse canonical correlation analysis. Medical Image Analysis, 18(6), 891–902. http://doi.org/10.1016/j.media.2013.10.010

Lin, D., Zhang, J., Li, J., Calhoun, V. D., Deng, H.-W., & Wang, Y.-P. (2013). Group sparse canonical correlation analysis for genomic data integration. BMC Bioinformatics, 14, 245. http://doi.org/10.1186/1471-2105-14-245

Lin, Z., & Altman, R. B. (2004). Finding Haplotype Tagging SNPs by Use of Principal Components Analysis. The American Journal of Human Genetics, 75(5), 850–861. http://doi.org/10.1086/425587

Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., & Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. Nature Methods, 8(10), 833–835. http://doi.org/10.1038/nmeth.1681

Li, T., Chen, C., Hu, X., Ball, D., Lin, S.-K., Chen, W., … Collier, D. A. (2004). Association analysis of the DRD4 and COMT genes in methamphetamine abuse. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics, 129B(1), 120–124. http://doi.org/10.1002/ajmg.b.30024

Liu, J., & Calhoun, V. D. (2014). A review of multivariate analyses in imaging genetics. Frontiers in Neuroinformatics, 8, 29. http://doi.org/10.3389/fninf.2014.00029

Liu, J., Yang, C., Shi, X., Li, C., Huang, J., Zhao, H., & Ma, S. (2013). A Penalized Multi-trait Mixed Model for Association Mapping in Pedigree-based GWAS. arXiv Preprint arXiv:1305.4413.

Liu, J. Z., Tozzi, F., Waterworth, D. M., Pillai, S. G., Muglia, P., Middleton, L., … Marchini, J. (2010). Meta-analysis and imputation refines the association of 15q25 with smoking quantity. Nature Genetics, 42(5), 436–440. http://doi.org/10.1038/ng.572

Liu, L., Zhang, D., Liu, H., & Arendt, C. (2013). Robust methods for population stratification in genome wide association studies. BMC Bioinformatics, 14, 132. http://doi.org/10.1186/1471-2105-14-132

Li, W., Freudenberg, J., Suh, Y. J., & Yang, Y. (2014). Using volcano plots and regularized-chi statistics in genetic association studies. Computational Biology and Chemistry, 48, 77–83. http://doi.org/10.1016/j.compbiolchem.2013.02.003

Long, J. C., Knowler, W. C., Hanson, R. L., Robin, R. W., Urbanek, M., Moore, E., … Goldman, D. (1998). Evidence for genetic linkage to alcohol dependence on chromosomes 4 and 11 from an autosome-wide scan in an American Indian population. American Journal of Medical Genetics, 81(3), 216–221.

Lohoff, F. W., Weller, A. E., Bloch, P. J., Nall, A. H., Ferraro, T. N., Kampman, K. M., … Berrettini, W. H. (2008). Association Between the Catechol-O-Methyltransferase Val158Met Polymorphism and Cocaine Dependence. Neuropsychopharmacology, 33(13), 3078–3084. http://doi.org/10.1038/npp.2008.126

Long, N., Gianola, D., Rosa, G. J. M., & Weigel, K. A. (2011). Application of support vector regression to genome-assisted prediction of quantitative traits. Theoretical and Applied Genetics, 123(7), 1065–1074. http://doi.org/10.1007/s00122-011-1648-y

Loos, R. J. F., & Yeo, G. S. H. (2014). The bigger picture of FTO – the first GWAS-identified obesity gene. Nature Reviews. Endocrinology, 10(1), 51–61. http://doi.org/10.1038/nrendo.2013.227

López, J. F., Akil, H., & Watson, S. J. (1999). Neural circuits mediating stress. Biological Psychiatry, 46(11), 1461–1471. http://doi.org/10.1016/S0006-3223(99)00266-8

Loth, E., Carvalho, F., & Schumann, G. (2011). The contribution of imaging genetics to the development of predictive markers for addictions. Trends in Cognitive Sciences, 15(9), 436–446.

Lu, A. T., Ogdie, M. N., Järvelin, M.-R., Moilanen, I. K., Loo, S. K., McCracken, J. T., … Smalley, S. L. (2008). Association of the cannabinoid receptor gene (CNR1) with ADHD and post-traumatic stress disorder. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics, 147B(8), 1488–1494. http://doi.org/10.1002/ajmg.b.30693

Lubke, G. H., Stephens, S. H., Lessem, J. M., Hewitt, J. K., & Ehringer, M. A. (2012). The CHRNA5/A3/B4 Gene Cluster and Tobacco, Alcohol, Cannabis, Inhalants and Other Substance Use Initiation: Replication and New Findings Using Mixture Analyses. Behavior Genetics, 42(4), 636–646. http://doi.org/10.1007/s10519-012-9529-y

Lull, M. E., Freeman, W. M., Vrana, K. E., & Mash, D. C. (2008). Correlating Human and Animal Studies of Cocaine Abuse and Gene Expression. *Annals of the New York Academy of Sciences*, *1141*(1), 58–75. https://doi.org/10.1196/annals.1441.013

Luna, F. de O. (2013). Population genetics and conservation strategies for the West Indian manatee (Trichechus manatus Linnaeaus,1758) in Brazil. Retrieved from http://repositorio.ufpe.br:8080/xmlui/handle/123456789/10681

Lynskey, M. T., Agrawal, A., Henders, A., Nelson, E. C., Madden, P. A. F., & Martin, N. G. (2012). An Australian Twin Study of Cannabis and Other Illicit Drug Use and Misuse, and Other Psychopathology. Twin Research and Human Genetics, 15(05), 631–641. http://doi.org/10.1017/thg.2012.41

Maher, B. S. (2015). Polygenic Scores in Epidemiology: Risk Prediction, Etiology, and Clinical Utility. Current Epidemiology Reports, 1–6.

Malinvaud, E. (1987). Data analysis in applied socio-economic statistics with special consideration of correspondence analysis, Marketing Science Conference, Jouy en Josas, France, 1987.

Mallard, T. T., Doorley, J., Esposito-Smythers, C. L., & McGeary, J. E. (2016). Dopamine D4 receptor VNTR polymorphism associated with greater risk for substance abuse among adolescents with disruptive behavior disorders: Preliminary results. The American Journal on Addictions, 25(1), 56–61. http://doi.org/10.1111/ajad.12320

Manor, O., & Segal, E. (2013). Predicting Disease Risk Using Bootstrap Ranking and Classification Algorithms. PLoS Comput Biol, 9(8), e1003200. http://doi.org/10.1371/journal.pcbi.1003200

Martinotti, G., Carli, V., Tedeschi, D., Di Giannantonio, M., Roy, A., Janiri, L., & Sarchiapone, M. (2009). Mono- and polysubstance dependent subjects differ on social factors, childhood trauma, personality, suicidal behaviour, and comorbid Axis I diagnoses. Addictive Behaviors, 34(9), 790–793. http://doi.org/10.1016/j.addbeh.2009.04.012

Matsumoto, M., Weickert, C. S., Akil, M., Lipska, B. K., Hyde, T. M., Herman, M. M., … Weinberger, D. R. (2003). Catechol O-methyltransferase mRNA expression in human and rat brain: evidence for a role in cortical neuronal function. Neuroscience, 116(1), 127–137.

McEachin, R. C., Saccone, N. L., Saccone, S. F., Kleyman-Smith, Y. D., Kar, T., Kare, R. K., … McInnis, M. G. (2010). Modeling complex genetic and environmental influences on comorbid bipolar disorder with tobacco use disorder. BMC Medical Genetics, 11, 14. http://doi.org/10.1186/1471-2350-11-14

McIntosh, A., Bookstein, F., Haxby, J., & Grady, C. (1996). Spatial pattern analysis of functional brain images using partial least squares. NeuroImage, 3, 143–157.

McIntosh, A., & Lobaugh, N. J. (2004). Partial least squares analysis of neuroimaging data: applications and advances. Neuroimage, 23, S250–S263.

McIntosh, A. R., & Mišić, B. (2013). Multivariate Statistical Analyses for Neuroimaging Data. Annual Review of Psychology, 64, 499–525.

McReynolds, J. R., Doncheck, E. M., Vranjkovic, O., Ganzman, G. S., Baker, D. A., Hillard, C. J., & Mantsch, J. R. (2015). CB1 receptor antagonism blocks stress-potentiated reinstatement of cocaine seeking in rats. Psychopharmacology, 1–11. http://doi.org/10.1007/s00213-015-4092-x

Meda, S. A., Jagannathan, K., Gelernter, J., Calhoun, V. D., Liu, J., Stevens, M. C., & Pearlson, G. D. (2010). A pilot multivariate parallel ICA study to investigate differential linkage between neural networks and genetic profiles in schizophrenia. NeuroImage, 53(3), 1007–1015. http://doi.org/10.1016/j.neuroimage.2009.11.052

Meda, S. A., Narayanan, B., Liu, J., Perrone-Bizzozero, N. I., Stevens, M. C., Calhoun, V. D., … Pearlson, G. D. (2012). A large scale multivariate parallel ICA method reveals novel imaging–genetic relationships for Alzheimer's disease in the ADNI cohort. NeuroImage, 60(3), 1608–1621. http://doi.org/10.1016/j.neuroimage.2011.12.076

Meda, S. A., Stevens, M. C., Potenza, M. N., Pittman, B., Gueorguieva, R., Andrews, M. M., … Pearlson, G. D. (2009). Investigating the behavioral and self-report constructs of impulsivity domains using principal component analysis. Behavioural Pharmacology, 20(5-6), 390–399. http://doi.org/10.1097/FBP.0b013e32833113a3

Mei, H., Chen, W., Dellinger, A., He, J., Wang, M., Yau, C., … Berenson, G. (2010). Principal-component-based multivariate regression for genetic association studies of metabolic syndrome components. BMC Genetics, 11(1), 100. http://doi.org/10.1186/1471-2156-11-100

Melroy-Greif, W. E., Stitzel, J. A., & Ehringer, M. A. (2016). Nicotinic acetylcholine receptors: upregulation, age-related effects and associations with drug use. Genes, Brain and Behavior, 15(1), 89-107.

Menozzi, P., Piazza, A., & Cavalli-Sforza, L. (1978). Synthetic maps of human gene frequencies in Europeans. Science, 201(4358), 786–792. http://doi.org/10.1126/science.356262

Merritt, L. L., Martin, B. R., Walters, C., Lichtman, A. H., & Damaj, M. I. (2008). The Endogenous Cannabinoid System Modulates Nicotine Reward and Dependence. The Journal of Pharmacology and Experimental Therapeutics, 326(2), 483–492. http://doi.org/10.1124/jpet.108.138321

Meule, A., Lutz, A. P. C., Vögele, C., & Kübler, A. (2014). Impulsive reactions to food-cues predict subsequent food craving. Eating Behaviors, 15(1), 99–105. http://doi.org/10.1016/j.eatbeh.2013.10.023

Meyer-Lindenberg, A., & Weinberger, D. R. (2006). Intermediate phenotypes and genetic mechanisms of psychiatric disorders. Nature Reviews Neuroscience, 7(10), 818–827. http://doi.org/10.1038/nrn1993

Michaelson, J., Alberts, R., Schughart, K., & Beyer, A. (2010). Data-driven assessment of eQTL mapping methods. BMC Genomics, 11(1), 502. http://doi.org/10.1186/1471-2164-11-502

Mick, E., Wozniak, J., Wilens, T. E., Biederman, J., & Faraone, S. V. (2009). Family-based association study of the BDNF, COMT and serotonin transporter genes and DSM-IV bipolar-I disorder in children. BMC Psychiatry, 9, 2. http://doi.org/10.1186/1471-244X-9-2

Miclaus, K., Wolfinger, R., & Czika, W. (2009). SNP selection and multidimensional scaling to quantify population structure. Genetic Epidemiology, 33(6), 488–496. http://doi.org/10.1002/gepi.20401

Millar, J. K., Wilson-Annan, J. C., Anderson, S., Christie, S., Taylor, M. S., Semple, C. A., … Porteous, D. J. (2000). Disruption of two novel genes by a translocation co-segregating with schizophrenia. Human Molecular Genetics, 9(9), 1415–1423.

Mitchell, M. R., & Potenza, M. N. (2014). Addictions and Personality Traits: Impulsivity and Related Constructs. Current Behavioral Neuroscience Reports, 1(1), 1–12. http://doi.org/10.1007/s40473-013-0001-y

Mitrovski, P., Heinze, D. A., Broome, L., Hoffmann, A. A., & Weeks, A. R. (2007). High levels of variation despite genetic fragmentation in populations of the endangered mountain pygmy-possum, Burramys parvus, in alpine Australia. Molecular Ecology, 16(1), 75–87. http://doi.org/10.1111/j.1365-294X.2006.03125.x

Mittag, F., Büchel, F., Saad, M., Jahn, A., Schulte, C., Bochdanovits, Z., … Sharma, M. (2012). Use of support vector machines for disease risk prediction in genome-wide association studies: Concerns and opportunities. Human Mutation, 33(12), 1708–1718. http://doi.org/10.1002/humu.22161

Mitteroecker, P., Cheverud, J. M., & Pavlicev, M. (2016). Multivariate Analysis of Genotype-Phenotype Association. Genetics, genetics.115.181339. http://doi.org/10.1534/genetics.115.181339

Miyajima, F., Quinn, J. P., Horan, M., Pickles, A., Ollier, W. E., Pendleton, N., & Payton, A. (2008). Additive effect of BDNF and REST polymorphisms is associated with improved general cognitive ability. Genes, Brain, and Behavior, 7(7), 714–719. http://doi.org/10.1111/j.1601-183X.2008.00409.x

Moeller, S. J., London, E. D., & Northoff, G. (2016). Neuroimaging markers of glutamatergic and GABAergic systems in drug addiction: relationships to resting-state functional connectivity. Neuroscience & Biobehavioral Reviews. http://doi.org/10.1016/j.neubiorev.2015.11.010

Moeller, S. J., Parvaz, M. A., Shumay, E., Beebe-Wang, N., Konova, A. B., Alia-Klein, N., … Goldstein, R. Z. (2013). Gene × Abstinence Effects on Drug Cue Reactivity in Addiction: Multimodal Evidence. The Journal of Neuroscience, 33(24), 10027–10036. http://doi.org/10.1523/JNEUROSCI.0695-13.2013

Moeller, S. J., Parvaz, M. A., Shumay, E., Wu, S., Beebe-Wang, N., Konova, A. B., … Goldstein, R. Z. (2014). Monoamine polygenic liability in health and cocaine dependence: Imaging genetics study of aversive processing and associations with depression symptomatology. Drug and Alcohol Dependence, 140, 17–24. http://doi.org/10.1016/j.drugalcdep.2014.04.019

Monteleone, P., Bifulco, M., Di Filippo, C., Gazzerro, P., Canestrelli, B., Monteleone, F., … Maj, M. (2009). Association of CNR1 and FAAH endocannabinoid gene polymorphisms with anorexia nervosa and bulimia nervosa: evidence for synergistic effects. Genes, Brain and Behavior, 8(7), 728–732. http://doi.org/10.1111/j.1601-183X.2009.00518.x

Monteleone, P., Bifulco, M., Maina, G., Tortorella, A., Gazzerro, P., Proto, M. C., … Maj, M. (2010). Investigation of CNR1 and FAAH endocannabinoid gene polymorphisms in bipolar disorder and major depression. Pharmacological Research, 61(5), 400–404. http://doi.org/10.1016/j.phrs.2010.01.002

Morrow, J. D., & Flagel, S. B. (2016). Neuroscience of resilience and vulnerability for addiction medicine: From genes to behavior. Progress in brain research, 223, 3-18.

Moser, G., Tier, B., Crump, R. E., Khatkar, M. S., & Raadsma, H. W. (2009). A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. Genetics Selection Evolution, 41(1), 56. http://doi.org/10.1186/1297-9686-41-56

Munafò, M. R., & Flint, J. (2011). Dissecting the genetic architecture of human personality. Trends in Cognitive Sciences, 15(9), 395–400. http://doi.org/10.1016/j.tics.2011.07.007

Munafò, M. R., Matheson, I. J., & Flint, J. (2007). Association of the DRD2 gene Taq1A polymorphism and alcoholism: a meta-analysis of case–control studies and evidence of publication bias. Molecular Psychiatry, 12(5), 454–461. http://doi.org/10.1038/sj.mp.4001938

Mus, L. V., Dravolina, O. A., Bespalov, A. Y., Käenmäki, M., Talka, R., Salminen, O., … Zvartau, E. E. (2013). Effects of Catechol-O-Methyltransferase Deficiency on the Reinforcing Effects of Cocaine (an experimental study). Neuroscience and Behavioral Physiology, 43(8), 913–917. http://doi.org/10.1007/s11055-013-9828-7

National Institute on Drug Addiction (2010). Report from the NIDA Science of Genetics Council Review Work Group. Retrieved April 1, 2012, from: http://www.drugabuse.gov/publications/infofacts/understanding-drug-abuse-addiction

National Institute on Mental Health (2016). Schizophrenia’s strongest known risk deconstructed. Retrieved March 22, 2016 from: http://www.nih.gov/news-events/news-releases/schizophrenias-strongest-known-genetic-risk-deconstructed

Naqvi, N. H., & Bechara, A. (2010). The insula and drug addiction: an interoceptive view of pleasure, urges, and decision-making. Brain Structure & Function, 214(0), 435–450. http://doi.org/10.1007/s00429-010-0268-7

Newton, P. M., Kim, J. A., McGeehan, A. J., Paredes, J. P., Chu, K., Wallace, M. J., … Messing, R. O. (2007). Increased response to morphine in mice lacking protein kinase C epsilon. *Genes, Brain and Behavior*, *6*(4), 329–338. https://doi.org/10.1111/j.1601-183X.2006.00261.x

Newton, P. M., & Messing, R. O. (2006). Intracellular signaling pathways that regulate behavioral responses to ethanol. *Pharmacology & Therapeutics*, *109*(1–2), 227–237. https://doi.org/10.1016/j.pharmthera.2005.07.004

Nikpay, M., Šeda, O., Tremblay, J., Petrovich, M., Gaudet, D., Kotchen, T. A., … Hamet, P. (2012). Genetic mapping of habitual substance use, obesity-related traits, responses to mental and physical stress, and heart rate and blood pressure measurements reveals shared genes that are overrepresented in the neural synapse. *Hypertension Research*, *35*(6), 585–591. https://doi.org/10.1038/hr.2011.233

Nikolova, Y. S., Ferrell, R. E., Manuck, S. B., & Hariri, A. R. (2011). Multilocus genetic profile for dopamine signaling predicts ventral striatum reactivity. Neuropsychopharmacology: Official Publication of the American College of Neuropsychopharmacology, 36(9), 1940–1947. http://doi.org/10.1038/npp.2011.82

Noël, X., Brevers, D., & Bechara, A. (2013). A neurocognitive approach to understanding the neurobiology of addiction. Current Opinion in Neurobiology, 23(4), 632–638. http://doi.org/10.1016/j.conb.2013.01.018

Nunes, S. O. V., de Castro, M. R. P., Watanabe, M. A. E., Guembarovski, R. L., Vargas, H. O., Reiche, E. M. V., … Berk, M. (2014). Genetic polymorphisms in glutathione-S-transferases are associated with anxiety and mood disorders in nicotine dependence. Psychiatric Genetics, 24(3), 87.

O’Donovan, M. C., & Owen, M. J. (2016). The implications of the shared genetics of psychiatric disorders. *Nature Medicine*, *advance online publication*. https://doi.org/10.1038/nm.4196

Okahisa, Y., Kodama, M., Takaki, M., Inada, T., Uchimura, N., Yamada, M., … Ujike, H. (2011). Association Study of Two Cannabinoid Receptor Genes, CNR1 and CNR2, with Methamphetamine Dependence. Current Neuropharmacology, 9(1), 183–189. http://doi.org/10.2174/157015911795017191

Oliveira, F. C. de, Borges, C. C. H., Almeida, F. N., Silva, F. F. e, Verneque, R. da S., Silva, M. V. G. da, & Arbex, W. (2014). SNPs selection using support vector regression and genetic algorithms in GWAS. BMC Genomics, 15(Suppl 7), S4. http://doi.org/10.1186/1471-2164-15-S7-S4

Onori, N., Turchi, C., Solito, G., Gesuita, R., Buscemi, L., & Tagliabracci, A. (2010). GABRA2 and Alcohol Use Disorders: No Evidence of an Association in an Italian Case–Control Study. Alcoholism: Clinical and Experimental Research, 34(4), 659–668. http://doi.org/10.1111/j.1530-0277.2009.01135.x

Onwuameze, O. E., Nam, K. W., Epping, E. A., Wassink, T. H., Ziebell, S., Andreasen, N. C., & Ho, B.-C. (2013). MAPK14 and CNR1 gene variant interactions: effects on brain volume deficits in schizophrenia patients with marijuana misuse. Psychological Medicine, 43(03), 619–631. http://doi.org/10.1017/S0033291712001559

O'Reilly, P. F., Hoggart, C. J., Pomyen, Y., Calboli, F. C. F., Elliott, P., Jarvelin, M.-R., & Coin, L. J. M. (2012). MultiPhen: Joint Model of Multiple Phenotypes Can Increase Discovery in GWAS. PLoS ONE, 7(5), e34861. http://doi.org/10.1371/journal.pone.0034861

Palmer, A. A., & de Wit, H. (2012). Translational genetic approaches to substance use disorders: bridging the gap between mice and humans. *Human Genetics*, *131*(6), 931–939. https://doi.org/10.1007/s00439-011-1123-5

Palmer, R. H. C., Brick, L., Nugent, N. R., Bidwell, L. C., McGeary, J. E., Knopik, V. S., & Keller, M. C. (2015). Examining the role of common genetic variants on alcohol, tobacco, cannabis and illicit drug dependence: genetics of vulnerability to drug dependence. Addiction (Abingdon, England), 110(3), 530–537. http://doi.org/10.1111/add.12815

Papenberg, G., Bäckman, L., Nagel, I. E., Nietfeld, W., Schröder, J., Bertram, L., … Li, S.-C. (2014). COMT polymorphism and memory dedifferentiation in old age. Psychology and Aging, 29(2), 374–383. http://doi.org/10.1037/a0033225

Papiol, S., Mitjans, M., Assogna, F., Piras, F., Hammer, C., Caltagirone, C., … Spalletta, G. (2014). Polygenic determinants of white matter volume derived from GWAS lack reproducibility in a replicate sample. Translational Psychiatry, 4(2), e362. http://doi.org/10.1038/tp.2013.126

Parekh, P. K., Ozburn, A. R., & McClung, C. A. (2015). Circadian clock genes: Effects on dopamine, reward and addiction. Alcohol, 49(4), 341–349. http://doi.org/10.1016/j.alcohol.2014.09.034

Parsons, L. H., & Hurd, Y. L. (2015). Endocannabinoid signalling in reward and addiction. Nature Reviews Neuroscience, 16(10), 579–594. http://doi.org/10.1038/nrn4004

Paul-Samojedny, M., Kowalczyk, M., Suchanek, R., Owczarek, A., Fila-Danilow, A., Szczygiel, A., & Kowalski, J. (2010). Functional Polymorphism in the Interleukin-6 and Interleukin-10 Genes in Patients with Paranoid Schizophrenia —A Case-Control Study. Journal of Molecular Neuroscience, 42(1), 112–119. http://doi.org/10.1007/s12031-010-9365-6

Paus, T., Keshavan, M., & Giedd, J. N. (2008). Why do many psychiatric disorders emerge during adolescence? *Nature Reviews Neuroscience*, *9*(12), 947–957. https://doi.org/10.1038/nrn2513

Pelayo-Terán, J. M., Pérez-Iglesias, R., Mata, I., Carrasco-Marín, E., Vázquez-Barquero, J. L., & Crespo-Facorro, B. (2010). Catechol-O-Methyltransferase (COMT) Val158Met variations and cannabis use in first-episode non-affective psychosis: Clinical-onset implications. Psychiatry Research, 179(3), 291–296. http://doi.org/10.1016/j.psychres.2009.08.022

Peres-Neto, P. R., Jackson, D. A., & Somers, K. M. (2005). How many principal components? stopping rules for determining the number of non-trivial axes revisited. Computational Statistics & Data Analysis, 49, 974–997.

Pérez-Enciso, M., Toro, M. A., Tenenhaus, M., & Gianola, D. (2003). Combining gene expression and molecular marker information for mapping complex trait genes: a simulation study. Genetics, 164(4), 1597–1606.

Perreau-Lenz, S., & Spanagel, R. (2015). Clock genes × stress × reward interactions in alcohol and substance use disorders. Alcohol, 49(4), 351–357. http://doi.org/10.1016/j.alcohol.2015.04.003

Perry, J., & Carroll, M. (2008). The role of impulsive behavior in drug abuse. Psychopharmacology, 200(1), 1–26. http://doi.org/10.1007/s00213-008-1173-0

Pfeifer, P., Sariyar, M., Eggermann, T., Zerres, K., Vernaleken, I., Tüscher, O., & Fehr, C. (2015). Alcohol Consumption in Healthy OPRM1 G Allele Carriers and Its Association with Impulsive Behavior. Alcohol and Alcoholism, agv019. http://doi.org/10.1093/alcalc/agv019

Pharo, H., Sim, C., Graham, M., Gross, J., & Hayne, H. (2011). Risky business: Executive function, personality, and reckless behavior during adolescence and emerging adulthood. Behavioral Neuroscience, 125(6), 970–978. http://doi.org/10.1037/a0025768

Potkin, S. G., Turner, J. A., Guffanti, G., Lakatos, A., Fallon, J. H., Nguyen, D. D., … FBIRN. (2009). A Genome-Wide Association Study of Schizophrenia Using Brain Activation as a Quantitative Phenotype. Schizophrenia Bulletin, 35(1), 96–108. http://doi.org/10.1093/schbul/sbn155

Powers, M. S., Breit, K. R., & Chester, J. A. (2015). Genetic Versus Pharmacological Assessment of the Role of Cannabinoid Type 2 Receptors in Alcohol Reward-Related Behaviors. Alcoholism: Clinical and Experimental Research, 39(12), 2438–2446. http://doi.org/10.1111/acer.12894

Proudnikov, D., Kroslak, T., Sipe, J. C., Randesi, M., Li, D., Hamon, S., … Kreek, M. J. (2010). Association of polymorphisms of the cannabinoid receptor (CNR1) and fatty acid amide hydrolase (FAAH) genes with heroin addiction: impact of long repeats of CNR1. The Pharmacogenomics Journal, 10(3), 232–242. http://doi.org/10.1038/tpj.2009.59

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., … Sham, P. C. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *American Journal of Human Genetics*, *81*(3), 559–575.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Rahman, S., Engleman, E. A., & Bell, R. L. (2016). Chapter Six-Recent Advances in Nicotinic Receptor Signaling in Alcohol Abuse and Alcoholism. Progress in molecular biology and translational science, 137, 183-201.

Ramaekers, J., van Wel, J., Spronk, D., Franke, B., Kenis, G., Toennes, S., … Verkes, R. (2015). Cannabis and cocaine decrease cognitive impulse control and functional corticostriatal connectivity in drug users with low activity DBH genotypes. Brain Imaging and Behavior, 1–10.

Ray, R., Ruparel, K., Newberg, A., Wileyto, E. P., Loughead, J. W., Divgi, C., … Lerman, C. (2011). Human Mu Opioid Receptor (OPRM1 A118G) Polymorphism Is Associated with Brain Mu-Opioid Receptor Binding Potential in Smokers. Proceedings of the National Academy of Sciences, 108(22), 9268–9273. http://doi.org/10.1073/pnas.1018699108

Read, J. P., Colder, C. R., Merrill, J. E., Ouimette, P., White, J., & Swartout, A. (2012). Trauma and posttraumatic stress symptoms predict alcohol and other drug consequence trajectories in the first year of college. Journal of Consulting and Clinical Psychology, 80(3), 426–439. http://doi.org/10.1037/a0028210

Reed, J. E., McCleery, R. A., Silvy, N. J., Smeins, F. E., & Brightsmith, D. J. (2014). Monk parakeet nest-site selection of electric utility structures in Texas. Landscape and Urban Planning, 129, 65–72. http://doi.org/10.1016/j.landurbplan.2014.04.016

Reitz, C., & Mayeux, R. (2009). Endophenotypes in normal brain morphology and Alzheimer's disease: a review. Neuroscience, 164(1), 174–190. http://doi.org/10.1016/j.neuroscience.2009.04.006

Rivero, O., Selten, M. M., Sich, S., Popp, S., Bacmeister, L., Amendola, E., … Lesch, K. P. (2015). Cadherin-13, a risk gene for ADHD and comorbid disorders, impacts GABAergic function in hippocampus and cognition. Translational Psychiatry, 5(10), e655. http://doi.org/10.1038/tp.2015.152

Robbins, T. W., Gillan, C. M., Smith, D. G., de Wit, S., & Ersche, K. D. (2012). Neurocognitive endophenotypes of impulsivity and compulsivity: towards dimensional psychiatry. Trends in Cognitive Sciences, 16(1), 81–91. http://doi.org/10.1016/j.tics.2011.11.009

Roberts, N. P., Roberts, P. A., Jones, N., & Bisson, J. I. (2015). Psychological interventions for post-traumatic stress disorder and comorbid substance use disorder: A systematic review and meta-analysis. Clinical Psychology Review, 38, 25–38. http://doi.org/10.1016/j.cpr.2015.02.007

Robinson, T. E., & Berridge, K. C. (2008). The incentive sensitization theory of addiction: some current issues. Philosophical Transactions of the Royal Society B: Biological Sciences, 363(1507), 3137–3146. http://doi.org/10.1098/rstb.2008.0093

Rodd, Z. A., Bertsch, B. A., Strother, W. N., Le-Niculescu, H., Balaraman, Y., Hayden, E., … Niculescu, A. B. (2006). Candidate genes, pathways and mechanisms for alcoholism: an expanded convergent functional genomics approach. *The Pharmacogenomics Journal*, *7*(4), 222–256. https://doi.org/10.1038/sj.tpj.6500420

Rosa, A., Peralta, V., Cuesta, M. J., Zarzuela, A., Serrano, F., Martínez-Larrea, A., & Fañanás, L. (2004). New Evidence of Association Between COMT Gene and Prefrontal Neurocognitive Function in Healthy Individuals From Sibling Pairs Discordant for Psychosis. American Journal of Psychiatry, 161(6), 1110–1112. http://doi.org/10.1176/appi.ajp.161.6.1110

Roshan, U., Chikkagoudar, S., Wei, Z., Wang, K., & Hakonarson, H. (2011). Ranking causal variants and associated regions in genome-wide association studies by the support vector machine and random forest. Nucleic Acids Research, 39(9), e62–e62. http://doi.org/10.1093/nar/gkr064

Rubens, M., Ramamoorthy, V., Attonito, J., Saxena, A., Appunni, S., Shehadeh, N., & Dévieux, J. G. (2015). A review of 5-HT transporter linked promoter region (5-HTTLPR) polymorphism and associations with alcohol use problems and sexual risk behaviors. Journal of Community Genetics, 1–10. http://doi.org/10.1007/s12687-015-0253-1

Ruocco, A. C., Rodrigo, A. H., Carcone, D., McMain, S., Jacobs, G., & Kennedy, J. L. (2016). Tryptophan hydroxylase 1 gene polymorphisms alter prefrontal cortex activation during response inhibition. Neuropsychology, 30(1), 18–27. http://doi.org/10.1037/neu0000237

Saccone, S. F., Hinrichs, A. L., Saccone, N. L., Chase, G. A., Konvicka, K., Madden, P. A. F., … Bierut, L. J. (2007). Cholinergic nicotinic receptor genes implicated in a nicotine dependence association study targeting 348 candidate genes with 3713 SNPs. Human Molecular Genetics, 16(1), 36–49. http://doi.org/10.1093/hmg/ddl438

Saccone, S. F., Saccone, N. L., Swan, G. E., Madden, P. A. F., Goate, A. M., Rice, J. P., & Bierut, L. J. (2008). Systematic biological prioritization after a genome-wide association study: an application to nicotine dependence. Bioinformatics (Oxford, England), 24(16), 1805–1811. http://doi.org/10.1093/bioinformatics/btn315

Sambataro, F., Reed, J. D., Murty, V. P., Das, S., Tan, H. Y., Callicott, J. H., … Mattay, V. S. (2009). Catechol-O-methyltransferase valine(158)methionine polymorphism modulates brain networks underlying working memory across adulthood. Biological Psychiatry, 66(6), 540–548. http://doi.org/10.1016/j.biopsych.2009.04.014

Samochowiec, J., Kucharska-Mazur, J., Grzywacz, A., Jabłoński, M., Rommelspacher, H., Samochowiec, A., … Pełka-Wysiecka, J. (2006). Family-based and case-control study of DRD2, DAT, 5HTT, COMT genes polymorphisms in alcohol dependence. Neuroscience Letters, 410(1), 1–5. http://doi.org/10.1016/j.neulet.2006.05.005

Sampaio, A. S., Hounie, A. G., Petribú, K., Cappi, C., Morais, I., Vallada, H., … Miguel, E. C. (2015). COMT and MAO-A Polymorphisms and Obsessive-Compulsive Disorder: A Family-Based Association Study. PLoS ONE, 10(3), e0119592. http://doi.org/10.1371/journal.pone.0119592

Saporta, G. (2011). Probabilités, analyse des données et statistique. Paris: Technip.

Schacht, J. P., Hutchison, K. E., & Filbey, F. M. (2012). Associations between cannabinoid receptor-1 (CNR1) variation and hippocampus and amygdala volumes in heavy cannabis users. Neuropsychopharmacology: Official Publication of the American College of Neuropsychopharmacology, 37(11), 2368–2376. http://doi.org/10.1038/npp.2012.92

Schellekens, A. F. A., Franke, B., Ellenbroek, B., Cools, A., de Jong, C. A. J., Buitelaar, J. K., & Verkes, R.-J. (2013). COMT Val158Met modulates the effect of childhood adverse experiences on the risk of alcohol dependence. Addiction Biology, 18(2), 344–356. http://doi.org/10.1111/j.1369-1600.2012.00438.x

Schifano, E. D., Li, L., Christiani, D. C., & Lin, X. (2013). Genome-wide Association Analysis for Multiple Continuous Secondary Phenotypes. American Journal of Human Genetics, 92(5), 744–759. http://doi.org/10.1016/j.ajhg.2013.04.004

Schizophrenia Working Group of the Psychiatric Genomics Consortium. (2014). Biological insights from 108 schizophrenia-associated genetic loci. Nature, 511(7510), 421–427. http://doi.org/10.1038/nature13595

Schlaepfer, I. R., Hoft, N. R., & Ehringer, M. A. (2008). The genetic components of alcohol and nicotine co-addiction: From genes to behavior. Current Drug Abuse Reviews, 1(2), 124–134.

Schlüter, T., Winz, O., Henkel, K., Eggermann, T., Mohammadkhani-Shali, S., Dietrich, C., … Vernaleken, I. (2016). MAOA-VNTR polymorphism modulates context-dependent dopamine release and aggressive behavior in males. NeuroImage, 125, 378–385. http://doi.org/10.1016/j.neuroimage.2015.10.031

Schwantes-An, T.-H., Zhang, J., Chen, L.-S., Hartz, S. M., Culverhouse, R. C., Chen, X., … Saccone, N. L. (2015). Association of the oprm1 variant rs1799971 (a118g) with non-specific liability to substance dependence in a collaborative de novo meta-analysis of european-ancestry cohorts. Behavior Genetics. http://doi.org/10.1007/s10519-015-9737-3

Sekar, A., Bialas, A. R., de Rivera, H., Davis, A., Hammond, T. R., Kamitaki, N., … McCarroll, S. A. (2016). Schizophrenia risk from complex variation of complement component 4. Nature, advance online publication. http://doi.org/10.1038/nature16549

Shahani, N., Seshadri, S., Jaaro-Peled, H., Ishizuka, K., Hirota-Tsuyada, Y., Wang, Q., … Sawa, A. (2014). DISC1 regulates trafficking and processing of APP and Aβ generation. Molecular Psychiatry. http://doi.org/10.1038/mp.2014.100

Shen, H., & Huang, J. Z. (2008). Sparse principal component analysis via regularized low rank matrix approximation. Journal of Multivariate Analysis, 99(6), 1015–1034. http://doi.org/10.1016/j.jmva.2007.06.007

Shen, L., Kim, S., Risacher, S. L., Nho, K., Swaminathan, S., West, J. D., … Saykin, A. J. (2010). Whole genome association study of brain-wide imaging phenotypes for identifying quantitative trait loci in MCI and AD: A study of the ADNI cohort. NeuroImage, 53(3), 1051–1063. http://doi.org/10.1016/j.neuroimage.2010.01.042

Sherva, R., Kranzler, H. R., Yu, Y., Logue, M. W., Poling, J., Arias, A. J., … Gelernter, J. (2010). Variation in Nicotinic Acetylcholine Receptor Genes is Associated with Multiple Substance Dependence Phenotypes. Neuropsychopharmacology, 35(9), 1921–1931. http://doi.org/10.1038/npp.2010.64

Shollenbarger, S. G., Price, J., Wieser, J., & Lisdahl, K. (2015). Poorer frontolimbic white matter integrity is associated with chronic cannabis use, FAAH genotype, and increased depressive and apathy symptoms in adolescents and young adults. NeuroImage: Clinical, 8, 117–125. http://doi.org/10.1016/j.nicl.2015.03.024

Sill, M., Saadati, M., & Benner, A. (2015). Applying Stability Selection to Consistently Estimate Sparse Principal Components in High-Dimensional Molecular Data. Bioinformatics, btv197. http://doi.org/10.1093/bioinformatics/btv197

Simon, G. (1977). Multivariate Generalization of Kendall's Tau with Application to Data Reduction. Journal of the American Statistical Association, 72(358), 367–376. http://doi.org/10.2307/2286801

Sinha, R. (2013). The Clinical Neurobiology of Drug Craving. Current Opinion in Neurobiology, 23(4), 649–654. http://doi.org/10.1016/j.conb.2013.05.001

Smith, C. T., Wallace, D. L., Dang, L. C., Aarts, E., Jagust, W. J., D'Esposito, M., & Boettiger, C. A. (2015). Modulation of Impulsivity and Reward Sensitivity in Intertemporal Choice by Striatal and Midbrain Dopamine Synthesis in Healthy Adults. Journal of Neurophysiology, jn.00261.2015. http://doi.org/10.1152/jn.00261.2015

Sokolowski, M., Wasserman, J., & Wasserman, D. (2015). An overview of the neurobiology of suicidal behaviors as one meta-system. *Molecular Psychiatry*, *20*(1), 56–71. https://doi.org/10.1038/mp.2014.101

Stein, J. L., Hua, X., Lee, S., Ho, A. J., Leow, A. D., Toga, A. W., … Thompson, P. (2010). Voxelwise genome-wide association study (vGWAS). NeuroImage, 53(3), 1160–1174. http://doi.org/10.1016/j.neuroimage.2010.02.032

Stephens, S. H., Franks, A., Berger, R., Palionyte, M., Fingerlin, T. E., Wagner, B., … Leonard, S. (2012). Multiple genes in the 15q13-q14 chromosomal region are associated with schizophrenia. Psychiatric Genetics, 22(1). http://doi.org/10.1097/YPG.0b013e32834c0c33

Stewart, S. H., Pihl, R. O., Conrod, P. J., & Dongier, M. (1998). Functional associations among trauma, PTSD, and substance-related disorders. Addictive Behaviors, 23(6), 797–812.

Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences, 100(16), 9440–9445. http://doi.org/10.1073/pnas.1530509100

Stranger, B. E., Stahl, E. A., & Raj, T. (2011). Progress and Promise of Genome-Wide Association Studies for Human Complex Trait. Genetics, 187(2), 367–383. http://doi.org/10.1534/genetics.110.120907

Strother, S. C., Anderson, J., Hansen, L. K., Kjems, U., Kustra, R., Sidtis, J., … Rottenberg, D. (2002). The quantitative evaluation of functional neuroimaging experiments: the NPAIRS data analysis framework. NeuroImage, 15(4), 747–771. http://doi.org/10.1006/nimg.2001.1034

Suchanek, R., Owczarek, A., Kowalczyk, M., Kucia, K., & Kowalski, J. (2011). Association Between C-281A and Val66met Functional Polymorphisms of BDNF Gene and Risk of Recurrent Major Depressive Disorder in Polish Population. Journal of Molecular Neuroscience, 43(3), 524–530. http://doi.org/10.1007/s12031-010-9478-y

Su, H., Tao, J., Zhang, J., Xie, Y., Han, B., Lu, Y., … He, J. (2015). The analysis of BDNF gene polymorphism haplotypes and impulsivity in methamphetamine abusers. Comprehensive Psychiatry, 59, 62–67. http://doi.org/10.1016/j.comppsych.2015.02.017

Sullivan, D., Pinsonneault, J. K., Papp, A. C., Zhu, H., Lemeshow, S., Mash, D. C., & Sadee, W. (2013). Dopamine transporter DAT and receptor DRD2 variants affect risk of lethal cocaine abuse: a gene–gene–environment interaction. Translational Psychiatry, 3(1), e222. http://doi.org/10.1038/tp.2012.146

Sun, L., Ji, S., Yu, S., & Ye, J. (2009, July). On the Equivalence between Canonical Correlation Analysis and Orthonormalized Partial Least Squares. In IJCAI (Vol. 9, pp. 1230-1235).

Sun, Y., Zhao, L. Y., Wang, G. B., Yue, W. H., He, Y., Shu, N., ... & Wang, H. M. (2016). ZNF804A variants confer risk for heroin addiction and affect decision making and gray matter volume in heroin abusers. Addiction biology, 21(3), 657-666.

TAG. (2010). Genome-wide meta-analyses identify multiple loci associated with smoking behavior. Nature Genetics, 42(5), 441–447. http://doi.org/10.1038/ng.571

Takane, Y., & Hwang, H. (2006). Regularized multiple correspondence analysis. Multiple Correspondence Analysis and Related Methods, 259–279.

Takane, Y., Hwang, H., & Abdi, H. (2008). Regularized multiple-set canonical correlation analysis. Psychometrika, 73(4), 753–775.

Takane, Y., & Jung, S. (2009). Regularized nonsymmetric correspondence analysis. Computational Statistics & Data Analysis, 53(8), 3159–3170. http://doi.org/10.1016/j.csda.2008.09.004

The Network and Pathway Analysis Subgroup of the Psychiatric Genomics Consortium. (2015). Psychiatric genome-wide association study analyses implicate neuronal, immune and histone pathways. Nature Neuroscience, 18(2), 199–209. http://doi.org/10.1038/nn.3922

Thompson, B. (2005). Canonical correlation analysis. Encyclopedia of statistics in behavioral science.

Thorgeirsson, T. E., Gudbjartsson, D. F., Surakka, I., Vink, J. M., Amin, N., Geller, F., … Stefansson, K. (2010). Sequence variants at CHRNB3-CHRNA6 and CYP2A6 affect smoking behavior. Nature Genetics, 42(5), 448–453. http://doi.org/10.1038/ng.573

Tian, C., Gregersen, P. K., & Seldin, M. F. (2008). Accounting for ancestry: population substructure and genome-wide association studies. Human Molecular Genetics, 17(R2), R143–R150. http://doi.org/10.1093/hmg/ddn268

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological), 58(1), 267–288.

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73(3), 273–282. http://doi.org/10.1111/j.1467-9868.2011.00771.x

Tiffany, S. T., & Wray, J. M. (2012). The clinical significance of drug craving. Annals of the New York Academy of Sciences, 1248, 1–17. http://doi.org/10.1111/j.1749-6632.2011.06298.x

Tikkanen, R., Tiihonen, J., Rautiainen, M. R., Paunio, T., Bevilacqua, L., Panarsky, R., … Virkkunen, M. (2015). Impulsive alcohol-related risk-behavior and emotional dysregulation among individuals with a serotonin 2B receptor stop codon. Translational Psychiatry, 5(11), e681. http://doi.org/10.1038/tp.2015.170

Timpano, K. R., Schmidt, N. B., Wheaton, M. G., Wendland, J. R., & Murphy, D. L. (2011). Consideration of the BDNF gene in relation to two phenotypes: Hoarding and obesity. Journal of Abnormal Psychology, 120(3), 700–707. http://doi.org/10.1037/a0024159

Tiwari, A. K., Zai, C. C., Likhodi, O., Lisker, A., Singh, D., Souza, R. P., … Müller, D. J. (2010). A Common Polymorphism in the Cannabinoid Receptor 1 (CNR1) Gene is Associated with Antipsychotic-Induced Weight Gain in Schizophrenia. Neuropsychopharmacology, 35(6), 1315–1324. http://doi.org/10.1038/npp.2009.235

Toseland, A., Daines, S. J., Clark, J. R., Kirkham, A., Strauss, J., Uhlig, C., … Mock, T. (2013). The impact of temperature on marine phytoplankton resource allocation and metabolism. Nature Climate Change, 3(11), 979–984. http://doi.org/10.1038/nclimate1989

Trendafilov, N. T. (2014). From simple structure to sparse components: a review. Computational Statistics, 29(3-4), 431–454. http://doi.org/10.1007/s00180-013-0434-5

Trendafilov, N. T., & Adachi, K. (2014). Sparse Versus Simple Structure Loadings. Psychometrika, 1–15. http://doi.org/10.1007/s11336-014-9416-y

Trucco, E. M., Villafuerte, S., Heitzeg, M. M., Burmeister, M., & Zucker, R. A. (2014). Rule breaking mediates the developmental association between GABRA2 and adolescent substance abuse. Journal of Child Psychology and Psychiatry, 55(12), 1372–1379. http://doi.org/10.1111/jcpp.12244

Tucker, G., Price, A. L., & Berger, B. (2014). Improving the Power of GWAS and Avoiding Confounding from Population Stratification with PC-Select. Genetics, 197(3), 1045–1049. http://doi.org/10.1534/genetics.114.164285

Tucker, L. R. (1944). A semi-analytical method of factorial rotation to simple structure. Psychometrika, 9(1), 43–68. http://doi.org/10.1007/BF02288713

Tucker, L. R. (1958). An inter-battery method of factor analysis. Psychometrika, 23, 111–136.

Tull, M. T., Gratz, K. L., Coffey, S. F., Weiss, N. H., & McDermott, M. J. (2013). Examining the interactive effect of posttraumatic stress disorder, distress tolerance, and gender on residential substance use disorder treatment retention. Psychology of Addictive Behaviors, 27(3), 763–773. http://doi.org/10.1037/a0029911

Tunbridge, E. M., Dunn, G., Murray, R. M., Evans, N., Lister, R., Stumpenhorst, K., … Freeman, D. (2015). Genetic moderation of the effects of cannabis: Catechol-O-methyltransferase (COMT) affects the impact of Δ9-tetrahydrocannabinol (THC) on working memory performance but not on the occurrence of psychotic experiences.

Journal of Psychopharmacology, 0269881115609073. http://doi.org/10.1177/0269881115609073

Tura, E., Turner, J. A., Fallon, J. H., Kennedy, J. L., & Potkin, S. G. (2008). Multivariate analyses suggest genetic impacts on neurocircuitry in schizophrenia. Neuroreport, 19(6), 603–607. http://doi.org/10.1097/WNR.0b013e3282fa6d8d

Turgeon, M., Oualkacha, K., Ciampi, A., Dehghan, G., Zanke, B. W., Benedet, A. L., … Initiative, A. D. N. (2016). Principal component of explained variance: an efficient and optimal data dimension reduction framework for association studies. bioRxiv, 036566. http://doi.org/10.1101/036566

Turner, S., Armstrong, L. L., Bradford, Y., Carlson, C. S., Crawford, D. C., Crenshaw, A. T., … Ritchie, M. D. (2011). Quality Control Procedures for Genome Wide Association Studies. Current Protocols in Human Genetics / Editorial Board, Jonathan L. Haines ... [et Al.], CHAPTER, Unit1.19. http://doi.org/10.1002/0471142905.hg0119s68

Tyndale, R. F. (2003). Genetics of alcohol and tobacco use in humans. *Annals of Medicine*, *35*(2), 94–121. https://doi.org/10.1080/07853890310010014

Uhl, G. R., Drgon, T., Johnson, C., Li, C.-Y., Contoreggi, C., Hess, J., … Liu, Q.-R. (2008). Molecular genetics of addiction and related heritable phenotypes: genome wide association approaches identify “connectivity constellation” and drug target genes with pleiotropic effects. *Annals of the New York Academy of Sciences*, *1141*, 318–381. https://doi.org/10.1196/annals.1441.018

van der Sluis, S., Posthuma, D., & Dolan, C. V. (2013). TATES: Efficient Multivariate Genotype-Phenotype Analysis for Genome-Wide Association Studies. PLoS Genet, 9(1), e1003235. http://doi.org/10.1371/journal.pgen.1003235

van Eekelen, J. A. M., Olsson, C. A., Ellis, J. A., Ang, W., Hutchinson, D., Zubrick, S. R., & Pennell, C. E. (2011). Identification and genetic determination of an early life risk disposition for depressive disorder: Atypical stress-related behaviour in early childhood. Australian Journal of Psychology, 63(1), 6–17. http://doi.org/10.1111/j.1742-9536.2011.00002.x

Vaske, J., Boisvert, D., Wright, J. P., & Beaver, K. M. (2013). A longitudinal analysis of the effects of a DRD4 polymorphism on marijuana use. Psychiatry Research, 210(1), 247–255. http://doi.org/10.1016/j.psychres.2013.04.022

Verbanck, M., Josse, J., & Husson, F. (2013). Regularised PCA to denoise and visualise data. Statistics and Computing, 25(2), 471–486. http://doi.org/10.1007/s11222-013-9444-y

Verdejo-García, A., Del Mar Sánchez-Fernández, M., Alonso-Maroto, L. M., Fernández-Calderón, F., Perales, J. C., Lozano, O., & Pérez-García, M. (2010). Impulsivity and executive functions in polysubstance-using rave attenders. Psychopharmacology, 210(3), 377–392. http://doi.org/10.1007/s00213-010-1833-8

Vergara, V. M., Ulloa, A., Calhoun, V. D., Boutte, D., Chen, J., & Liu, J. (2014). A Three-way Parallel ICA Approach to Analyze Links among Genetics, Brain Structure and Brain Function. NeuroImage. http://doi.org/10.1016/j.neuroimage.2014.04.060

Vilaça, S. T., Vargas, S. M., Lara-Ruiz, P., Molfetti, É., Reis, E. C., Lôbo-Hajdu, G., … Santos, F. R. (2012). Nuclear markers reveal a complex introgression pattern among marine turtle species on the Brazilian coast. Molecular Ecology, 21(17), 4300–4312. http://doi.org/10.1111/j.1365-294X.2012.05685.x

Villalba, K., Attonito, J., Mendy, A., Devieux, J. G., Gasana, J., & Dorak, T. M. (2015). A meta-analysis of the associations between the SLC6A4 promoter polymorphism (5HTTLPR) and the risk for alcohol dependence. Psychiatric Genetics, 25(2), 47–58.

Visscher, P. M., Brown, M. A., McCarthy, M. I., & Yang, J. (2012). Five Years of GWAS Discovery. The American Journal of Human Genetics, 90(1), 7–24. http://doi.org/10.1016/j.ajhg.2011.11.029

Vogel, S., Fernández, G., Joëls, M., & Schwabe, L. (in press, 2016). Cognitive Adaptation under Stress: A Case for the Mineralocorticoid Receptor. Trends in Cognitive Sciences. http://doi.org/10.1016/j.tics.2015.12.003

Voineskos, A. N., Felsky, D., Wheeler, A. L., Rotenberg, D. J., Levesque, M., Patel, S., … Malhotra, A. K. (2015). Limited Evidence for Association of Genome-Wide Schizophrenia Risk Variants on Cortical Neuroimaging Phenotypes. Schizophrenia Bulletin, sbv180. http://doi.org/10.1093/schbul/sbv180

Voisey, J., Swagell, C. D., Hughes, I. P., van Daal, A., Noble, E. P., Lawford, B. R., … Morris, C. P. (2012). A DRD2 and ANKK1 haplotype is associated with nicotine dependence. Psychiatry Research, 196(2–3), 285–289. http://doi.org/10.1016/j.psychres.2011.09.024

Volkow, N. D., Fowler, J. S., Wang, G.-J., Telang, F., Logan, J., Jayne, M., … Swanson, J. M. (2010). Cognitive control of drug craving inhibits brain reward regions in cocaine abusers. NeuroImage, 49(3), 2536–2543. http://doi.org/10.1016/j.neuroimage.2009.10.088

Volkow, N. D., Koob, G. F., & McLellan, A. T. (2016). Neurobiologic Advances from the Brain Disease Model of Addiction. New England Journal of Medicine, 374(4), 363–371. http://doi.org/10.1056/NEJMra1511480

Volkow, N. D., & Muenke, M. (2012). The genetics of addiction. Human Genetics, 131(6), 773–777. http://doi.org/10.1007/s00439-012-1173-3

Volkow, N. D., Wang, G.-J., Fowler, J. S., & Tomasi, D. (2012). Addiction Circuitry in the Human Brain. Annual Review of Pharmacology and Toxicology, 52, 321–336. http://doi.org/10.1146/annurev-pharmtox-010611-134625

Volkow, N. D., Wang, G.-J., Fowler, J. S., Tomasi, D., & Telang, F. (2011). Addiction: Beyond dopamine reward circuitry. Proceedings of the National Academy of Sciences, 108(37), 15037–15042. http://doi.org/10.1073/pnas.1010654108

Volkow, N. D., Wang, G.-J., Fowler, J. S., Tomasi, D., Telang, F., & Baler, R. (2010). Addiction: Decreased reward sensitivity and increased expectation sensitivity conspire to overwhelm the brain's control circuit. BioEssays, 32(9), 748–755. http://doi.org/10.1002/bies.201000042

Vormfelde, S. V., & Brockmöller, J. (2007). On the value of haplotype-based genotype–phenotype analysis and on data transformation in pharmacogenetics and -genomics. Nature Reviews Genetics, 8(12). http://doi.org/10.1038/nrg1916-c1

Vounou, M., Janousova, E., Wolz, R., Stein, J. L., Thompson, P. M., Rueckert, D., & Montana, G. (2012). Sparse reduced-rank regression detects genetic associations with voxel-wise longitudinal phenotypes in Alzheimer's disease. NeuroImage, 60(1), 700–716. http://doi.org/10.1016/j.neuroimage.2011.12.029

Vounou, M., Nichols, T. E., & Montana, G. (2010). Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach. NeuroImage, 53(3), 1147–1159. http://doi.org/10.1016/j.neuroimage.2010.07.002

Vrieze, S. I., Hicks, B. M., Iacono, W. G., & McGue, M. (2012). Decline in Genetic Influence on the Co-Occurrence of Alcohol, Marijuana, and Nicotine Dependence Symptoms From Age 14 to 29. American Journal of Psychiatry, 169(10), 1073–1081. http://doi.org/10.1176/appi.ajp.2012.11081268

Wang, D., Sun, Y., Stang, P., Berlin, J. A., Wilcox, M. A., & Li, Q. (2009). Comparison of methods for correcting population stratification in a genome-wide association study of rheumatoid arthritis: principal-component analysis versus multidimensional scaling. BMC Proceedings, 3(Suppl 7), S109.

Wang, K.-S., Zuo, L., Pan, Y., Xie, C., & Luo, X. (2015). Genetic variants in the CPNE5 gene are associated with alcohol dependence and obesity in Caucasian populations. Journal of Psychiatric Research, 71, 1–7. http://doi.org/10.1016/j.jpsychires.2015.09.008

Wang, L., Liu, X., Luo, X., Zeng, M., Zuo, L., & Wang, K.-S. (2013). Genetic Variants in the Fat Mass- and Obesity-Associated (FTO) Gene are Associated with Alcohol Dependence. Journal of Molecular Neuroscience, 51(2), 416–424. http://doi.org/10.1007/s12031-013-0044-2

Wang, T., Ho, G., Ye, K., Strickler, H., & Elston, R. C. (2009). A partial least-square approach for modeling gene-gene and gene-environment interactions when multiple markers are genotyped. Genetic Epidemiology, 33(1), 6–15. http://doi.org/10.1002/gepi.20351

Ware, J. J., & Munafò, M. R. (2014). Determining the Causes and Consequences of Nicotine Dependence: Emerging Genetic Research Methods. Current Psychiatry Reports, 16(10), 1–6. http://doi.org/10.1007/s11920-014-0477-5

Watanabe, Y., Shibuya, M., & Someya, T. (2015). DRD2 Ser311Cys polymorphism and risk of schizophrenia. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics, 168(3), 224-228.

Weinberger, D. R., Glick, I. D., & Klein, D. F. (2015). Whither Research Domain Criteria (RDoC)?: The Good, the Bad, and the Ugly. JAMA Psychiatry, 72(12), 1161–1162.

Weiner, M. P., & Hudson, T. J. (2002). Introduction to SNPs: discovery of markers for disease. BioTechniques, Suppl, 4–7, 10, 12–13.

Wellenreuther, M., & Hansson, B. (2016). Detecting Polygenic Evolution: Problems, Pitfalls, and Promises. Trends in Genetics, 32(3), 155–164. http://doi.org/10.1016/j.tig.2015.12.004

Wetherill, L., Agrawal, A., Kapoor, M., Bertelsen, S., Bierut, L. J., Brooks, A., … Foroud, T. (2015). Association of substance dependence phenotypes in the COGA sample. Addiction Biology, 20(3), 617–627. http://doi.org/10.1111/adb.12153

Whelan, R., Conrod, P. J., Poline, J.-B., Lourdusamy, A., Banaschewski, T., Barker, G. J., … Garavan, H. (2012). Adolescent impulsivity phenotypes characterized by distinct brain networks. Nat Neurosci, 15(6), 920–925. http://doi.org/10.1038/nn.3092

Williams, H. J., Glaser, B., Williams, N. M., Norton, N., Zammit, S., Macgregor, S., … O'Donovan, M. C. (2005). No Association Between Schizophrenia and Polymorphisms in COMT in Two Large Samples. American Journal of Psychiatry, 162(9), 1736–1738. http://doi.org/10.1176/appi.ajp.162.9.1736

Wilson, J., Markie, D., & Fitches, A. (2012). Cholecystokinin system genes: Associations with panic and other psychiatric disorders. *Journal of Affective Disorders*, *136*(3), 902–908. https://doi.org/10.1016/j.jad.2011.09.011

Winham, S. J., Cuellar-Barboza, A. B., McElroy, S. L., Oliveros, A., Crow, S., Colby, C. L., … Biernacka, J. M. (2014). Bipolar disorder with comorbid binge eating history: A genome-wide association study implicates APOB. Journal of Affective Disorders, 165, 151–158. http://doi.org/10.1016/j.jad.2014.04.026

Witt, E. A., Hopwood, C. J., Morey, L. C., Markowitz, J. C., McGlashan, T. H., Grilo, C. M., … Brent, M. (2010). Psychometric characteristics and clinical correlates of NEO-PI-R fearless dominance and impulsive antisociality in the Collaborative Longitudinal Personality Disorders Study. Psychological Assessment, 22(3), 559–568. http://doi.org/10.1037/a0019617

Witten, D. M., Tibshirani, R., & Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics, kxp008. http://doi.org/10.1093/biostatistics/kxp008

Wold, H. (1975). Soft modelling by latent variables: the non-linear iterative partial least squares (NIPALS) approach. Perspectives in Probability and Statistics, In Honor of MS Bartlett, 117–144.

Wold, S., Ruhe, A., Wold, H., & Dunn, III, W. (1984). The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. SIAM Journal on Scientific and Statistical Computing, 5(3), 735–743.

Xie, P., Kranzler, H. R., Krystal, J. H., Farrer, L. A., Zhao, H., & Gelernter, J. (2014). Deep resequencing of 17 glutamate system genes identifies rare variants in DISC1 and GRIN2B affecting risk of opioid dependence. *Addiction Biology*, *19*(5), 955–964. https://doi.org/10.1111/adb.12072

Xu, X., Clark, U. S., David, S. P., Mulligan, R. C., Knopik, V. S., McGeary, J., … Sweet, L. H. (2014). The Effects of Nicotine Deprivation and Replacement on BOLD-fMRI Response to Smoking Cues as a Function of DRD4 VNTR Genotype. Nicotine & Tobacco Research, ntu010. http://doi.org/10.1093/ntr/ntu010

Xu, Y., Hu, W., Yang, Z., & Xu, C. (2016). A multivariate partial least squares approach to joint association analysis for multiple correlated traits. The Crop Journal. http://doi.org/10.1016/j.cj.2015.11.001

Yang, Z., Seneviratne, C., Wang, S., Ma, J. Z., Payne, T. J., Wang, J., & Li, M. D. (2013). Serotonin transporter and receptor genes significantly impact nicotine dependence through genetic interactions in both European American and African American smokers. Drug and Alcohol Dependence, 129(3), 217–225. http://doi.org/10.1016/j.drugalcdep.2012.12.007

Yan, J., Du, L., Kim, S., Risacher, S. L., Huang, H., Moore, J. H., … for the ADNI (2014). Transcriptome-guided amyloid imaging genetic analysis via a novel structured sparse learning algorithm. Bioinformatics, 30(17), i564–i571. http://doi.org/10.1093/bioinformatics/btu465

Yan, J., Du, L., Kim, S., Risacher, S. L., Huang, H., Moore, J. H., … Shen, L. (2014). Transcriptome-guided amyloid imaging genetic analysis via a novel structured sparse learning algorithm. Bioinformatics, 30(17), i564–i571. http://doi.org/10.1093/bioinformatics/btu465

Yan, J., Du, L., Risacher, S. L., Shen, L., & Saykin, A. J. (2016). Identification of diagnosis related imaging genomics associations through outcome-guided sparse CCA: An Alzheimer's disease study. Poster, 12th International Imaging Genetics Conference.

Yan, J., Huang, H., Risacher, S. L., Kim, S., Inlow, M., Moore, J. H., ... & Shen, L. (2013, September). Network-guided sparse learning for predicting cognitive outcomes from MRI measures. In International Workshop on Multimodal Brain Image Analysis (pp. 202-210). Springer International Publishing.

Yan, J., Zhang, H., Du, L., Wernert, E., Saykin, A. J., & Shen, L. (2014). Accelerating Sparse Canonical Correlation Analysis for Large Brain Imaging Genetics Data. In Proceedings of the 2014 Annual Conference on Extreme Science and Engineering Discovery Environment (pp. 4:1–4:7). New York, NY, USA: ACM. http://doi.org/10.1145/2616498.2616515

Yan, Q., Weeks, D. E., Celedón, J. C., Tiwari, H. K., Li, B., Wang, X., … Liu, N. (2015). Associating Multivariate Quantitative Phenotypes with Genetic Variants in Family Samples with a Novel Kernel Machine Regression Method. Genetics, 201(4), 1329–1339. http://doi.org/10.1534/genetics.115.178590

Yanai, H., Takeuchi, K., & Takane, Y. (2011). Projection matrices, generalized inverse matrices, and singular value decomposition. New-York: Springer-Verlag.

Yarosh, H. L., Meda, S. A., Wit, H. de, Hart, A. B., & Pearlson, G. D. (2015). Multivariate analysis of subjective responses to d-amphetamine in healthy volunteers finds novel genetic pathway associations. Psychopharmacology, 1–14. http://doi.org/10.1007/s00213-015-3914-1

Yee, C. M., Javitt, D. C., & Miller, G. A. (2015). Replacing DSM Categorical Analyses With Dimensional Analyses in Psychiatry Research: The Research Domain Criteria Initiative. JAMA Psychiatry, 72(12), 1159–1160.

Zapala, M. A., & Schork, N. J. (2006). Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables. Proceedings of the National Academy of Sciences, 103(51), 19430 –19435. http://doi.org/10.1073/pnas.0609333103

Zhang, L., Kendler, K. S., & Chen, X. (2006). The μ-opioid receptor gene and smoking initiation and nicotine dependence. Behavioral and Brain Functions, 2(1), 1–6. http://doi.org/10.1186/1744-9081-2-28

Zeiger, J. S., Haberstick, B. C., Schlaepfer, I., Collins, A. C., Corley, R. P., Crowley, T. J., … Ehringer, M. A. (2008). The Neuronal Nicotinic Receptor Subunit Genes (CHRNA6 and CHRNB3) Are Associated with Subjective Responses to Tobacco. Human Molecular Genetics, 17(5), 724–734. http://doi.org/10.1093/hmg/ddm344

Zhang, J.-P., Lencz, T., & Malhotra, A. K. (2010). D2 Receptor Genetic Variation and Clinical Response to Antipsychotic Drug Treatment: A Meta-Analysis. American Journal of Psychiatry, 167(7), 763–772. http://doi.org/10.1176/appi.ajp.2009.09040598

Zhang, X.-B., Zhao, Z.-H., Chen, H.-Y., Wang, J.-C., Qian, J., Yang, Y.-J., … Lu, D.-R. (2011). Human chromosome 8p11 (CHRNB3-CHRNA6) region gene polymorphisms and susceptibility to lung cancer in Chinese Han population. Yi Chuan = Hereditas / Zhongguo Yi Chuan Xue Hui Bian Ji, 33(8), 886–894.

Zhou, X., & Stephens, M. (2014). Efficient Algorithms for Multivariate Linear Mixed Models in Genome-wide Association Studies. Nature Methods, 11(4), 407–409. http://doi.org/10.1038/nmeth.2848

Zhou, Y.-H., & Wright, F. A. (2015). The projack: a resampling approach to correct for ranking bias in high-throughput studies. Biostatistics, kxv022. http://doi.org/10.1093/biostatistics/kxv022

Zhu, X., Dutta, N., Helton, S. G., Schwandt, M., Yan, J., Hodgkinson, C. A., ... & Phillips, M. (2015). Resting-state functional connectivity and presynaptic monoamine signaling in Alcohol Dependence. Human Brain Mapping, 36(12), 4808-4818.

Zhu, X., Need, A. C., Petrovski, S., & Goldstein, D. B. (2014). One gene, many neuropsychiatric disorders: lessons from Mendelian diseases. Nature Neuroscience, 17(6), 773–781. http://doi.org/10.1038/nn.3713

Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse Principal Component Analysis. Journal of Computational and Graphical Statistics, 15(2), 265–286. http://doi.org/10.1198/106186006X113430

Zuvich, R. L., Armstrong, L. L., Bielinski, S. J., Bradford, Y., Carlson, C. S., Crawford, D. C., … Ritchie, M. D. (2011). Pitfalls of Merging GWAS Data: Lessons Learned in the eMERGE Network and Quality Control Procedures to Maintain High Data Quality. Genetic Epidemiology, 35(8), 887–898. http://doi.org/10.1002/gepi.20639

BIOGRAPHICAL SKETCH

Derek Beaton was born in Massachusetts where he grew up. He received his BS and MS in Computer and Information Sciences from the University of Massachusetts Dartmouth (UMD), and during that time, worked for a mathematics education research center. During his time at UMD he received a National Science Foundation international fellowship (0812995). He has also received his MS and PhD in Cognition and Neuroscience from The University of Texas at Dallas. During his PhD he was awarded a National Research Service Award from the NIH (F31DA035039). He is now a post-doctoral fellow at the Rotman Research Institute, and a post-doctoral scholar of the Ontario Neurodegenerative Disease Research Initiative. His interests and research are at the intersection of neurological and psychiatric disorders, genetics, and development of statistical techniques. Some of his representative work includes Beaton, D., Dunlop, J., ADNI, & Abdi, H. (2016), Beaton, D., Abdi, H., & Filbey, F. M. (2014), and Beaton, D., Chin Fatt C.R., & Abdi, H. (2014). Over the course of his studies he has taught or helped teach a variety of courses that range from introduction to programming through multivariate statistics.

CURRICULUM VITAE

UPDATED: 3.25.2017

Google Scholar Page

Derek Beaton
Rotman Research Institute
Baycrest Health Sciences
3560 Bathurst Street
Toronto, Ontario
Canada M6A 2E1
http://www.derekbeaton.com

Education & Training

(**Post-doc**) Rotman Research Institute, Baycrest Health Sciences. Aug. 2016-present.
Advisor: Stephen Strother, Ph.D.
(**Ph.D.**) Cognition and Neuroscience, The University of Texas at Dallas. February 2017.
Advisor: Hervé Abdi, Ph.D.
(**M.S.**) Applied Cognition and Neuroscience, The University of Texas at Dallas. May 2011.
Advisor: Hervé Abdi, Ph.D.
(**M.S.**) Computer and Information Sciences, University of Massachusetts Dartmouth. May 2008.
Advisor: Iren Valova, Ph.D.
(**B.S.**) Computer and Information Sciences, University of Massachusetts Dartmouth. December 2005.
Advisor: Iren Valova, Ph.D.

Expertise

In no particular order: Statistical analyses, factor analyses, genetics/genomics, substance use disorders, neuroimaging, bioinformatics, image processing/analysis, artificial intelligence, data mining, machine learning.

Publications

Journal Articles

2016

Beaton, D., Dunlop, J. D., ADNI, Abdi, H. (2016). Partial Least Squares Correspondence Analysis: A Framework to Simultaneously Analyze Behavioral and Genetic Data. *Psychological Methods*.

Allen, G.I., Amoroso, N., Anghel, C., Balagurusamy, V., Bare, C.J., **Beaton, D.**, …, ADNI. (2016). Crowdsourced estimation of cognitive decline and resilience in Alzheimer's disease. *Alzheimer's & Dementia.*

2015

Schwarz, A.L., van Kleeck, A., **Beaton, D.**, Horne, E., MacKenzie, H., Abdi, H. (accepted, 2015). A Read-Aloud Storybook Selection System for Pre-Readers at the Preschool Language Level: A Pilot Study. *Journal of Speech, Language, and Hearing Research*. **58**

2014

Cioli C., Abdi H., **Beaton D.**, Burnod Y., Mesmoudi S. (2014) Differences in Human Cortical Gene Expression Match the Temporal Properties of Large-Scale Functional Networks. *PLoS ONE* 9(12): e115913. doi:10.1371/journal.pone.0115913

Beaton, D., Abdi, H., & Filbey, F. M. (2014). Unique aspects of impulsive traits in substance use and overeating: specific contributions of common assessments of impulsivity. *The American Journal of Drug and Alcohol Abuse*, *40*(6), 463–475. doi:10.3109/00952990.2014.937490

Beaton, D., Chin Fatt C.R., & Abdi, H. (2014). An ExPosition of multivariate analysis with the Singular Value Decomposition in R. *Computational Statistics & Data Analysis*. *72*(0), 176 – 189. doi:http://dx.doi.org/10.1016/j.csda.2013.11.006

2012

Abdi, H., Williams, L.J., **Beaton, D.,** Posamentier, M., Harris, T.S., Krishnan, A., & Devous, M.D. (2012). Analysis of regional cerebral blood flow data to discriminate among Alzheimer's disease, fronto-temporal dementia, and elderly controls: A multi-block barycentric discriminant analysis (MUBADA) methodology. *Journal of Alzheimer's Disease*. 31 (S3). S189-S201

Pinkham, A. E., Sasson, N. J., **Beaton, D.,** Abdi, H., Kohler, C. G., Penn, D. L. (2012). Qualitatively Distinct Factors Contribute to Elevated Rates of Paranoia in Autism and Schizophrenia. *Journal of Abnormal Psychology*. 121(3). 767-777.

2011

Beaton, D., Valova, I., & MacLean, D. (2011). TurSOM: A paradigm bridging Turing's unorganized machines and self-organizing maps demonstrating dual self-organization. *Neurocomputing*, 74(17), 3125 – 3141.

2010

Beaton, D., Valova, I., & MacLean, D. (2010). CQoCO: A measure for comparative quality of coverage and organization for self-organizing maps. *Neurocomputing*. 73(10-12), 2147-2159.

Valova, I., **Beaton, D.**, Buer, A., & MacLean, D. (2010). Fractal initialization for high-quality mapping with self-organizing maps. *Neural Computing and Applications*. 19(7), 953-966.

Valova, I., **Beaton, D.**, MacLean, D., & Hammond, J. (2010). NIPSOM: Parallel Architecture and Implementation of a Growing SOM. *The Computer Journal*, 53(6), 753-771.

Books

To appear

Abdi, H., & **Beaton, D.** (to appear). Principal Component and Correspondence Analyses Using R. New York: Springer Verlag.

Chapters

2017

Abdi, H., Guillemot, V., Eslami, A., **Beaton, D.** (to appear, 2017). Canonical Correlation Analysis. In Alhajj, H., & Rokne, J. (Eds.), *Encyclopedia of Social Network Analysis and Mining*. New York: Springer Verlag.

2016

Beaton, D., Kriegsman, M., ADNI, Dunlop, J., Filbey, F. M., & Abdi, H. (to appear, 2016). Partial Least Squares for mixed-data types: An application for imaging genetics. In Abdi, H., Esposito Vinzi, V., Saporta, G., Russolillo, G., & Trinchera, L. (Eds.), *The Multiple Facets of Partial Least Squares Methods.* New York: Springer Verlag.

2013

Beaton, D., Filbey, F., & Abdi H. (2013). Integrating partial least squares correlation and correspondence analysis for nominal data. In Abdi, H., Chin, W., Esposito Vinzi, V., Russolillo, G., & Trinchera, L. (Eds.), *New Perspectives in Partial Least Squares and Related Methods.* New York: Springer Verlag. pp.81-94.

Kovacevic, N., Abdi, H., **Beaton, D.**, Alzheimer's Disease Neuroimaging Initiative, & McIntosh, A.R. (2013). Revisiting PLS resampling: Comparing significance vs. reliability across range of simulations. In Abdi, H., Chin, W., Esposito Vinzi, V., Russolillo, G., & Trinchera, L. (Eds.), *New Perspectives in Partial Least Squares and Related Methods.* New York: Springer Verlag.

2010

Valova, I., **Beaton, D.**, & MacLean, D. (2010). Self-organizing Maps for Machine Learning. In Machine Learning. IN-TECH. [link].

Press

2014

Beaton, D. (2014, December). DFW's Favorite Local Beer, As Proven By An Actual Scientist. *Dallas Observer* (online article). http://blogs.dallasobserver.com/cityofate/2014/12/dfws_favorite_local_beer_proven_by_a_real_phd_scientist.php

Refereed Conference Proceedings/Short papers

2009

Beaton, D., Valova, I., & MacLean, D. (2009). The Use of TurSOM for Color Image Segmentation. Presented at the IEEE International Conference on Systems, Man and Cybernetics, San Antonio, TX, USA.

Beaton, D., Valova, I., & MacLean, D. (2009). TurSOM: A Turing inspired Self-Organizing Map. In Neural Networks, IEEE - INNS - ENNS International Joint Conference on (Vol. 0, pp. 288-295). Los Alamitos, CA, USA: IEEE Computer Society.

Beaton, D., Valova, I., & MacLean, D. (2009). Growing mechanisms and cluster identification with TurSOM. In Neural Networks, IEEE - INNS - ENNS International Joint Conference on (Vol. 0, pp. 280-287). Los Alamitos, CA, USA: IEEE Computer Society.

Valova, I., MacLean, D., & **Beaton, D.** (2009). ParaSOM: An Efficient Self-Organizing Map for Parallel Multidimensional Input Processing and Clustering. In ASME Press Series on Intelligent Engineering Systems Through Artificial Neural Networks. Presented at the ANNIE 2009, St. Louis, MO, USA.

Hegedus, S., Dalton, S., Brookstein, A., Tapper, J., & **Beaton, D.** (2009). Relationships between motivation and student performance in a technology-rich classroom environment. In Proceedings of the Thirty First Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education (Vol. 5, pp. 185-192). Presented at the PME-NA 2009, Atlanta, GA: Georgia State University.

2008

Valova, I., **Beaton, D.**, & MacLean, D. (2008). Role of Initialization in SOM networks - study of self-similar curve topologies. In ASME Press Series on Intelligent Engineering Systems Through Artificial Neural Networks. Presented at the ANNIE 2008, St. Louis, MO, USA.

Valova, I., MacLean, D., & **Beaton, D.** (2008). Identification of Patterns via Region-Growing Parallel SOM Neural Network. In Machine Learning and Applications, Fourth International Conference on (Vol. 0, pp. 853-858). Los Alamitos, CA, USA: IEEE Computer Society.

2007

Beaton, D., & Valova, I. (2007). RADDACL: A Recursive Algorithm for Clustering and Density Discovery on Non-linearly Separable Data. In Neural Networks, 2007 (pp. 1633-1638). Presented at the International Joint Conference on Neural Networks, 2007. IJCNN 2007.

2006

Beaton, D., & Hegedus, S. (2006). Constructing an Architecture for an Interactive Education Database - Issues of Design and Implementation. Presented at the IADIS Web Applications and Research.

Beaton, D., Valova, I., MacLean, D., & Hammond, J. (2006). Convergence and Optimization Study of a Growing Parallel SOM Through a Genetic Algorithm. In 25th Digital Avionics Systems Conference, 2006 IEEE/AIAA (pp. 1-9). Presented at the 25th Digital Avionics Systems Conference, 2006 IEEE/AIAA.

2005

Beaton, D., & Valova, I. (2005). Alzheimer's detection using neural network techniques and enhanced EEG measurements. In Circuits, Signals, and Systems. Presented at the IASTED International Conference on Circuits, Signals and Systems, Marina Del Rey, CA, USA.

Technical Reports

2009

Hegedus, S., Dalton, S., Brookstein, A., **Beaton, D.**, Moniz, R., Fishman, B., Roschelle, J., Penuel, W. (2009). Scaling Up SimCalc Project: Diffusion of a Research-Based Initiative in Terms of Sustainability and Spread (Technical Report No. 02). N. Dartmouth, MA, USA: James J. Kaput Center for Research and Innovation in STEM Education.

Fishman, B., Penuel, W., Hegedus, S., Moniz, R., Dalton, S., Brookstein, A., **Beaton, D.**, et al. (2009). What Happens When The Research Ends? Factors Related to Sustainability of Research-based Initiative (Technical Report No. 04). Menlo Park, CA, USA: SRI International.

Theses & Dissertations

2017

Dissertation

Beaton, D. (2017). Implementing Appropriate Multivariate Methods for Higher Quality Results from Genome-wide Association Studies in Substance Abuse Populations. Dissertation in Cognition and Neuroscience, Behavioral and Brain Sciences, The University of Texas at Dallas. Committee: **H. Abdi**, **F.M. Filbey**, K. Rodrigue, R. Golden.

2012

Qualifying thesis

Beaton, D. (2012). Partial Least Squares-Correspondence Analysis Reveals the Genetic Correlates of Impulse and Addiction. Qualifying Thesis in Cognition and Neuroscience, Behavioral and Brain Sciences, The University of Texas at Dallas. Committee: **H. Abdi**, **F.M. Filbey**, M.D. Rugg.

2008

Master's thesis

Beaton, D. (2008). Bridging Turing Unorganized Machines and Self-organizing Maps for Cognitive Replication. Master's Thesis in Computer Science, University of Massachusetts at Dartmouth. Committee: **I. Valova**, H. Xu, X. Zhang

2007

Master's project

Beaton, D. (2007). A neural network based-system for classification of EEG data correlating to neural or cognitive impairment. Master's Project in Computer Science, University of Massachusetts at Dartmouth. Committee: **I. Valova**, L. Shen, X. Zhang

Refereed Abstracts

2016

Beaton, D., Dunlop, J., Abdi, H., Filbey, F.M. (2016). Endocannabinoid, GABA-ergic, and dopaminergic contributions to neural response of different reward types. Presented at International Imaging Genetics Conference, Irvine, CA. -- **Awarded 3rd Prize.**

Beaton, D., Palombo, D.J., Bacopulos, A., Todd, R.M., Mueller, D.J., Anderson, A.K., Abdi, H., Levine, B. (2016). Genetic associations of objective and subjective measures of autobiographical memory. Presented at Cognitive Neuroscience Society, New York, New York.

Alhazmi, F., **Beaton, D.,** Abdi, H. (2016). Identifying latent semantic groups of studies and their corresponding brain regions from the NeuroSynth database. Presented at Cognitive Neuroscience Society, New York, New York.

Dutcher, A., Kmiecik, M.J., **Beaton, D.** (2016). Analyzing multi-block data: A tutorial of Multiple Factor Analysis in R. Presented at Southwest Psychological Association, Dallas, TX.

Beaton, D., Abdi, H. (2016). Reproducible components through split-half resampling: another look at stopping rules. Presented at Southwest Psychological Association, Dallas, TX.

2015

Abdi, H., **Beaton, D.,** Saporta, G. (2015). Generalizing partial least squares and correspondence analysis methods to predict categorical and heterogeneous data. Presented at Correspondence Analysis and Related Methods, Naples, Italy.

Beaton, D., Dunlop, J., Abdi, H., Filbey, F.M. (2015). Specific effects of FAAH and CNR on neural response to different reward types. Presented at Organization for Human Brain Mapping. Honolulu, HI.

Cioli, C., Abdi, H., **Beaton**, **D.,** Burnod, Y., Mesmoudi, S. (2015). Human cortical gene expression matches the properties of functional networks:a hierarchical approach. Presented at Organization for Human Brain Mapping. Honolulu, HI.

Williams, L.J., Kim, H., Fitzpatrick, K., **Beaton, D.**, Abdi, H., Bjornson, B. (2015). Ascertaining atypical individual patterns from group mental rotation networks in epileptic children. Presented at Organization for Human Brain Mapping. Honolulu, HI.

Faso. D., **Beaton., D.** (2015). Complementary and alternative approaches to *p*-values: Intervals, Effect Sizes, and Practical Advice. Presented at Southwestern Psychological Association. Wichita, KS.

Beaton, D., Faso, D., Sasson, N., Abdi, H. (2015). An Introduction to Permutation and Bootstrap Resampling. Presented at Southwestern Psychological Association. Wichita, KS.

2014

Krishnan, A., et al., (2014). Meta-analysis of neuroimaging data in a multivariate framework: A barycentric hellinger discriminant analysis (BAHEDA) approach. Presented at Society for Neuroscience 2014. Washington, D.C.

Krishnan, A., Atzil, S., Satpute, A., **Beaton, D.,** Ruzic, L., Abdi, H., Wager., T., Barret, L. (2014). MULTIVARIATE META-ANALYSIS: SEARCH FOR CONSISTENT MIND-BRAIN CORRESPONDENCE ACROSS THE NEUROIMAGING LITERATURE. Presented at Society for Affective Science 2014.

Cioli, C., Mesmoudi, S., **Beaton, D.**, Rudrauf, D., Abdi, H., Burnod, Y. (2014). Integration of functional cerebral networks and genetic expression: the dual intertwined rings architecture of the cerebral cortex. Presented at Computational Neuroscience 2014.

Beaton, D., Dunlop, J., Filbey, F., M., Abdi, H. (2014). ConcatPLS: PLSC for one continuous data set and one categorical data set**.** Presented at Partial Least Squares and Related Methods 2014. Paris, France.

Beaton, D., Bernard, A., Abdi, H., Saporta., G. (2014). An integration of partial least squares and sparsified multiple correspondence analysis for genetic data. Presented at Partial Least Squares and Related Methods 2014. Paris, France.

Beaton, D., Dunlop, J., Abdi, H. (2014). An ExPosition of Bootstrap and Permutation for Principal Components Analyses. Presented at the SWPA 2014. San Antonio, TX.

Faso, D., **Beaton, D.,** Sasson, N., M., Abdi, H. (2014). An introduction to ANOVA in R. Presented at the SWPA 2014. San Antonio, TX.

Beaton, D., Dunlop, J., Filbey, F., M., Abdi, H. (2014). Partial Least Squares for mixed-data analysis in Imaging Genetics. Presented at the Tenth International Imaging Genetics Conference, Irvine, CA.

2013

Levine, B., Bacopulos, A., Anderson, N.D., Black, S.E., Davidson, P.S.R., Fitneva, S.A., McAndrews, M.P., Spaniol, J., Jeyakumar, N., Abdi, H., **Beaton, D.,** Owen, A.M., & Hampshire, A. (2013). Validation of a Novel Computerized Test Battery for Automated Testing. Poster presented at Canadian Stroke Congress, October 17-20, 2013, Montreal, Quebec.

Beaton, D., Filbey, F., M., Abdi, H. (2013). Partial Least Squares Correspondence Analysis Reveals Genetic Correlates of Impulsivity and Addiction. Presented at Southwestern Psychological Association Annual Meeting, Ft. Worth, TX. **Awarded first prize in Society for Applied Multivariate Research**.

Faso, D., **Beaton, D.,** Abdi, H., Sasson, N. J., Pinkham, A. E. (2013). Distinct Visual Patterns in Clinical Populations Identified by Partial Least Squares. Presented at Southwestern Psychological Association Annual Meeting, Ft. Worth, TX. **Awarded third prize in Society for Applied Multivariate Research**.

Rieck, J. R., **Beaton, D**., McDonough, I., Abdi, H, Park, D. C. (2013) Dissociating age-related patterns of neural activity along the ventral visual stream during face and object processing. Presented at Twentieth Cognitive Neuroscience meeting, San Francisco, CA.

Wong, J., de Chastelaine, M., **Beaton, D.,** Abdi, H., Rugg, M. D. (2013). Dissociation between item-item and item-context memory associations. Presented at Twentieth Cognitive Neuroscience meeting, San Francisco, CA.

Beaton, D., Dunlop, J., Filbey, F. M., Abdi, H. (2013). Connecting Connectivity: Revealing the genetic influences of brain networks in a substance abuse population. Presented at the Ninth International Imaging Genetics Conference, Irvine, CA.

2012

Beaton, D., Filbey, F. M., Abdi, H. (2012). Integrating genetic, neuroimaging and behavioral data with correspondence analysis: An illustration in addictive populations. Presented at the Eighth International Imaging Genetics Conference, Irvine, CA. [PDF]

Rieck, J. R., **Beaton, D.**, Krishnan, A., Abdi, H., & Park, D. C. (2012). Older adults show less neural differentiation for processing human, primate, and cat faces in ventral temporal cortex. Poster presented at the Cognitive Aging Conference, Atlanta, GA.

Chin Fatt, C. R., **Beaton, D.,** Abdi, H. (2012). DISTATIS: Three-way metric multidimensional scaling and its extensions. Presented at the Classification Society Annual Meeting 2012, Pittsburgh, PA.

Schwarz, A.L., van Kleeck, A., **Beaton, D.,** Horne, E., Ahn, L., MacKenzie, H. (2012). Capturing Decision-Making: SLPs' Criteria for Selecting Preschool Read-Aloud Storybooks. Presented at American Speech-Language-Hearing Association Annual Meeting 2012, Atlanta, GA.

2011

Rieck, J., **Beaton, D.,** Krishnan, A., Abdi, H., Park, D. (2011). Identifying Patterns of Neural Activity During Visual Category Processing in Young and Old Adults with Multi-Block Barycentric Discriminant Analysis. Presented at Society for Neuroscience 2011, Washington, D.C.

Beaton, D., Abdi, H. (2011). Partial Least Squares-Correspondence Analysis (PLS-CA): A New Method to Analyze Common Patterns in Measures of Cognition and Genetics. Presented at NeuroInformatics 2011, Boston, MA. [PDF]

Beaton, D., Abdi, H. (2011). Partial Least Squares-Correspondence Analysis (PLS-CA): A new approach to link measures of cognition and genetics. Presented at MathPsych 2011, Boston, MA. [PDF].

Beaton, D., Abdi, H. (2011). Integrating Partial-least Squares and Correspondence Analysis for Genetics-based Cognition Research. Presented at the Seventh International Imaging Genetics Conference, Irvine, CA. [PDF]

Sasson, N. J., Dichter, G. S., **Beaton, D.**, Bodfish, J. W. (2011). Adults with and without Autism Differ In Their Emotional Responses to Non-Social Images Related to Circumscribed Interests. Presented at International Meeting for Autism Research 2011, San Diego, CA.

2010

Beaton, D., Abdi, H. (2010). Predicting behavior from genetics with correspondence analysis. Presented at the 43rd Annual Conference of the Society for Mathematical Psychology, Portland, OR. [PDF]

2009

Beaton, D., Valova, I., & MacLean, D. (2009). Color Objects Distinction with TurSOM. Presented at the International Conference on Cognitive and Neural Systems, Boston, MA.

Keenan, B., **Beaton, D.**, Le, V., Akbari, Y., Hee Lee, J., Parekh, T., Antes, T., et al. (2009). MicroRNA Profiling Can Distinguish Normal From Neoplastic Urothelium. Presented at the 78th Annual Meeting of the New England Section of the American Urological Association, DC.

Le, V., Juan, D., **Beaton, D.**, Keenan, B., Akbari, Y., Hee Lee, J., Parekh, T., et al. (2009). Evaluating miRNA Expression Integrity in FFPE Samples with qRT-PCR by Using Patient-matched Formalin-fixed and Fresh-frozen Tissues from Renal Cell Carcinoma (RCC) Patients. Presented at the 78th Annual Meeting of the New England Section of the American Urological Association, Washington, DC.

Akbari, Y., **Beaton, D.**, Keenan, B., Le, V., Parekh, T., Hee Lee, J., Antes, T., et al. (2009). Assessing the Potential of the Htert Cell Line as a Model of Normal Bladder Urothelium via Analysis of miRNA Expression. Presented at the 78th Annual Meeting of the New England Section of the American Urological Association, Washington, DC.

Unpublished/Not Refereed Abstracts & Posters

2013

Levine, B., Bacopulos, A., Anderson, N.D., Black, S.E., Davidson, P.S.R., Fitneva, S.A., McAndrews, M.P., Spaniol, J., Jeyakumar, N., Abdi, H., **Beaton, D.,** Owen, A.M., & Hampshire, A. (2013). Validation of a Novel Computerized Test Battery for Automated Testing. Poster presented at the 23rd Annual Neuroscience Conference: Brain Plasticity and Neurorehabilitation, March 2-4, 2013, Toronto, Ontario.

2012

Beaton, D., Dunlop, J., Filbey, F. M., Abdi, H. (2012). Connecting Connectivity: A Brief Overview of Methods to Combine Multiple Data Types in Connectivity Analyses. Neural Computation 2012, Hanover, NH.

2011

Beaton, D., Abdi, H., Dunlop, J., Krishnan, A., Buschbaum, B. (2011). An ExPosition into Neural Computation with the SVD in R. Neural Computation 2011, Hanover, NH. [PDF]

Beaton, D., Abdi, H. (2011). Partial Least Squares-Correspondence Analysis (PLS-CA): A New Method to Reveal Common Patterns in Measures of Memory and Genetics. Neuroscience of Stress and Memory Conference, Dallas, TX. [PDF]

Beaton, D., Abdi, H. (2011). Linking Genetics and Cognition with Partial-Least Squares and Correspondence Analysis: A Study of the ADNI Cohort. Dallas Aging and Cognition Conference, Dallas, TX. [PDF]

2010

Beaton, D., Abdi, H. (2010). Integrating Partial-least Squares and Correspondence Analysis for Genetics-based Cognition Research. Presented at the Greater Dallas Human Brain Imaging Retreat, Dallas, TX. [PDF]

2009

Dalton, S., Brookstein, A., Hegedus, S., & **Beaton, D.** (2009). Democratizing Access to Core

Mathematics Across Grades 9-12: A Longitudinal Study. Presented at the Sigma Xi 2009, Poster, University of Massachusetts at Dartmouth.

2005

Beaton, D., & Valova, I. (2005). Alzheimer's Detection with EEG and ICA. Presented at the Sigma Xi 2005, Poster, University of Massachusetts at Dartmouth.

Software

R

Beaton, D. (2013). prettyGraphs: publication style graphics (Version 2.1.5) [Software]. Available from http://cran.r-project.org/web/packages/prettyGraphs/index.html or http://code.google.com/p/exposition-family/source/browse/#svn%2FPackages

Beaton, D., Chin Fatt, C., R., Abdi, H. (2013). ExPosition: Exploratory analysis with the singular value decomposition (Version 2.8.19) [Software]. Available from http://cran.r-project.org/web/packages/ExPosition/index.html or http://code.google.com/p/exposition-family/source/browse/#svn%2FPackages

Beaton, D., Dunlop, J., Abdi, H. (2013). InPosition: Inference tests for ExPosition (Version 0.12.7) [Software]. Available from http://cran.r-project.org/web/packages/ExPosition/index.html or http://code.google.com/p/exposition-family/source/browse/#svn%2FPackages

Beaton, D., Rieck, J., Chin Fatt, C., R., Abdi, H. (2013). TExPosition: Two-table ExPosition (Version 2.6.10) [Software]. Available from http://cran.r-project.org/web/packages/TExPosition/index.html or http://code.google.com/p/exposition-family/source/browse/#svn%2FPackages

Beaton, D., Rieck, J., Abdi, H. (2013). TInPosition: Inference tests for TExPosition (Version 0.13.6) [Software]. Available from http://cran.r-project.org/web/packages/TExPosition/index.html or http://code.google.com/p/exposition-family/source/browse/#svn%2FPackages

Chin Fatt, C., R., **Beaton, D.,** Abdi, H. (2013). MExPosition: Multi-table ExPosition (Version 2.0.3) [Software]. Available from http://cran.r-project.org/web/packages/MExPosition/index.html or http://code.google.com/p/exposition-family/source/browse/#svn%2FPackages

Beaton, D., Chin Fatt, C., R., Abdi, H. (2013). DistatisR: DiSTATIS Three Way Metric Multidimensional Scaling (Version 1.0) [Software]. Available from http://http://cran.r-project.org/web/packages/DistatisR/index.html or http://code.google.com/p/exposition-family/source/browse/#svn%2FPackages

Awards

Funding

2016-2018

Ontario Neurodegenerative Disease Research Initiative (ONDRI) Postdoctoral Scholar.

ONDRI Scholars Announced

2013-2016

NIH NRSA #F31DA035039-01A1.

Implementing Appropriate Multivariate Methods for Higher Quality Results from Genome-wide Association Studies in Substance Abuse Populations.

2014

Neuroimaging Training Program (NITP) Summer Fellowship

UCLA. Course details.

Invited Guest: Department of Psychology, University of Chester

Invited guest on behalf of Dr. Ljubica Damjanovic (Santander Visiting Fellowship Award).

March 08-16, 2014 -- **CANCELLED due to funding limitations.**

2011

Student Travel Award(s):

NeuroInformatics 2011. [link]

Society for Mathematical Psychology: MathPsych 2011

2010

Student Travel Award:

Society for Mathematical Psychology: MathPsych 2010

2008

NSF Grant #0812995: East Asia Pacific Summer Institute Fellow. [link]

Graduate Institute of Network Learning Technology, National Central University, Jhong-Li, Taiwan. Host advisor: Tak-Wai Chan, Ph.D.

http://www.nsf.gov/awardsearch/showAward.do?AwardNumber=0812995

Paper/presentation/misc. Awards

2017

Merit Abstract Award: OHBM 2017

Beaton D., Bartha R., Black S.E., Casaubon L., Dowlatshahi D., Hassan A., Kwan D., Levine B., Mandzia J., McLaughlin P., Orange J., Peltsch A., Ramirez J., Roberts A., Sahlas D., Saposnik G., Swartz R., Symons S., Troyer A., Strother S.C., ONDRI Investigators (2017). Vascular cognitive impairment subgroups show distinct aspects of preserved cognition. To be presented at OHBM 2017, Vancouver, BC, Canada.

Most promising project: BrainHack Global Toronto

Rieck, J., **Beaton, D.**: MARINeR: *M*ultivariate *A*nalysis and *R*esampling *I*nference for *Ne*uroimaging in *R*. https://github.com/derekbeaton/BrainHack_TO_2017/tree/master/MARINeR

2016

Third prize in poster competition

Beaton, D., Dunlop, J., Abdi, H., Filbey, F.M. (2016). Endocannabinoid, GABA-ergic, and dopaminergic contributions to neural response of different reward types. Presented at International Imaging Genetics Conference, Irvine, CA.

2013

First prize in Society for Applied Multivariate Research.

Beaton, D., Filbey, F. M., Abdi, H. (2013). Partial Least Squares Correspondence Analysis Reveals Genetic Correlates of Impulsivity and Addiction. Presented at Southwestern Psychological Association Annual Meeting, Ft. Worth, TX.

Third prize in Society for Applied Multivariate Research.

Faso, D., **Beaton, D.,** Abdi, H., Sasson, N. J., Pinkham, A. E., (2013). Distinct Visual Patterns in Clinical Populations Identified by Partial Least Squares. Presented at Southwestern Psychological

Association Annual Meeting, Ft. Worth, TX.

2012

Meritorious Poster Submission (1 of 48 Meritorious Posters out of 1138 posters)
Schwarz, A. L., van Kleeck, A., **Beaton, D.,** Horne, E., Ahn, L., MacKenzie, H. (2012). Capturing Decision-Making: SLPs' Criteria for Selecting Preschool Read-Aloud Storybooks. Presented at American Speech-Language-Hearing Association Annual Meeting 2012, Atlanta, GA.

2008

Best Paper
Valova, I., **Beaton, D.**, MacLean, D. (2008). Role of Initialization in SOM networks - study of self-similar curve topologies. In ASME Press Series on Intelligent Engineering Systems Through Artificial Neural Networks. Presented at the ANNIE 2008, St. Louis, MO, USA.

Professional Relationships and Duties

Committees

- 2017 Program Chair for Society for Applied Multivariate Research (at Southwestern Psychological Association).
- 2016 Program Chair for Society for Applied Multivariate Research (at Southwestern Psychological Association).
- 2015 Program Chair for Society for Applied Multivariate Research (at Southwestern Psychological Association).

Ad-hoc Reviewer

- The American Journal of Drug and Alcohol Abuse
- Research in Autism Spectrum Disorders
- Food Quality and Preference
- IEEE Transactions on Neural Networks and Learning Systems
- Symposium Series on Computational Intelligence (SSCI) 2013
- International Joint Conference on Neural Networks (IJCNN) 2010-2014

Member

- Organization for Human Brain Mapping (2015-Present)
- Cognitive Neuroscience Society (2015-Present)
- Foundation for Open Access Statistics (FOAS; 2013-Present)
- Society for Applied Multivariate Research (SAMR; 2013-Present)
- Southwestern Psychological Association (SWPA; 2012-Present)
- Society for Mathematical Psychology (SMP; 2010-Present)
- International Neural Network Society (INNS; 2007-2009)
- National Council of Teachers of Mathematics (NCTM; 2006-2008)

Mentorship

- Rachelle Akpanumoh (linear models for single nucleotide polymorphisms; Summer 2014)
- Fahd Alhazmi (univariate vs. multivariate tests for single nucleotide polymorphisms, multivariate mega-meta semantic analyses of neuroimaging databases; Summer 2014-Present)
- Libby Damjanovic (multivariate and non-parametric analyses for surveys; September 02-12, 2014)

Public service

Wikipedia

- ○ Substantial update to Generalized Singular Value Decomposition
- ○ Substantial update to Correspondence Analysis
- ○ Minor additions to Principal Component Analysis
- ○ Minor clarification to Rv coefficient

CURRICULUM VITAE

UPDATED: 3.25.2017 Google Scholar Page

Professional Experience

2016-present
Postdoctoral Fellow, Rotman Research Institute, Baycrest Health Sciences and Ontario Neurodegenerative Disease Research Initiative.

2013-2016:
NRSA Predoctoral Fellow, School of Behavioral and Brain Sciences, The University of Texas at Dallas.

2009-2017:
Graduate Assistant, School of Behavioral and Brain Sciences, The University of Texas at Dallas (Hervé Abdi, Ph.D.)
Responsibilities: Development of software tools, statistical modeling/analysis, pattern classification, pattern recognition, functional neuroimaging, behavioral measures, genomics.

2009-2013:
Teaching Assistant, School of Behavioral and Brain Sciences, The University of Texas at Dallas (Dana Roark, Ph.D.; Daniel Krawczyk, Ph.D.; Hervé Abdi, Ph.D.; Joseph Dunlop, M.S.; Noah Sasson, Ph.D.; Gail Tillman, Ph.D.)
Responsibilities: Grading, material reviews, class lectures, course content.

2008-2012:
Contract Biostatistician/Bioinformatics Analyst, General Urology Lab, Boston University Medical Center (Louis Liou, MD)
Responsibilities: Statistical analyses, development of software tools.

2008-2009:
Research Associate, James J. Kaput Center for Research and Innovation in Mathematics Education (Stephen Hegedus, Ph.D.)
Responsibilities: Data collection, teacher training, software development/testing, statistical analysis.

2007-2008:
Research Assistant, James J. Kaput Center for Research and Innovation in Mathematics Education (Stephen Hegedus, Ph.D.)
Responsibilities: Data collection, teacher training, software development/testing.

2006-2008:
Research Assistant, Neural and Adaptive Systems Lab (Iren Valova, Ph.D.)
Responsibilities: Development and analysis of neural algorithms, authorship.

Teaching Assistant, Computer and Information Sciences Department, University of Massachusetts at Dartmouth (Iren Valova, Ph.D.; Li Shen, Ph.D.; Ryan Robidoux, MS)
Responsibilities: Grading, curriculum design, teaching.

2004-2007:
Research Assistant, SimCalc Projects (James Kaput, Ph.D.; Stephen Hegedus, Ph.D.)
Responsibilities: Software development/testing, data collection, teacher training.

CURRICULUM VITAE

UPDATED: 3.25.2017 Google Scholar Page

Teaching Experience

2015

Teaching Assistant/Co-instructor: Univariate and Multivariate Analysis in R (Graduate course).
Instructor: Hervé Abdi, Ph.D.
Guest Lectures: Research Methods in Behavioral and Brain Sciences III (Graduate course).
Topics: PCA, PLS, Discriminant Methods, Clustering, etc...
Instructor: Hervé Abdi, Ph.D.

2014

Guest Lectures: Research Methods in Behavioral and Brain Sciences II (Graduate course).
Instructor: Hervé Abdi, Ph.D.
Guest Lectures: Cognitive Science (Graduate course).
Topic: Alan Turing and his contributions to cognitive science.
Instructor: Alice O'Toole

2013

Teaching Assistant/Co-instructor: Univariate and Multivariate Analysis in R (Graduate course).
Instructor: Hervé Abdi, Ph.D.
Guest Lectures: Univariate and Multivariate Analysis in R (Graduate course).
Topics: R, PCA, MDS, PLS, Discriminant Methods, Multi-block methods, Clustering, etc...
Instructor: Hervé Abdi, Ph.D.
Teaching Assistant: Research Methods in Behavioral and Brain Sciences III (Graduate course).
Instructor: Hervé Abdi, Ph.D.
Guest Lectures: Research Methods in Behavioral and Brain Sciences III (Graduate course).
Topics: PCA, PLS, Discriminant Methods, Clustering, etc...
Instructor: Hervé Abdi, Ph.D.
Teaching Assistant: Research Methods in Behavioral and Brain Sciences II (Graduate course).
Instructor: Hervé Abdi, Ph.D.
Guest Lectures: Cognitive Science (Graduate course).
Topic: Alan Turing and his contributions to cognitive science.
Instructor: Alice O'Toole

2012

Teaching Assistant: Research & Evaluation Methods (Undergraduate course)
Instructor: Gail Tillman, Ph.D.
Guest Lectures: Research & Evaluation Methods (Undergraduate course)
Topics: t-tests, ANOVAs
Instructor: Gail Tillman, Ph.D.
Teaching Assistant/Co-instructor: Univariate and Multivariate Analysis in R (Graduate course).
Instructor: Hervé Abdi, Ph.D.
Guest Lectures: Univariate and Multivariate Analysis in R (Graduate course).
Topics: R, PCA, MDS, PLS, Discriminant Methods, Multi-block methods, Clustering, etc...
Instructor: Hervé Abdi, Ph.D.
Teaching Assistant: Research Methods in Behavioral and Brain Sciences III (Graduate course).
Instructor: Hervé Abdi, Ph.D.
Guest Lectures: Research Methods in Behavioral and Brain Sciences III (Graduate course).
Topics: PCA, PLS, Discriminant Methods, Clustering, etc...
Instructor: Hervé Abdi, Ph.D.
Teaching Assistant: Research Methods in Behavioral and Brain Sciences II (Graduate course).

Instructor: Hervé Abdi, Ph.D.
Guest Lectures: Research Methods in Behavioral and Brain Sciences II (Graduate course).
Topics: Correlation, regression, ANOVA
Instructor: Hervé Abdi, Ph.D.
Guest Lectures: Cognitive Science (Graduate course).
Topic: Alan Turing and his contributions to cognitive science.
Instructor: Alice O'Toole

2011

Teaching Assistant: Research Design & Analysis (Undergraduate course).
Instructor: Noah Sasson, Ph.D.
Guest Lectures: Research Design & Analysis (Undergraduate course)
Topic: Communicating Research in Research Design & Analysis
Instructor: Noah Sasson, Ph.D.
Teaching Assistant/Co-instructor: Univariate and Multivariate Analysis in R (Graduate course).
Instructor: Hervé Abdi, Ph.D.
Guest Lectures: Univariate and Multivariate Analysis in R (Graduate course).
Topics: Intro to R, ANOVA, Bootstrap, Permutation Tests, PCA in R Course
Instructor: Hervé Abdi, Ph.D.
Teaching Assistant: Research Methods in Behavioral and Brain Sciences III (Graduate course).
Instructor: Hervé Abdi, Ph.D.
Guest Lectures: Research Methods in Behavioral and Brain Sciences III (Graduate course).
Topics: Multi-block Analysis, STATIS, Clustering
Instructor: Hervé Abdi, Ph.D.
Teaching Assistant: Research Methods in Behavioral and Brain Sciences II (Graduate course).
Instructor: Hervé Abdi, Ph.D.
Guest Lectures: Research Methods in Behavioral and Brain Sciences II (Graduate course).
Topics: Correlation, regression, ANOVA
Instructor: Hervé Abdi, Ph.D.
Guest Lectures: Cognitive Science (Graduate course).
Topic: Alan Turing and his contributions to cognitive science.
Instructor: Alice O'Toole

2010

Teaching Assistant: Research Methods in Behavioral and Brain Sciences III (Graduate course).
Instructor: Hervé Abdi, Ph.D.
Guest Lectures: Research Methods in Behavioral and Brain Sciences III (Graduate course).
Topics: Statistical Analyses Using R.
Instructor: Hervé Abdi, Ph.D.
Teaching Assistant: Research Methods in Behavioral and Brain Sciences I (Graduate course).
Instructor: Joseph Dunlop, MS
Teaching Assistant: Cognitive Psychology (Undergraduate course).
Instructor: Daniel Krawczyk, Ph.D.

2009

Teaching Assistant: Experimental Projects (Undergraduate course).
Instructor: Dana Roark, Ph.D.
Guest Lectures: Advanced Research Methods: Statistical Analysis Using R.
Topics: PCA, FactoMineR, Clustering
Instructor: Hervé Abdi, Ph.D.
Guest Lectures: Research Methods in Behavioral and Brain Sciences III.

Topics: Clustering
Instructor: Hervé Abdi, Ph.D.

2008

Teaching Assistant/Lab Instructor: Introduction to C Programming (Undergraduate course).
Instructor: Ryan Robidoux, M.S.
Teaching Assistant: Video Game Design (Undergraduate course).
Instructor: Iren Valova, Ph.D.
Guest Lectures: Video Game Design (Undergraduate course).
Topic: Proprioception and Haptics in Gaming
Instructor: Iren Valova, Ph.D.
Teaching Assistant: Artificial Intelligence (Undergraduate course).
Instructor: Iren Valova, Ph.D.
Guest Lectures: Artificial Intelligence (Undergraduate course).
Topics: Neural Computation
Instructor: Iren Valova, Ph.D.

2007

Teaching Assistant/Lab Instructor: Introduction to C Programming (Undergraduate course).
Instructor: Ryan Robidoux, M.S.
Teaching Assistant: Artificial Intelligence (Undergraduate course).
Instructor: Iren Valova, Ph.D.
Guest Lectures: Artificial Intelligence (Undergraduate course).
Topics: Neural Computation
Instructor: Iren Valova, Ph.D.

2006

Teaching Assistant/Lab Instructor: Procedural Programming with C (Undergraduate course).
Instructor: Li Shen, Ph.D.
Guest Lectures: Data Mining and Knowledge Discovery (Undergraduate course).
Topic: Hierarchical clustering
Instructor: Iren Valova, Ph.D.

Conferences, Workshops, & Presentations

2016

- Twelfth International Imaging Genetics Conference, 2016
  - Poster presentation -- awarded 3rd prize.

2015

- Organization for Human Brain Mapping, 2015
  - Three posters (one first author)
- Southwest Psychological Association, 2015
  - Workshop: An Introduction to Permutation and Bootstrap Resampling.
  - Workshop: Complementary and alternative approaches to *p*-values: Intervals, Effect Sizes, and Practical Advice.

2014

- Partial Least Squares and Related Methods, 2014
  - Two talks
- Southwest Psychological Association, 2014 (workshop materials)
  - Workshop: An ExPosition of Bootstrap and Permutation tests for Principal Components

    - Analyses.
    - Workshop: An Introduction to ANOVA in R.
- Tenth International Imaging Genetics Conference, 2014
    - Poster presentation

2013

- Southwest Psychological Association, 2013
    - Poster presentation
- Ninth International Imaging Genetics Conference, 2013
    - Poster presentation

2012

- Neural Computation, 2012
    - Poster Presentation
- Partial Least Squares and Related Methods, 2012
    - Talk
- Eighth International Imaging Genetics Conference, 2012
    - Poster presentation

2011

- Society for Neuroscience, 2011
    - Poster Presentation
- NeuroInformatics, 2011
    - Poster Presentation
- Neural Computation, 2011
    - Poster Presentation
- Society for Mathematical Psychology, 2011
    - Talk
- Neuroscience of Stress and Memory Conference, 2011
    - Data Blitz
- Dallas Aging and Cognition Conference, 2011
    - Poster presentation
- Seventh International Imaging Genetics Conference, 2011
    - Poster presentation

2010

- First Annual Greater Dallas Human Brain Imaging Retreat, 2010
    - Poster presentation
- Society for Mathematical Psychology, 2010
    - Poster presentation

2009

- International Joint Conference on Neural Networks, 2009 - Atlanta, GA
    - Two poster presentations
- International Conference on Cognitive and Neural Systems, 2009 - Boston, MA
    - Poster presentation

2007

- Executive Briefing Center, Apple Inc, Cupertino, CA
    - Presentation of SimCalc materials, software and curriculum
- International Joint Conference on Neural Networks, 2007 - Orlando, FL
    - Poster presentation
- National Council of Teachers of Mathematics 2007 – Atlanta, GA
    - Several workshops (SimCalc software)

- T3 ("T-Cubed", Teachers Teaching with Technology) – Chicago, IL.
  - SimCalc booth operator

2006

- International Association for the Development of the Information Society: Web Applications and Research
  - Paper presentation
- SRI TexTeam Scale Up 2006 (SimCalc representative)
  - Technology/Curriculum Instructor (June – Edinburg, TX; August – Ft. Worth, TX)

2005

- International Association of Science and Technology for Development: Circuits, Signals and Systems 2005 – Marina Del Rey, CA.
  - Paper presentation
- National Council of Teachers of Mathematics 2005 – Denver, CO
  - SimCalc workshop
- T3 ("T-Cubed", Teachers Teaching with Technology) 2005 – Washington, D.C.
  - SimCalc booth operator

2004

- Association of Teachers of Mathematics in New England 2004 – Providence, RI
  - SimCalc workshop

Relevant Courses Taken

Graduate

- Face Processing (UT Dallas)
- Neuroimaging Training Program (UCLA)
- Perception (UT Dallas)
- Cognitive Neuroscience of Human Memory (UT Dallas)
- Cognitive Development (UT Dallas)
- Research Methods in Behavioral and Brain Sciences III (UT Dallas)
- Advanced Research Methods: Statistical Analysis Using R (UT Dallas)
- Fundamentals of Functional Neuroimaging (UT Dallas)
- Complex Systems (New England Complex Systems Institute)
- Neural Computing (UMass Dartmouth)
- Bioinformatics (UMass Dartmouth)
- Artificial Intelligence (UMass Dartmouth)
- Data Mining and Knowledge Discovery (UMass Dartmouth)
- Computational Theory (UMass Dartmouth)
- Computer Graphics (UMass Dartmouth)

Undergraduate

- Bioinformatics (UMass Dartmouth)
- Image Processing and Analysis (UMass Dartmouth)
- Artificial Intelligence (UMass Dartmouth)