Psychological Methods

© 2015 American Psychological Association
1082-989X/15/$12.00 http://dx.doi.org/10.1037/met0000053

Partial Least Squares Correspondence Analysis: A Framework to Simultaneously Analyze Behavioral and Genetic Data

Derek Beaton and Joseph Dunlop
The University of Texas at Dallas

Alzheimer's Disease Neuroimaging Initiative

Hervé Abdi
The University of Texas at Dallas

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 by the availability of genotyping technologies (e.g., microarrays) that provide new genetic data, such as single nucleotide polymorphisms (SNPs). 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, but this coding scheme can make difficult the multivariate analysis of their relationships with behavioral measurements, because most multivariate techniques developed for the analysis between sets of variables are designed for quantitative variables. To palliate this problem, we present a generalization of partial least squares—a technique used to extract the information common to 2 different data tables measured on the same observations—called *partial least squares correspondence analysis*—that is specifically tailored for the analysis of categorical and mixed ("heterogeneous") data types. Here, we formally define and illustrate—in a tutorial format—how partial least squares correspondence analysis extends to various types of data and design problems that are particularly relevant for psychological research that include genetic data. We illustrate partial least squares correspondence analysis with genetic, behavioral, and neuroimaging data from the Alzheimer's Disease Neuroimaging Initiative. R code is available on the Comprehensive R Archive Network and via the authors' websites.

Keywords: partial least squares, correspondence analysis, chi square, genetics, resampling

Supplemental materials: http://dx.doi.org/10.1037/met0000053.supp

Derek Beaton, School of Behavioral and Brain Sciences, The University of Texas at Dallas; Joseph Dunlop, Center for BrainHealth and School for Behavioral and Brain Sciences, The University of Texas at Dallas; Alzheimer's Disease Neuroimaging Initiative; Hervé Abdi, School of Behavioral and Brain Sciences, The University of Texas at Dallas.

Joseph Dunlop is now with the SAS Institute Inc., Cary, NC. Derek Beaton is currently supported via training grant by the National Institute on Drug Abuse (F31DA035039). Hervé Abdi would like to acknowledge the support of an EURIAS fellowship at the Paris Institute for Advanced Studies (France), with the support of the European Union's 7th Framework Program for research, and from a funding from the French State managed by the "Agence Nationale de la Recherche (program: Investissements d'avenir, ANR-11-LABX-0027-01 Labex RFIEA+)." Data collection and sharing for this project was funded by the Alzheimer's disease Neuroimaging Initiative (ADNI; National Institutes of Health Grant U01 AG024904) and Department of Defense ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Health care; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. For up-to-date information and data see: http://www.adni-info.org and http://adni.loni.usc.edu/. Data used in preparation of this article were obtained from the ADNI database (http://adni.loni.usc.edu/). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete list of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf. Many thanks are due to two anonymous reviewers, the editor, and associate editor of *Psychological Methods* for their help and constructive feedback. We would also like to thank Jenny Rieck for expert comments and feedback on the manuscript.

Correspondence concerning this article should be addressed to Derek Beaton or Hervé Abdi, School of Behavioral and Brain Sciences, The University of Texas at Dallas, MS: GR4.1, 800 West Campbell Road, Richardson, TX 75080. E-mail: derekbeaton@utdallas.edu or herve@utdallas.edu

Interests in the genetic contributions to psychological traits date back over a century, and these interests fostered advancements in measurement and assessment of genetic contributions to behavior. Thorndike (1905) initiated some of the earliest efforts "to make modern statistical methods current in psychology" (Sanford, 1908, p. 142) as well as in genetics, and he was soon followed by Fisher (1919), who favored more formal approaches. Over a decade later, Thurstone (1934) suggested using "factor analysis" to understand cognitive abilities and personality traits. Thurstone favored these methods, in part, because he firmly believed "that the isolation of the mental abilities will turn out to be essentially a problem in genetics" (p. 32).

These same beliefs, interests, questions, and advancements have been revived in recent years with the advent of technology such as microarrays and genome-wide association studies (GWASs). Today, 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). SNPs can provide insight into numerous diseases (Weiner & Hudson, 2002), and many studies have shown associations of specific SNPs with particular diseases (Lakatos et al., 2010), disorders (Clarke et al., 2014; Filbey, Schacht, Myers, Chavez, & Hutchison, 2010; Lantieri, Glessner, Hakonarson, Elia, & Devoto, 2010; Romanos et al., 2008), personality (Munafò & Flint, 2011), phenotypes (Cho et al., 2011), and traits (Hamidovic, Dlugos, Palmer, & de Wit, 2010; Miyajima et al., 2008).

And now again—just as in the early days of Thorndike and Fisher—new statistical methods are being developed to understand how genetics may contribute to behaviors (or traits). Most of these modern statistical techniques are multivariate in nature. Some approaches include derivatives of independent components analysis (Liu et al., 2009; Yang, Liu, Sui, Pearlson, & Calhoun, 2010), sparse multivariate regression approaches (Vounou, Nichols, Montana, & Alzheimer's Disease Neuroimaging Initiative, 2010), distance matrix regression (Zapala & Schork, 2006), and path modeling (Franić et al., 2013). Further, some of these recent techniques are designed to simultaneously analyze behavioral and genetic data (Bloss, Schiabor, & Schork, 2010), in part in order to increase statistical power to detect genetic contributions to traits, behaviors, and phenotypes (Schifano, Li, Christiani, & Lin, 2013; Seoane, Campbell, Day, Casas, & Gaunt, 2014; van der Sluis, Posthuma, & Dolan, 2013). This boost in power can be particularly relevant in psychological research in which sample sizes can rarely reach "standard" sizes for genome-wide studies (i.e., $N \approx 5{,}000$).

One versatile family of multivariate techniques—the partial least squares (PLS) family—is particularly suited for analyzing two tables of data measured on the same observations, and has been developed into regression (Abdi, 2010; Tenenhaus, 1998), correlation (Krishnan, Williams, McIntosh, & Abdi, 2011; McIntosh & Lobaugh, 2004), and path-modeling approaches (Esposito Vinzi, Chin, Henseler, & Wang, 2010; Tenenhaus, Esposito Vinzi, Chatelin, & Lauro, 2005). Further, PLS approaches have been used in studies on genetics and genomics (Michaelson, Alberts, Schughart, & Beyer, 2010; Moser, Tier, Crump, Khatkar, & Raadsma, 2009; Wang, Ho, Ye, Strickler, & Elston, 2009), integrating genetic and brain data (Le Floch et al., 2012; Tura, Turner, Fallon, Kennedy, & Potkin, 2008), or genetics and clinical status (Chun, Ballard, Cho, & Zhao, 2011; Sullivan et al., 2008). However, nearly all current approaches to analyzing SNPs (including PLS approaches) require SNPs to be treated as numerical data—often because of particular inheritance models—even though SNPs themselves are intrinsically categorical. This preference for methods based on numerical assumptions is likely because, in part, of the lack of multivariate methods designed specifically for the analysis of categorical—or heterogeneous—data.

SNP Coding Problem and Solution

Table 1 provides an overview of common analytical approaches and inheritance models used in association studies. As noted in Table 1, the majority of approaches are naturally categorical—in which the goal is to test presence versus absence of particular genotypes. However, for practical, analytical, and statistical purposes, SNPs in association studies are (most often) represented through an allelic counting scheme (as found, e.g., in PLINK; Purcell et al., 2007). With this coding scheme, a SNP is represented either by its number of minor alleles (e.g., as in de Leon et al., 2008: 0, 1, or 2 minor alleles) or by its number of major alleles (e.g., as in Cruchaga et al., 2010: 0, 1, or 2 major alleles). However, this counting scheme relies on two unrealistic assumptions about how SNPs might contribute to traits, behaviors, and diagnoses. With the first assumption, the (statistical) emphasis is always placed on the same allele (e.g., minor alleles) across all SNPs. The second—and more problematic—assumption states that the effect of SNPs is *uniform and linear*. A simple counterexample shows the limitations of strictly linear additive assumptions: The risk of Alzheimer's disease due to ApoE genotypes is neither uniform nor linear (see, e.g., Table 2 in Genin et al., 2011). Furthermore, "ApoE E4" gene alleles are considered risk factors for Alzheimer's disease, whereas "ApoE E2" gene alleles are considered protective against Alzheimer's disease (Corder et al., 1994). By contrast, "ApoE E2" is considered a risk factor for the inability to break down fats, and this inability could, in turn, lead to certain types of vascular diseases or obesity (Koopal, van der Graaf, Asselbergs, Westerink, & Visseren, 2014). Thus, directionality of genetic effects becomes dependent on the field of study.

Furthermore, there are two other specific cases in which the [0, 1, 2] (and similar) coding scheme is problematic, especially when the "risk" allele is computed empirically (i.e., as a frequency of, say, the minor allele). First, the minor allele for a given SNP in one sample or cohort is not guaranteed to be the same minor allele in a separate sample or cohort. This discrepancy is more likely to occur in smaller samples such as candidate gene studies. Thus, an empirical "2" in one study could be an empirical "0" in another. Second, studies that employ aggregate scores, such as profile scores or multilocus scores (e.g., Blum, Oscar-Berman, Demetrovics, Barh, & Gold, 2014; Nikolova, Ferrell, Manuck, & Hariri, 2011; van Eekelen et al., 2011)—essentially, the sum or average of the risk alleles across multiple SNPs—could in fact miss effects, if the risk is empirically defined and the emphasis is in the same direction (i.e., minor homozygote as "2"). In fact, ApoE is an

Table 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 (three levels)
Genotypic[a]	AA	Aa	aa	Categorical (three 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 (three levels)
Genotypic	[1 0 0]	[0 1 0]	[0 0 1]	Categorical (three 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).

example of such an effect. ApoE is genotyped by the two SNPs—rs7412 and rs429358, in which the ApoE E4/E4 genotype (the major risk factor for Alzheimer’s disease) is produced by a major homozygote (i.e., 0) from rs7412, and a minor homozygote (i.e., 2) from rs429358.[1]

When risk alleles and true inheritance patterns are unknown, we (of course) do not know which allele should be emphasized as the risk factor, nor whether an effect should be tested with linear assumptions. Furthermore, if more than one SNP contributes to an effect in complex ways (e.g., a haplotype) sometimes, the linear additive [0, 1, 2] coding can completely miss effects (Vormfelde & Brockmöller, 2007). So, applying these [0, 1, 2] values to all SNPs—and subsequently testing each SNP independently—could either miss an effect or even misrepresent the true inheritance pattern. This problem occurs because allelic counts do not represent how much a SNP actually contributes to an effect, but rather codes only for a presumed effect (i.e., a linear contrast applied to an ANOVA design) about each genotype (e.g., AA, AG, or GG). Thus, the most parsimonious approach to genetic association studies—especially when we do not know the inheritance pattern, the direction, or the size of the effect—is to treat each genotype (e.g., AA, AG, or GG) as a level of a categorical variable (e.g., rs6859). This representation closely matches the genotypic or codominant models of inheritance and is a more general approach to testing. In fact, prior work has shown that the codominant model (see Table 1 and Appendix A) is the best choice when choosing a single model if the true inheritance pattern is unknown (Lettre, Lange, & Hirschhorn, 2007). Furthermore, because the genotypic or codominant models are the most general, there is no requirement for additional testing of other inheritance models (i.e., additive, dominant, and recessive) on the same data set (which would require corrections for multiple tests).

Here, we present partial least squares correspondence analysis (PLSCA; Beaton, Filbey, & Abdi, 2013)—a derivative of PLS—and several of its novel extensions tailored for the particular data and design issues often confronted in genetic association studies within the psychological, cognitive, and neurological sciences. PLSCA (like traditional PLS) is designed to simultaneously analyze two tables of data. However, PLSCA is more versatile than PLS because it can analyze diverse types of data, such as two categorical data sets, one categorical and one quantitative data set, or two data sets in which each is a mixture of categorical and quantitative data within each table. This versatility of PLSCA makes it an ideal method for genetic association studies because (a) as a multivariate method, PLSCA can detect how multiple genotypes contribute to many various phenotypes (traits) simultaneously; (b) phenotypes (traits) can be either categorical (e.g., diagnosis, sex) or quantitative (e.g., summary score, brain size); and (c) PLSCA—when used with a fully categorical coding scheme to represent SNPs—is flexible enough to detect linear and nonlinear relationships within (e.g., genotypic) and between (genotypic-phenotypic) data sets.

Because a uniform and linear assumption could misrepresent the expected contribution of genetic data to traits, behaviors, and diagnoses, we present PLSCA with an explicit categorical coding scheme for SNPs that represents each genotype. However, it is worth noting that—with a proper coding scheme—PLSCA can also analyze other inheritance models or even analyze mixed inheritance models within a single analysis (see Appendix A [and R code] for a detailed discussion on how to apply PLSCA with other inheritance models). Therefore, instead of using allelic

[1] For a simple explanation of the ApoE genotype and haplotype, see Nyholt, Yu, and Visscher (2009). See also the National Institutes of Health’s minor allele frequencies at http://www.ncbi.nlm.nih.gov/SNP/snp_ref.cgi?rs=7412 and http://www.ncbi.nlm.nih.gov/SNP/snp_ref.cgi?rs=429358. Furthermore, prior work (Bennet et al., 2010) suggests that only rs429358 is a risk factor for Alzheimer’s disease.

Table 2
Nominal, Additive, 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
Additive		
Subject 1	1	1
Subject 2	2	1
Subject *i*	1	2
Subject *I*	0	0

	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, additive, and disjunctive coding of illustrative SNPs referred to as SNP 1 and SNP 2. Here, both illustrative SNPs are presented generally where A = major allele; a = minor allele. The major homozygote, heterozygote, and minor homozygote are denoted ‘AA’, ‘Aa’, and ‘aa’, respectively.

counts (i.e., 0, 1, and 2), we present PLSCA with SNPs as categorical variables (see Table 2). While this approach is rarely used in multivariate analyses, there is some precedence for analyzing genetic data as categorical (Greenacre & Degos, 1977; Park, Lee, & Kim, 2007). Additionally, multivariate contingency analyses of genetic data exist in some domains, such as ecology and biology (Dray, 2014; Edelaar et al., 2015; Frantz et al., 2013; Gasi et al., 2013; Kocovsky, Sullivan, Knight, & Stepien, 2013).

PLSCA is a derivative technique of both correspondence analysis (CA; Abdi & Williams, 2010a; Greenacre, 2010) and partial least squares correlation (PLSC; Krishnan et al., 2011; McIntosh & Lobaugh, 2004). CA is a multivariate technique designed specifically to analyze categorical data (as found, e.g., in a contingency table) in a χ^2 framework—a framework particularly well suited for identifying rare occurrences (Greenacre, 2007, 1984). PLSC is a family of methods—often used in neuroimaging (Krishnan et al., 2011; McIntosh & Lobaugh, 2004)—designed to analyze two tables of variables measured on the same observations (Abdi, Chin, Esposito Vinzi, Russolillo, & Trinchera, 2013). Interestingly, PLSC has recently been shown to be the best approach—among several canonical correlation or covariance analyses—to detect genetic associations in high dimensional neuroimaging data sets (Grellmann et al., 2015). Thus, PLSCA—by leveraging features of CA and PLSC—is particularly suited for analyzing complex genetic association studies (e.g., candidate gene or genome-wide). PLSCA can analyze genetic data in conjunction with data that are well defined and robust (e.g., clinical measures), and can do so for very high-dimensional data (even when the variables outnumber the observations).

The remainder of this article is outlined as follows. We first formally define PLSCA, and then we explain how to use particular nonparametric inference methods (e.g., tests via permutation and bootstrap) with PLSCA. Next we present a tutorial on—and introduce several variants of—PLSCA. This tutorial is broken into several studies that aim to tease apart the specific genetic contributions to: depression, dementia, and differences in brain structures in Alzheimer’s disease (via the Alzheimer’s Disease Neuroimaging Initiative [ADNI] study). Each study highlights a particular variation of PLSCA. Finally, we discuss the findings from the illustrative studies and highlight the methodological contributions of PLSCA.

A Précis of PLSC

In this section, we present a brief summary of PLSC—sometimes also called Tucker’s interbattery analysis (Tucker, 1958), singular value decomposition (SVD) of the covariance between two fields (Bretherton, Smith, & Wallace, 1992), or coinertia analysis (Dray, 2014)—in order to (a) provide the background, and (b) establish the concepts and notations we need for PLSCA.

Notation

Matrices are denoted with uppercase bold letters (e.g., $\mathbf{X}$), and vectors with lowercase bold letters (e.g., $\mathbf{x}$); scalars are denoted by uppercase italic letters (e.g., I), and indices by lowercase 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 “×.” Further, we reserve some letters to denote vectors and matrices that represent specific, and common, features of these methods. These reserved letters are listed in Table 3.

Table 3
Reserved Notation Descriptions

Reserved notation	Definition
$\mathbf{Z}_*$	Centered and (usually) normalized matrices
$\mathbf{L}_*$	Latent variables scores
$\mathbf{F}_*$	Component (or factor) scores
$\mathbf{m}_*$	Vectors of masses
$\mathbf{W}_*$	Diagonal matrices of weights
$\mathbf{O}_*$	Observed values (under χ^2 assumptions)
$\mathbf{E}_*$	Expected values (under χ^2 assumptions)

Note. Notation letters that are reserved for specific uses and their corresponding definitions. In each case, the asterisk (*) is replaced by the name of a matrix (e.g., **X**) or by an index (e.g., *J*) to denote where the matrix originates from. For example, (a) $\mathbf{Z_X}$ would mean a centered and normalized version of **X**; (b) $\mathbf{W_Y}$ is a matrix of weights derived from the **Y**; and (c) $\mathbf{F}_J$ are the component scores associated with the *J* columns of a matrix.

PLSC

PLSC analyzes the relationship between two data matrices of sizes (respectively) *I* by *J* and *I* by *K*, denoted (respectively) **X** and **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 **X** and **Y** are denoted $\mathbf{Z_X}$ and $\mathbf{Z_Y}$. The common information between these two data tables is represented by the matrices computed as

$$\mathbf{R} = \mathbf{X}^{\mathsf{T}}\mathbf{Y} \text{ and } \mathbf{Z_R} = \mathbf{Z_X^{\mathsf{T}}}\mathbf{Z_Y}. \tag{1}$$

This multiplication produces a *J* by *K* cross-product matrix (**R**) or correlation matrix ($\mathbf{Z_R}$). In PLSC (Abdi & Williams, 2013; Bookstein, 1994; Krishnan et al., 2011; McIntosh, Bookstein, Haxby, & Grady, 1996), the variables are, in general, centered and normalized (i.e., matrices $\mathbf{Z_X}$ and $\mathbf{Z_Y}$ are used), and therefore the matrix $\mathbf{Z_R}$ is used for further analysis. The matrix $\mathbf{Z_R}$ is decomposed with the SVD (see Appendix B) as

$$\mathbf{Z_R} = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}^{\mathsf{T}}, \tag{2}$$

where (a) *L* is the rank of $\mathbf{Z_R}$, (b) **U** is the *J* by *L* orthonormal matrix of left singular vectors, (c) **V** is the *K* by *L* orthonormal matrix of right singular vectors, and (d) **Δ** is an *L* by *L* diagonal matrix (i.e., the off-diagonal elements of **Δ** are all 0), in which the elements of the vector diag{**Δ**} 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 **U** and **V** are also called *saliences* (Bookstein, 1994; Krishnan et al., 2011; McIntosh & Lobaugh, 2004). The matrices **UΔ** and **VΔ** are akin to component (a.k.a. *factor*) scores for principal components analysis (PCA) (Abdi & Williams, 2010a) and CA (Abdi & Béra, 2014). In this article, we tend to use the more ubiquitous nomenclature from PCA and CA rather than the more specialized nomenclature from PLSC.

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 Abdi & Williams, 2010b; Krishnan et al., 2011). The coefficients—which play a role analogous to loadings in PCA—of these linear combinations are given by the (respectively left and right) singular vectors of $\mathbf{Z_R}$. The latent variables for $\mathbf{Z_X}$ and $\mathbf{Z_Y}$ are computed as

$$\mathbf{L_X} = \mathbf{Z_X}\mathbf{U} \text{ and } \mathbf{L_Y} = \mathbf{Z_Y}\mathbf{V}. \tag{3}$$

What Does PLSC Maximize?

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

$$\delta = \arg\max(\mathbf{l_X^{\mathsf{T}}}\mathbf{l_Y}) = \arg\max \operatorname{cov}(\mathbf{l_X}, \mathbf{l_Y}), \tag{4}$$

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

$$\mathbf{u}_l^{\mathsf{T}}\mathbf{u}_l = 1 = \mathbf{v}_l^{\mathsf{T}}\mathbf{v}_l. \tag{5}$$

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:

$$\mathbf{l}_{\mathbf{X},l}^{\mathsf{T}}\mathbf{l}_{\mathbf{Y},l'} = 0 \text{ when } l \neq l'. \tag{6}$$

The coefficients of the successive linear transformations (stored in matrices $\mathbf{L_X}$ and $\mathbf{L_Y}$) are obtained from the SVD of $\mathbf{Z_R}$ (see Equation 2), as shown by

$$\mathbf{L_X^{\mathsf{T}}}\mathbf{L_Y} = \mathbf{U}^{\mathsf{T}}\mathbf{Z_X^{\mathsf{T}}}\mathbf{Z_Y}\mathbf{V} = \mathbf{U}^{\mathsf{T}}\mathbf{Z_R}\mathbf{V} = \mathbf{U}^{\mathsf{T}}\mathbf{U}\boldsymbol{\Delta}\mathbf{V}^{\mathsf{T}}\mathbf{V} = \boldsymbol{\Delta}. \tag{7}$$

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 are orthogonal (as defined by Equation 6) to the first pair of latent variables. This property holds for each subsequent value of *l* (for proofs, see, e.g., Bookstein, 1994, and Tucker, 1958).

PLS for Categorical Data Types

The properties of PLSC hold when matrices **X** and **Y** contain quantitative variables (and therefore $\mathbf{Z_R}$ is a correlation matrix). However, SNPs and many types of behavioral data (e.g., surveys, clinical assessments, and diagnostic groups) are inherently categorical. We now present a new PLSC method—called partial least squares correspondence analysis (PLSCA)—designed specifically to analyze two tables of categorical data (Beaton, Filbey, et al., 2013). We have implemented PLSCA (and several of its derivatives) in the R package TExPosition (Beaton, Chin Fatt, & Abdi, 2014; Beaton, Rieck, Fatt, & Abdi, 2013). Additional code—which illustrates the formalization of PLSCA and several examples—can be found on the authors' websites.[2]

[2] https://code.google.com/p/exposition-family/source/browse/Publications/PsyMet_2015 and http://www.utd.edu/~herve/PsyMet_2015.

Formalization of PLSC for Categorical Data

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., Greenacre, 1984; Lebart, Morineau, & Warwick, 1984), as illustrated in Table 2. 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*:

$$\mathbf{m_X} = \left(\mathbf{1}^{\mathrm{T}}\mathbf{X}\mathbf{1}\right)^{-1} \times \left(\mathbf{1}^{\mathrm{T}}\mathbf{X}\right) \text{ and } \mathbf{m_Y} = \left(\mathbf{1}^{\mathrm{T}}\mathbf{Y}\mathbf{1}\right)^{-1} \times \left(\mathbf{1}^{\mathrm{T}}\mathbf{Y}\right) \quad (8)$$

(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

$$\mathbf{W_X} = \mathrm{diag}\{\mathbf{m_X}\}^{-1} \text{ and } \mathbf{W_Y} = \mathrm{diag}\{\mathbf{m_Y}\}^{-1}. \quad (9)$$

As in PLSC, the disjunctive data matrices **X** and **Y** are, in general, preprocessed to have zero mean and unitary norm. Here, centered and normalized matrices are denoted $\mathbf{Z_X}$ and $\mathbf{Z_Y}$, 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 $\mathbf{Z_X}$ and $\mathbf{Z_Y}$ are computed as

$$\mathbf{Z_X} = \left(\mathbf{X} - \left(\mathbf{1}\left(\mathbf{1}^{\mathrm{T}}\mathbf{X} \times I^{-1}\right)\right)\right) \times \left(N_{\mathbf{X}} I^{\frac{1}{2}}\right)^{-1} \quad (10)$$

and

$$\mathbf{Z_Y} = \left(\mathbf{Y} - \left(\mathbf{1}\left(\mathbf{1}^{\mathrm{T}}\mathbf{Y} \times I^{-1}\right)\right)\right) \times \left(N_{\mathbf{Y}} I^{\frac{1}{2}}\right)^{-1}. \quad (11)$$

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

$$\mathbf{Z_R} = \mathbf{Z_X}^{\mathrm{T}}\mathbf{Z_Y}. \quad (12)$$

Then we decompose $\mathbf{Z_R}$ with the generalized SVD (GSVD; see Appendix B) as

$$\mathbf{Z_R} = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}^{\mathrm{T}} \text{ with } \mathbf{U}^{\mathrm{T}}\mathbf{W_X}\mathbf{U} = \mathbf{I} = \mathbf{V}^{\mathrm{T}}\mathbf{W_Y}\mathbf{V}. \quad (13)$$

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

$$\mathbf{L_X} = \mathbf{Z_X}\mathbf{W_X}\mathbf{U} \text{ and } \mathbf{L_Y} = \mathbf{Z_Y}\mathbf{W_Y}\mathbf{V}, \quad (14)$$

where—by analogy with PLSC—$\mathbf{W_X}\mathbf{U}$ and $\mathbf{W_Y}\mathbf{V}$ are called saliences. PLSCA performs a maximization similar PLSC, namely, that the first pair of latent variables have maximum covariance evaluated just as in Equation 4, except under the constraints that **u** and **v** each have unit $\mathbf{W_X}$-norm and $\mathbf{W_Y}$-norm, respectively:

$$\mathbf{u}_l^{\mathrm{T}}\mathbf{W_X}\mathbf{u}_l = 1 = \mathbf{v}_l^{\mathrm{T}}\mathbf{W_Y}\mathbf{v}_l. \quad (15)$$

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 (as defined in Equation 6). The coefficients of the successive linear transformations (stored in matrices $\mathbf{L_X}$ and $\mathbf{L_Y}$) are obtained from the GSVD of $\mathbf{Z_R}$:

$$\begin{aligned}\mathbf{L_X}^{\mathrm{T}}\mathbf{L_Y} &= \mathbf{U}^{\mathrm{T}}\mathbf{W_X}\mathbf{Z_X}^{\mathrm{T}}\mathbf{Z_Y}\mathbf{W_Y}\mathbf{V} = \mathbf{U}^{\mathrm{T}}\mathbf{W_X}\mathbf{Z_R}\mathbf{W_Y}\mathbf{V} \\ &= \mathbf{U}^{\mathrm{T}}\mathbf{W_X}\mathbf{U}\boldsymbol{\Delta}\mathbf{V}^{\mathrm{T}}\mathbf{W_Y}\mathbf{V} = \boldsymbol{\Delta}. \quad (16)\end{aligned}$$

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,

$$\mathrm{diag}\left\{\mathbf{L_X}^{\mathrm{T}}\mathbf{L_Y}\right\} = \mathrm{diag}\{\boldsymbol{\Delta}\}. \quad (17)$$

Links to CA

PLSCA can be seen as a generalization of PLSC for two categorical data tables, but also as an extension of CA (Abdi & Williams, 2010a; Greenacre, 1984; Lebart et al., 1984). CA, 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

$$\mathbf{R} = \mathbf{X}^{\mathrm{T}}\mathbf{Y}. \quad (18)$$

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

$$\mathbf{O_R} = \mathbf{R} \times (\mathbf{1}^{\mathrm{T}}\mathbf{R}\mathbf{1})^{-1}, \quad (19)$$

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 Equation 8):

$$\mathbf{E_R} = \mathbf{m_X}\mathbf{m_Y}^{\mathrm{T}}. \quad (20)$$

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

$$\mathbf{Z_R} = \mathbf{O_R} - \mathbf{E_R}, \quad (21)$$

a formula which is equivalent to Equation 12, and thus $\mathbf{Z_R}$ can be decomposed according to Equation 13. In CA, the component scores for the rows and the columns of a matrix (the *J* and *K* elements of **R**) are computed as

$$\mathbf{F}_J = \mathbf{W_X}\mathbf{U}\boldsymbol{\Delta} \text{ and } \mathbf{F}_K = \mathbf{W_Y}\mathbf{V}\boldsymbol{\Delta}. \quad (22)$$

Like in CA and PCA, several additional indices can be computed from the component scores. These indices are called *contributions*, *direction 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; 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. Equation 19):

$$\mathbf{O_X} = \mathbf{X} \times (\mathbf{1}^\mathrm{T}\mathbf{X}\mathbf{1})^{-1} \text{ and } \mathbf{O_Y} = \mathbf{Y} \times (\mathbf{1}^\mathrm{T}\mathbf{Y}\mathbf{1})^{-1}, \quad (23)$$

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

$$\mathbf{F_X} = \mathbf{O_X}\mathbf{F}_J\mathbf{\Delta}^{-1} = \mathbf{O_X}\mathbf{W_X}\mathbf{U}\mathbf{\Delta}\mathbf{\Delta}^{-1} = \mathbf{O_X}\mathbf{W_X}\mathbf{U}, \quad (24)$$

$$\mathbf{F_Y} = \mathbf{O_Y}\mathbf{F}_K\mathbf{\Delta}^{-1} = \mathbf{O_Y}\mathbf{W_Y}\mathbf{V}\mathbf{\Delta}\mathbf{\Delta}^{-1} = \mathbf{O_Y}\mathbf{W_Y}\mathbf{V}. \quad (25)$$

Finally, we compute the latent variables—which are proportional to the supplementary projections obtained by rescaling the component scores (in Equations 24 and 25):

$$\mathbf{L_X} = \mathbf{F_X} \times I^{\frac{1}{2}} \text{ and } \mathbf{L_Y} = \mathbf{F_Y} \times I^{\frac{1}{2}}. \quad (26)$$

Equivalently, the latent variables could be directly computed as

$$\mathbf{L_X} = \mathbf{Z_X}\mathbf{F}_J\mathbf{\Delta}^{-1} = \mathbf{Z_X}\mathbf{W_X}\mathbf{U} \text{ and } \mathbf{L_Y} = \mathbf{Z_Y}\mathbf{F}_K\mathbf{\Delta}^{-1} = \mathbf{Z_Y}\mathbf{W_Y}\mathbf{V}. \quad (27)$$

Thus, in conclusion—as the name *partial least squares correspondence analysis* indicates—the computations and rationale of the technique can be interpreted either as a generalization of PLSC or an extension of CA.

Nonparametric Inference in PLSCA

PLSCA by itself is a multivariate descriptive (i.e., fixed-effect) technique. However, its results can be complemented by a variety of inference tests. These tests are computed from nonparametric resampling methods such as permutation and bootstrap resampling. Resampling methods generate a large number (e.g., thousands) of new data sets 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."

Permutation Tests

The permutation approach creates new samples by permuting the original data according to the null hypothesis to evaluate (Berry, Johnston, & Mielke, 2011). When all possible permutations are computed, this procedure creates an *exact test* for the null hypothesis (when only a random sample of permutations is used, the permutation test is asymptotically exact). In PLSCA, observations from **X** are reordered, whereas **Y** remains static. That is, a (random) permutation breaks the relationship between the two sets. Several tests can be performed with permutation resampling.

Omnibus. The original χ^2 value of the entire table (**R**) is tested against a generated set of χ^2 values generated from permutations. To note, the sum of all the cells the contingency table **R** multiplied by the sum of the eigenvalues (called the "inertia") from CA is equal to the χ^2 computed directly from the contingency table. As a result, an alternate test could simply use the sum of the eigenvalues. A rare value of the observed χ^2 or total inertia (i.e., it is among the α largest values) of the permuted samples indicates a significant *omnibus* effect.

Components. In SVD-based techniques, it is often difficult to identify the components to interpret (Raîche, Walls, Magis, Riopel, & Blais, 2013), especially because large data sets generate large numbers of components. Permutation resampling can also be used to identify the stable components from PLSCA. While there are numerous methods of testing components (Malinvaud, 1987; Peres-Neto, Jackson, & Somers, 2005; Saporta, 2011), we use a simple, but conservative, method (see "lambda-test" in Peres-Neto et al., 2005). PLSCA is computed, in full, on each permuted version of **R** (Equation 13) to generate distributions of singular values for each component.

Bootstrap

The bootstrap is a resampling *with replacement* technique (Chernick, 2008; Efron & Tibshirani, 1993). In PLSCA, observations are assumed to represent a population of interest. New samples are generated by resampling (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 to resample within *a priori* groups (e.g., Alzheimer's, mild cognitive impairment, control). The bootstrap is used to derive two different types of inferential statistics: bootstrap ratios (BSRs) and confidence intervals.

BSR tests. BSRs come from 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 $\alpha = .05$, that is, $P\ (|t| > 2) \approx .05$, 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, that is, $P\ (|t| > 3) \approx .0013$ or $\mathrm{P}(|t| > 4) \approx 3.17 \times 10^{-5}$).

Confidence intervals. Confidence intervals are created from percentile cutoffs 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., Alzheimer's group, mildly cognitive impaired group, control group). Confidence intervals can be displayed on component maps as peeled convex hulls (Greenacre, 2007) or as ellipsoids (Abdi, Dunlop, & Williams, 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).

Interpreting PLSCA and Its Extensions

Here, we illustrate PLSCA—and its extensions—with subsets of data from clinical, neuroimaging, and genome-wide data from the ADNI (Phase 1 data). First, we describe the ADNI project and the data used in this article. Next, we describe how PLSCA can be tailored for various experimental designs and research questions, as well as different data structures and even different data types.

We chose a very specific subset of the ADNI data to illustrate PLSCA for two reasons. First, the ADNI data set is available to qualified researchers, and thus what we present via PLSCA can be easily replicated. Second, the ADNI data are well studied and this allows us to highlight how PLSCA provides more insight from previous studies using the ADNI data.

ADNI and Data

Data used in the preparation of this article come from Phase 1 of the ADNI database (adni.loni.usc.edu). ADNI was launched in 2003 as a public–private funding partnership and includes public funding by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and the Food and Drug Administration. The primary goal of ADNI has been to test a wide variety of measures to assess the progression of mild cognitive impairment and early Alzheimer's disease. The ADNI project is the result of efforts of many coinvestigators from a broad range of academic institutions and private corporations. Michael W. Weiner (VA Medical Center, and University of California–San Francisco) is the ADNI Principal Investigator. Subjects have been recruited from over 50 sites across the United States and Canada (for up-to-date information, see www.adni-info.org).

For this article, we used several types of data from the ADNI project. First, we extracted the diagnostic criteria (i.e., clinical classification) of individuals: Alzheimer's Disease, mild cognitive impairment, or control. To note, we refer to our specific participant groups as the abbreviations AD (Alzheimer's disease), MCI (mild cognitive impairment), and CON (control). For example, "AD" refers to the Alzhiemer's disease group in this article, while "Alzheimer's disease" refers to the disease itself. We also extracted volumetric (mm^3) brain data (collected at baseline) from the following regions: ventricles, hippocampus, entorhinal, fusiform, and medial temporal. Volumetric brain data and diagnostic information were extracted from the R package ADNIMERGE.

Next, we extracted three behavioral measures that are widely used in Alzheimer's research and clinical assessment: the Mini-Mental State Exam (MMSE), the Clinical Dementia Rating (CDR), and the Geriatric Depression Scale (GDS). The MMSE is a brief measure designed to capture aspects such as memory, attention, and orientation (time, place). The CDR assesses behavioral aspects of dementia. The GDS is an assessment of depression for older populations.

Finally, we extracted a specific subset of candidate SNPs from the ADNI genome-wide association data. According to the literature, these SNPs are strongly linked to various behavioral, neurological, and physiological aspects of Alzheimer's disease and dementia. We refer to this approach of selecting SNPs *a priori* as a "candidate GWAS," as it can be seen as a compromise between genome-wide association and candidate gene studies. The general goal of our "candidate GWAS" is to determine whether SNPs (and genes) routinely associated with Alzheimer's disease are also associated with some of the diagnostic criteria used for Alzheimer's disease. All data and the ADNIMERGE package can be acquired from the ADNI project database.

Data Preprocessing

ADNI Phase 1 data (for this study) included 757 total subjects from three groups: AD ($n = 337$), MCI ($n = 210$), and CON ($n = 210$). Individuals were only used in these analyses if they had mostly complete data (i.e., $\geq$90%). Any missing datum (e.g., a SNP) was imputed to the mean of the entire sample. Behavioral, diagnostic, and genetic (i.e., SNP) data were recoded into disjunctive format (see Table 2). Volumetric brain data were transformed into Z scores before further processing (see Table 4).

The ADNI's genome-wide data set comprises 620,901 SNPs. ADNI's clinical data include separate genotyping for ApoE (see ADNI's clinical data). Because ApoE status is defined by two SNPs (rs429358 and rs7412), we can account for these SNP genotypes in the genome-wide data: Thus, we obtain a total of 620,903 SNPs. PLINK was used for genome-wide data preprocessing (Purcell et al., 2007). We used the following quality control (QC) criteria: Participant and SNP call rates (i.e., completeness of data) $\geq$90%, minor allele frequency $\geq$5%, Hardy-Weinberg equilibrium[3] $p \leq 10^{-6}$, in addition to gender and relation checks. SNP-based QC criteria were computed from the CON subjects. After QC, 756 participants (AD = 336, MCI = 210, CON = 210) and 531,339 SNPs remained.

From the genome-wide preprocessed data, we selected, for the studies of this article, 145 candidate SNPs because these SNPs—or their associated genes—are well-established candidates (e.g., ApoE) or have recently been implicated in Alzheimer's disease and related conditions. These SNPs have been reported in Hollingworth et al. (2011), Kauwe et al. (2008), Potkin et al. (2009), Shen et al. (2010), and Wijsman et al. (2011), or from an aggregate source: AlzGene (www.alzgene.org; Bertram, McQueen, Mullin, Blacker, & Tanzi, 2007). Two studies from this list (i.e., Potkin et al., 2009, and Shen et al., 2010) were also conducted on earlier versions of the ADNI data set.

Studies

We present five studies that illustrate PLSCA and its four different extensions. In general, these extensions map to some of the current variants of PLSC (Krishnan et al., 2011; McIntosh et al., 1996; McIntosh & Lobaugh, 2004), which often aim to simultaneously analyze neuroimaging and behavioral (or experimental design) related data.

We use "simple" PLSCA (as defined in PLS for Categorical Data) to illustrate the relationship between two categorical data sets: genetic (SNPs) and behavioral (MMSE, CDR, and GDS). Each extension of PLSCA is tailored to a particular problem of data type, data structure, or experimental design. The first extension solves the issue of data structured in blocks (i.e., MMSE, CDR, and GDS are separate "blocks" of variables within the set of behavioral data). Next, we define a discriminant version of PLSCA, called "mean-centered" PLSCA, which maximally distinguishes data based on a grouping

[3] This is a test of the Hardy-Weinberg principle, which states that allele frequencies follow a very particular probability distribution to be considered in "equilibrium." This statistical hypothesis is tested via χ^2.

Table 4
Schematic of Escofier-Style Transformation

	$\mathbf{x}_1$		$\mathbf{x}_2$	
Continuous data				
Subject 1	$x_{1,1}$		$x_{1,2}$	
Subject 2	$x_{2,1}$		$x_{2,2}$	
Subject *i*	$x_{\mathrm{i},1}$		$x_{\mathrm{i},2}$	
Subject *I*	$x_{\mathrm{I},1}$		$x_{\mathrm{I},2}$	
	$\mathbf{x}_1$		$\mathbf{x}_2$	
	−	+	−	+
Escofier-style transform				
Subject 1	$\frac{1-x_{1,1}}{2}$	$\frac{1+x_{1,1}}{2}$	$\frac{1-x_{1,2}}{2}$	$\frac{1+x_{1,2}}{2}$
Subject 2	$\frac{1-x_{2,1}}{2}$	$\frac{1+x_{2,1}}{2}$	$\frac{1-x_{2,2}}{2}$	$\frac{1+x_{2,2}}{2}$
Subject *i*	$\frac{1-x_{i,1}}{2}$	$\frac{1+x_{i,1}}{2}$	$\frac{1-x_{i,2}}{2}$	$\frac{1+x_{i,2}}{2}$
Subject *I*	$\frac{1-x_{I,1}}{2}$	$\frac{1+x_{I,1}}{2}$	$\frac{1-x_{I,2}}{2}$	$\frac{1+x_{I,2}}{2}$

Note. Example of the Escofier-style coding of continuous data. This transform is used to perform principal components analysis of continuous data via correspondence analysis. $\mathbf{x}_j$ denotes the *j*th vector from a Matrix **X**, where $x_{i,j}$ denotes a specific values at row *i* and column *j*.

variable (i.e., the three clinical groups). Similarly, we define “seed” PLSCA as a method to specifically to maximize the interrelationships *within* a particular set of categorical data—for example, we want to identify novel candidate SNPs from well-established candidate markers such as ApoE). Finally, we define a particularly useful version of PLSCA: “Mixed-data” PLSCA, which can analyze the relationship between one categorical data set (i.e., SNPs) and one quantitative (continuous, interval scale) data set (i.e., volumetric data from brain imaging).

Simple PLSCA

Simple PLSCA was used to analyze the relationship between the behavioral data set and the SNP data set. A permutation test indicates that the omnibus effect was significant ($\chi^2_{\text{omnibus}} = 23150.41$, $p_{\text{perm}} < .001$). For the components, only Components 1 and 3 (out of 68) reached significance (30.85%, $p_{\text{perm}} < .001$, and 6.25%, $p_{\text{perm}} = .017$, respectively), and were therefore kept for the analysis. Components 1 and 3 are presented in PLSC style (with normalized saliences as bar plots) in Figure 1 and CA style (with normalized saliences on component maps) in Figure 2. As a noteworthy feature, PLSCA provides information for *each level* of all variables. This can be seen with the contrast of, for example, “correct” versus “incorrect” results on the MMSE (see Figures 2a and 2b; horizontal axis). Likewise, the SNPs are coded in the same fashion, and provide information *per genotype*.

Component 1. With PLSCA, it is often easier to begin interpretation with the behavioral results, as they provide context for the genetics results. Component 1 (Figure 1a, Figure 2a, and Figure 2b, horizontal axis) shows a contrast for the MMSE and the CDR. On the left side, component scores for MMSE items are associated with “incorrect” responses and CDR items range from “moderate” to “severe” (dementia); therefore, any genetic markers on the left side (Figure 1c, Figure 2c, and Figure 2d) are more associated with deficits on the MMSE and dementia than any other behaviors.

The left side of Component 1 includes two very notable markers: the heterozygote and minor homozygote of rs429358 (a SNP used to assign the genotype for ApoE). We also see the minor homozygotes of rs7910977 and rs7526034, which have been associated, respectively, with β-amyloid degradation (Kauwe et al., 2008) and structural changes in brain regions in Alzheimer’s disease (Shen et al., 2010).

Additionally, PLSCA shows that the heterozygote of rs2075650 and the *major homozygote* of rs157580 are also on the left side of Component 1, and therefore are also associated with dementia and deficits on the MMSE. Both rs2075650 and rs157580 are associated with the TOMM40 gene—a gene strongly linked to ApoE—and have been associated with Alzheimer’s disease in previous studies (Potkin et al., 2009; Roses et al., 2010; Shen et al., 2010). However, this conclusion runs counter to expectations because it suggests that the genetic contributions to pathological behaviors are not strictly because of standard “risk” (i.e., minor) alleles. Such a conclusion could be missed in a standard analysis using “additive coding” (see Table 2), but becomes explicit with a disjunctive coding scheme.

In contrast to the left side of Component 1, the right side of Component 1 is associated with “healthy” responses on the MMSE and CDR. This level of detail is advantageous because we can identify possibly protective genetic markers. Some of the alleles with the strongest effects are the heterozygote of rs157580 and the major homozygote of rs2075650 (TOMM40), as well as the major homozygote of rs429358 (ApoE). We can conclude that Component 1 is a “pathological versus healthy” dimension, and thus we can associate specific genotypes of SNPs (as opposed to entire SNPs) with pathological or healthy features.

Component 3. In general, the GDS items contribute a substantial amount of variance to the top side of Component 3 (see

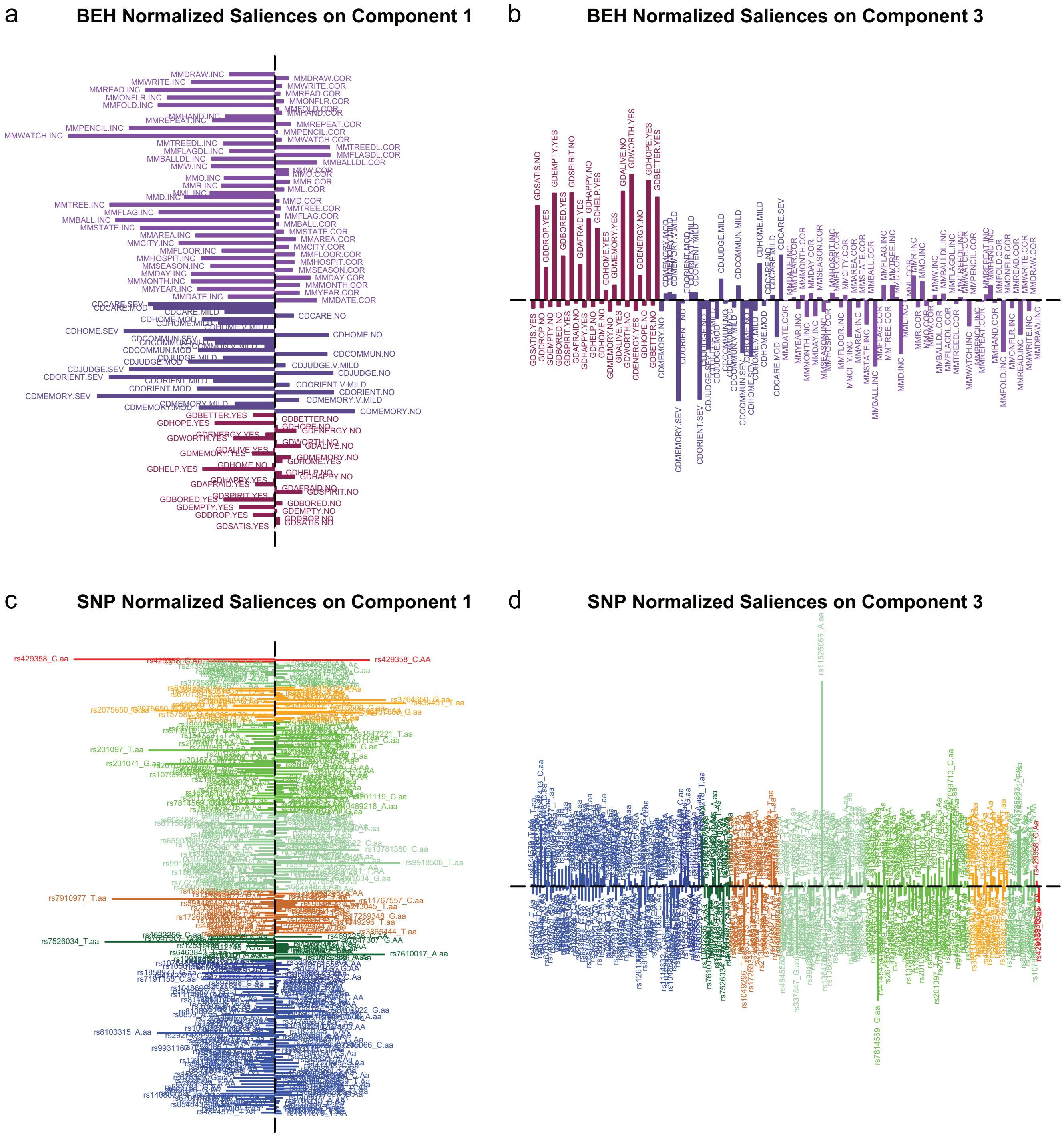

Figure 1. Data presented as saliences, per component, as in partial least squares correlation. Top (a, b) figures show the behavioral data on Components 1 and 3. Bottom (c, d) show the SNPs data on Components 1 and 3. The longer the bar associated with an item, the more variance the item contributes to the respective component. SNP = single nucleotide polymorphism; BEH = behavioral data. See the online article for the color version of this figure.

Figure 1b and Figure 2a and b). All of these items are associated with geriatric depression (e.g., Q: "Are you basically satisfied with your life?" A: "No"; Q: "Do you feel that your life is empty?" A: "Yes"). Therefore, Component 3 can be described as a "depression" dimension. This component, therefore, suggests that some genes may contribute to geriatric depression and that this effect can be seen in early stages of cognitive decline (e.g., in MCI). For example, the pattern displayed by the SNPs

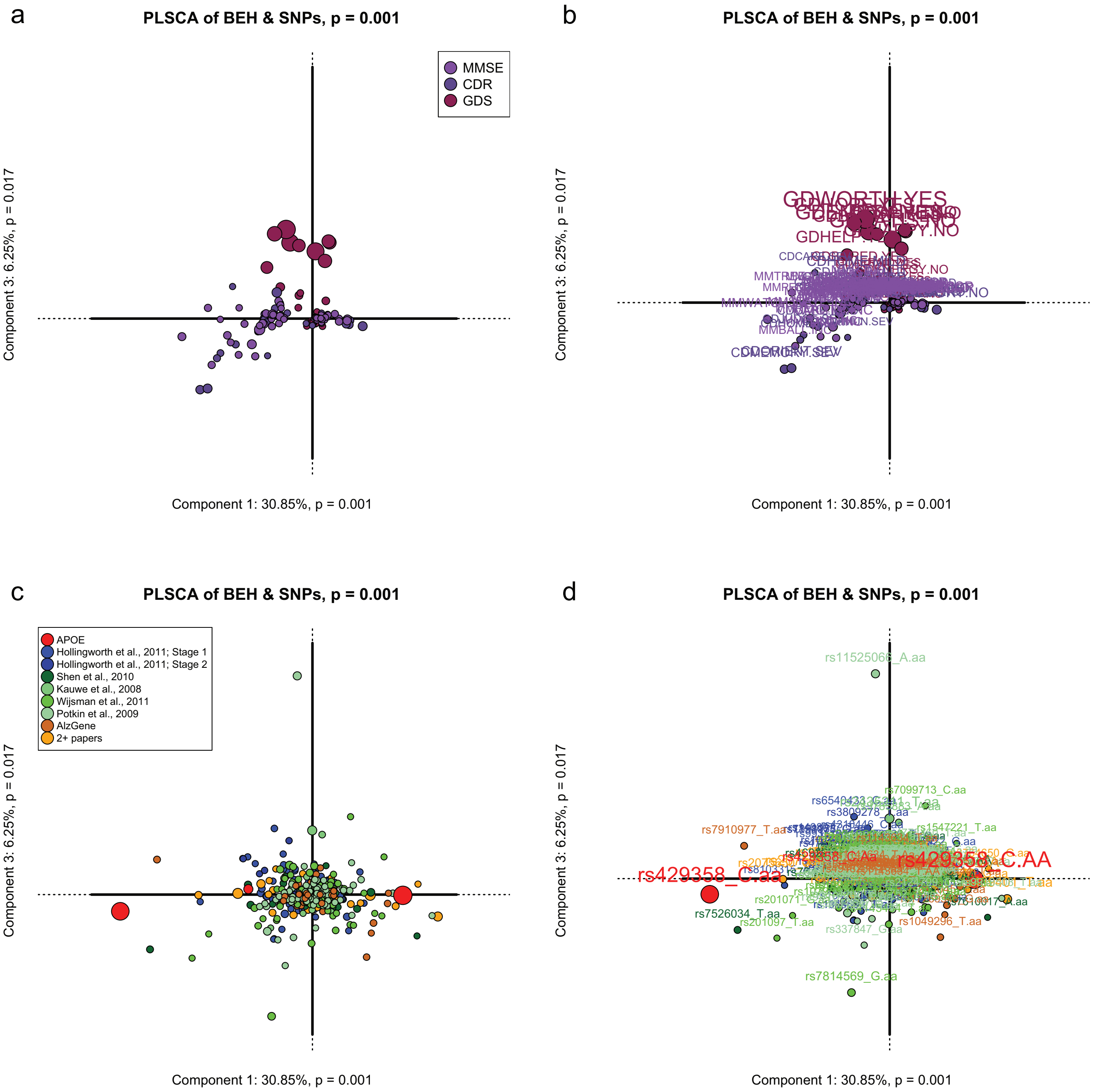

Figure 2. Component maps presented as in correspondence analysis. Top and bottom rows are the same figure without and with labels (for the items), respectively. Top (a, b): Behavioral measures for Components 1 and 3. Bottom (c, d): SNPs data on Components 1 and 3. Top-row maps have no labels on the behavioral and SNP markers. Instead, a legend is provided to indicate which measures they are (BEH = behavioral; a, b) or the selected source material (SNPs = genetic data; c, d). SNPs labeled as "2+ papers" are referred to in at least two of the sources cited for the selected SNPs. Bottom-row maps have labels for each item in the analysis. The further from the origin an item is, the more variance it contributes to the visualized components. PLSCA = partial least squares correspondence analysis; SNP = single nucleotide polymorphism; BEH = behavioral data; MMSE = Mini-mental state exam; CDR = Clinical dementia rating; GDS = geriatric depression scale. See the online article for the color version of this figure.

(Figure 1d, Figure 2b, and Figure 2d) reveals a strong effect of the minor allele of rs11525066, which has been previously associated with hippocampal atrophy (Potkin et al., 2009).

Bootstrap. BSRs can be used to identify important (e.g., "significant") variables, and can be graphically represented on component maps (see Figure 3) or as bar plots (see Figure 4). Many of the significant behavioral variables for Component 1 correspond to either incorrect responses on the MMSE, or "moderate" to "severe" responses on the CDR (Figure 4a). The complete list of SNPs associated with either pathological (left side of Component 1) or beneficial (right side of Component 1) behaviors can be found in Supplemental Materials (Table S1). Component 3 seems to oppose depression with mild impairment (top of Component 3) to depression with severe impairment (bottom of Component 3). The complete list of genetic markers associated with depression versus impairment can be found in Table S1 of the online supplemental materials.

Latent variables and participants. In PLSCA, just as in PLSC, observations can be represented by the values of their latent variables (see Equations 14, 26, and 27), and this provides maps in which the coordinates for one observation are its values of the latent variables for **X** and **Y** (see Figure 5 for PLSC style, and Figure 6 for CA style). These maps can be interpreted as scatterplots and the dispersion of the observations reflects their relationship (a perfect relationship will show the observations positioned on a line). Groups of observations can also be represented in these maps by averaging the component scores of the observations in a group and bootstrap resampling can provide confidence intervals for these groups (see Figure 5b and d). Figure 5 indicates a clear distinction of individuals within groups for behavior versus genetics on Component 1. Furthermore, the groups are significantly different from one another because their confidence intervals do not overlap (Figure 5b). Component 1 reflects a continuum for the groups: from CON to MCI to AD. These results indicate that there are specific behavioral and genetic signatures for each group. Component 3 (see Figure 5 and Figure 6a) appears to show individual (as opposed to group) effects of geriatric depression.

Multiblock PLSCA

Multiblock PLSCA is a simple extension of PLSCA that includes multiblock projections (Abdi, Williams, Beaton, et al., 2012; Abdi, Williams & Valentin, 2013; Krishnan et al., 2011). Often, the variables of these data sets are structured into related, but distinct, blocks (also called tables or subtables). In our example, the MMSE, CDR, and GDS are all distinct blocks of variables within the behavioral data set. It is worth noting that genome-wide data have several possible block structures (e.g., genes, haplotypes, and chromosomes). Multiblock PLSCA can be used to answer the following question: How does each behavioral measure *uniquely* contribute to the results?

The goal of multiblock PLSCA is to provide both an overall and a "block" point of view of the participants' latent variable scores. In general, the data are structured as

$$\mathbf{Z}_{\mathbf{X}} = [\mathbf{Z}_{\mathbf{X}_1}, \ldots, \mathbf{Z}_{\mathbf{X}_b}, \ldots, \mathbf{Z}_{\mathbf{X}_B}] \tag{28}$$

where the columns of **X** (and subsequently $\mathbf{Z}_{\mathbf{X}}$) are partitioned into blocks. This block structure propagates to the rows of the matrix $\mathbf{Z}_{\mathbf{R}}$ and to its GSVD (cf. Equation 13):

$$\mathbf{Z}_{\mathbf{R}} = [\mathbf{U}_1, \ldots, \mathbf{U}_b, \ldots, \mathbf{U}_B]\boldsymbol{\Delta}\mathbf{V}^{\mathrm{T}}. \tag{29}$$

The partition of **U** into blocks also propagates to the latent variables of **X** (see Equation 14):

$$\mathbf{L}_{\mathbf{X}_b} = \mathbf{Z}_{\mathbf{X}_b}\mathbf{W}_{\mathbf{X}_b}\mathbf{U}_b. \tag{30}$$

Multiblock analyses were conducted within the same analysis as "simple" PLSCA. Figure 7 shows the participants' latent variable scores, separately, for the MMSE (Figure 7a), the CDR (Figure 7b), and the GDS (Figure 7c). The multiblock projection shows that the CDR and MMSE—as indicated before—constitute strong sources of variance for Component 1 (Figure 7a and b). Finally, with the multiblock projections, the interpretation of Component 3 becomes clearer. The upper part of Component 3 is associated with strong features of depression (Figure 7c), whereas the lower part is associated with severe dementia (Figure 7b).

Mean-Centered PLSCA

Mean-centered PLSCA is a simple form of PLSCA also known as barycentric discriminant correspondence analysis (DiCA; Abdi, 2007a; William, Abdi, French, & Orange, 2010). The goal of mean-centered PLSCA to maximize the separation between a set of a priori defined groups. Alternatively, this can be thought of as finding the best set of variables that can be used to assign observations to a priori defined groups. In mean-centered PLSCA, a disjunctive design matrix for the a priori groups is **X** and the disjunctive data matrix (e.g., SNPs) is **Y**. Mean-centered PLSCA maximizes the separation between groups, and thus detects which SNPs patterns (because of participant groups) can be used to assign individual participants to clinical groups (i.e., CON, MCI, and AD).

In mean-centered PLSCA, the omnibus test is significant (χ^2 = 785.60, $p_{\text{perm}} < .001$). Because we have three groups, we have only two components: Component 1 explains 67% of the variance and is significant ($p_{\text{perm}} < .001$), whereas Component 2 explains 33% of the variance but is not significant ($p_{\text{perm}} = .521$). Figure 8a shows the groups (large dots) and individuals (small dots), and suggests a separation of all groups. Figure 8b shows the bootstrap confidence intervals around the groups. None of the intervals overlap; thus, all groups significantly differ from each other.

The genotypes closest to CON are more associated with this group than with any of the other groups. In Figure 8c, only significant SNPs (according to their BSR) are colored. In the lower part of the map in Figure 8c, we see the minor homozygote of rs12610605, which has been previously linked to Alzheimer's disease (Hollingworth et al., 2011). However, mean-centered PLSCA shows that the minor homozygote of rs12610605 is much more associated with the MCI group than with any other group. Furthermore, we highlight a particular effect in Figure 8d, with two TOMM40 SNPs (rs157580, rs2075650) and one ApoE SNP (rs429358). Both the rs2075650 and rs429358 show a similar structure: an effect from left (major homozygote) to right (minor homozygote) that is somewhere between a linear and a multiplicative effect on Component 1. However, rs157580 appears to show a different inheritance pattern—a dominance

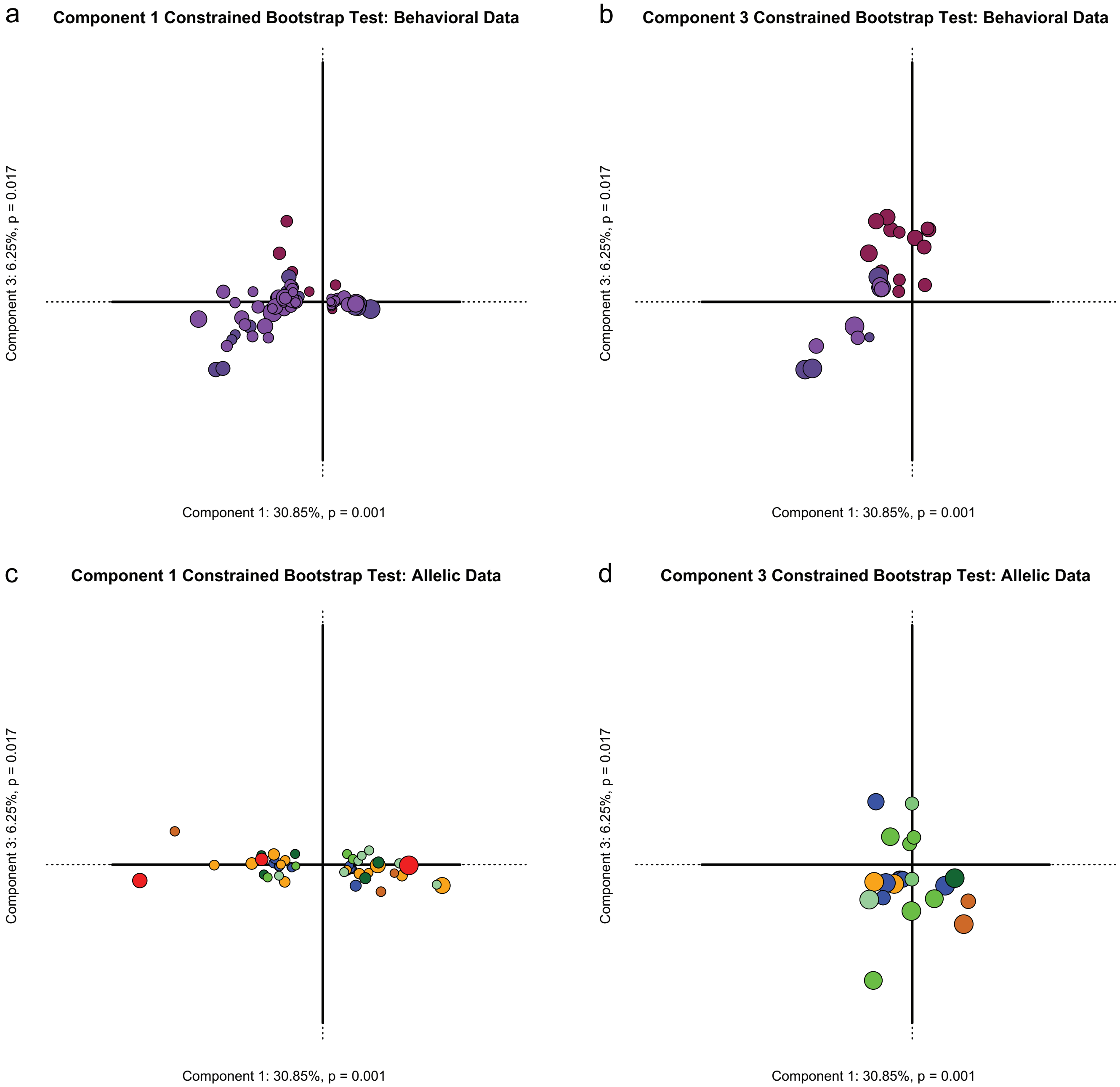

Figure 3. Only items with a bootstrap ratio with a magnitude greater than 2 are plotted. Behavioral markers are displayed in the top figures (a, b), and genetic data are displayed in the bottom figures (c, d). Items close to each other are highly correlated; items on the opposing sides of the map are negatively correlated. See the online article for the color version of this figure.

effect—and *in the opposite direction* of both rs2075650 and rs429358: The minor homozygote and heterozygote of rs157580 are positioned closely to the major homozygotes of rs2075650 and rs429358 on Component 1. This finding highlights how PLSCA—with categorical coding—*detects* a variety of effects (without the need to explicitly model each of them). The effects observed on Component 1 could suggest a complex haplotype of the ApoE and TOMM40 SNPs that better describes AD risk than any single genotype.

Thus, mean-centered PLSCA indicates that (a) the genotypes on the right side of Component 1 are more associated with AD than with any other group, (b) the genotypes on the bottom of Component 2 are more associated with MCI than with any other group, and arguably the most important aspect, (c) the genotypes on the left of Component 1 are more associated with the CON group than with any other group. This level of detail can only be obtained from PLSCA and disjunctive coding. A complete set of significant SNP markers—for mean-centered

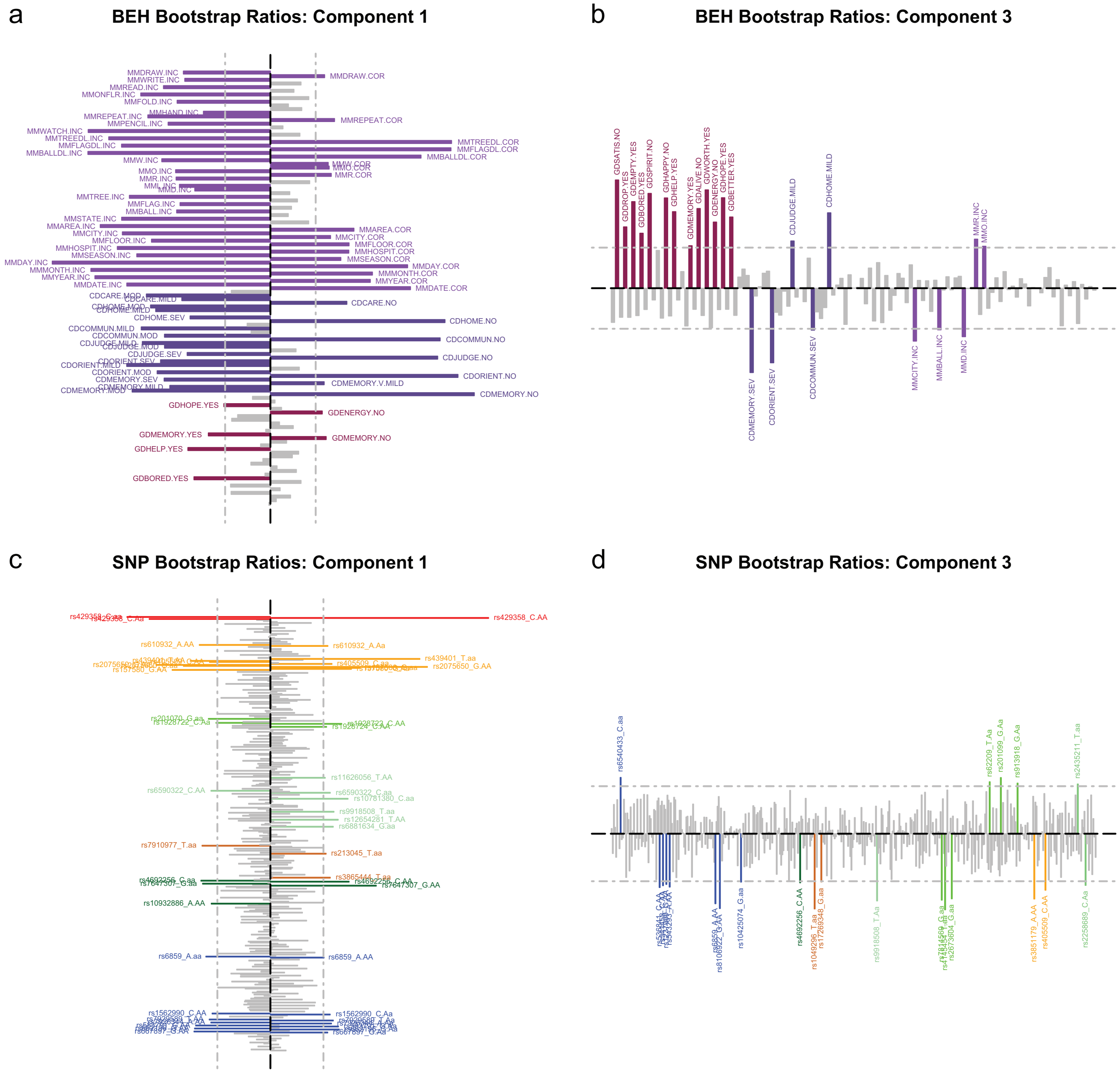

Figure 4. Top row (a, b): Behavioral markers are plotted on the bar graph as a function of their bootstrap ratio (BSR) value. Bottom row (c, d): SNPs are plotted on the bar graph as a function of their BSR value. In all figures, the longer the bar of an item, the bigger the (absolute value) BSR of this item. The dashed gray lines indicate the threshold for significance (i.e., −2 and +2). Gray items did not reach the level for significance. SNP = single nucleotide polymorphism; BEH = behavioral data. See the online article for the color version of this figure.

PLSCA results—is provided in Table S2 of the online supplemental materials.

Seed PLSCA

The term *seed partial least squares* is commonly used in functional brain connectivity analyses (McIntosh, Nyberg, Bookstein, & Tulving, 1997). In a "seed" analysis, there is only one data set, but a specific portion of that data set is extracted and treated as a second data set (i.e., a seed). In this case, the SNPs are still referred to as **Y** and can also be thought of as having a block structure (as in Equation 28):

$$\mathbf{Y} = [\mathbf{Y}_1, \ldots, \mathbf{Y}_s, \ldots, \mathbf{Y}_S] \tag{31}$$

where $\mathbf{Y}_s$ denotes one SNP in disjunctive coding (see Table 2).

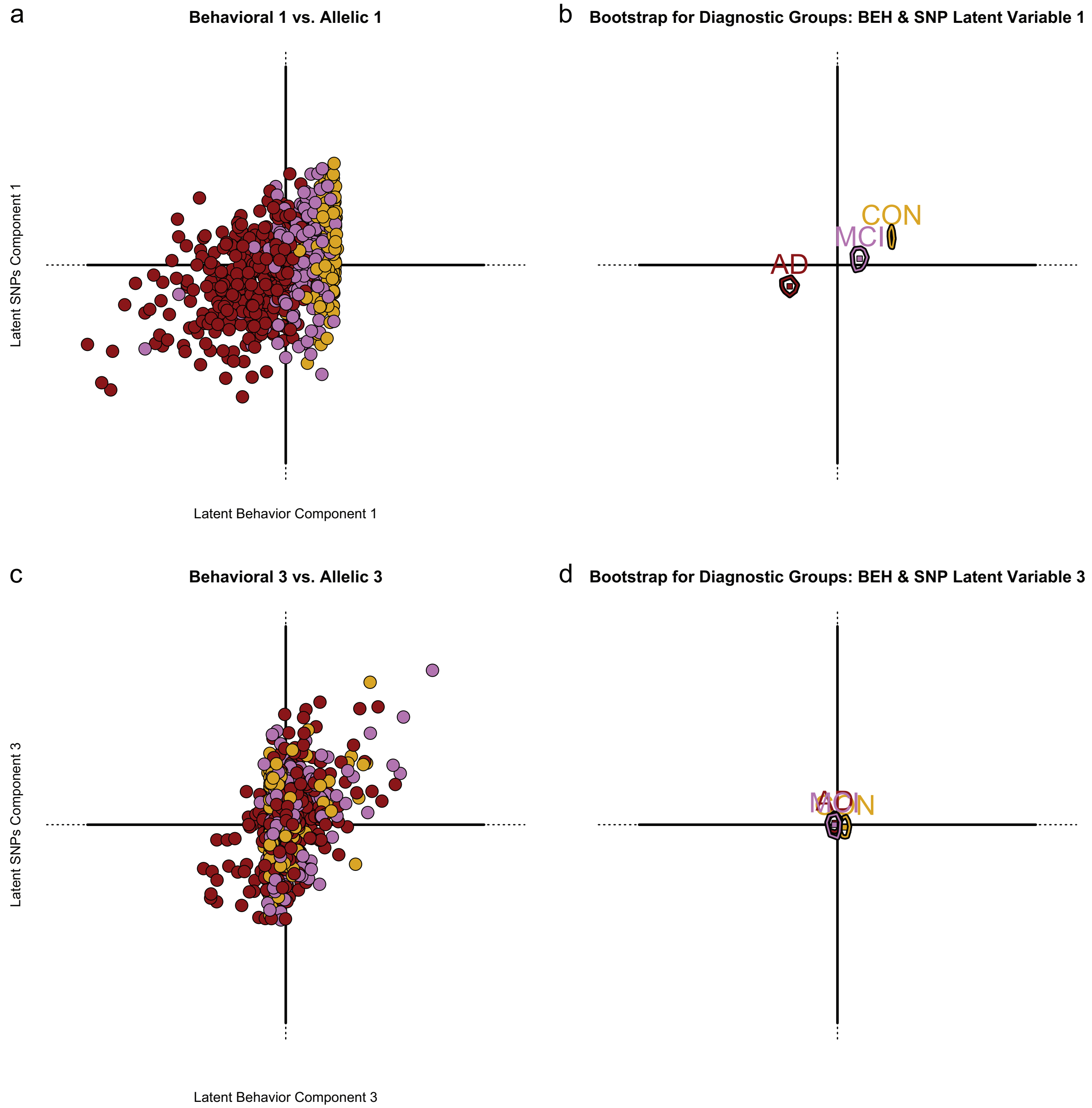

Figure 5. Partial least squares correlation (PLSC)-style latent variables. In PLSC, it is common to project the latent variables of one data set (e.g., behavior) against the latent variables of the other set (e.g., SNPs). The top (a, b) show the latent vectors for behavioral and SNP data for Component 1. The bottom (c, d) show the latent vectors for behavioral and SNP data for Component 3. SNP = single nucleotide polymorphism; BEH = behavioral data; AD = Alzheimer's Disease group; MCI = mild cognitive impairment group; CON = control group. See the online article for the color version of this figure.

From $\mathbf{Y}$, a particular set of SNPs are selected as "seeds" and removed from that matrix. The remaining SNPs are left in $\mathbf{Y}$. If, for example, the first two SNPs are "seeds," $\mathbf{X}$ and $\mathbf{Y}$ become

$$\mathbf{X} = [\mathbf{Y}_1, \mathbf{Y}_2] \qquad (32)$$

$$\mathbf{Y} = [\mathbf{Y}_3, \mathbf{Y}_4, \mathbf{Y}_5, \ldots, \mathbf{Y}_s, \ldots, \mathbf{Y}_S]. \qquad (33)$$

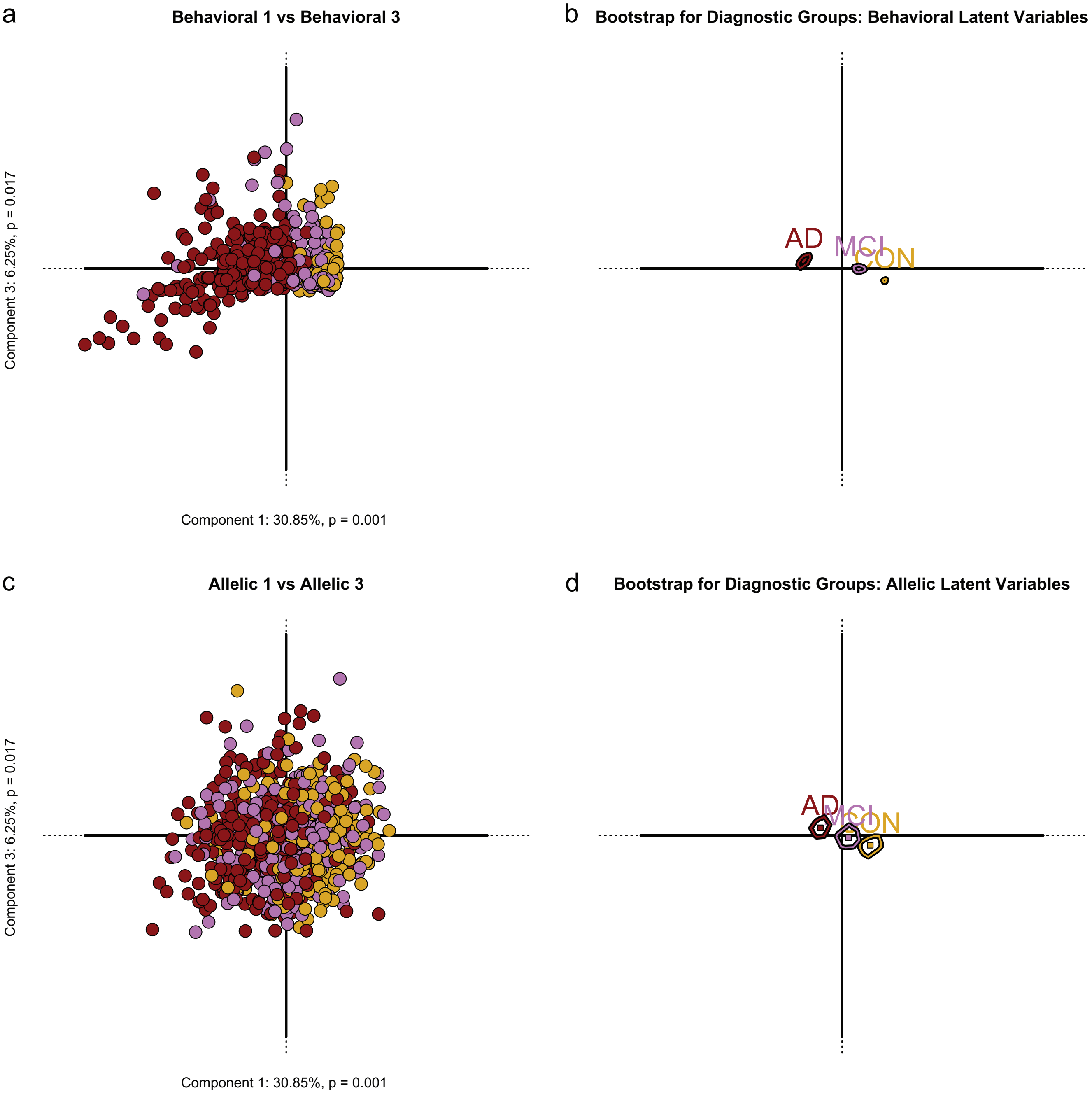

Figure 6. Plots of latent variables in the style of correspondence analysis (CA). In CA, it is more common to represent the latent variables of one data set (e.g., behavior) together. AD = Alzheimer's disease group; MCI = mild cognitive impairment group; CON = control group. See the online article for the color version of this figure.

The seed approach reveals similarity (and dissimilarity) of genotypes across SNPs. Seed PLSCA is a useful approach to find novel candidate markers. Seed PLSCA can be used to reveal which genotypes are most similar to "candidate" genotypes (e.g., SNP or gene alleles). This example uses rs429358 (APOE) and rs2075650 (TOMM40) as "seeds."

Seed PLSCA benefits from using an "asymmetric" visualization, in which the component scores of the "seed set" (i.e., **X**) are (cf. Equation 22) computed as

$$\mathbf{F}_{\mathbf{X}} = \mathbf{W}_{\mathbf{X}}\mathbf{U}, \tag{34}$$

because the weighted saliences define a simplex, which gives

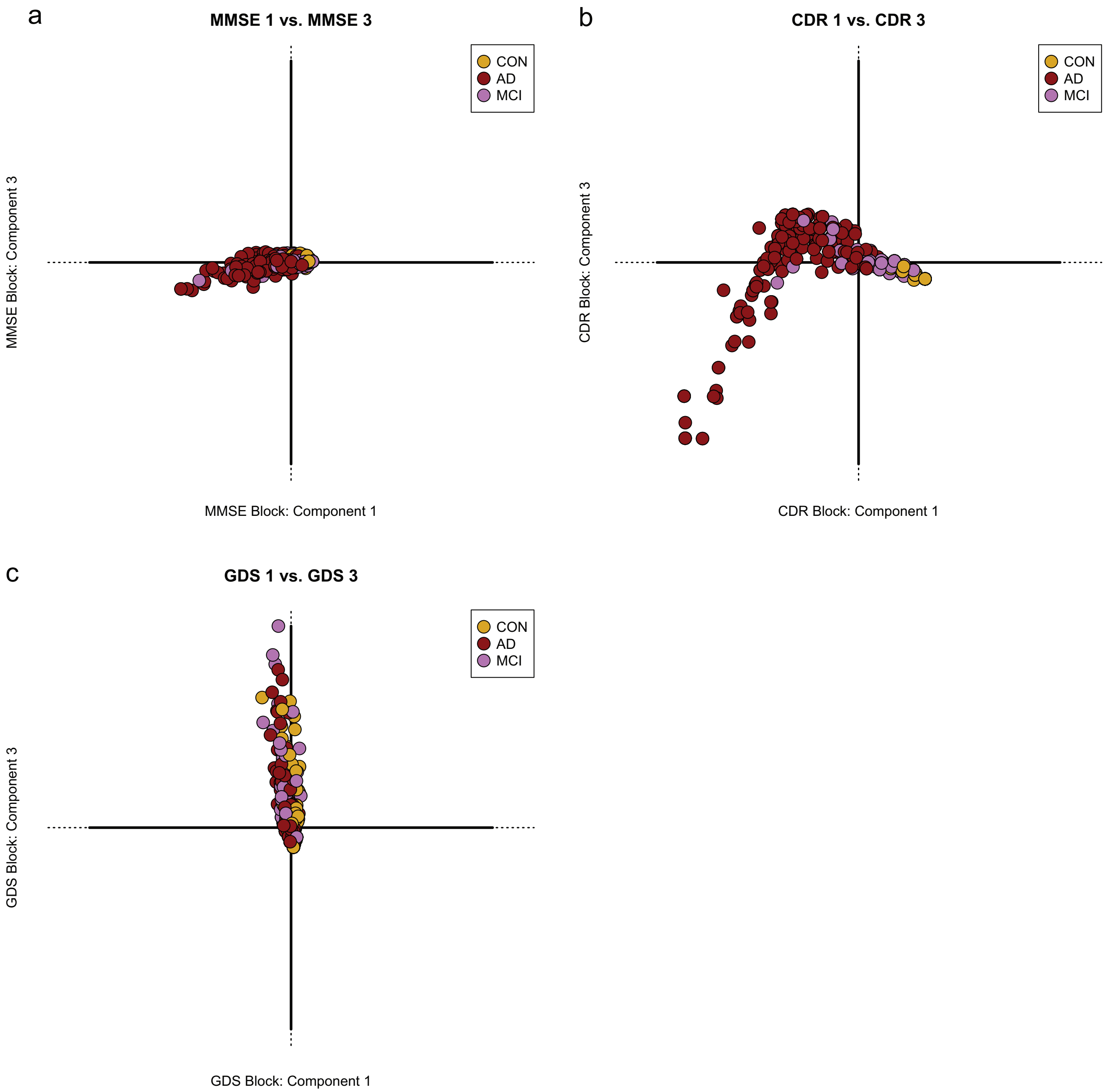

Figure 7. Multiblock projection of each of the behaviors from partial least squares correspondence analysis: Mini-Mental State Exam (MMSE) (a), Clinical Dementia Rating (CDR) (b), and Geriatric Depression Scale (GDS) (c). AD = Alzheimer's disease group; MCI = mild cognitive impairment group; CON = control group. See the online article for the color version of this figure.

the boundary of the component space (see Abdi & Béra, 2014, Abdi & Williams, 2010b, and Greenacre, 1984, 2007, for more details). Asymmetric visualization could be used for any form of PLSCA in which one data set is "privileged" over the other. In seed PLSC, we use the asymmetric visualization for the seed SNPs. Thus, in seed PLSCA genotypes across sets that are close to one another co-occur frequently, genotypes that are in the exact same position co-occur precisely with one another, and genotypes that are far apart co-occur infrequently or in opposition.

Figure 9a shows the plot of the first two components of seed PLSCA. Components 1 and 2 explain, respectively, 72% and 20% of the total variance. Note that the heterozygotes (upper left of Figure 9a) and minor homozygotes (lower left of Figure

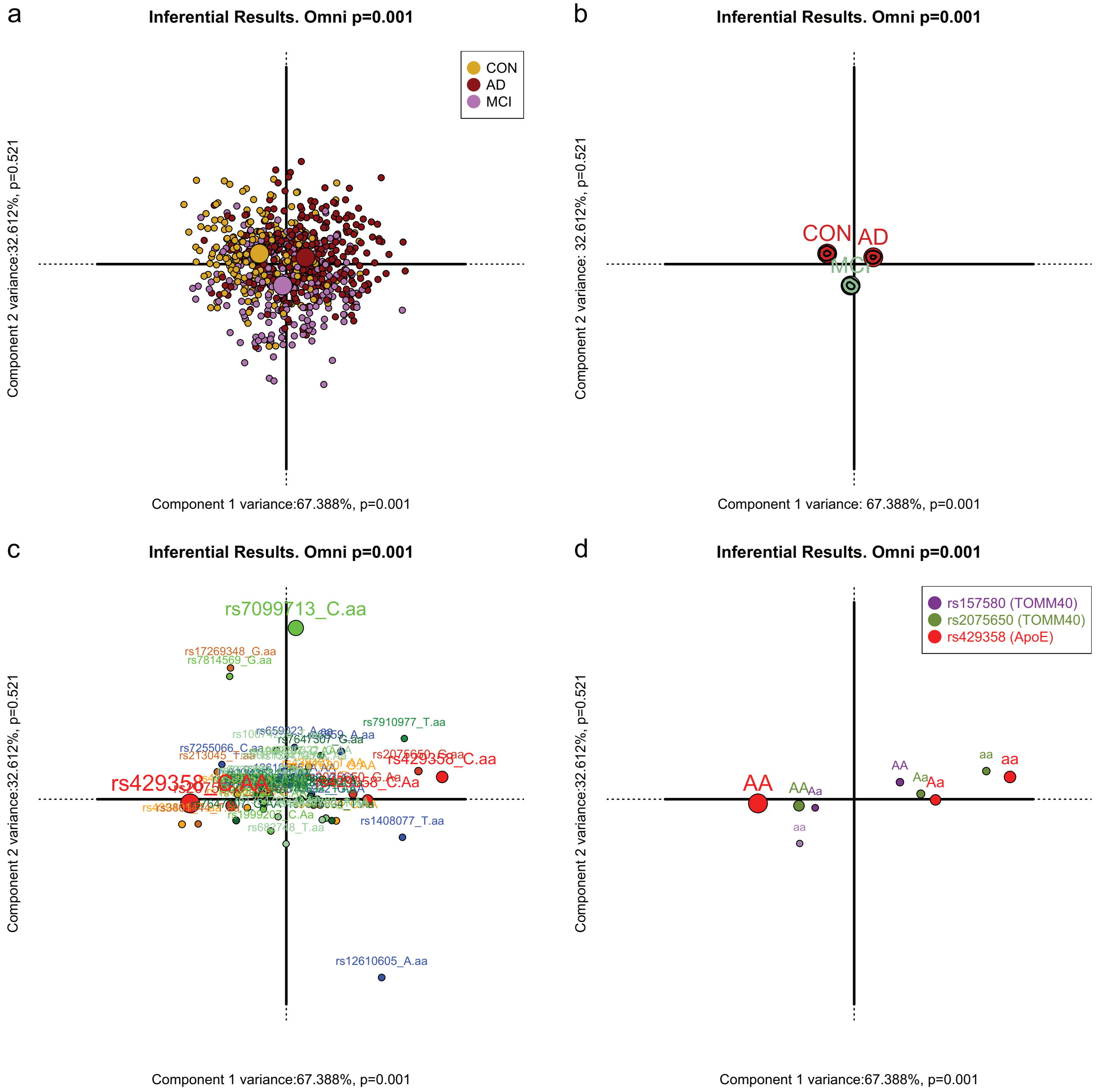

Figure 8. (a) Individuals and groups. (b) Groups and 95% confidence intervals. Because the confidence intervals do not overlap, all groups significantly differ from one another. (c) Component scores for the genotypes; only significant genotypes are shown. (d) Genotypes (in generic AA, Aa, aa format) of one ApoE SNP (rs429358) and two TOMM40 SNPs (rs157580, rs2075650). The distribution of SNP rs157580 is opposite to the distribution of the other two SNPs. SNP = single nucleotide polymorphism; AD = Alzheimer's disease group; MCI = mild cognitive impairment group; CON = control group. See the online article for the color version of this figure.

9a) of rs429358 and rs2075650 are distant from nearly all other genotypes. That is, the minor allele-based markers of the "seeds" are not close to any other genotypes. Furthermore, the major homozygotes of rs429358 and rs2075650 are very close to other genotypes (lower right of Figure 9a). This indicates that major homozygotes of these SNPs co-occur very frequently, or exactly, with many other major homozygotes. Furthermore, Figure 9b shows the latent variables, via the distribution of participants with respect to their group. This figure shows that AD is more associated with the minor alleles than the other

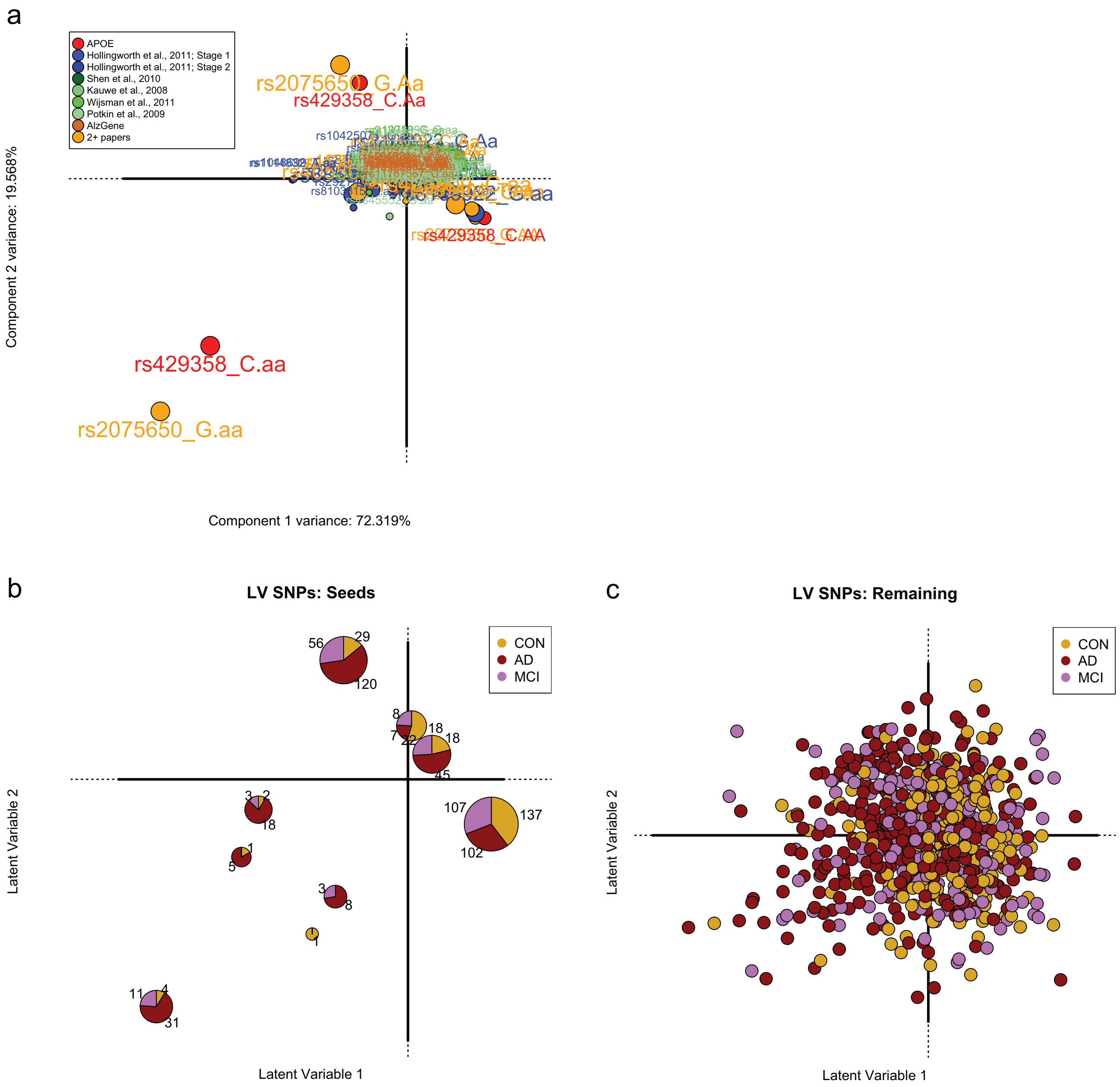

Figure 9. Top left image (a) shows component scores (from seed partial least squares correspondence analysis, with APOE and TOMM40 as seeds). The configuration of APOE and TOMM40 shows that they are relatively unique, because of the unique component scores for their heterozygotes and minor homozygotes on the component map. Bottom images (b, c) show the observations expressed as their latent variables scores. Bottom left (b) are the latent variables with respect to the "seeds." Because there are few possible combinations of these two markers, there are only a few possible scores. The number of observations, per group, are expressed with pie charts at each unique latent variable score. Latent scores for the seed set appear to show some effect of group, in which the Alzheimer's Disease (AD) group is most associated with the minor allele. Bottom right (c) shows latent variable (LV) scores for the nonseed set. No discernible pattern exists for the individuals. MCI = mild cognitive impairment group; CON = control group. See the online article for the color version of this figure.

groups, whereas Figure 9c shows no real discernible pattern to separate individuals.

The "seed" analysis suggests that although the major homozygotes are highly similar to other major homozygotes, the heterozygotes and minor homozygotes of rs429358 and rs2075650 express a relatively unique pattern. This uniqueness suggests that there is no novel candidate from this set of SNPs to match the ApoE and TOMM40 minor homozygotes and heterozygotes.

Table 5
Example of Escofier-Style Transformation

	Hippocampus	Ventricles
(a) Example of continuous data		
Subject 1	4,581	40,559
Subject 2	7,090	26,125
Subject *i*	5,732	57,383
Subject *I*	7,463	27,759
(b) *Z* scores of example		
Subject 1	−1.239	.180
Subject 2	.662	−.818
Subject *i*	−.367	1.343
Subject *I*	.844	−.705

	Hippocampus		Ventricles	
	−	+	−	+
(c) Escofier-style of example				
Subject 1	1.1195	−.1195	.4100	.5900
Subject 2	.1690	.8310	.9090	.0910
Subject *i*	.6835	.3165	−.1715	1.1715
Subject *I*	.0280	.9720	.8525	.1475

Note. Example of "Escofier-style" coding with actual values. Here, (a) shows the volumetric data (in mm^3) of each region, where (b) shows the *Z*-score transformation of the data in (a). Part (c) shows the Escofier-style transform of the "bipolar" values (above and below a mean) in (b).

PLSCA With Heterogeneous Data

Though we present PLSCA as a method for two categorical tables, PLSCA can be generalized to include continuous variables, because with a simple preprocessing (i.e., coding) approach, PLSCA can easily analyze mixtures of categorical and continuous variables. Specifically, in 1979, Brigitte Escofier proposed such an approach for analyzing mixtures of data with CA (Escofier, 1979), and showed that a quantitative variable, noted **x** (i.e., a column from **X**) that is centered and scaled (i.e., it has mean of zero and norm one) can be used with CA if it is re-expressed as two variables computed as $\frac{1-\mathbf{x}}{2}$ and $\frac{1+\mathbf{x}}{2}$. Applying CA (following Equations 19 to 22) to an "Escofier-transformed" matrix (of a continuous data set; see Table 5) produces results identical to a PCA of **X** because this table is analogous to a complete disjunctive table (see Table 2).[4]

As in the previous examples, mixed-data PLSCA provides component scores for each set of variables, and latent variable scores for participants. However, one noticeable difference is that the continuous data set component scores—here, volumetric data of brain regions—are "duplicated" (also called "bipolar," because each variable is represented by two poles; see Greenacre, 1984, 2014), and are denoted with "+" and "–" (see Figure 10).

In this analysis, there was a significant omnibus effect ($p_{\text{perm}} <$ .001), but only Component 1 was significant ($p_{\text{perm}} <$ .001). Figure 10 shows the component maps for both the genotypes (Figure 10a and b) and the brain regions (Figure 10c and d). Significant genotypes are presented in Table S3 of the online supplemental materials. In Figure 10, we can see that there are particular genotypes associated with particular patterns of brain region volumes. More specifically, Figures 10b and 10d suggest that certain genotypes are associated with brain region volumes above the grand mean (right side of Figure 10b, and Figure 10d) but that other genotypes are associated with brain region values below the grand mean (left side of Figure 10b, and Figure 10d).

In particular, Figure 11 shows that brain region volumes above the grand mean (Figure 11a) are more associated with the CON group (Figure 11b) than with other groups. Further, we see that numerous major homozygotes—such as rs429358 (ApoE), rs7647307, and rs2075650—are more associated with brain region volumes above the grand mean (Figure 11c). Interestingly, however, some heterozygotes and minor homozygotes are also associated with brain region volumes above the grand mean. These include rs439401, rs667897, and rs337847. To note, although these SNPs have been previously implicated in Alzheimer's disease, this implication has been almost exclusively via the assumptions of an "additive" model in which, typically, the minor alleles are assumed to be associated with risk factors. For example, rs337847 has been associated with hippocampal atrophy (Potkin et al., 2009). However, we show that the minor homozygote of rs337847 is more related to brain region volume above the grand mean than below the grand mean, a pattern that suggests the *major* allele of rs337847 is likely to be the *risk factor*.

Discussion

PLSCA was designed to address several challenges (e.g., the categorical nature of SNPs, rare occurrences, multivariate analysis) while simultaneously analyzing genetic data and data traditionally associated with the psychological, cognitive, and neuro-

[4] To note, if there are two continuous data tables, **X** and **Y**, transformed via Escofier's approach, and PLSCA is applied to the cross product of the transformed **X** and **Y**, the results are identical (within a scaling factor) to standard PLSC.

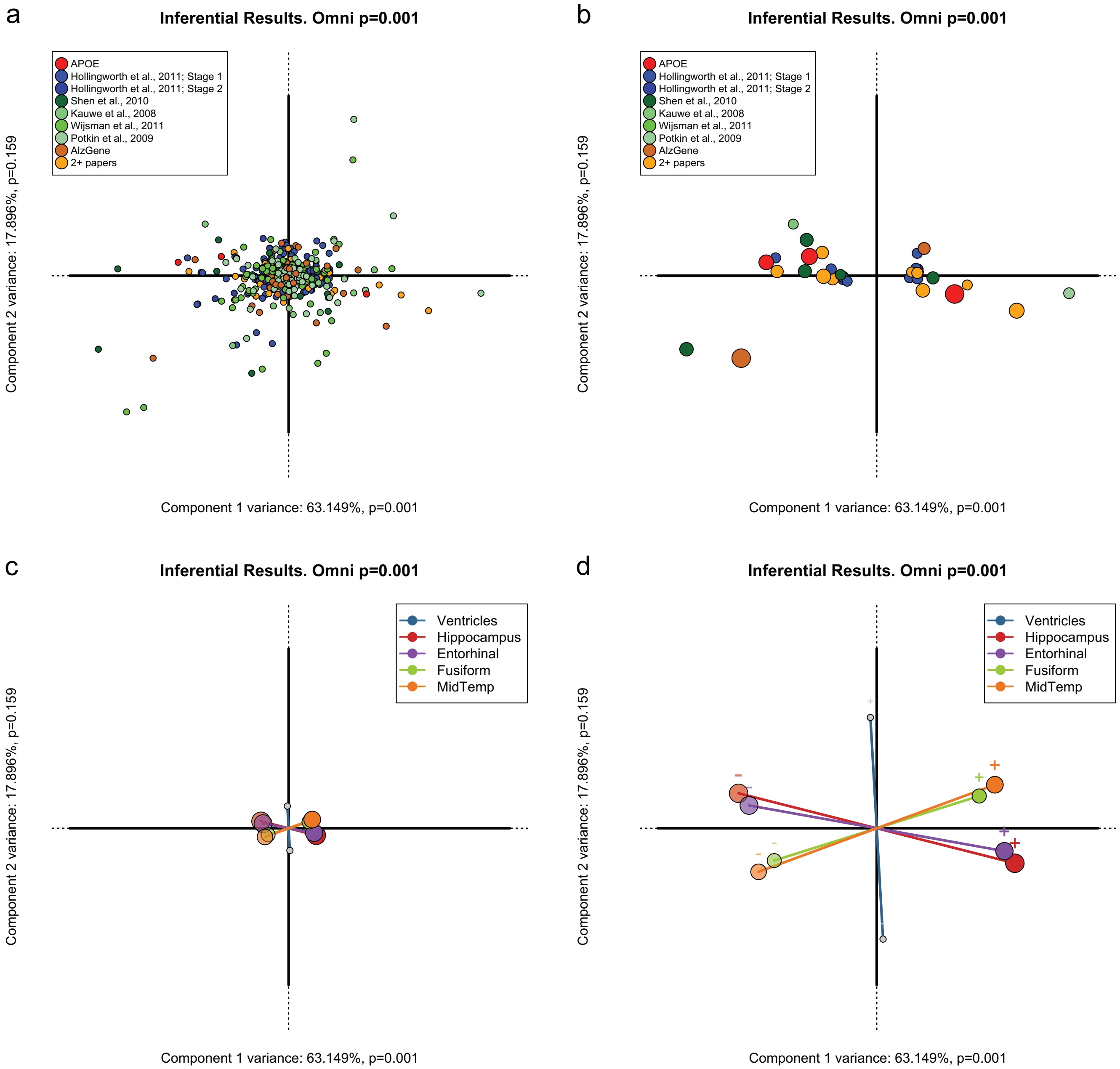

Figure 10. (a) Component map for Components 1 and 2 for all genotypes. (b) Component scores only for the significant genotypes for Component 1 (via a bootstrap ratio test). (c) Component map on Components 1 and 2 for brain regions. Because only Component 1 was significant, the ventricles are denoted with gray circles. This is because the ventricles do not significantly contribute to Component 1. Note that each region is duplicated, with “Escofier” recoding. (d) A *zoomed in* version of the component map (as seen in c) provides a closer look at the brain regions. MidTemp = Medial Temporal. See the online article for the color version of this figure.

logical sciences. In the following sections, we highlight some of the affordances PLSCA provides.

Findings Conclusions

The analyses we performed, with each extension of PLSCA, added a unique perspective on the genetic contribution to Alzheimer’s disease and associated traits. First, across several of our analyses (e.g., simple PLSCA and mixed-data PLSCA), we have shown that the additive risk model approach (e.g., emphasizing minor alleles as the risk factor) may be an inappropriate assumption to explain, for example, complex behaviors, traits, and diagnoses. In several of our analyses, a number of major homozygotes were associated with AD, MCI, and traits often associated with those groups (such as memory and attention issues; Figures 3 and

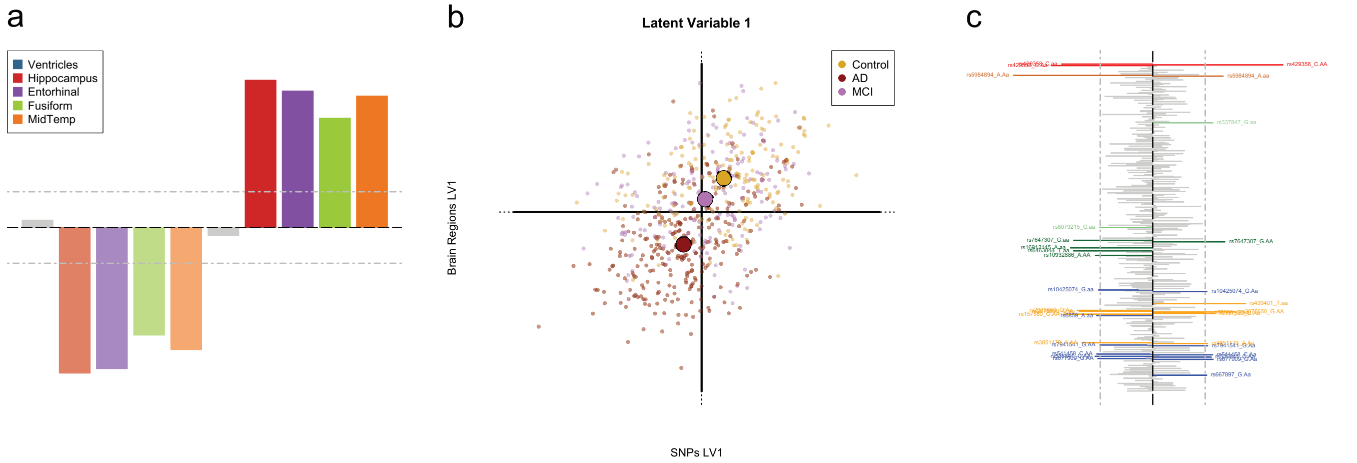

Figure 11. The left image (a) shows bootstrap ratios for the brain regions on Component 1. Note that, with the "Escofier" recoding scheme, each region is duplicated. The ventricles are gray because they do not significantly contribute to Component 1. The gray lines indicate the threshold for significance (−2 and +2). The center image (b) plots the latent variables (LV; participants) from each data set on map of the LVs projected against one another. Only LV 1 is shown because only Component 1 is significant. The horizontal axis represents the SNPs and the vertical axis represents the brain regions. (c) Bootstrap ratios for all the genotypes in this analysis. Gray genotypes do not significantly contribute to Component 1. The gray lines indicate the threshold for significance (−2 and +2). AD = Alzheimer's disease group; MCI = mild cognitive impairment group; MidTemp = Medial temporal. See the online article for the color version of this figure.

4). Furthermore, some minor homozygotes were associated with CON participants and nonpathological—and arguably "healthy"—traits (e.g., the minor homozygote of rs337847 is not associated with smaller regional brain volume, but rather regional sizes *above* the mean).

The results from both mean-centered and simple PLSCA suggest that there are unique genetic signatures per group. First, this provides a more distinct view of genetic risk factors for AD. Further, our results suggest that there are particular genetic (and behavioral) profiles of the CON and MCI groups as well. Given these unique profiles, the genetic markers associated with the CON group may suggest the existence of protective (against AD) genetic factors.

However, we also show unique genetic and behavioral profiles for the MCI group. This group is somewhat associated with behavioral measures of geriatric depression. Because mild cognitive impairment is considered part of a pathological trajectory (Sperling et al., 2011), this result is somewhat surprising. These findings suggest that the MCI group is not a homogeneous group necessarily on a path toward Alzheimer's disease or other dementia. Rather, the MCI group could consist of (at least) two undifferentiable populations of individuals: (a) those on the path toward further cognitive decline (Dotson, Beydoun, & Zonderman, 2010), and (b) those who suffer from depression with specific incidents of cognitive complaints or impairment (Richard et al., 2013).

Finally, we also show, via mixed-data PLSCA, that some genotypes are more associated with brain region volumes above or below the grand mean of the sample. However, we show that typical Alzheimer's disease risk—atrophy or shrinkage of brain regions—is not exclusively related to minor alleles. In fact, our results suggest that there is a complex interaction between SNPs—in which some major and some minor alleles contribute to protective or risk factors associated with Alzheimer's disease.

Methods Conclusions

One issue of growing interest is, simply, how to concurrently analyze sets of behavioral and genetic variables. The joint analysis of behavioral and genetic data could boost power to detect genetic effects (Allison et al., 1998; Schifano et al., 2013; Seoane et al., 2014). Additionally, these data are typically high-dimension (i.e., large number of variables) and low-sample-size (HDLSS). SVD (and thus eigen) based techniques tend to be reliable under particular conditions in HDLSS (Chi, 2012; Jung & Marron, 2009), even in GWAS (Duan et al., 2013). Thus, PLSCA is well suited for multivariate analyses of behavioral and genetic data.

Further, because PLSCA is a multivariate technique, it also has several other advantages over many currently used methods. Multivariate approaches are more beneficial than univariate approaches (Schmitz, Cherny, & Fulker, 1998), especially with respect to multiple testing issues (Nyholt, 2004). Additionally, PLSCA incorporates several nonparametric resampling methods for inference tests. Resampling methods are particularly relevant for data that are noisy or non-normal (Hesterberg, 2011), or when conservative estimates should be made (Chernick, 2008), all of which are conditions especially relevant in genetics (Gao, Becker, Becker, Starmer, & Province, 2010).

In addition, PLSCA analyzes levels (genotypes) of SNPs, and does so in a χ^2 framework (i.e., CA), which is already used in many aspects of genetic analyses such as the (a) χ^2 tests of association or Hardy-Weinberg disequilibrium, and (b) explicit genotypic tests and the codominant model of inheritance. Furthermore, PLSCA generalizes PLSC, which has been shown to be robust in association studies with high dimensional data (Grellmann et al., 2015).

Because of the features, properties, and categorical approach of PLSCA, interpretation of specific allelic and genotypic effects becomes more apparent. The more general approach (i.e., genotypic model coding) of PLSCA can be more advantageous than relying on any one model or testing all possible models (and then correcting for multiple tests), because if a particular inheritance model is apparent, PLSCA will reveal it on a component-by-component basis. For example, if we look at two of the analyses presented here, "simple" PLSCA and "mixed-data" (a.k.a. "heterogeneous", or "mixed-modality") PLSCA, we see evidence for two different genetic models for how rs429358 (an ApoE SNP) may contribute to different aspects of AD: (a) a linear additive effect with respect to diagnostic criteria for Alzheimer's disease (Figure 2, horizontal axis), and (b) a dominance effect (presence of a minor allele) with respect to volume of brain regions (Figure 10, horizontal axis). Finally, because components from (within each of) these analyses are orthogonal, they can be interpreted independently. For example, in "seed" PLSCA (see Figure 9) we can see that the "seeds" are roughly equidistant to the major homozygotes and heterozygotes, and we see, again, the same pattern for the heterozygotes and the minor homozygotes for both ApoE and TOMM40 on the first component (Figure 9, horizontal axis). However, the second component (Figure 9, vertical axis) shows a contrast of the heterozygote to both the homozygotes. Components configured in this way show two different genetic models of these two SNPs (in ApoE and TOMM40) with respect to the other SNPs: The first component implies a linear model, but the second component implies a heterozygous effect. Overall, the inheritance of ApoE and TOMM40 markers appears to be quite unique when compared to other risk markers.

Because of the features of PLSCA, we have shown a much more detailed perspective, within a limited set of SNPs, of the genetic contributions to AD and associated behaviors. We were able to associate very specific genotypic contributions—contrary to usual risk assumptions—with standard and robust measures often used in clinical and diagnostic settings for Alzheimer's disease. The level of detail that PLSCA provides could shed more light onto the genetic contributions to complex traits and diseases, such as identifying possibly protective genetic markers. Future work includes combining PLSCA with (a) regularization and sparsification techniques, and (b) predictive and path-modeling approaches—to create PLS-based regression and path-modeling approaches designed for categorical data.

References

Abdi, H. (2007a). Discriminant correspondence analysis. In N. Salkind (Ed.), *Encyclopedia of measurement and statistics* (pp. 270–275). Thousand Oaks, CA: Sage. http://dx.doi.org/10.4135/9781412952644.n140

Abdi, H. (2007b). Singular value decomposition (SVD) and generalized singular value decomposition (GSVD). In N. J. Salkind (Ed.), *Encyclopedia of measurement and statistics* (pp. 907–912). Thousand Oaks, CA: Sage.

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

Abdi, H. (2010). Partial least squares regression and projection on latent structure regression (PLS regression). *Wiley Interdisciplinary Reviews: Computational Statistics, 2,* 97–106. http://dx.doi.org/10.1002/wics.51

Abdi, H., & Béra, M. (2014). Correspondence analysis. In R. Alhajj & J. Rokne (Eds.), *Encyclopedia of social networks and mining* (pp. 275–284). New York, NY: Springer-Verlag.

Abdi, H., Chin, W. W., Esposito Vinzi, V., Russolillo, G., & Trinchera, L. (Eds.). (2013). *New perspectives in partial least squares and related methods*. New York, NY: Springer-Verlag. http://dx.doi.org/10.1007/978-1-4614-8283-3

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. http://dx.doi.org/10.1016/j.neuroimage.2008.11.008

Abdi, H., & Williams, L. J. (2010a). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics, 2,* 433–459. http://dx.doi.org/10.1002/wics.101

Abdi, H., & Williams, L. J. (2010b). Correspondence analysis. In N. J. Salkind, D. M. Dougherty, & B. Frey (Eds.), *Encyclopedia of research design* (pp. 267–278). Thousand Oaks, CA: Sage. http://dx.doi.org/10.4135/9781412961288.n83

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* (pp. 549–579). New York, NY: Springer-Verlag. http://dx.doi.org/10.1007/978-1-62703-059-5_23

Abdi, H., Williams, L. J., Beaton, D., Posamentier, M. T., Harris, T. S., Krishnan, A., & Devous, M. D., Sr. (2012). Analysis of regional cerebral blood flow data to discriminate among Alzheimer's disease, frontotemporal dementia, and elderly controls: A multi-block barycentric discriminant analysis (MUBADA) methodology. *Journal of Alzheimer's Disease, 31*(Suppl. 3), S189–S201.

Abdi, H., Williams, L. J., & Valentin, D. (2013). Multiple factor analysis: Principal component analysis for multitable and multiblock data sets. *Wiley Interdisciplinary Reviews: Computational Statistics, 5,* 149–179. http://dx.doi.org/10.1002/wics.1246

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. *American Journal of Human Genetics, 63,* 1190–1201. http://dx.doi.org/10.1086/302038

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. http://dx.doi.org/10.1016/j.csda.2013.11.006

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, NY: Springer-Verlag. http://dx.doi.org/10.1007/978-1-4614-8283-3_4

Beaton, D., Rieck, J., Fatt, C. R. C., & Abdi, H. (2013). TExPosition: Two-table exploratory analysis with the singular value decomposition [Computer software manual]. (R package version 2.0.1). Retrieved from https://cran.r-project.org/web/packages/TExPosition/TExPosition.pdf

Bennet, A. M., Reynolds, C. A., Gatz, M., Blennow, K., Pedersen, N. L., & Prince, J. A. (2010). Pleiotropy in the presence of allelic heterogeneity: Alternative genetic models for the influence of APOE on serum LDL, CSF amyloid-β42, and dementia. *Journal of Alzheimer's Disease, 22,* 129–134.

Berry, K. J., Johnston, J. E., & Mielke, P. W., Jr. (2011). Permutation methods. *Wiley Interdisciplinary Reviews: Computational Statistics, 3,* 527–542. http://dx.doi.org/10.1002/wics.177

Bertram, L., McQueen, M. B., Mullin, K., Blacker, D., & Tanzi, R. E. (2007). Systematic meta-analyses of Alzheimer disease genetic association studies: The *AlzGene* database. *Nature Genetics, 39,* 17–23. http://dx.doi.org/10.1038/ng1934

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,* 177–188. http://dx.doi.org/10.1016/j.brainresbull.2010.04.012

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,* 765–796.

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

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. http://dx.doi.org/10.1175/1520-0442(1992)005<0541:AIOMFF>2.0.CO;2

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

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

Cho, S.-C., Kim, J.-W., Kim, H.-W., Kim, B.-N., Shin, M.-S., Cho, D.-Y., . . . Son, J.-W. (2011). Effect of ADRA2A and BDNF gene-gene interaction on the continuous performance test phenotype. *Psychiatric Genetics, 21,* 132–135. http://dx.doi.org/10.1097/YPG.0b013e328341a389

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,* 479–486.

Clarke, T.-K., Weiss, A. R., 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,* 33–39. http://dx.doi.org/10.1111/ahg.12046

Corder, E. H., Saunders, A. M., Risch, N. J., Strittmatter, W. J., Schmechel, D. E., Gaskell, P. C., Jr., . . . Pericak-Vance, M. A. (1994). Protective effect of apolipoprotein E type 2 allele for late onset Alzheimer disease. *Nature Genetics, 7,* 180–184. http://dx.doi.org/10.1038/ng0694-180

Cruchaga, C., Kauwe, J. S., Mayo, K., Spiegel, N., Bertelsen, S., Nowotny, P., . . . Goate, A. M., & the Alzheimer's Disease Neuroimaging Initiative. (2010). SNPs associated with cerebrospinal fluid phospho-tau levels influence rate of decline in Alzheimer's disease. *PLoS Genetics, 6*(9), e1001101. http://dx.doi.org/10.1371/journal.pgen.1001101

de Leon, J., Correa, J. C., Ruaño, G., Windemuth, A., Arranz, M. J., & Diaz, F. J. (2008). Exploring genetic variations that may be associated with the direct effects of some antipsychotics on lipid levels. *Schizophrenia Research, 98,* 40–46. http://dx.doi.org/10.1016/j.schres.2007.10.003

Dotson, V. M., Beydoun, M. A., & Zonderman, A. B. (2010). Recurrent depressive symptoms and the incidence of dementia and mild cognitive impairment. *Neurology, 75,* 27–34. http://dx.doi.org/10.1212/WNL.0b013e3181e62124

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* (pp. 289–300). Boca Raton, FL: CRC Press.

Duan, F., Ogden, D., Xu, L., Liu, K., Lust, G., Sandler, J., . . . Zhang, Z. (2013). Principal component analysis of canine hip dysplasia phenotypes and their statistical power for genome-wide association mapping. *Journal of Applied Statistics, 40,* 235–251. http://dx.doi.org/10.1080/02664763.2012.740617

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*. Advance online publication.

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap* (Vol. 57). Boca Raton, FL: Chapman & Hall/CRC. http://dx.doi.org/10.1007/978-1-4899-4541-9

Escofier, B. (1979). Traitement simultané de variables qualitatives et quantitatives en analyse factorielle [Simultaneous analysis of qualitative and quantitative variables in factor analysis]. *Les Cahiers de l'Analyse des Donnees, 4,* 137–146.

Esposito Vinzi, V., Chin, W. W., Henseler, J., & Wang, H. (Eds.). (2010). *Handbook of partial least squares: Concepts, methods and applications*. New York, NY: Springer. http://dx.doi.org/10.1007/978-3-540-32827-8

Filbey, F. M., Schacht, J. P., Myers, U. S., Chavez, R. S., & Hutchison, K. E. (2010). Individual and additive effects of the CNR1 and FAAH genes on brain response to marijuana cues. *Neuropsychopharmacology, 35,* 967–975. http://dx.doi.org/10.1038/npp.2009.200

Fisher, R. A. (1919). The genesis of twins. *Genetics, 4,* 489–499.

Franić, S., Dolan, C. V., Borsboom, D., Hudziak, J. J., van Beijsterveldt, C. E., & Boomsma, D. I. (2013). Can genetics help psychometrics? Improving dimensionality assessment through genetic factor modeling. *Psychological Methods, 18,* 406–433. http://dx.doi.org/10.1037/a0032755

Frantz, A. C., Zachos, F. E., Kirschning, J., Cellina, S., Bertouille, S., Mamuris, Z., . . . Burke, T. (2013). Genetic evidence for introgression between domestic pigs and wild boars (*Sus scrofa*) in Belgium and Luxembourg: A comparative approach with multiple marker systems. *Biological Journal of the Linnaean Society, 110,* 104–115. http://dx.doi.org/10.1111/bij.12111

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,* 100–105.

Gasi, F., Kurtovic, M., Kalamujic, B., Pojskic, N., Grahic, J., Kaiser, C., & Meland, M. (2013). Assessment of European pear (Pyrus communis l.) genetic resources in Bosnia and Herzegovina using microsatellite markers. *Scientia Horticulturae, 157,* 74–83. http://dx.doi.org/10.1016/j.scienta.2013.04.017

Genin, E., Hannequin, D., Wallon, D., Sleegers, K., Hiltunen, M., Combarros, O., . . . Campion, D. (2011). APOE and Alzheimer disease: A major gene with semi-dominant inheritance. *Molecular Psychiatry, 16,* 903–907. http://dx.doi.org/10.1038/mp.2011.52

Greenacre, M. J. (1984). *Theory and applications of correspondence analysis*. London, UK: Academic Press.

Greenacre, M. J. (2007). *Correspondence analysis in practice*. Boca Raton, FL: CRC Press. http://dx.doi.org/10.1201/9781420011234

Greenacre, M. J. (2010). Correspondence analysis. *Wiley Interdisciplinary Reviews: Computational Statistics, 2,* 613–619. http://dx.doi.org/10.1002/wics.114

Greenacre, M. J. (2014). Data doubling and fuzzy coding. In J. Blasius & M. Greenacre (Eds.), *Visualization and verbalization of data* (pp. 239–270). Boca Raton: CRC Press.

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

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://dx.doi.org/10.1016/j.neuroimage.2014.12.025

Hamidovic, A., Dlugos, A., Palmer, A. A., & de Wit, H. (2010). Polymorphisms in dopamine transporter (*SLC6A3*) are associated with stimulant effects of *D*-amphetamine: An exploratory pharmacogenetic study using healthy volunteers. *Behavior Genetics, 40,* 255–261. http://dx.doi.org/10.1007/s10519-009-9331-7

Hesterberg, T. (2011). Bootstrap. *Wiley Interdisciplinary Reviews: Computational Statistics, 3,* 497–526. http://dx.doi.org/10.1002/wics.182

Hollingworth, P., Harold, D., Sims, R., Gerrish, A., Lambert, J.-C., Carrasquillo, M. M., . . . EADI1 consortium. (2011). Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease. *Nature Genetics, 43,* 429–435. http://dx.doi.org/10.1038/ng.803

Jolliffe, I. (2002). *Principal component analysis*. New York, NY: Springer-Verlag.

Jung, S., & Marron, J. S. (2009). PCA consistency in high dimension, low sample size context. *Annals of Statistics, 37,* 4104–4130. http://dx.doi.org/10.1214/09-AOS709

Kauwe, J. S. K., Cruchaga, C., Mayo, K., Fenoglio, C., Bertelsen, S., Nowotny, P., . . . Goate, A. M. (2008). Variation in MAPT is associated with cerebrospinal fluid tau levels in the presence of amyloid-beta deposition. *PNAS Proceedings of the National Academy of Sciences of the United States of America, 105,* 8050–8054. http://dx.doi.org/10.1073/pnas.0801227105

Kocovsky, P. M., Sullivan, T. J., Knight, C. T., & Stepien, C. A. (2013). Genetic and morphometric differences demonstrate fine-scale population substructure of the yellow perch Perca flavescens: Need for redefined management units. *Journal of Fish Biology, 82,* 2015–2030. http://dx.doi.org/10.1111/jfb.12129

Koopal, C., van der Graaf, Y., Asselbergs, F. W., Westerink, J., & Visseren, F. L. (2014). Influence of APOE-2 genotype on the relation between adiposity and plasma lipid levels in patients with vascular disease. *International Journal of Obesity (2005), 39,* 265–269.

Krishnan, A., Williams, L. J., McIntosh, A. R., & Abdi, H. (2011). Partial least squares (PLS) methods for neuroimaging: A tutorial and review. *NeuroImage, 56,* 455–475. http://dx.doi.org/10.1016/j.neuroimage.2010.07.034

Lakatos, A., Derbeneva, O., Younes, D., Keator, D., Bakken, T., Lvova, M., . . . Alzheimer's Disease Neuroimaging Initiative. (2010). Association between mitochondrial DNA variations and Alzheimer's disease in the ADNI cohort. *Neurobiology of Aging, 31,* 1355–1363.

Lantieri, F., Glessner, J. T., Hakonarson, H., Elia, J., & Devoto, M. (2010). Analysis of GWAS top hits in ADHD suggests association to two polymorphisms located in genes expressed in the cerebellum. *American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics: The Official Publication of the International Society of Psychiatric Genetics, 153B,* 1127–1133.

Lebart, L., Morineau, A., & Warwick, K. M. (1984). *Multivariate descriptive statistical analysis: Correspondence analysis and related techniques for large matrices*. London, UK: Wiley.

Le Floch, E., Guillemot, V., Frouin, V., Pinel, P., Lalanne, C., Trinchera, L., . . . Duchesnay, E. (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,* 11–24. http://dx.doi.org/10.1016/j.neuroimage.2012.06.061

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,* 358–362. http://dx.doi.org/10.1002/gepi.20217

Liu, J., Pearlson, G., Windemuth, A., Ruano, G., Perrone-Bizzozero, N. I., & Calhoun, V. (2009). Combining fMRI and SNP data to investigate connections between brain function and genetics using parallel ICA. *Human Brain Mapping, 30,* 241–255. http://dx.doi.org/10.1002/hbm.20508

Malinvaud, E. (1987). Data analysis in applied socio-economic statistics with consideration of correspondence analysis. Paper presented at Marketing Science Conference, Jouy-en-Josas, France, June, 1987.

McIntosh, A. R., Bookstein, F. L., Haxby, J. V., & Grady, C. L. (1996). Spatial pattern analysis of functional brain images using partial least squares. *NeuroImage, 3,* 143–157. http://dx.doi.org/10.1006/nimg.1996.0016

McIntosh, A. R., & Lobaugh, N. J. (2004). Partial least squares analysis of neuroimaging data: Applications and advances. *NeuroImage, 23*(Suppl. 1), S250–S263. http://dx.doi.org/10.1016/j.neuroimage.2004.07.020

McIntosh, A. R., Nyberg, L., Bookstein, F. L., & Tulving, E. (1997). Differential functional connectivity of prefrontal and medial temporal cortices during episodic memory retrieval. *Human Brain Mapping, 5,* 323–327. http://dx.doi.org/10.1002/(SICI)1097-0193(1997)5:4<323::AID-HBM20>3.0.CO;2-D

Michaelson, J. J., Alberts, R., Schughart, K., & Beyer, A. (2010). Data-driven assessment of eQTL mapping methods. *BMC Genomics, 11,* 502. http://dx.doi.org/10.1186/1471-2164-11-502

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 & Behavior, 7,* 714–719. http://dx.doi.org/10.1111/j.1601-183X.2008.00409.x

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,* 56. http://dx.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,* 395–400.

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,* 1940–1947. http://dx.doi.org/10.1038/npp.2011.82

Nyholt, D. R. (2004). A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *American Journal of Human Genetics, 74,* 765–769. http://dx.doi.org/10.1086/383251

Nyholt, D. R., Yu, C.-E., & Visscher, P. M. (2009). On Jim Watson's APOE status: Genetic information is hard to hide. *European Journal of Human Genetics, 17,* 147–149. http://dx.doi.org/10.1038/ejhg.2008.198

Paige, C. C., & Saunders, M. A. (1981). Towards a generalized singular value decomposition. *SIAM Journal on Numerical Analysis, 18,* 398–405. http://dx.doi.org/10.1137/0718026

Park, M., Lee, J. W., & Kim, C. (2007). Correspondence analysis approach for finding allele associations in population genetic study. *Computational Statistics & Data Analysis, 51,* 3145–3155. http://dx.doi.org/10.1016/j.csda.2006.09.002

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. http://dx.doi.org/10.1016/j.csda.2004.06.015

Potkin, S. G., Guffanti, G., Lakatos, A., Turner, J. A., Kruggel, F., Fallon, J. H., . . . Alzheimer's Disease Neuroimaging Initiative. (2009). Hippocampal atrophy as a quantitative trait in a genome-wide association study identifying novel susceptibility genes for Alzheimer's disease. *PLoS ONE, 4*(8), e6501. http://dx.doi.org/10.1371/journal.pone.0006501

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., 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,* 559–575. http://dx.doi.org/10.1086/519795

Raîche, G., Walls, T. A., Magis, D., Riopel, M., & Blais, J.-G. (2013). Non-graphical solutions for Cattell's scree test. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 9,* 23–29. http://dx.doi.org/10.1027/1614-2241/a000051

Richard, E., Reitz, C., Honig, L. H., Schupf, N., Tang, M. X., Manly, J. J., . . . Luchsinger, J. A. (2013). Late-life depression, mild cognitive impairment, and dementia. *Journal of the American Medical Association Neurology, 70,* 383–389. http://dx.doi.org/10.1001/jamaneurol.2013.603

Romanos, M., Freitag, C., Jacob, C., Craig, D. W., Dempfle, A., Nguyen, T. T., . . . Lesch, K. P. (2008). Genome-wide linkage analysis of ADHD using high-density SNP arrays: Novel loci at 5q13.1 and 14q12. *Molecular Psychiatry, 13,* 522–530. http://dx.doi.org/10.1038/mp.2008.12

Roses, A. D., Lutz, M. W., Amrine-Madsen, H., Saunders, A. M., Crenshaw, D. G., Sundseth, S. S., . . . Reiman, E. M. (2010). A TOMM40 variable-length polymorphism predicts the age of late-onset Alzheimer's disease. *The Pharmacogenomics Journal, 10,* 375–384. http://dx.doi.org/10.1038/tpj.2009.69

Sanford, E. C. (1908). Review of measurements of twins by Edward L. Thorndike. *The American Journal of Psychology, 19,* 142–143.

Saporta, G. (2011). *Probabilités, analyse des données et statistique* [Probability, data analysis, and statistics]. Paris, France: Technip.

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,* 744–759. http://dx.doi.org/10.1016/j.ajhg.2013.04.004

Schmitz, S., Cherny, S. S., & Fulker, D. W. (1998). Increase in power through multivariate analyses. *Behavior Genetics, 28,* 357–363. http://dx.doi.org/10.1023/A:1021669602220

Seoane, J. A., Campbell, C., Day, I. N. M., Casas, J. P., & Gaunt, T. R. (2014). Canonical correlation analysis for gene-based pleiotropy discovery. *PLoS Computational Biology, 10*(10), e1003876. http://dx.doi.org/10.1371/journal.pcbi.1003876

Shen, L., Kim, S., Risacher, S. L., Nho, K., Swaminathan, S., West, J. D., . . . Alzheimer's Disease Neuroimaging Initiative. (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,* 1051–1063. http://dx.doi.org/10.1016/j.neuroimage.2010.01.042

Sperling, R. A., Aisen, P. S., Beckett, L. A., Bennett, D. A., Craft, S., Fagan, A. M., . . . Phelps, C. H. (2011). Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association, 7,* 280–292. http://dx.doi.org/10.1016/j.jalz.2011.03.003

Sullivan, P. F., Lin, D., Tzeng, J.-Y., van den Oord, E., Perkins, D., Stroup, T. S., . . . Close, S. L. (2008). Genomewide association for schizophrenia in the CATIE study: Results of stage 1. *Molecular Psychiatry, 13,* 570–584. http://dx.doi.org/10.1038/mp.2008.25

Tenenhaus, M. (1998). *La régression pls: Théorie et pratique* [PLS regression: Theory and practice]. Paris: Technip.

Tenenhaus, M., Esposito Vinzi, V., Chatelin, Y., & Lauro, C. (2005). PLS path modeling. *Computational Statistics & Data Analysis, 48,* 159–205. http://dx.doi.org/10.1016/j.csda.2004.03.005

Thorndike, E. L. (1905). *Measurements of twins*. New York, NY: Science Press.

Thurstone, L. L. (1934). The vectors of mind. *Psychological Review, 41,* 1–32. http://dx.doi.org/10.1037/h0075959

Tucker, L. R. (1958). An inter-battery method of factor analysis. *Psychometrika, 23,* 111–136. http://dx.doi.org/10.1007/BF02289009

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: For Rapid Communication of Neuroscience Research, 19,* 603–607. http://dx.doi.org/10.1097/WNR.0b013e3282fa6d8d

van der Sluis, S., Posthuma, D., & Dolan, C. V. (2013). TATES: Efficient multivariate genotype-phenotype analysis for genome-wide association studies. *PLoS Genetics, 9*(1), e1003235. http://dx.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,* 6–17. http://dx.doi.org/10.1111/j.1742-9536.2011.00002.x

Van Loan, C. F. (1976). Generalizing the singular value decomposition. *SIAM Journal on Numerical Analysis, 13,* 76–83. http://dx.doi.org/10.1137/0713009

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*. Advance online publication. http://dx.doi.org/10.1038/nrg1916-c1

Vounou, M., Nichols, T. E., Montana, G., & Alzheimer's Disease Neuroimaging Initiative. (2010). Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach. *NeuroImage, 53,* 1147–1159. http://dx.doi.org/10.1016/j.neuroimage.2010.07.002

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,* 6–15. http://dx.doi.org/10.1002/gepi.20351

Weiner, M. P., & Hudson, T. J. (2002, June). Introduction to SNPs: Discovery of markers for disease. *BioTechniques* (Suppl.), 4–7, 10, 12–13.

Wijsman, E. M., Pankratz, N. D., Choi, Y., Rothstein, J. H., Faber, K. M., Cheng, R., . . . NIA-LOAD/NCRAD Family Study Group. (2011). Genome-wide association of familial late-onset Alzheimer's disease replicates BIN1 and CLU and nominates CUGBP2 in interaction with APOE. *PLoS Genetics, 7*(2), e1001308.

Williams, L. J., Abdi, H., French, R., & Orange, J. B. (2010). A tutorial on multiblock discriminant correspondence analysis (MUDICA): A new method for analyzing discourse data from clinical populations. *Journal of Speech, Language, and Hearing Research, 53,* 1372–1393. http://dx.doi.org/10.1044/1092-4388(2010/08-0141)

Yanai, H., Takeuchi, K., & Takane, Y. (2011). *Projection matrices, generalized inverse matrices, and singular value decomposition*. New York, NY: Springer-Verlag. http://dx.doi.org/10.1007/978-1-4419-9887-3

Yang, H., Liu, J., Sui, J., Pearlson, G., & Calhoun, V. D. (2010). A hybrid machine learning method for fusing fMRI and genetic data: Combining both improves classification of schizophrenia. *Frontiers in Human Neuroscience, 4,* 192. http://dx.doi.org/10.3389/fnhum.2010.00192

Zapala, M. A., & Schork, N. J. (2006). Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables. *PNAS Proceedings of the National Academy of Sciences of the United States of America, 103,* 19430–19435. http://dx.doi.org/10.1073/pnas.0609333103

(*Appendices follow*)

Appendix A

Alternative and Additional Strategies With Partial Least Squares Correspondence Analysis (PLSCA): Testing Other Genetic Models

We present a multivariate strategy that is both parsimonious and, arguably, the most general approach (i.e., genotypic/codominant) to detecting associations between many single nucleotide polymorphisms (SNPs) and traits (or behaviors, diagnoses, etc.). However, the completely categorical approach that we present is not the only approach to association analyses. In this appendix, we briefly discuss how to use PLSCA with other models of inheritance (e.g., additive, dominant, recessive).

Sometimes it might be advantageous to test a specific genetic model of inheritance, especially if the model is known (Lettre, Lange, & Hirschhorn, 2007). As in Table 1, we have noted a number of formats taken by genetic data for a variety of analyses and inheritance models. Generally, there are six widely used models: codominant (or genotypic), dominant, recessive, heterozygous, additive, and multiplicative. While the additive model is one of the most commonly used models, the codominant and genotypic models are the most general, quite powerful, and the best fit when an inheritance model is unknown (Lettre et al., 2007). While multiple models can be tested in succession, this procedure becomes impractical with large data sets for two reasons: computational expense and correcting for multiple tests (especially because the standard genome-wide association studies threshold is already very conservative: $p < .05 \times 10^{-8}$). Thus, we recommend the fully categorical approach (à la the genotypic model), especially because in the psychological, cognitive, and neurological sciences, we may not know which model is most appropriate in order to explain complex behaviors, traits, and diagnoses.

We briefly illustrate how data could be recoded for a variety of formats, in addition to recommendations for how to approach PLSCA with different inheritance models. To note, most formats discussed are easily exported from most genetic association software (e.g., PLINK). The disjunctive recoding and analyses can be achieved with the ExPosition and TExPosition packages (Beaton et al., 2014; Beaton, Rieck, Fatt, & Abdi, 2013). Example data and code are available at https://code.google.com/p/exposition-family/source/browse/Publications/PsyMet_2015 and http://www.utd.edu/~herve/PsyMet_2015

Categorical Models

Four of the six commonly used genetic models are categorical in nature. These models are (a) codominant, (b) dominant, (c) recessive, and (d) heterozygous. Because PLSCA is presented in a format that aligns with codominant (and genotypic) model(s), we forgo that discussion here and focus on the other three categorical models. The categorical models can be used in any of the PLSCA approaches with another categorical data set (e.g., surveys, diagnosis). If the genetic data are categorical and the trait data are quantitative, the mixed-modality version of PLSCA must be used.

Table A1 illustrates how to code for dominant, recessive, and heterozygous models. Both the nominal genotype and (standard) additive approach are provided for reference (Table A1, a and b). SNPs are presented in a general format in which “A” is the major allele and “a” the minor allele. The major homozygote, heterozygote, and minor homozygote are presented as AA, Aa, and aa, respectively.

Dominant. The dominant model assumes that just the presence of a minor allele is considered a risk factor. Essentially, genotypes become dichotomized between the major homozygote vs. the other genotypes ([AA] vs. [Aa + aa]). Categorical and disjunctive coding for the dominant model can be found in Table A1, c and d.

Recessive. The recessive model assumes that only the minor homozygote is the risk factor. Essentially, genotypes become dichotomized between the minor homozygote vs. the other genotypes ([AA + Aa] vs. [aa]). Categorical and disjunctive coding for the recessive model can be found in Table A1, e and f.

Heterozygous. The heterozygous model assumes that just the heterozygote confers either a risk or protective factor. Essentially, genotypes become dichotomized between the heterozygote vs. the homozygotes ([AA + aa] vs. [Aa]). Categorical and disjunctive coding for the heterozygous model can be found in Table A1, g and h.

In sum, any of these categorical coding approaches are easily exported by most genetic association software, and can be used trivially with PLSCA.

Quantitative Models

In general, there are two quantitative genetic models: (a) additive, and (b) multiplicative. Here, we outline how to use PLSCA with these models in two different formats: representation of each genotype with a quantitative value in a disjunctive format (i.e., a compromise between quantitative and co-dominant coding), in addition to representation of just SNPs (i.e., strictly quantitative).

A compromise between quantitative and categorical. Additive coding expects equal risk between ordinal pairs of nucleotides. That is, the risk value between the major homozygote and the heterozygote should be equal to the risk value between the heterozygote and the minor homozygote (e.g., 0, 1, 2). In Table A2, Part c, we provide an example of how to code the additive model in the disjunctive format. In disjunctive format, there are three possible configurations: [1 0], [0.5 0.5], and [0 1], which represent (respectively) the major homozygote, heterozygote, and minor homozygote.

(*Appendices continue*)

Table A1

Nominal, Additive (for Reference), Dominant, Recessive, Heterozygous, and Disjunctive Formats of SNP Data

	SNP1	SNP2
(a) Nominal		
Subject 1	Aa	Aa
Subject 2	aa	Aa
Subject *i*	Aa	aa
Subject *I*	AA	AA
(b) Additive		
Subject 1	1	1
Subject 2	2	1
Subject *i*	1	2
Subject *I*	0	0
(c) Dominant		
Subject 1	D	D
Subject 2	D	D
Subject *i*	D	D
Subject *I*	N	N

	SNP1		SNP2	
	D	N	D	N
(d) Disjunctive for dominant				
Subject 1	1	0	1	0
Subject 2	1	0	1	0
Subject *i*	1	0	1	0
Subject *I*	0	1	0	1

	SNP1	SNP2
(e) Recessive		
Subject 1	N	N
Subject 2	R	N
Subject *i*	N	R
Subject *I*	N	N

	SNP1		SNP2	
	R	N	R	N
(f) Disjunctive for recessive				
Subject 1	0	1	0	1
Subject 2	1	0	0	1
Subject *i*	0	1	1	0
Subject *I*	0	1	0	1

	SNP1	SNP2
(g) Heterozygous		
Subject 1	H	H
Subject 2	N	H
Subject *i*	H	N
Subject *I*	N	N

(*Appendices continue*)

Table A1 (*continued*)

	SNP1		SNP2	
	H	N	H	N
(h) Disjunctive for heterozgous				
Subject 1	1	0	1	0
Subject 2	0	1	1	0
Subject *i*	1	0	0	1
Subject *I*	0	1	0	1

Note. Here, 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. Example of (a) nominal, (b) additive, followed by (c) dominant (dominant genotypes marked as "D"; nondominant marked as "N") with (d) the disjunctive form of dominant, followed by (e) recessive (recessive genotypes marked as "R"; nonrecessive marked as "N") with (f) disjunctive form of recessive, followed by (g) heterozygous (heterozygous genotypes marked as 'H', non heterozygous marked as 'N') with the (h) disjunctive form of heterozygous. Disjunctive data for dominant, recessive, and heterozygous models are essentially dichotomized, and thus presented as disjunctive code for two categories.

Table A2

Nominal, Additive (for Reference), and Compromise Between Quantitative and Disjunctive Formats of SNP Data

	SNP1	SNP2
(a) Nominal		
Subject 1	Aa	Aa
Subject 2	aa	Aa
Subject *i*	Aa	aa
Subject *I*	AA	AA
(b) Additive		
Subject 1	1	1
Subject 2	2	1
Subject *i*	1	2
Subject *I*	0	0

	SNP1		SNP2	
	A	a	A	a
(c) Compromise between linear additive and disjunctive				
Subject 1	.5	.5	.5	.5
Subject 2	0	1	.5	.5
Subject *i*	.5	.5	0	1
Subject *I*	1	0	1	0
(d) Compromise between multiplicative and disjunctive with emphasis on the minor allele				
Subject 1	.25	.75	.25	.75
Subject 2	0	1	.25	.75
Subject *i*	.25	.75	0	1
Subject *I*	1	0	1	0

Note. Here, SNPs are presented generally where "A" is the major allele and "a" the minor allele. The major homozygote, heterozygote, and minor homozygote are presented as "AA", "Aa", and "aa", respectively. Example of (a) nominal, (b) additive, followed by (c) a compromise between linear additive and disjunctive. Here, coding exists only for the presence of a major allele ('A') or a minor allele ('a'). If a subject has the heterozygote, they receive equal parts major and minor homozygote, and (d) a compromise between multiplicative and disjunctive. Here, coding exists only for the presence of a major allele ('A') or a minor allele ('a'). If a subject has the heterozygote, there is a larger emphasis on the minor allele than the major allele.

(*Appendices continue*)

However, if the quantitative inheritance model should place more emphasis on a particular allele, then only one set of these values changes: those associated with the heterozygote. If, for example, there should be more emphasis on the minor allele when it is present (e.g., the heterozygote and minor homozygote; analogous to a dominance model), then the heterozygote can be coded as different values that sum to 1 (e.g., [0.25 0.75], [0.1, 0.9]) that best describes the risk of a minor allele (see Table A2d). To note, as long as the sum of the columns that span a single variable (e.g., SNP1) sum to 1, this pattern meets the criteria for disjunctive coding. When the values are within the interval of [0 1], this is often called "fuzzy coding" or "bipolar coding" (Greenacre, 2014; Lebart, Morineau, & Warwick, 1984). This format is still considered a categorical table and the same rules apply as in the other categorical coding formats. If the genetic and trait data are both categorical, standard PLSCA applies. If the genetic data are categorical and the trait data are quantitative, the mixed-modality version of PLSCA must be used.

Quantitative models as they are. If a researcher wishes to use SNPs in a quantitative format, represented by single columns, as one of their data sets (see, e.g., any of the "additive" examples in Tables 1, 2, Table A1, and Table A2), then there are only two options in the PLSCA framework. First, if the SNPs are quantitative but the trait data are categorical or dummy coded (e.g., diagnostic group; see mean-centered PLSCA), then the mixed-modality PLSCA approach applies, wherein the SNPs are treated as quantitative (and thus duplicated, as in Table 4). To note, there is one particular circumstance that essentially reduces to a simple regression: If the SNPs are quantitative and the categorical data simply designate whether a subject is either a "case" or a "control," this will produce only one component that best separates "case" from "control" groups.

Second, if the SNPs *and* the traits are quantitative, then both data sets must be recoded as in Table 4. However—because PLSCA generalizes PLSC—this simply reduces to a standard PLSC (within a constant scaling factor) between the SNPs and the traits. If both sets are treated as quantitative, it is best to just perform PLSC because it reduces the required computational time and memory (i.e., there is no duplication of columns).

Appendix B

The Main Tool: The Generalized Singular Value Decomposition

The singular value decomposition (SVD)—and, by extension, the generalized SVD (GSVD)—is the core tool for many techniques such as principal components analysis, correspondence analysis, partial least squares, and numerous related techniques. The SVD is a generalization of the eigenvalue decomposition (EVD; see Abdi 2007c). The EVD decomposes square, symmetric tables, whereas the SVD decomposes rectangular tables (Yanai, Takeuchi, & Takane, 2011).

The SVD produces orthogonal components (sometimes called dimensions, axes, principal axes, or factors). Components are new variables computed as linear combinations of the original variables of the original data matrix. 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. 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 for each component, called *component scores*. The values reflect how much an observation contributes to the variance of each component. Additionally, component scores of observations or measures can be plotted, with respect to components, to produce component maps (much akin to scatterplots). Component maps show the spatial relationship between observations, between measures, and between the two sets (Greenacre, 1984): Items close to each other are similar, and items far apart differ.

The SVD

The SVD is the core of most linear multivariate techniques (see Abdi, 2007b). The SVD decomposes a data matrix $\mathbf{R}$—with J rows and K columns—into three matrices:

$$\mathbf{R} = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}^{\mathrm{T}}, \tag{B1}$$

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 $\boldsymbol{\Delta}$ is an L by L diagonal matrix in which diag$\{\boldsymbol{\Delta}\}$ stores the singular values (and diag$\{\boldsymbol{\Delta}^2\}$ stores the eigenvalues). Furthermore, $\mathbf{U}$ and $\mathbf{V}$ are orthonormal matrices such that

$$\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I} = \mathbf{V}^{\mathrm{T}}\mathbf{V}. \tag{B2}$$

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

$$\mathbf{F}_J = \mathbf{U}\boldsymbol{\Delta} \text{ and } \mathbf{F}_K = \mathbf{V}\boldsymbol{\Delta}, \tag{B3}$$

and can be plotted—often with two components at a time—to produce component maps.

(*Appendices continue*)

The Generalized SVD (GSVD)

The GSVD[5] differs from the SVD in that there are constraints placed upon the rows and columns. The constraints are represented by positive definite matrices of sizes J by J and K by K applied to the rows and columns, respectively. These constraints matrices are often diagonal matrices, and when this is 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 Equation B1, with the following constraints:

$$\mathbf{U}^{\mathrm{T}}\mathbf{W}_J\mathbf{U} = \mathbf{I} = \mathbf{V}^{\mathrm{T}}\mathbf{W}_K\mathbf{V}, \tag{B4}$$

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

$$\mathbf{F}_J = \mathbf{W}_J\mathbf{U}\boldsymbol{\Delta} \text{ and } \mathbf{F}_K = \mathbf{W}_K\mathbf{V}\boldsymbol{\Delta}. \tag{B5}$$

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

[5] There are actually two techniques called the generalized singular value decomposition (GSVD). One described by Van Loan (1976), and extended by Paige and Saunders (1981), was designed for decomposition of two matrices that share the same columns, in which, typically, one matrix is ill-conditioned. The other, which is the one we refer to, is described as an approach to the SVD with constraints placed on the left and right singular vectors, as described by Greenacre (1984), Lebart et al. (1984), and Abdi (2007b).

Received December 3, 2014
Revision received June 11, 2015
Accepted June 17, 2015 ■