Chapter 6
Imaging Genetics with Partial Least Squares for Mixed-Data Types (MiMoPLS)

Derek Beaton, Michael Kriegsman, ADNI*, Joseph Dunlop, Francesca M. Filbey, and Hervé Abdi

In H. Abdi, V. Esposito Vinzi, G. Russolillo,
S. Saporta, & L. Trinchera, (Eds.), (2016).
The Multiple Facets of Partial Least Squares and Related Methods.
New York: Springer Verlag.

Abstract "Imaging genetics" studies the genetic contributions to brain structure and function by finding correspondence between genetic data—such as single nucleotide polymorphisms (SNPs)—and neuroimaging data—such as diffusion tensor imaging (DTI). However, genetic and neuroimaging data are heterogenous data types, where neuroimaging data are quantitative and genetic data are (usually) categorical. So far, methods used in imaging genetics treat all data as quantitative, and this sometimes requires unrealistic assumptions about the nature of genetic data. In this article we present a new formulation of Partial Least Squares Correlation (PLSC)—called Mixed-modality Partial Least Squares (MiMoPLS)—specifically tailored for heterogeneous (mixed-) data types. MiMoPLS integrates features of PLSC and Correspondence Analysis (CA) by using special properties of quantitative data and Multiple Correspondence Analysis (MCA). We illustrate MiMoPLS with an example data set from the Alzheimer's Disease Neuroimaging Initiative (ADNI) with DTI and SNPs.

Key words: Imaging genetics, MiMoPLS, Alzheimer Disease, (Multiple) Correspondence Analysis, Burt's stripe, SNPs

---

Derek Beaton
School of Behavioral and Brain Sciences, The University of Texas at Dallas, e-mail: derekbeaton@utdallas.edu

Michael Kriegsman
School of Behavioral and Brain Sciences, The University of Texas at Dallas, e-mail: michael.kriegsman@utdallas.edu

for the Alzheimer's Disease Neuroimaging Initiative
*Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (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 listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/\\how_to_apply/ADNI_Acknowledgement_List.pdf

Joseph Dunlop
SAS Institute Inc., Cary NC. e-mail: joseph.dunlop@gmail.com

Francesca M. Filbey
Center for BrainHealth and School of Behavioral and Brain Sciences, The University of Texas at Dallas e-mail: francesca.filbey@utdallas.edu

Hervé Abdi
School of Behavioral and Brain Sciences, The University of Texas at Dallas, e-mail: herve@utdallas.edu

6.1 Introduction

Imaging genetics (and "imaging genomics") combines two scientific disciplines: neuroimaging—often from the cognitive neuroscience perspective—and genetics—often from the genomics perspective [1, 2]. Imaging genetics integrates neuroimaging and genetic data to understand how genetics contributes to brain structure and function—often with respect to diagnostic criteria or complex behavior and traits (such as personality). Usually, the data sets in imaging genetics are very large: neuroimaging data (measured in number of voxels) can comprise up to one million variables, whereas genetic data (often genome-wide with single nucleotide polymorphisms [SNPs]) can comprise more than three million variables. With such large data sets it is often impractical to use mass-univariate statistics, simply because the corrections for multiple comparisons become then too drastic.

So, instead of using mass-univariate approaches, imaging genetics researchers often turn to multivariate methods [3] such as sparse reduced rank regression [4], distance matrix regression [5], independent components analysis [6,7], Canonical Correlation Analysis (CCA) [8], or Partial Least Squares (PLS) [16]. Because the goal of imaging genetics is to understand the relationships between imaging and genetics, researchers often turn to multivariate techniques designed to conjointly analyze two tables of data (e.g., imaging and genetics). However, nearly all implementations of CCA, PLS, and most other multivariate techniques are designed for quantitative data and this can be problematic because many types of genetic data—especially SNPs—are categorical data.

6.1.1 Ambiguity with Allelic Coding

With the advent of genome-wide technology, many biological, medical, and psychological disciplines conduct genome-wide association (GWA) studies. Typically, genome-wide data consist in single nuclear polymorphisms (SNPs) [17]. A SNP is expressed by the two nucleotide letters that exist at a particular genomic location. These two letters can be, for example, AA, AT, or TT. For a given SNP, each letter can be a major allele—say A—or a minor allele—say T. For analyses, SNPs are often recoded into an allelic count; typically, SNPs emphasize the minor allele. Thus our example—AA, AT, and TT—would be recoded respectively as the numbers 0, 1, or 2 (because AA has 0 minor allele, and TT has 2 minor alleles). This $\{0,1,2\}$ coding scheme is often called an "additive" model. In biological, medical, and psychological studies with SNPs, the minor allele is usually assumed to be associated with risk for diseases and disorders [18, 19].

This allelic count makes several unrealistic assumptions. First, the $\{0,1,2\}$ scheme is an implicit contrast—which, in GWA studies, emphasizes the minor allele for hundreds of thousands or even millions of SNPs. Second, this contrast is linear even though many risk factors are non-linear (e.g., risk of Alzheimer's Disease from ApoE) [20]. Finally, because the minor allele frequency is usually computed *per study sample*, there is a possibility that a separate sample would detect a different minor allele, and so the "2" in one study would be a "0" in another study (and this could create problems with replication); thus the only unambiguous zygote—across different samples and populations—is the heterozygote marked as "1" (e.g., AT in our example).

To avoid these measurement assumptions, SNPs can be expressed in a purely categorical format that preserves exactly the alleles found without presuming a linear contrast effect. However, there exists only a few statistical methods (e.g., Multiple Factor Analysis, [21]) designed to simultaneously analyze heterogeneous data such as SNPs (categorical) and neuroimaging (continuous). In this paper, we provide a new formulation of PLS designed for heterogeneous data types that allows both SNPs and imaging data to remain in

their natural formats (categorical, and continuous, respectively). This approach—called"mixed-modality" PLS (MiMoPLS)—generalizes PLS for use with data sets that comprise both quantitative and categorical variables.

6.2 Notation and Prerequisites

This section presents the notations and a sketch of the main prerequisite methods: the singular value decomposition and its generalization, principal components analysis, (multiple) correspondence analysis, partial least squares correlation, and partial least squares correspondence analysis.

6.2.1 Notation

Uppercase bold letters denote matrices (e.g., $\mathbf{X}$) and lower case bold letters denote vectors (e.g., $\mathbf{x}$). The transpose operation is denoted $^{\mathrm{T}}$, the inverse operation $^{-1}$, and the diagonal operation—which turns a vector into a diagonal matrix, or extracts the diagonal as a vector from a diagonal matrix—is denoted $\mathsf{diag}\,\{\}$. The identity matrix is denoted $\mathbf{I}$, an identity matrix of a specific size is denoted $\mathbf{I}_a$ where a indicates the size (i.e., the number of rows and columns) of $\mathbf{I}$; $\mathbf{1}_a$ is a vector of ones of length a. Matrices denoted as $\mathbf{Z}_*$ are centered and normalized (i.e., each column of $\mathbf{Z}_*$ has mean 0 and norm 1). Italic or bold subscripts of a matrix denote its relationship with an index or another matrix (e.g., matrix $\mathbf{Z}_\mathbf{Y}$ is centered and normalized $\mathbf{Y}$, matrix $\mathbf{W}_K$ denotes the "weights" matrix derived from the K set).

6.2.2 The singular value decomposition

The singular value decomposition (SVD) of a $J \times K$ matrix $\mathbf{R}$ of rank L (with $L \leq \min(J,K)$) is expressed as

$$\mathbf{R} = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}^{\mathrm{T}}, \text{ where } \mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}_L = \mathbf{V}^{\mathrm{T}}\mathbf{V}, \tag{6.1}$$

where $\mathbf{U}$ is the $J \times L$ matrix of the left singular vectors, $\mathbf{V}$ the $K \times L$ matrix of the right singular vectors, and $\boldsymbol{\Delta}$ is an $L \times L$ diagonal matrix whose diagonal contains the singular values (ordered from the largest to the smallest). When squared, the singular values become eigenvalues and so $\boldsymbol{\Lambda} = \mathsf{diag}\,\{\boldsymbol{\Delta}\}^2$ is a diagonal matrix of eigenvalues. The first singular value and pair of singular vectors are solution of the following optimization problem:

$$\delta = \underset{\mathbf{u},\mathbf{v}}{\arg\max} \left(\mathbf{u}^{\mathrm{T}}\mathbf{R}\mathbf{v}\right) \quad \text{under the constraints } \mathbf{u}^{\mathrm{T}}\mathbf{u} = \mathbf{v}^{\mathrm{T}}\mathbf{v} = 1. \tag{6.2}$$

The other pairs of singular vectors are solutions of the same optimization problem with the additional constraint that right (respectively left) singular vectors are orthogonal to all other right (respectively left) singular vectors associated with a larger singular value (see [9–11], for details).

6.2.3 The generalized singular value decomposition

The generalized singular value decomposition (GSVD) generalizes the SVD by imposing, on the left and right singular vectors, orthogonality constraints (also called "metrics") expressed by positive-definite matrices denoted $\mathbf{\Omega}$ and $\mathbf{\Phi}$ [9–11]. The GSVD of a $J \times K$ matrix $\mathbf{R}$ of rank L (with $L \leq \min(J,K)$) is expressed as

$$\mathbf{R} = \mathbf{U\Delta V}^{\mathrm{T}}, \text{ with } \mathbf{U}^{\mathrm{T}}\mathbf{\Omega U} = \mathbf{I}_L = \mathbf{V}^{\mathrm{T}}\mathbf{\Phi V}. \tag{6.3}$$

The first generalized singular value and pair of generalized singular vectors are solution of the following optimization problem (cf. 6.2):

$$\delta = \underset{\mathbf{u},\mathbf{v}}{\arg\max} \left(\mathbf{u}^{\mathrm{T}}\mathbf{Rv}\right) \quad \text{under the constraints } \mathbf{u}^{\mathrm{T}}\mathbf{\Omega u} = \mathbf{v}^{\mathrm{T}}\mathbf{\Phi v} = 1. \tag{6.4}$$

The other pairs of singular vectors are solutions of the same optimization problem with the additional constraint that right (respectively left) singular vectors are $\mathbf{\Omega}$-orthogonal (respectively $\mathbf{\Phi}$–orthogonal) to all other right (respectively left) singular vectors associated with a larger singular value.

In a multivariate framework, factor and component scores are obtained as:

$$\mathbf{F}_J = \mathbf{\Omega U\Delta} \text{ and } \mathbf{F}_K = \mathbf{\Phi V\Delta}. \tag{6.5}$$

Often, the GSVD is expressed via the compact "triplet notation" [12–14] and, for example, with this notation, the GSVD of Equation 6.3 is presented as the analysis of the triplet $(\mathbf{R}, \mathbf{\Phi}, \mathbf{\Omega})$.

6.2.3.1 Principal Components Analysis

PCA analyzes a quantitative data matrix $\mathbf{X}$ with I rows (observations) and J columns (variables) [15]. The matrix $\mathbf{X}$ is first pre-processed such that columns are centered and often normalized (i.e., the sum of squares of each column equals 1). With the centered and normed matrix denoted $\mathbf{Z}_\mathbf{X}$, PCA is then defined as the analysis of the triplet $(\mathbf{Z}_\mathbf{X}, \mathbf{I}_J, \mathbf{I}_I)$.

6.2.3.2 Correspondence Analysis

Correspondence Analysis (CA) is analogous to a PCA but for—typically—contingency tables (i.e., the cross product of two disjunctive data tables; see Table 6.1) [10, 11, 22, 23]. CA requires specific pre-processing and constraints prior to the GSVD step. First, for a matrix $\mathbf{R}$ of size J by K we compute a matrix of *observed* values:

$$\mathbf{O}_\mathbf{R} = N^{-1}\mathbf{R} \tag{6.6}$$

where N is the total sum of $\mathbf{R}$. The row (respectively column) constraint matrix $\mathbf{M}$ (respectively $\mathbf{W}$)) is defined as:

$$\mathbf{m} = \mathbf{O}_\mathbf{R}\mathbf{1}_J \text{ and } \mathbf{M} = \mathsf{diag}\left\{\mathbf{m}\right\}, \tag{6.7}$$

and (respectively) as

$$\mathbf{w} = \mathbf{1}_K\mathbf{O}_\mathbf{R} \text{ and } \mathbf{W} = \mathsf{diag}\left\{\mathbf{w}\right\} \tag{6.8}$$

where $\mathbf{m}$ (respectively $\mathbf{w}$) is the vector of the row (respectively column) sums of $\mathbf{O_R}$. Next, we compute a matrix of *expected*:

$$\mathbf{E_R} = \mathbf{m}\mathbf{w}^{\mathrm{T}}. \tag{6.9}$$

Finally, we compute the matrix of deviations:

$$\mathbf{Z_R} = \mathbf{O_R} - \mathbf{E_R}, \tag{6.10}$$

Finally, the CA of $\mathbf{R}$ is performed from the analysis of the triplet $\left(\mathbf{Z_R}, \mathbf{W}^{-1}, \mathbf{M}^{-1}\right)$.

Table 6.1: Example of nominal data table, and its disjunctive counterpart.

(a) Nominal

	Variable 1	...	Variable *J*
Subj.1	A	...	A
Subj.2	A	...	A
...	...	...	...
Subj.I-1	B	...	C
Subj.I	C	...	B

(b) Disjunctive

	Variable 1			...	Variable *J*		
	A	B	C		A	B	C
Subj.1	1	0	0	...	1	0	0
Subj.2	1	0	0	...	1	0	0
...	...	...	...	...	...	...	...
Subj.I-1	0	1	0	...	0	0	1
Subj.I	0	0	1	...	0	1	0

6.2.3.3 Multiple Correspondence Analysis

Multiple correspondence analysis (MCA) is a specific version of CA applied to a single disjunctive data table (see Table 6.1). MCA can be carried out by following the steps of CA as outlined in Section 6.2.3.2. However, there are several ways to define MCA as a centered, non-normalized, and weighted PCA [21]. Here, we provide another alternative MCA formulation.

Given a matrix $\mathbf{N}$ with I rows as observations and N nominal columns (see Table 6.1a.), we then transform $\mathbf{N}$ into the disjunctive formated (see Table 6.1b.) matrix $\mathbf{R}$, which has I rows and J columns. First, we define the constraints, where R is the sum of $\mathbf{R}$:

$$\mathbf{M} = I\mathbf{I}_I \text{ and } \mathbf{m} = \mathsf{diag}\left\{\mathbf{M}\right\} \tag{6.11}$$

$$\mathbf{w} = R^{-1}\left(\mathbf{1}_J\mathbf{R}\right) \text{ and } \mathbf{W} = \mathsf{diag}\left\{\mathbf{w}\right\}. \tag{6.12}$$

MCA can now be performed as the analysis of triplet: $\left(R^{-1}\mathbf{Z_R}, \mathbf{W}^{-1}, \mathbf{M}\right)$, where $\mathbf{Z_R}$ is the centered non-normalized version of $\mathbf{R}$.

PCA and MCA are equivalent when all variables have *exactly* two levels [10, 11]. For example "yes" vs. "no" which would be coded as [1 0] and [0 1], respectively. The equivalence holds in the following case. Traditional MCA—as performed via CA—would be applied to the complete disjunctive matrix (which represents all levels), whereas PCA would be applied to a *strictly binary* table where each variable is repre-

sented by only 1 column. In this case, for example, "yes" is denoted with a 1 whereas "no" is denoted with a 0 (essentially, just half of the usual table for MCA).

6.2.4 Partial Least Squares Correlation

Partial Least Squares Correlation (PLSC) [24–27] exists under a wide varieties of other appellations such as the SVD of two covariance fields [28], PLS-SVD [29], canonical covariance analysis [30], or co-inertia analysis [13], but is probably best traced back to Tucker's inter-battery factor analysis [31]—a method that analyzes the information common to two data tables measured on the same set of observations. Given two matrices, $\mathbf{X}$ and $\mathbf{Y}$, each containing I rows (observations) with (respectively) J columns ($\mathbf{X}$'s variables) and K columns ($\mathbf{Y}$'s variables), the matrices $\mathbf{Z_X}$ and $\mathbf{Z_Y}$ are the centered and unitary normed versions of $\mathbf{X}$ and $\mathbf{Y}$. With $\mathbf{Z_R} = \mathbf{Z_X}^{\mathrm{T}}\mathbf{Z_Y}$, PLSC is then defined as the analysis of the triplet $(\mathbf{Z_R}, \mathbf{W_Y}, \mathbf{W_X})$ where $\mathbf{W_X} = \mathbf{I}_J$ and $\mathbf{W_Y} = \mathbf{I}_K$, PLSC extracts the information common to $\mathbf{X}$ and $\mathbf{Y}$ by computing two sets of latent variables defined as:

$$\mathbf{L_X} = \mathbf{Z_X}\mathbf{W_X}\mathbf{U} \text{ and } \mathbf{L_Y} = \mathbf{Z_Y}\mathbf{W_Y}\mathbf{V} \tag{6.13}$$

In PLSCA, associated latent variables have maximal covariance. Specifically, call $\mathbf{u}_\ell$ and $\mathbf{v}_\ell$ the linear transformation coefficients for $\mathbf{Z_X}$ and $\mathbf{Z_Y}$ respectively. A latent variable for each matrix is defined as $\mathbf{l_X} = \mathbf{Z_X}\mathbf{W_X}\mathbf{u}_\ell$ and $\mathbf{l_Y} = \mathbf{Z_Y}\mathbf{W_Y}\mathbf{v}_\ell$ where

$$\underset{\mathbf{u},\mathbf{v}}{\arg\max}\left(\mathbf{l_X}^{\mathrm{T}}\mathbf{l_Y}\right) = \underset{\mathbf{u},\mathbf{v}}{\arg\max}\,\mathrm{cov}\left(\mathbf{l_X},\mathbf{l_Y}\right), \tag{6.14}$$

under the constraints that $\mathbf{u}_\ell$ and $\mathbf{v}_\ell$ have unit norm:

$$\mathbf{u}_\ell{}^{\mathrm{T}}\mathbf{W_X}\mathbf{u}_\ell = 1 = \mathbf{v}_\ell{}^{\mathrm{T}}\mathbf{W_Y}\mathbf{v}_\ell. \tag{6.15}$$

After the ℓ-th pair of latent variables are extracted, the subsequent ones are extracted under the additional constraint of orthogonality:

$$\mathbf{l_X}_\ell{}^{\mathrm{T}}\mathbf{l_Y}_{\ell'} = 0 \text{ when } \ell \neq \ell'. \tag{6.16}$$

Each successive $\mathbf{l_X}$ and $\mathbf{l_Y}$ is stored in $\mathbf{L_X}$ and $\mathbf{L_Y}$, respectively, where

$$\mathbf{L_X}^{\mathrm{T}}\mathbf{L_Y} = \mathbf{U}^{\mathrm{T}}\mathbf{W_X}\mathbf{Z_X}^{\mathrm{T}}\mathbf{Z_Y}\mathbf{W_Y}\mathbf{V} = \mathbf{U}^{\mathrm{T}}\mathbf{W_X}\mathbf{Z_R}\mathbf{W_Y}\mathbf{V} = \mathbf{U}^{\mathrm{T}}\mathbf{W_X}\mathbf{U}\boldsymbol{\Delta}\mathbf{V}^{\mathrm{T}}\mathbf{W_Y}\mathbf{V} = \boldsymbol{\Delta}, \tag{6.17}$$

because $\mathbf{U}^{\mathrm{T}}\mathbf{W_X}\mathbf{U} = \mathbf{I}_L = \mathbf{V}^{\mathrm{T}}\mathbf{W_Y}\mathbf{V}$ (where L is the rank of $\mathbf{Z_R}$). The latent variables of PLSC maximize the covariance as expressed by the singular values (for proofs, see [27, 31]).

6.2.5 Partial Least Squares-Correspondence Analysis

Recently, we presented a PLSC method designed specifically for the analysis of two categorical data matrices: Partial Least Squares-Correspondence Analysis (PLSCA)—a technique that combines features of PLSC and CA [32]. PLSCA can be expressed as follows: $\mathbf{X}$ and $\mathbf{Y}$ are disjunctive matrices where $\mathbf{R} = \mathbf{X}^{\mathrm{T}}\mathbf{Y}$ is a contingency table. CA, as defined in Section 6.2.3.2, is applied to $\mathbf{R}$. The latent variables in PLSCA are

computed according to Equation 6.13, where:

$$\mathbf{Z_X} = I^{\frac{1}{2}} X^{-1} \mathbf{X} \tag{6.18}$$

$$\mathbf{Z_Y} = I^{\frac{1}{2}} Y^{-1} \mathbf{Y} \tag{6.19}$$

where X and Y are (respectively) the sums of $\mathbf{X}$ and $\mathbf{Y}$, and where $\mathbf{W_X}$ and $\mathbf{W_Y}$ are computed from Equations 6.7 and 6.8 (i.e., $\mathbf{M}$ and $\mathbf{W}$).

6.3 PLSC for mixed data types

Here we establish a framework for PLSC that applies to mixed data types. We formalize this approach with respect to one table of continuous data and one table of categorical data. Categorical data can be treated as continuous data and analyzed with PCA to produce identical results to a MCA (see Section 6.2.3.3).

6.3.1 Escofier-style transform for PCA

In 1979, Brigitte Escofier presented a technique to analyze continuous data with CA to produce the same results as PCA (within a scaling factor) [33]. Escofier showed that a quantitative variable, say $\mathbf{x}$ (i.e., a column from the matrix $\mathbf{X}$) that is centered with unitary norm, can be analyzed with CA if it is expressed as two vectors: $\frac{1-\mathbf{x}}{2}$ and $\frac{1+\mathbf{x}}{2}$ (see Table 6.2). Incidentally, dividing each set by 2 with this Escofier-style coding is superfluous when using the stochastic version of CA (see Section 6.2.3.2).

Table 6.2: Example of Escofier's coding scheme of continuous data to perform a CA on continuous data. $\mathbf{x}_j$ denotes the j vector from a matrix $\mathbf{X}$ where $x_{i,j}$ denotes a specific value at row i and column j. This coding scheme is similar to the thermometer coding scheme often used for ordinal data in MCA.

(a) Continuous data

	$\mathbf{x}_1$	…	$\mathbf{x}_J$
Subj.1	$x_{1,1}$	…	$x_{1,J}$
Subj.2	$x_{2,1}$	…	$x_{2,J}$
…	…	…	…
Subj.I-1	$x_{I-1,1}$	…	$x_{I-1,J}$
Subj.I	$x_{I,1}$	…	$x_{I,J}$

(b) Escofier-style transform

	$-\mathbf{x}_1$	$+\mathbf{x}_1$	…	$-\mathbf{x}_J$	$+\mathbf{x}_J$
Subj.1	$\frac{1-x_{1,1}}{2}$	$\frac{1+x_{1,1}}{2}$	…	$\frac{1-x_{1,J}}{2}$	$\frac{1+x_{1,J}}{2}$
Subj.2	$\frac{1-x_{2,1}}{2}$	$\frac{1+x_{2,1}}{2}$	…	$\frac{1-x_{2,J}}{2}$	$\frac{1+x_{2,J}}{2}$
…	…	…	…	…	…
Subj.I-1	$\frac{1-x_{I-1,1}}{2}$	$\frac{1+x_{I-1,1}}{2}$	…	$\frac{1-x_{I-1,J}}{2}$	$\frac{1+x_{I-1,J}}{2}$
Subj.I	$\frac{1-x_{I,1}}{2}$	$\frac{1+x_{I,1}}{2}$	…	$\frac{1-x_{I,J}}{2}$	$\frac{1+x_{I,J}}{2}$

Call $\mathbf{Z_X}$ the centered and unitary norm version of $\mathbf{X}$ with I rows and J observations, where $\mathbf{B}_- = 1 - \mathbf{Z_R}$ and $\mathbf{B}_+ = 1 + \mathbf{Z_R}$ where

$$\mathbf{B} = \left[\mathbf{B}_- \; \mathbf{B}_+\right], \tag{6.20}$$

The matrix $\mathbf{B}$ can then be analyzed with CA (as in Section 6.2.3.2) which is equivalent to a PCA via the analysis of the triplet: $\left(\frac{1}{IJ}\mathbf{Z_X}, \mathbf{1}_J, \mathbf{1}_I\right)$.

There is one exception to the equivalence between these two methods: in the Escofier-style approach, the number of columns in $\mathbf{B}$ is $2J$, where J is the number of columns in $\mathbf{X}$. Each variable from $\mathbf{X}$ has essentially been duplicated in $\mathbf{B}$ much like "thermometer coding" (a.k.a. doubling or fuzzy coding [34]) for ordinal data analysis with MCA. Thermometer coding expresses each variable by two points that are equidistant from 0 (i.e., the mean).

6.3.2 Escofier-style transform for PLSC

To formalize PLSC for mixed data types, we first, define PLSC approach for 2 continuous data matrices—$\mathbf{X}$ and $\mathbf{Y}$—but in the Escofier framework (Section 6.3.1 and also see Table 6.2). Let us call $\mathbf{B_X}$ the Escofier-style transform of $\mathbf{X}$ and $\mathbf{B_Y}$ the Escofier-style transform of $\mathbf{Y}$. If we use the standard form of PLSC, we decompose $\mathbf{B_R} = \mathbf{B_X^T B_Y}$, where:

$$\mathbf{B_R} = \begin{bmatrix} \left(\mathbf{B}_{\mathbf{X}-}^{\mathrm{T}}\mathbf{B}_{\mathbf{Y}-}\right) & \left(\mathbf{B}_{\mathbf{X}-}^{\mathrm{T}}\mathbf{B}_{\mathbf{Y}+}\right) \\ \left(\mathbf{B}_{\mathbf{X}+}^{\mathrm{T}}\mathbf{B}_{\mathbf{Y}-}\right) & \left(\mathbf{B}_{\mathbf{X}+}^{\mathrm{T}}\mathbf{B}_{\mathbf{Y}+}\right) \end{bmatrix}. \tag{6.21}$$

Because $\mathbf{B_X}$ and $\mathbf{B_Y}$ are each in the Escofier-style (i.e., pseudo-categorical), this problem can be treated as one tailored for PLSCA (i.e., PLSC for the two *categorical* matrices; see Section 6.2.5). The PLSCA of $\mathbf{B_X^T B_Y}$ is equivalent to the PLSC (see Section 6.2.4) of $\mathbf{Z_X}$ and $\mathbf{Z_Y}$ (within scaling factors). There are three items used to define equivalence between these approaches: (1) singular values, (2) component scores (for both rows and columns), and (3) latent variables.

Call (respectively) $\mathbf{\Delta}_{Z_R}$ and $\mathbf{\Delta}_{B_R}$ the singular values from a standard PLSC (of $\mathbf{Z_X}$ and $\mathbf{Z_Y}$) and the singular values from an Escofier-style PLSCA (of $\mathbf{B_X}$ and $\mathbf{B_Y}$). We use the Escofier style approach as the preferred method because, as an extension of CA, it provides a natural dual representation of the rows and columns. To transition between the two approaches, we do the following:

$$\mathbf{\Delta}_{B_R} = \frac{1}{I\sqrt{JK}}\mathbf{\Delta}_{Z_R}. \tag{6.22}$$

The transition between component scores is also defined as follows:

$$\mathbf{F}_{J\mathbf{B_R}} = \frac{1}{\frac{I}{J}\sqrt{J^2K}}\left[-\mathbf{F}_{J\mathbf{Z_R}} \quad \mathbf{F}_{J\mathbf{Z_R}}\right] \tag{6.23}$$

$$\mathbf{F}_{K\mathbf{B_R}} = \frac{1}{\frac{I}{K}\sqrt{K^2J}}\left[-\mathbf{F}_{K\mathbf{Z_R}} \quad \mathbf{F}_{K\mathbf{Z_R}}\right]. \tag{6.24}$$

And finally, the transition between latent variables are:

$$\mathbf{L}_{\mathbf{B_X}} = \sqrt{IJ}\,\mathbf{L}_{\mathbf{Z_X}} \tag{6.25}$$

$$\mathbf{L}_{\mathbf{B_Y}} = \sqrt{IK}\,\mathbf{L}_{\mathbf{Z_Y}}, \tag{6.26}$$

where the latent variables for the "standard" approach are defined as in Section 6.13, and the computation of latent variables for the Escofier-approach are defined as those for PLSCA in Section 6.2.5.

We have to duplicate the component scores from the standard PLSC and multiply by -1 because the Escofier-style transform is a "thermometer" style coding of the data (equidistant above and below 0; see Table 6.2). Given these properties, we can compute the standard PLSC with equivalence to the PLSCA via the GSVD as follows. First define $\mathbf{Z}_{\mathbf{X}*}$ and $\mathbf{Z}_{\mathbf{Y}*}$:

$$\mathbf{Z}_{\mathbf{X}*} = J^{-1}\sqrt{\frac{1}{I}}\,\mathbf{Z}_{\mathbf{X}} \tag{6.27}$$

$$\mathbf{Z}_{\mathbf{Y}*} = K^{-1}\sqrt{\frac{1}{I}}\,\mathbf{Z}_{\mathbf{Y}} \tag{6.28}$$

$$\mathbf{Z}_{\mathbf{R}*} = \mathbf{Z}_{\mathbf{X}*}^{\mathrm{T}}\mathbf{Z}_{\mathbf{Y}*}, \tag{6.29}$$

where (1) $\mathbf{Z}_{\mathbf{X}}$ and $\mathbf{Z}_{\mathbf{Y}}$ are centered and normed matrices of (respectively) $\mathbf{X}$ and $\mathbf{Y}$, (2) I are the number of rows (observations) in $\mathbf{X}$ and $\mathbf{Y}$, and (3) J and K are the number of columns (variables) for $\mathbf{X}$ and $\mathbf{Y}$. To produce the same results as the Escofier-style PLSC approach, the GSVD is described by the triplet: $(\mathbf{Z}_{\mathbf{R}*}, K\mathbf{I}_K, J\mathbf{I}_J)$. Thus, for continuous data, we can transition between the standard approach to PLSC (see Section 6.2.4) and the Escofier-style approach to PLSCA (see Section 6.2.5).

6.3.3 Mixed Data and PLSC

The Escofier-style transformed matrix (see Table 6.2) is similar to a fully disjunctive matrix; and, because PLSC and PLSCA are equivalent when using Escofier-style pseudo-categorical matrices, we can use PLSCA to analyze mixed data types (i.e., one matrix of continuous data and one matrix of categorical data).

Call $\mathbf{B}_{\mathbf{Y}}$ the Escofier-style transform of a continuous data matrix $\mathbf{Y}$ and call $\mathbf{X}$ a fully disjunctive data matrix (as in Table 6.1). Because both matrices are in a categorical or pseudo-categorical format, we can define $\mathbf{R} = \mathbf{X}^{\mathrm{T}}\mathbf{B}_{\mathbf{Y}}$ as a pseudo-contingency table since this $\mathbf{R}$ is the cross-product between a categorical matrix and a pseudo-categorical matrix. In fact, $\mathbf{R}$ expresses some of the properties we would expect from a contingency table but maintains the properties of $\mathbf{X}$ and $\mathbf{B}_{\mathbf{Y}}$: the column sums of $\mathbf{R}$ are equal to one another—just as in $\mathbf{B}_{\mathbf{Y}}$ and are also proportional to the column sums of $\mathbf{B}_{\mathbf{Y}}$. This is also true for the row sums of $\mathbf{R}$ and the column sums of $\mathbf{X}$. Thus, the relationship between $\mathbf{X}$ and $\mathbf{B}_{\mathbf{Y}}$ can be analyzed with PLSCA (see Section 6.2.5) and the properties that define PLSC still hold (see Sections 6.2.4 and 6.2.5).

However, there is a minor drawback to this approach: The continuous data matrix, $\mathbf{Y}$, represents each variable twice in $\mathbf{B}_{\mathbf{Y}}$ (see Section 6.2) and this could be problematic for very large data sets (e.g., neuroimaging, genomics). Thus, we now define a mixed data approach to PLS closer to PLSC, but that keeps key properties of CA (i.e., dual representation, distributional equivalence, emphasis on rare occurrences). Call $\mathbf{Y}$ a data matrix, with J columns, of continuous data where $\mathbf{Z}_{\mathbf{Y}}$ is centered and normalized. Call $\mathbf{X}$ a fully disjunctive matrix, with K columns from N variables, where $\mathbf{Z}_{\mathbf{X}}$ is centered but not normalized. Both $\mathbf{X}$ and $\mathbf{Y}$ have I rows (i.e., observations). First we define the data matrices derived from $\mathbf{X}$ and $\mathbf{Y}$:

$$\mathbf{Z}_{\mathbf{X}*} = N^{-1}\sqrt{\frac{1}{I}}\,\mathbf{Z}_{\mathbf{X}} \tag{6.30}$$

$$\mathbf{Z}_{\mathbf{Y}*} = K^{-1}\sqrt{\frac{1}{I}}\,\mathbf{Z}_{\mathbf{Y}}. \tag{6.31}$$

Next, we define weights associated to each set (where X is the sum of $\mathbf{X}$):

$$\mathbf{w}_{\mathbf{X}} = X^{-1}\mathbf{1}_I\mathbf{X} \text{ and } \mathbf{W}_{\mathbf{X}} = \mathsf{diag}\left\{\mathbf{w}_{\mathbf{X}}\right\}, \tag{6.32}$$

and $\mathbf{W}_{\mathbf{Y}} = K\mathbf{I}_K$. PLSC can then be performed on $\mathbf{Z}_{\mathbf{Y}*}$ and $\mathbf{Z}_{\mathbf{X}*}$ where $\mathbf{W}_{\mathbf{X}}$ and $\mathbf{W}_{\mathbf{Y}}$ are constraints for the GSVD. The GSVD step of PLSC in this case would analyze the triplet: $\left(\mathbf{R}, \mathbf{W}_{\mathbf{Y}}, \mathbf{W}_{\mathbf{X}}^{-1}\right)$ with $\mathbf{R} = \mathbf{Z}_{\mathbf{X}*}^{\mathrm{T}}\mathbf{Z}_{\mathbf{Y}*}$. This approach is derived, in part, from MCA where MCA is treated as a centered, non-normalized, weighted PCA (see Section 6.2.3.3) and the standard approach to PLSC (see Section 6.2.4). We also imposed particular constraints on this formulation so that the results here would be equivalent to those done on $\mathbf{X}$ and $\mathbf{B}_{\mathbf{Y}}$ obtained with PLSCA. However, there is also a drawback to this reformulation: supplemental projections are more difficult to compute than in the CA approach. Therefore, we define MiMoPLS in one, final, way that combines the simplicity of the PLSCA approach with the minimally required data in the PLSC approach.

First, $\mathbf{X}$ is the complete disjunctive matrix where $\mathbf{B}_{\mathbf{Y}+} = \mathbf{Z}_{\mathbf{Y}} + 1$ (see Eq. 6.21 and Section 6.3.3), and $\mathbf{R} = \mathbf{X}^{\mathrm{T}}\mathbf{B}_{\mathbf{Y}+}$. The total sum of $\mathbf{B}_{\mathbf{Y}+}$ is equal to IK, where I is the number of observations and K is the number of columns in $\mathbf{Y}$ and we then use CA (see Section 6.2.3.2) where both $\mathbf{w}_{\mathbf{X}}$ and $\mathbf{W}_{\mathbf{X}}$ are obtained from Equation 6.32, and where $\mathbf{W}_{\mathbf{Y}} = K^{-1}\mathbf{I}_K$ where $\mathbf{w}_{\mathbf{Y}} = \mathsf{diag}\left\{\mathbf{W}_{\mathbf{Y}}\right\}$. Next we define the *observed*, *expected*, and *deviations* matrices (with R being the sum of all elements of $\mathbf{R}$):

$$\mathbf{O}_{\mathbf{R}} = R^{-1}\mathbf{R} \tag{6.33}$$

$$\mathbf{E}_{\mathbf{R}} = \mathbf{w}_{\mathbf{X}}\mathbf{w}_{\mathbf{Y}}^{\mathrm{T}} \tag{6.34}$$

$$\mathbf{Z}_{\mathbf{R}} = \mathbf{O}_{\mathbf{R}} - \mathbf{E}_{\mathbf{R}}. \tag{6.35}$$

The GSVD step then correspond to the analysis of the triplet $\left(\mathbf{Z}_{\mathbf{R}}, \mathbf{W}_{\mathbf{Y}}^{-1}, \mathbf{W}_{\mathbf{X}}^{-1}\right)$. Finally, the latent variables are computed as:

$$\mathbf{L}_{\mathbf{X}} = \left(I^{\frac{1}{2}}X^{-1}\mathbf{X}\right)\mathbf{W}_{\mathbf{X}}^{-1}\mathbf{U} \tag{6.36}$$

$$\mathbf{L}_{\mathbf{X}} = \left(I^{\frac{1}{2}}B^{-1}\mathbf{Z}_{\mathbf{Y}}\right)\mathbf{W}_{\mathbf{Y}}^{-1}\mathbf{V}, \tag{6.37}$$

where $\mathbf{Z}_{\mathbf{Y}}$ is the column centered and normalized version of $\mathbf{Y}$, and where X and B are (respectively) the sums of $\mathbf{X}$ and $\mathbf{B}_{\mathbf{Y}+}$. Recall that X is equal to IN, where N is the number of variables in $\mathbf{X}$, and B is equal to IK and this makes Equations. 6.36 analogous to the computation of the "observed" values in CA (see Section 6.2.3.2).

We now have an approach of analyzing mixed data types that (1) is in the PLSC fashion, (2) maintains the properties of PLSCA and CA (e.g., dual representation, simple supplemental projections), and (3) does not duplicate the representation of the continuous data matrix.

6.4 An Application to Alzheimer's Disease

We illustrate MiMoPLS with a data set—from the Alzheimer's Disease Neuroimaging Initiative (ADNI)—that contains brain imaging data obtained from diffusion tensor imaging (DTI)—as measured with fractional anisotropy (FA)—and genetic data obtained from single nuclear polymorphisms (SNPs). These data 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 AD. The ADNI project is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations. Michael W. Weiner, MD (VA Medical Center and University of California San Francisco) is the ADNI PI. Subjects have been recruited from over 50 sites across the U.S. and Canada (for up-to-date information, see `www.adni-info.org`).

Participants include 29 individuals from the ADNI2 cohort classified into 4 groups: control ($N = 9$; CON), early mild cognitive impairment ($N = 11$; eMCI), late mild cognitive impairment ($N = 4$; ℓMCI), and Alzheimer's Disease ($N = 5$; AD). All participants were genotyped with genome-wide SNPs (Illumina HumanOmniExpress). SNPs underwent standard preprocessing (SNP & participant call rates were $\geq 90\%$, Hardy-Weinberg disequilibrium $\leq 1 \times 10^{-6}$, and minor allele frequency $\leq 5\%$). From the genome-wide data, we extracted 386 SNPs that, according the literature and aggregate sources [35], should be associated with AD. We also extracted 35,062 voxels, which contained FA values, from 48 white matter tracts according to the JHU-ICBM-DTI-81 mask (see Figure 6.1) [36]. We analyzed these data to identify the genetic contributions to white matter changes in an AD related population.

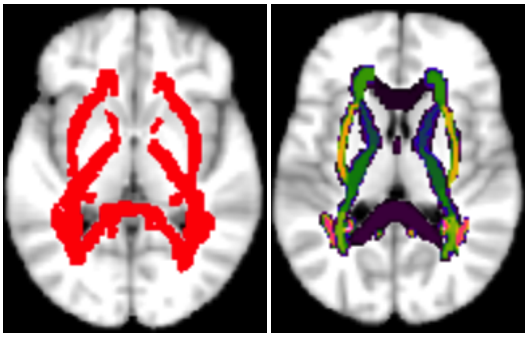

Fig. 6.1 Masks to identify white matter regions in a common (MNI) space. The left figure illustrates all the voxels included, whereas the right figure illustrates the separate tracts within this mask.

We present the analysis first with the descriptive component maps (Figure 6.2). For illustrative purposes, we limit discussion to only the first two components. We can note that there is a higher variability of SNP-zygotes (top left; Figure 6.2) than the FA values (top right; Figure 6.2). Interpretation of these maps are done as they would be in CA: a SNP-zygote that is close to particular voxels is considered more related to those voxels than is the average SNP-zygotes.

The latent variables suggest two interpretations of the components. First, Component 1 largely reflects the differences between ℓMCI (left side of Component 1) and AD (right side of Component 1), whereas Component 2 is characterized by {CON & eMCI} vs. {ℓMCI & AD}. This pattern suggests that Component 1 separates real AD pathology from possible misdiagnoses, whereas Component 2 appears to characterize non-pathological to pathological features. Further, we can interpret the latent variables (bottom; Figure 6.2)

as we would in both CA and in PLSC. Participants whose scores are closer to particular SNP-zygotes or FA values are more associated with those features than the average participant. Furthermore, we can include more meaningful information (e.g., group averages) to better understand the relationship between SNP-zygotes and white matter integrity. Doing so indicates that the CON group is associated with the upper left quadrant, the AD group is associated with the lower right quadrant, the eMCI group is associated with the upper right quadrant, and the ℓMCI group is associated with the lower left quadrant. Thus, we can infer that particular SNP-zygotes and voxels are more associated to these groups than others. However, given that there are so many SNP-zygotes and voxels, we use inferential methods to eliminate non-significant SNP-zygotes and voxels.

One approach is to use the bootstrap [37, 38] and to compute bootstrap ratio values (BSRs) [24] which are *t* statistics computed from the mean and standard deviation of the bootstrap distribution. With BSRs, we can reduce the number of items to interpret by selecting only the items that significantly contribute to the component structure (see Figure 6.3): Here we only show items whose BSR magnitude is larger than 2.50. SNPs are labeled by the Gene with which they are most associated, the voxels are plotted in standard MNI brain maps.

We first interpret the brain images (because more is known about white matter integrity than genetics in clinical populations); they provide a baseline from which a genetic relationship can be inferred. Component 1 (Figure 6.4; lower left) shows small clusters in bilateral superior corona radiata and posterior internal capsule (blue colored voxels), whereas there are large clusters throughout anterior white matter tracts (i.e., genu and body of corpus callosum, internal and external capsule, and corona radiata; denoted with red voxels). Component 2, generally only has negative BSR values. The voxels (denoted in red) trace a path from lateral temporal lobe, to longitudinal tracts leading to frontal regions (i.e., internal and external capsuale, and corona radiata). Taken in context with the latent variables (Figure 6.2), changes in white matter in anterior tracts are more associated with AD, whereas longitudinal tracts are more associated with ℓMCI. This pattern suggests that early biomarkers indicate the progression from ℓMCI to AD and, overall, as indicated by Figure 6.4 that particular markers are associated with specific clinical groups: For example, UCK2 heterozygotes are more associated with ℓMCI whereas UCK2 major homozygotes are more associated with AD. Component 2 identifies fiber paths that interconnect temporal, parietal, and frontal regions—all regions often implicated in the progression of Alzheimer's pathology. Taken in context with the latent variables (Figure 6.2), this pattern suggest that there are substantial changes in these regions in late stage (ℓMCI) and pathological (AD) groups. Finally Figure 6.4 shows that the heterozygote SNP associated with ZNF423 and the minor homozygote of a SNP associated with APOE—a pattern that confirms the importance of these two genes routinely associated with AD.

6.5 Conclusion

This article presents a new approach to PLS that integrates mixed-data types. Our presentation included continuous (brain imaging) and categorical (SNPs) data, but the method can be easily extended to ordinal data (via thermometer coding, see Section 6.3.1). Though we present MiMoPLS via PLSC, MiMoPLS can easily be extended to other PLS approaches (e.g., regression, path-modeling). Future work includes regularization and sparsification designed specifically for block-wise categorical data [39] and two-way sparsification of the SVD [40].

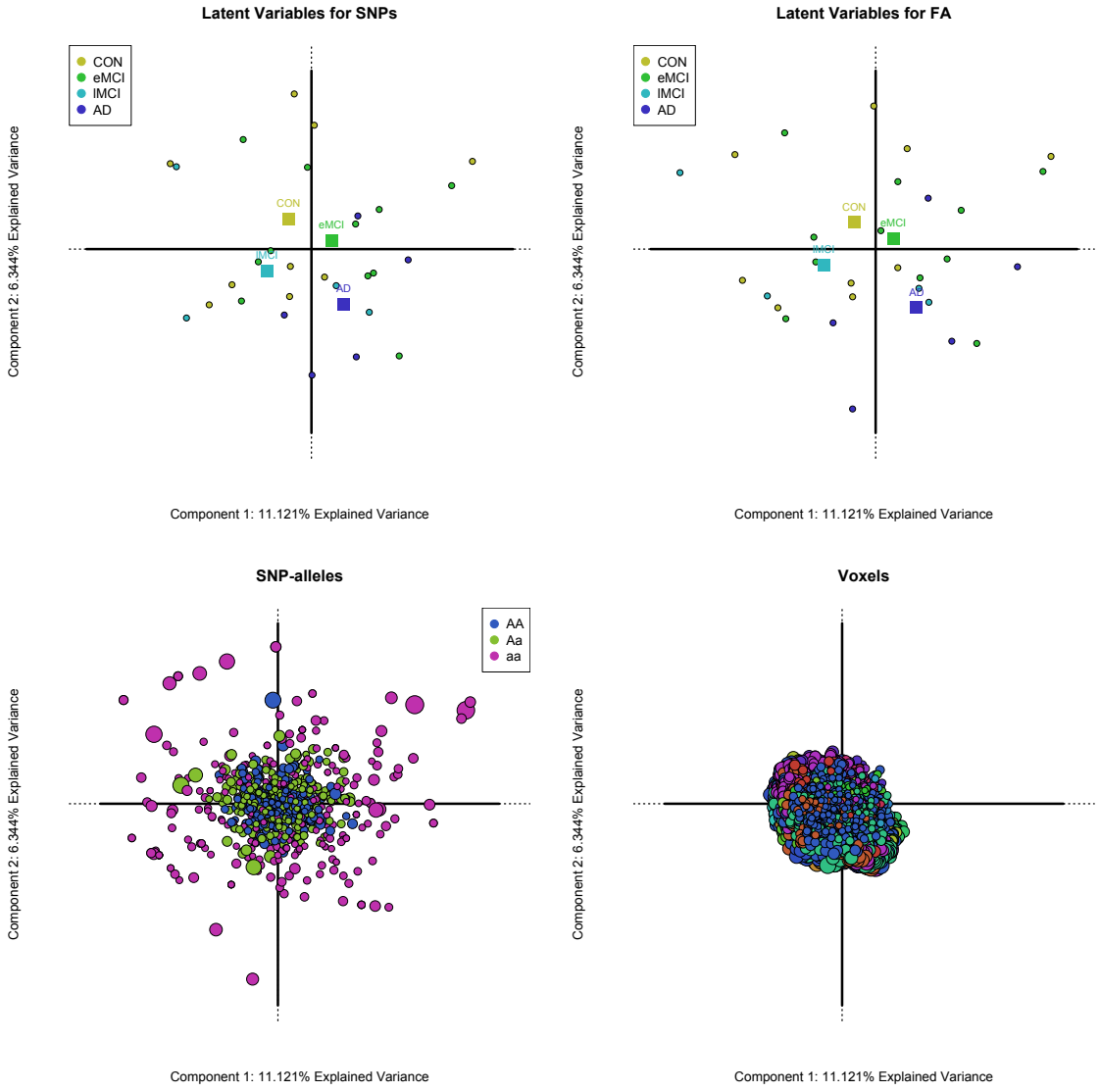

Fig. 6.2: Top figures show the individual participants' scores (latent variables) with respect to the SNP-zygotes (left) and FA values (right). The average of each participant group is labeled with a large square, whereas participants are labeled with small circles. Bottom figures show the component scores of the SNP-zygotes (left; colored by zygote), and the voxels (right; colored by tract).

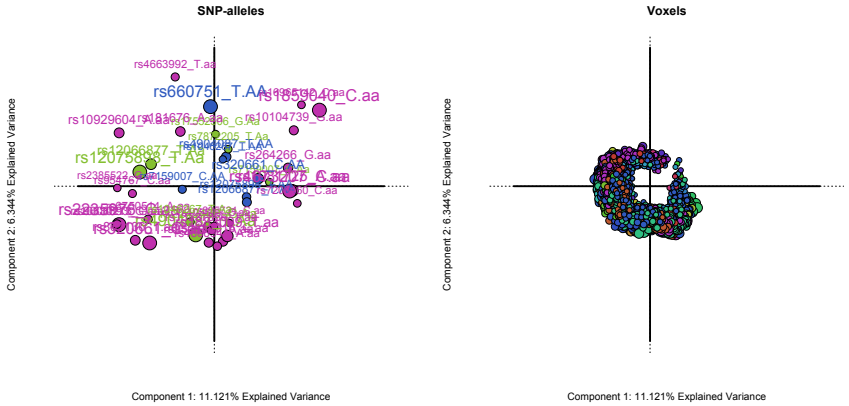

Fig. 6.3: Bootstrap ratios identify items that significantly contribute to the component structure.

Acknowledgements **DB** is currently supported via training grant by the NIH and National Institute on Drug Abuse (F31DA035039).
FMF is currently supported by the NIH and National Institute on Drug Abuse (R01DA030344).
HA 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+)."
ADNI: Data collection and sharing for this project was funded by the ADNI (NIH Grant U01 AG024904) and DOD ADNI (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 Healthcare; ; 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.

References

1. A. Meyer-Lindenberg, "The future of *f*MRI and genetics research," *NeuroImage* **62**, pp. 1286–1292, 2012.
2. P.M. Thompson, N.G. Martin and M.J. Wright, "Imaging genomics," *Current Opinion in Neurology* **23**, pp. 368–373, 2010.
3. J. Liu and V.D. Calhoun, "A review of multivariate analyses in imaging genetics," *Frontiers in Neuroinformatics* **8**, pp. 29, 2014.
4. M. Vounou, T.E. Nichols, and G. Montana, "Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach," *NeuroImage* **53**, pp. 1147–1159, 2010.
5. M.A. Zapala and N.J. Schork, "Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables," *Proceedings of the National Academy of Sciences* **103**, pp. 19430–19435, 2006.
6. S.A. Meda, K. Jagannathan, J. Gelernter, V.D. Calhoun, J. Liu, M.C. Stevens, and G.D. Pearlson, "A pilot multivariate parallel ICA study to investigate differential linkage between neural networks and genetic profiles in schizophrenia," *NeuroImage* **53**, pp. 1007–1015, 2010.

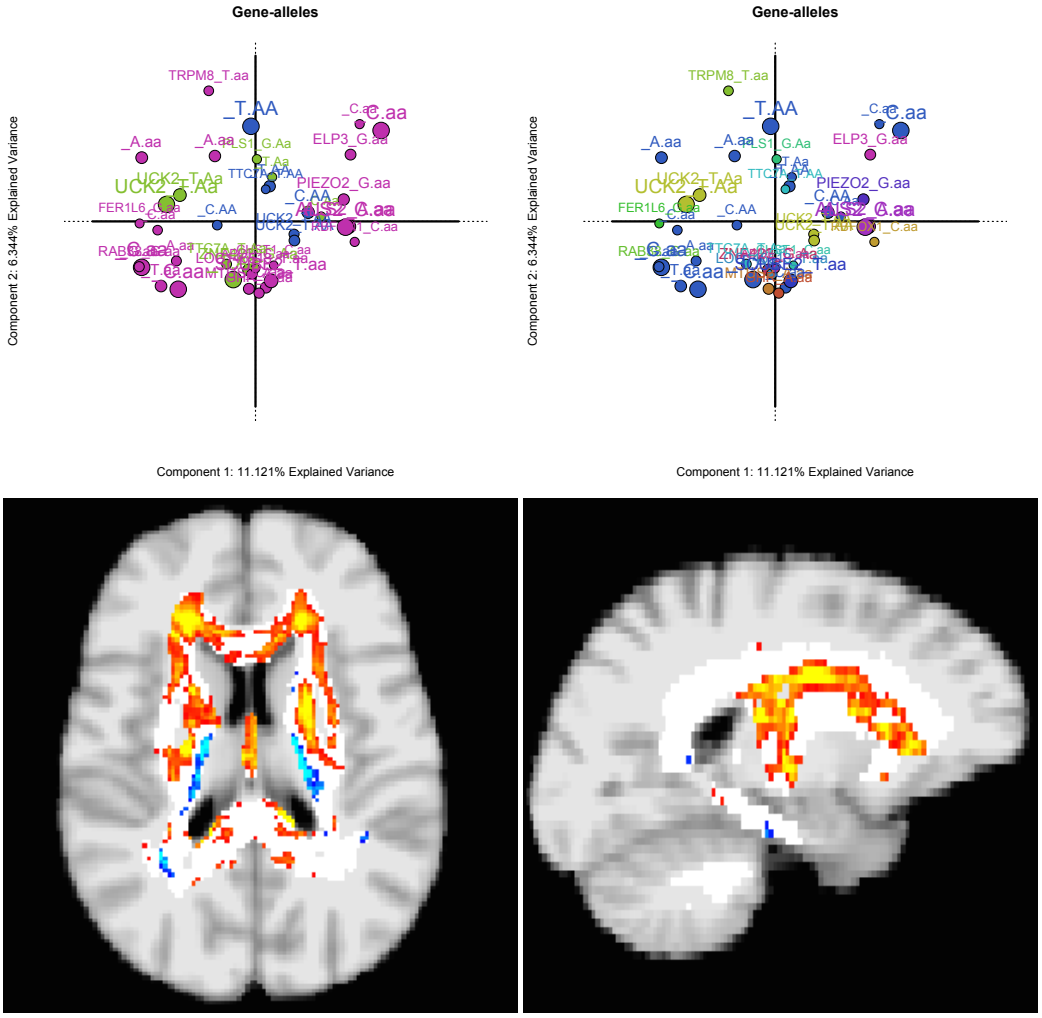

Fig. 6.4: All SNP-zygotes have been renamed to show which gene they are most associated with. SNP-zygotes colored by their zygote (top left) and their respective genes (top right). Bootstrap ratio values are plotted in the voxels (bottom) to indicate their location and the strength of their contribution.

7. J. Liu, G. Pearlson, A. Windemuth, G. Ruano, N.I. Perrone-Bizzozero, and V. Calhoun, "Combining *f*MRI and SNP data to investigate connections between brain function and genetics using parallel ICA," *Human Brain Mapping* **30**, pp. 241–255, 2009.
8. J. Sheng, S., Kim, J. Yan, J. Moore, A. Saykin, and L. Shen, "Data synthesis and method evaluation for brain imaging genetics," *In 2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*, pp. 1202–1205, 2014.
9. H. Abdi, "Singular value decomposition (SVD) and generalized singular value decomposition (GSVD)," in *Encyclopedia of Measurement and Statistics*, N. Salkind, ed., pp. 907–912, Sage, Thousand Oaks, (CA), 2007.
10. M. J. Greenacre, *Theory and Applications of Correspondence Analysis*, London: Academic Press, 1984.
11. L. Lebart, A. Morineau, and K.M. Warwick, *Multivariate Descriptive Statistical Analysis: Correspondence Analysis and Related Techniques for Large Matrices*, New York: Wiley, 1984.

12. Y. Escoufier, "Operators related to a data matrix: a survey," in *COMPSTAT: 17th symposium proceedings in computational statistics (Rome, Italy, 2006)*, A. Rizzi and M. Vichi, eds., pp 285–297. New York: Physica Verlag, 2006.
13. S. Dray, "Analyzing a pair of tables: co-inertia analysis and duality diagrams," in *Visualization and Verbalization of Data*, in J. Blasius and M. Greenacre, eds., London: CRC Press, pp. 289–300, 2014.
14. O. De la Cruz, and S.P. Holmes, "The duality diagram in data analysis: examples of modern applications," *Annals of Applied Statistics* **5**, pp. 2266–2277, 2010.
15. H. Abdi and L.J Williams,(2010). "Principal component analysis," *Wiley Interdisciplinary Reviews: Computational Statistics* **2**, pp. 433–459, 2010.
16. E. Le Floch, V. Guillemot, V. Frouin, P. Pinel, C. Lalanne, L. Trinchera, ... É. Duchesnay, "Significant correlation between a set of genetic polymorphisms and a functional brain network revealed by feature selection and sparse Partial Least Squares," *NeuroImage* **63**, pp. 11–24, 2012.
17. M.P. Weiner, and T.J.Hudson, "Introduction to SNPs: discovery of markers for disease," *BioTechniques*, pp. 4–7, 10, 12–13, 2002.
18. R.M. Cantor, K. Lange, and J.S. Sinsheimer, "Prioritizing GWAS results: A review of statistical methods and recommendations for their application," *The American Journal of Human Genetics* **86**, pp. 6–22, 2010.
19. P.M. Visscher, M.A. Brown, M.I. McCarthy, and J. Yang, "Five years of GWAS discovery," *The American Journal of Human Genetics* **90**, pp. 7–24, 2012.
20. E. Genin, D. Hannequin, D. Wallon, K. Sleegers, M. Hiltunen, O. Combarros, ... D. Campion, "APOE and Alzheimer disease: a major gene with semi-dominant inheritance," *Molecular Psychiatry* **16**, pp. 903–907, 2012.
21. M. Bécue-Bertaut and J. Pagès, "Multiple factor analysis and clustering of a mixture of quantitative, categorical and frequency data," *Computational Statistics & Data Analysis* **52**, pp. 3255–3268, 2008.
22. H. Abdi and L. J. Williams, "Correspondence analysis," in *Encyclopedia of Research Design*, N. Salkind, ed., pp. 267–278. Sage, Thousand Oaks, (CA), 2010.
23. H. Abdi and M. Béra, "Correspondence analysis," in *Encyclopedia of Social Networks and Mining*, R. Alhajj and J. Rokne, eds., pp. 275–284, New York: Springer Verlag, 2014.
24. A. Krishnan, L.J. Williams, A.R. McIntosh, and H. Abdi, "Partial least squares (PLS) methods for neuroimaging: A tutorial and review," *NeuroImage* **56**, pp. 455–475, 2011.
25. H. Abdi, and L.J. Williams, "Partial least squares methods: Partial least squares correlation and partial least square regression," in *Methods in Molecular Biology: Computational Toxicology*, B. Reisfeld and A. Mayeno, eds., New York: Springer Verlag, pp. 549–579, 2013.
26. A.R. McIntosh, F.S. Bookstein, J. Haxby, and C. Grady, "Spatial pattern analysis of functional brain images using partial least squares," *NeuroImage* **3**, pp. 143–157, 1996.
27. F. Bookstein, "Partial least squares: a dose–response model for measurement in the behavioral and brain sciences," *Psycoloquy* **5**(23), 1994.
28. C.S. Bretherton, C. Smith, and J.M. Wallace, "An intercomparison of methods for finding coupled patterns in climate data," *Journal of Climate* **5**, pp. 541–560, 1992.
29. J.A. Wegelin, "A survey of partial least squares (PLS) methods, with emphasis on the two-block case," *Technical Report*, University of Washington, 2000.
30. A. Tishler, D. Dvir, A. Shenhar, and S. Lipovetsky, "Identifying critical success factors in defense development projects: A multivariate analysis," *Technological Forecasting and Social Change* **51**, pp. 151–171, 1996.
31. L. R. Tucker, "An inter-battery method of factor analysis," *Psychometrika* **23**, pp. 111–136, 1958.
32. D. Beaton, F. M. Filbey, and H. Abdi "Integrating partial least squares correlation and correspondence analysis for nominal data," in *New Perspectives in Partial Least Squares and Related Methods*, H. Abdi, W.W. Chin, V. Esposito Vinzi, G. Russolillo, and L. Trinchera, eds., pp. 81–94. New York (NY): Springer-Verlag, 2013.
33. B. Escofier, "Traitement simultané de variables qualitatives et quantitatives en analyse factorielle," *Les Cahiers de l'Analyse Des Données* **4**, pp.137–146, 1979.
34. M. Greenacre, "Data doubling and fuzzy coding," in *Visualization and Verbalization of Data*, J. Blasius and M. Greenacre, eds., London: CRC Press, pp. 239–253, 2014.
35. L. Bertram, M.B. McQueen, K. Mullin, D. Blacker, R.E. Tanzi, "Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database", *Nature Genetics* **39**, pp. 17–23, 2007.
36. K. Oishi, K. Zilles, K. Amunts, A. Faria, H. Jiang, H., X. Li, ... S. Mori, "Human brain white matter atlas: identification and assignment of common anatomical structures in superficial white matter," *NeuroImage* **43**, pp. 447–457, 2008.
37. B. Efron "Bootstrap methods: another look at the Jackknife," *The Annals of Statistics* **7**, pp. 1–26, 1979.
38. T. Hesterberg, "Bootstrap," *Wiley Interdisciplinary Reviews: Computational Statistics* **3**, pp. 497–526, 2011.
39. Y. Takane, and H. Hwang, "Regularized multiple correspondence analysis," in *Multiple Correspondence Analysis and Related Methods*, M. Greenacre and J. Blasius, eds., pp. 259–279, London: Academic Press, 2006.
40. G. I. Allen "Sparse and Functional Principal Components Analysis," *arXiv preprint arXiv:1309.2895*, 2013.