A generalization of partial least squares regression and correspondence analysis for categorical and mixed data: An application with the ADNI data

Derek Beaton
Rotman Research Institute, Baycrest Health Sciences
and
ADNI [*]
ADNI
and
Gilbert Saporta
Conservatoire National des Arts et Metiers
and
Herve Abdi
Behavioral and Brain Sciences, The University of Texas at Dallas

October 1, 2019

Abstract

The present and future of large scale studies of human brain and behavior—in typical and disease populations—is "mutli-omics", "deep-phenotyping", or other types of multi-source and multi-domain data collection initiatives. These massive studies rely on highly interdisciplinary teams that collect extremely diverse types of data across numerous systems and scales of measurement (e.g., genetics, brain structure, behavior, and demographics). Such large, complex, and heterogeneous data requires relatively simple methods that allow for flexibility in analyses without the loss of the inherent properties of various data types. Here we introduce a method designed

[*] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (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 listing of ADNI investigators can be found at http://adni.loni.ucla.edu/wpcontent/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

specifically to address these problems: partial least squares-correspondence analysis-regression (PLS-CA-R). PLS-CA-R generalizes PLS regression for use with virtually any data type (e.g., continuous, ordinal, categorical, non-negative values). Though the primary emphasis here is on a PLS-regression approach generalized for data types, we also show that PLS-CA-R leads to additional generalizations of many routine "two-table" multivariate techniques and their respective algorithms, such as various PLS approaches, canonical correlation analysis, and redundancy analysis (a.k.a. reduced rank regression).

Keywords: generalized singular value decomposition, latent models, genetics, neuroimaging, canonical correlation analysis

1 Introduction

Today's large scale and multi-site studies, such as the UK BioBank (`https://www.ukbiobank.ac.uk/`) and the Rotterdam study (`http://www.erasmus-epidemiology.nl/`), collect population level data across numerous types and modalities, including but not limited to genetics, neurological, and various behavioral, clinical, and laboratory measures. Similarly, other types of large scale studies—typically those that emphasize diseases and disorders—collect more "depth" of data for each participant: many measures and modalities on smaller samples. Some such studies include the Ontario Neurodegenerative Disease Research Initiative (ONDRI) (Farhan et al. 2016) which includes genetics, multiple types of magnetic resonance brain imaging (Duchesne et al. 2019), a wide array of behavioral, cognitive, clinical, and laboratory batteries, as well as many modalities "between" those, such as ocular imaging, gait & balance (Montero-Odasso et al. 2017), eye tracking, and neuropathology. Though large samples (e.g., UK BioBank) and depth of data (e.g., ONDRI) are necessary to understand typical and disordered samples and populations, few statistical and machine learning approaches exist that easily accomodate such large (whether "big" or "wide"), complex, and heterogeneous data sets that also respect the inherent properties of such data, while also accomodating numerous issues such as numerous predictors and responses, latent effects, high collinearity, and rank deficiency.

In many cases, the mixture of data types results in the sacrifices of information and inference, due in part because of transformations or assumptions that may be inappropriate or incorrect. For example, to analyze categorical and continuous data together, a typical—but inappropriate—strategy is to recode the continous data into categories such as dichotomization, trichotomization, or other (often arbitrary) binning strategies. Furthermore, ordinal and Likert scale data—such as responses on many cognitive, behavioral, clinical, and survey instruments—are often incorrectly treated as metric or continuous values (Bürkner & Vuorre n.d.). And when it comes to genetic data, such as single nucleotide polymorphims (SNPs), there is almost exclusive use of the additive model based on the minor homozygote: 0 for the most major homozygote, 1 for the heterozygote, and 2 for the minor homozygote. The additive model holds as nearly exclusive even though other models (e.g., dominant, recessive) or more general models (i.e., genotypic) exist and perform better

(Lettre et al. 2007). Furthermore, {0, 1, 2} recoding of genotypes (1) presumes additive and linear effects based on the minor homozygote and (2) are often treated as metric/continuous values (as opposed to categorical or ordinal), even when known effects of risk are neither linear nor additive, such as haplotypic effects (Vormfelde & Brockmöller 2007) nor exclusively based on the minor homozygotes, such as ApoE in Alzheimer's Disease (Genin et al. 2011).

Here we introduce partial least squares-correspondence analysis-regression (PLS-CA-R): a regression modeling and latent variable approach better suited for the complex data sets of today's studies. We first show PLS-CA-R as a generalization of PLS regression (Wold 1975, Wold et al. 1984, Tenenhaus 1998, Abdi 2010), CA (Greenacre 1984, 2010, Lebart et al. 1984), and PLS-CA (Beaton et al. 2016)—the last of which is the "correlation" companion to, and basis of, the proposed PLS-CA-R method. We then illustrate PLS-CA-R as a data-type general PLS regression method. PLS-CA-R combines the features of CA to allow for flexibility of data types with the features of PLS-R as a regression method designed to replace OLS when we cannot meet the assumptions or requirements of ordinary least squares (OLS). Both PLS-R and CA—and thus PLS-CA-R—are latent variable approaches by way of components via the generlized singular value decomposition (GSVD). We show multiple variants of PLS-CA-R, that address a variety of approaches and data types, on data from the Alzheimer's Disease Neuroimaging Initiative (ADNI). Data for these problems span diagnosis (mutually exclusive categories), SNPs (genotypes are categorical), multiple behavioral and clinical instruments (that could be ordinal, categorical, or continuous), and several neuroimaging measures and indices (generally either continous or non-negative). PLS-CA-R came about because it is clearly necessary: we need a method to accomodate these data types in a predictive or fitting framework, capitalize on the ability to regress (residualize) confounds out of mixed and likely collinear data, reveal latent variables, and most importantly, do all of these things simply and within a well-established framework.

After we formalize and illustrate PLS-CA-R, we then show that the core of PLS-CA-R also works as a much more generalized framework, and we then go into detail on how PLS-CA-R provides the basis of more generalized PLS framework spans: multiple data types, various optmizations (e.g., covariance as in PLS or correlation as in canonical correlation), various transformations for alternate metrics, and ridge-like regularization. Furthermore this framework spans multiple PLS algorithms. We place particular emphasis on three of the most commonly used PLS algorithms: (1) PLS "correlation" decomposition (Krishnan et al. 2011, Bookstein 1994, McIntosh et al. 1996)—also known as PLSSVD (Tenenhaus 1998) or Tucker's Interbattery Factor Analysis (Tucker 1958), (2) PLS "canonical" decompostion (Tenenhaus 1998), and (3) PLS "regression" (Wold 1975, Wold et al. 1984, 2001). PLS correlation is a symmetric method where neither data table plays a privileged (or predictive) role, and is performed with a single pass of the singular value decomposition (SVD). PLS canonical is also symmetric, but makes use of the SVD iteratively and deflates both data tables in each iteration. PLS regression is an asymmetric method where one data table is privileged ("predictors") and also makes use of the SVD iteratively with deflation of both data tables. We discuss the particularities of these algorithms later in the paper.

This paper is orgnized as follows. In Section 2 we introduce sufficient background for, and then formalize, PLS-CA-R. Next, in Section 3, we illustrate PLS-CA-R on the TADPOLE challenge (`https://tadpole.grand-challenge.org/`) and additional genetics data from ADNI across three examples: a simple discriminant example with entirely categorical data, a mixed data example that requires residualiation (i.e., adusting for confounds), and the finally a larger example of multiple genetic markers and whole brain tracer uptake (non-negative values). Finally in Section 4 we discuss PLS-CA-R, but then provide further details on how PLS-CA-R naturally leads to a much broader generalized PLS framework that spans multiple optimizations, algorithms, metrics, and ridge-like regularization.

2 Partial least squares-correspondence analysis-regression

Here we present the generalization of partial least square-regression (PLS-R) to multiple correspondence analysis (MCA) and correspondence analysis (CA)-like problems that generally apply to categorical (nominal) data. Via CA, we can also generalize to other data types including mixed types (e.g., categorical, ordinal, continuous, contingency). We use a mixture of nomenclature associated with χ^2-analyses, CA, and PLS-R.

Notation is as follows. Bold uppercase letters denotes matrices (e.g., $\mathbf{X}$), bold lowercase letters denote vectors (e.g., $\mathbf{x}$), and italic lowercase letters denote specific elements (e.g., x).

Upper case italic letters denote cardinality, size, or length (e.g., I) where a lower case italic denotes a specific index (e.g., i). A generic element of $\mathbf{X}$ would be denoted as $x_{i,j}$. Common letters of varying type faces, for example $\mathbf{X}$, $\mathbf{x}$, $x_{i,j}$, come from the same data struture. A preprocessed or transformed version of a matrix $\mathbf{X}$ will be denoted as $\mathbf{Z_X}$. Vectors are assumed to be column vectors unless otherwise specified. Two matrices side-by-side denotes standard matrix multiplication (e.g., $\mathbf{XY}$), where $\odot$ denotes element-wise (Hadamard) multiplication where $\oslash$ denotes element-wise (Hadamard) division. The matrix $\mathbf{I}$ denotes the identity matrix. Superscript ^{T} denotes the transpose operation, superscript $^{-1}$ denotes standard matrix inversion, and superscript $^+$ denotes the Moore-Penrose pseudo-inverse. The diagonal operation, diag{}, transforms a vector into a diagonal matrix, or extracts the diagonal of a matrix and produces a vector.

2.1 The SVD, GSVD, CA, and GPLSSVD

Assume we have a matrix $\mathbf{X}$ with I rows and J columns, where $\mathbf{X}$ is preprocessed in some way as $\mathbf{Z_X}$. The singular value decomposition (SVD) decomposes $\mathbf{Z_X}$ as $\mathbf{Z_X} = \mathbf{U\Delta V}^T$ where $\mathbf{U}^T\mathbf{U} = \mathbf{I} = \mathbf{V}^T\mathbf{V}$, where $\mathbf{Z_X}$ is of rank A and $\mathbf{\Delta}$ is an $A \times A$ diagonal matrix of singular values, and $\mathbf{\Lambda} = \mathbf{\Delta}^2$ is a $A \times A$ diagonal matrix of eigenvalues. $\mathbf{U}$ and $\mathbf{V}$ are referred to as the left and right singular vectors, respectively. From the SVD we can compute component (a.k.a. factor) scores as $\mathbf{F}_I = \mathbf{U\Delta}$ and $\mathbf{F}_J = \mathbf{V\Delta}$ for the I rows and J columns of $\mathbf{X}$, respectively.

The GSVD generalizes the SVD wherein the GSVD decomposes $\mathbf{Z_X}$ as $\mathbf{Z_X} = \mathbf{P\Delta Q}^T$ under the constraints of $\mathbf{P}^T\mathbf{M_X}\mathbf{P} = \mathbf{I} = \mathbf{Q}^T\mathbf{W_X}\mathbf{Q}$, where $\mathbf{Z_X}$ is of rank A and $\mathbf{\Delta}$ is a $A \times A$ diagonal matrix of singular values. With the GSVD, $\mathbf{P}$ and $\mathbf{Q}$ are referred to as the *generalized* singular vectors. Practically, the GSVD is performed through the SVD as $\widetilde{\mathbf{Z}}_\mathbf{X} = \mathbf{M}_\mathbf{X}^{\frac{1}{2}}\mathbf{Z_X}\mathbf{W}_\mathbf{X}^{\frac{1}{2}} = \mathbf{U\Delta V}^T$, where the generalized singular vectors are computed from the singular vectors as $\mathbf{P} = \mathbf{M}_\mathbf{X}^{-\frac{1}{2}}\mathbf{U}$ and $\mathbf{Q} = \mathbf{W}_\mathbf{X}^{-\frac{1}{2}}\mathbf{V}$. From the weights, generalized singular vectors, and singular values we can obtain component (a.k.a. factor) scores as $\mathbf{F}_I = \mathbf{M_X P\Delta}$ and $\mathbf{F}_J = \mathbf{W_X Q\Delta}$ for the I rows and J columns of $\mathbf{X}$, respectively. For simplicity and brevity we will refer to the use of the GSVD through "triplet notation" (Holmes 2008) but in the form of GSVD($\mathbf{M_X}, \mathbf{Z_X}, \mathbf{W_X}$), which is akin to how the multiplication steps work (see also Beaton et al. 2018). The standard SVD is the GSVD but with identity matrices as the weights: GSVD$(\mathbf{I}, \mathbf{X}, \mathbf{I})$ (see also Takane 2003), and thus if $\mathbf{Z}_{\mathbf{X}}$ were a column-wise centered and/or normalized version of $\mathbf{X}$ then GSVD$(\mathbf{I}, \mathbf{Z}_{\mathbf{X}}, \mathbf{I})$ is principal components analysis (PCA).

Correspondence analysis (CA) is a technique akin to PCA but initially designed for contingency and nominal data, and operates under the assumption of indepdence (i.e., akin to χ^2). See Greenacre (1984), Greenacre (2010), and Lebart et al. (1984) for detailed English explanations of CA and see Escofier-Cordier (1965) and Benzécri (1973) for the origins of CA. Assume the matrix $\mathbf{X}$ is some $I \times J$ matrix comprised of non-negative data, generally counts (co-occurences between rows and columns) or categorical data transformed into disjunctive format (see, e.g., "SEX" in Table 1). CA is performed with the GSVD as follows. First we define the *observed* matrix $\mathbf{O}_{\mathbf{X}} = \mathbf{X} \times (\mathbf{1}^T\mathbf{X}\mathbf{1})^{-1}$. Next we compute the marginal probabilities from the observed matrix as $\mathbf{m}_{\mathbf{X}} = \mathbf{O}_{\mathbf{X}}\mathbf{1} =$ and $\mathbf{w}_{\mathbf{X}} = (\mathbf{1}^T\mathbf{O}_{\mathbf{X}})^T$. We then define the *expected* (under the assumption of independence, i.e., χ^2) matrix as $\mathbf{E}_{\mathbf{X}} = \mathbf{m}_{\mathbf{X}}\mathbf{w}_{\mathbf{X}}^T$. We then compute the *deviation* (from independence) matrix as $\mathbf{Z}_{\mathbf{X}} = \mathbf{O}_{\mathbf{X}} - \mathbf{E}_{\mathbf{X}}$. Finally, we perform CA as GSVD$(\mathbf{M}_{\mathbf{X}}^{-1}, \mathbf{Z}_{\mathbf{X}}, \mathbf{W}_{\mathbf{X}}^{-1})$ with the weights of $\mathbf{M}_{\mathbf{X}} = \text{diag}\{\mathbf{m}_{\mathbf{X}}\}$ and $\mathbf{W}_{\mathbf{X}} = \text{diag}\{\mathbf{w}_{\mathbf{X}}\}$.

We now introduce an extension of the GSVD and its triplet concept for PLS, called the "GPLSSVD sextuplet" to decompose the the relationship between two matrices each with I rows: $\mathbf{X}$ with J columns and $\mathbf{Y}$ with K columns, where $\mathbf{Z}_{\mathbf{X}}$ and $\mathbf{Z}_{\mathbf{Y}}$ are preprocessed versions of $\mathbf{X}$ and $\mathbf{Y}$, respectively. The "GPLSSVD sextuplet" takes the form of GPLSSVD$(\mathbf{M}_{\mathbf{X}}, \mathbf{Z}_{\mathbf{X}}, \mathbf{W}_{\mathbf{X}}, \mathbf{M}_{\mathbf{Y}}, \mathbf{Z}_{\mathbf{Y}}, \mathbf{W}_{\mathbf{Y}})$. Like with the GSVD let us refer to $\widetilde{\mathbf{Z}}_{\mathbf{X}} = \mathbf{M}_{\mathbf{X}}^{\frac{1}{2}}\mathbf{Z}_{\mathbf{X}}\mathbf{W}_{\mathbf{X}}^{\frac{1}{2}}$ and $\widetilde{\mathbf{Z}}_{\mathbf{Y}} = \mathbf{M}_{\mathbf{Y}}^{\frac{1}{2}}\mathbf{Z}_{\mathbf{Y}}\mathbf{W}_{\mathbf{Y}}^{\frac{1}{2}}$. The GPLSSVD makes use of the SVD wherein $\widetilde{\mathbf{Z}}_{\mathbf{R}} = \widetilde{\mathbf{Z}}_{\mathbf{X}}^T\widetilde{\mathbf{Z}}_{\mathbf{Y}} = (\mathbf{M}_{\mathbf{X}}^{\frac{1}{2}}\mathbf{Z}_{\mathbf{X}}\mathbf{W}_{\mathbf{X}}^{\frac{1}{2}})^T(\mathbf{M}_{\mathbf{Y}}^{\frac{1}{2}}\mathbf{Z}_{\mathbf{Y}}\mathbf{W}_{\mathbf{Y}}^{\frac{1}{2}}) = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}^T$. The GPLSSVD generalized singular vectors and component scores are computed as $\mathbf{P} = \mathbf{W}_{\mathbf{X}}^{-\frac{1}{2}}\mathbf{U}$ and $\mathbf{F}_J = \mathbf{W}_{\mathbf{X}}\mathbf{P}\boldsymbol{\Delta}$ for the J columns of $\mathbf{X}$, and $\mathbf{Q} = \mathbf{W}_{\mathbf{Y}}^{-\frac{1}{2}}\mathbf{V}$ and $\mathbf{F}_K = \mathbf{W}_{\mathbf{Y}}\mathbf{Q}\boldsymbol{\Delta}$ for the K columns of $\mathbf{Y}$. Like with the SVD $\mathbf{U}^T\mathbf{U} = \mathbf{I} = \mathbf{V}^T\mathbf{V}$, and like with the GSVD $\mathbf{P}^T\mathbf{W}_{\mathbf{X}}\mathbf{P} = \mathbf{I} = \mathbf{Q}^T\mathbf{W}_{\mathbf{Y}}\mathbf{Q}$. The GPLSSVD also produces scores for the I rows of each matrix—usually called latent variables—as $\mathbf{L}_{\mathbf{X}} = \widetilde{\mathbf{Z}}_{\mathbf{X}}\mathbf{U}$ and $\mathbf{L}_{\mathbf{Y}} = \widetilde{\mathbf{Z}}_{\mathbf{Y}}\mathbf{V}$ where $\mathbf{L}_{\mathbf{X}}^T\mathbf{L}_{\mathbf{Y}} = \boldsymbol{\Delta}$. By its definition, the GPLSSVD maximization of the latent variables—i.e., $\mathbf{L}_{\mathbf{X}}^T\mathbf{L}_{\mathbf{Y}} = \boldsymbol{\Delta}$—is the PLS correlation decomposition (Krishnan et al. 2011, Bookstein 1994, McIntosh et al. 1996), also known as PLSSVD (Tenenhaus 1998) and originally as Tucker's interbattery factor analysis (Tucker 1958). Specifically, if $\mathbf{X}$ and $\mathbf{Y}$ were each column-wise centered and/or normalized, then GPLSSVD($\mathbf{I}, \mathbf{Z}_{\mathbf{X}}, \mathbf{I}, \mathbf{I}, \mathbf{Z}_{\mathbf{Y}}, \mathbf{I}$) is PLS correlation (a.k.a. PLSSVD or Tucker's approach).

Finally, we also introduce a small modification of the "triplet" and "sextuplet" notations as a "quadruplet" and a "septuplet" that indicate the desired rank to be returned by the GSVD or GPLSSVD. For example, if we want only one component from either approach we would indicate the desired rank to return as GSVD($\mathbf{M}_{\mathbf{X}}, \mathbf{Z}_{\mathbf{X}}, \mathbf{W}_{\mathbf{X}}, 1$) and GPLSSVD($\mathbf{M}_{\mathbf{X}}, \mathbf{Z}_{\mathbf{X}}, \mathbf{W}_{\mathbf{X}}, \mathbf{M}_{\mathbf{Y}}, \mathbf{Z}_{\mathbf{Y}}, \mathbf{W}_{\mathbf{Y}}, 1$). Both the GSVD and GPLSSVD in these cases would return only one set of singular vectors, generalized singular vectors, and component scores, and one singular value; for GPLSSVD it would return only one pair of latent variables.

Table 1: An example of disjunctive (SEX) and pseudo-disjunctive (AGE, EDU) coding through the fuzzy or Escofier transforms. For disjunctive an pseudo-disunctive data, each variable has a row-wise sum of 1 across its respective columns, and thus the row sums across the table are the number of original variables.

	Original coding			Disjunctive and pseudo-disjunctive coding					
				SEX		AGE		EDU	
	SEX	AGE	EDU	Male	Female	AGE-	AGE+	EDU-	EDU+
SUBJ 1	Male	64.8	16	1	0	1.03	-0.03	0.33	0.67
SUBJ 2	Female	63.6	18	0	1	1.11	-0.11	0.17	0.83
SUBJ 3	Female	76.4	18	0	1	0.24	0.76	0.17	0.83
SUBJ 4	Male	66.0	18	1	0	0.95	0.05	0.17	0.83
SUBJ 5	Female	61.9	14	0	1	1.23	-0.23	0.50	0.50
SUBJ 6	Female	66.7	14	0	1	0.90	0.10	0.50	0.50

2.2 PLS-CA-R

For simplicity assume in the following formulation that $\mathbf{X}$ and $\mathbf{Y}$ are both complete disjunctive tables as seen in Table 1 (see SEX columns) or Table 2. This formulation also applies generally to non-negative data (see later sections). We define observed matrices for $\mathbf{X}$ and $\mathbf{Y}$ as

$$\begin{aligned}\mathbf{O}_{\mathbf{X}} &= \mathbf{X} \times (\mathbf{1}^T\mathbf{X}\mathbf{1})^{-1},\\ \mathbf{O}_{\mathbf{Y}} &= \mathbf{Y} \times (\mathbf{1}^T\mathbf{Y}\mathbf{1})^{-1}\end{aligned} \tag{1}$$

Next we compute marginal probabilities for the rows and columns. We compute row probabilities as

$$\mathbf{m}_{\mathbf{X}} = \mathbf{O}_{\mathbf{X}}\mathbf{1} \text{ and } \mathbf{m}_{\mathbf{Y}} = \mathbf{O}_{\mathbf{Y}}\mathbf{1}, \tag{2}$$

which are the row sums of the observed matrices in Eq. 1. Then we compute column probabilities as

$$\mathbf{w}_{\mathbf{X}} = (\mathbf{1}^T\mathbf{O}_{\mathbf{X}})^T \text{ and } \mathbf{w}_{\mathbf{Y}} = (\mathbf{1}^T\mathbf{O}_{\mathbf{Y}})^T, \tag{3}$$

which are the column sums of the observed matrices in Eq. 1. We then define expected matrices as

$$\mathbf{E}_{\mathbf{X}} = \mathbf{m}_{\mathbf{X}}\mathbf{w}_{\mathbf{X}}^T \text{ and } \mathbf{E}_{\mathbf{Y}} = \mathbf{m}_{\mathbf{Y}}\mathbf{w}_{\mathbf{Y}}^T, \tag{4}$$

and deviation matrices as

$$\mathbf{Z}_{\mathbf{X}} = \mathbf{O}_{\mathbf{X}} - \mathbf{E}_{\mathbf{X}} \text{ and } \mathbf{Z}_{\mathbf{Y}} = \mathbf{O}_{\mathbf{Y}} - \mathbf{E}_{\mathbf{Y}}, \tag{5}$$

For PLS-CA-R we have two matrices, $\mathbf{Z}_{\mathbf{X}}$ which is $I \times J$ and $\mathbf{Z}_{\mathbf{Y}}$ which is $I \times K$, and their respective row and column weights of $\mathbf{M}_{\mathbf{X}} = \text{diag}\{\mathbf{m}_{\mathbf{X}}\}$, $\mathbf{M}_{\mathbf{Y}} = \text{diag}\{\mathbf{m}_{\mathbf{Y}}\}$, $\mathbf{W}_{\mathbf{X}} = \text{diag}\{\mathbf{w}_{\mathbf{X}}\}$, and $\mathbf{W}_{\mathbf{Y}} = \text{diag}\{\mathbf{w}_{\mathbf{Y}}\}$. PLS-CA-R makes use of the rank 1 GPLSSVD iteratively—GPLSSVD($\mathbf{M}_{\mathbf{X}}^{-1}, \mathbf{Z}_{\mathbf{X}}, \mathbf{W}_{\mathbf{X}}^{-1}, \mathbf{M}_{\mathbf{Y}}^{-1}, \mathbf{Z}_{\mathbf{Y}}, \mathbf{W}_{\mathbf{Y}}^{-1}, 1$)—and works as follows.

First we have $\widetilde{\mathbf{Z}}_{\mathbf{X}} = \mathbf{M}_{\mathbf{X}}^{-\frac{1}{2}}\mathbf{Z}_{\mathbf{X}}\mathbf{W}_{\mathbf{X}}^{-\frac{1}{2}}$ and $\widetilde{\mathbf{Z}}_{\mathbf{Y}} = \mathbf{M}_{\mathbf{Y}}^{-\frac{1}{2}}\mathbf{Z}_{\mathbf{Y}}\mathbf{W}_{\mathbf{Y}}^{-\frac{1}{2}}$. Then we compute the cross-product between $\widetilde{\mathbf{Z}}_{\mathbf{X}}$ and $\widetilde{\mathbf{Z}}_{\mathbf{Y}}$ as $\mathbf{Z}_{\mathbf{R}} = \widetilde{\mathbf{Z}}_{\mathbf{X}}^T\widetilde{\mathbf{Z}}_{\mathbf{Y}} = (\mathbf{M}_{\mathbf{X}}^{-\frac{1}{2}}\mathbf{Z}_{\mathbf{X}}\mathbf{W}_{\mathbf{X}}^{-\frac{1}{2}})^T(\mathbf{M}_{\mathbf{Y}}^{-\frac{1}{2}}\mathbf{Z}_{\mathbf{Y}}\mathbf{W}_{\mathbf{Y}}^{-\frac{1}{2}})$, where

$$\mathbf{Z}_{\mathbf{R}} = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}^T. \tag{6}$$

Because we make use of the rank 1 solution iteratively, we only retain the first vectors and values from Eq. 6. Thus we distinguish what we retain as $\tilde{\delta}$, $\tilde{\mathbf{u}}$, $\tilde{\mathbf{v}}$, $\tilde{\mathbf{p}}$, $\tilde{\mathbf{q}}$, and $\tilde{\mathbf{f}}_J$, and $\tilde{\mathbf{f}}_K$. We then compute the latent variables as

$$\boldsymbol{\ell}_{\mathbf{X}} = \widetilde{\mathbf{Z}}_{\mathbf{X}}\tilde{\mathbf{u}} \text{ and } \boldsymbol{\ell}_{\mathbf{Y}} = \widetilde{\mathbf{Z}}_{\mathbf{Y}}\tilde{\mathbf{v}}. \tag{7}$$

Next we compute $\mathbf{t}_{\mathbf{X}} = \boldsymbol{\ell}_{\mathbf{X}} \times ||\boldsymbol{\ell}_{\mathbf{X}}||^{-1}$, $b = \boldsymbol{\ell}_{\mathbf{Y}}^T\mathbf{t}_{\mathbf{X}}$, and $\widehat{\mathbf{u}} = \mathbf{t}_{\mathbf{X}}^T(\mathbf{M}_{\mathbf{X}}^{-\frac{1}{2}}\mathbf{Z}_{\mathbf{X}}\mathbf{W}_{\mathbf{X}}^{-\frac{1}{2}})$. We use $\mathbf{t}_{\mathbf{X}}$, b, and $\widehat{\mathbf{u}}$ to compute rank 1 "predicted" versions of $\mathbf{Z}_{\mathbf{X}}$ and $\mathbf{Z}_{\mathbf{Y}}$ as

$$\begin{aligned} \widehat{\mathbf{Z}}_{\mathbf{X},1} &= \mathbf{M}_{\mathbf{X}}^{\frac{1}{2}}(\mathbf{t}_{\mathbf{X}}\widehat{\mathbf{u}}^T)\mathbf{W}_{\mathbf{X}}^{\frac{1}{2}} \text{ and} \\ \widehat{\mathbf{Z}}_{\mathbf{Y},1} &= \mathbf{M}_{\mathbf{Y}}^{\frac{1}{2}}(b\mathbf{t}_{\mathbf{X}}\tilde{\mathbf{v}}^T)\mathbf{W}_{\mathbf{Y}}^{\frac{1}{2}}. \end{aligned} \tag{8}$$

Finally, we deflate $\mathbf{Z}_{\mathbf{X}}$ and $\mathbf{Z}_{\mathbf{Y}}$ as $\mathbf{Z}_{\mathbf{X}} = \mathbf{Z}_{\mathbf{X}} - \widehat{\mathbf{Z}}_{\mathbf{X},1}$ and $\mathbf{Z}_{\mathbf{Y}} = \mathbf{Z}_{\mathbf{Y}} - \widehat{\mathbf{Z}}_{\mathbf{Y},1}$. We then repeat the iterative procedure with these deflated $\mathbf{Z}_{\mathbf{X}}$ and $\mathbf{Z}_{\mathbf{Y}}$: GPLSSVD($\mathbf{M}_{\mathbf{X}}^{-1}, \mathbf{Z}_{\mathbf{X}}, \mathbf{W}_{\mathbf{X}}^{-1}, \mathbf{M}_{\mathbf{Y}}^{-1}, \mathbf{Z}_{\mathbf{Y}}, \mathbf{W}_{\mathbf{Y}}^{-1}, 1$). The computations outlined above are performed for C iterations where: (1) C is some pre-specified number of intended latent variables where $C < A$ where where A is the rank of $\widetilde{\mathbf{Z}}_{\mathbf{X}}$, (2) $C = A$, or (3) when $\mathbf{Z}_{\mathbf{X}} = \mathbf{0}$ or $\mathbf{Z}_{\mathbf{Y}} = \mathbf{0}$ where $\mathbf{0}$ is a null matrix. Upon the stopping condition we would have C components, and would have collected any vectors into corresponding matrices. Those matrices are

- two $C \times C$ diagonal matrices $\mathbf{B}$ and $\widetilde{\boldsymbol{\Delta}}$ with each b and $\tilde{\delta}$ on the diagonal with zeros off-diagonal,
- the $I \times C$ matrices $\mathbf{L}_{\mathbf{X}}$, $\mathbf{L}_{\mathbf{Y}}$, and $\mathbf{T}_{\mathbf{X}}$,
- the $J \times C$ matrices $\widetilde{\mathbf{U}}$, $\widehat{\mathbf{U}}$, $\widetilde{\mathbf{P}}$, and $\widetilde{\mathbf{F}}_J$, and
- the $K \times C$ matrices $\widetilde{\mathbf{V}}$, $\widetilde{\mathbf{Q}}$, $\widetilde{\mathbf{F}}_K$.

The algorithm for PLS-CA-R is presented in Algorithm 2 in Section 4. We present Algorithm 2 as a "generalized partial least squares regression" by way of the GPLSSVD sextuplet. We discuss the generalized aspects of the algorithm in more detail in Section 4.

2.3 Maximization in PLS-CA-R

PLS-CA-R maximizes the common information between $\mathbf{Z}_{\mathbf{X}}$ and $\mathbf{Z}_{\mathbf{Y}}$ such that

$$\begin{aligned}\underset{\boldsymbol{\ell}_{\mathbf{X}},\boldsymbol{\ell}_{\mathbf{Y}}}{\arg\max} = \boldsymbol{\ell}_{\mathbf{X}}^{T}\boldsymbol{\ell}_{\mathbf{Y}} = (\mathrm{M}_{X}^{-\frac{1}{2}}\mathbf{Z}_{\mathbf{X}}\mathbf{W}_{X}^{-1}\widetilde{\mathbf{p}})^{T}(\mathrm{M}_{Y}^{-\frac{1}{2}}\mathbf{Z}_{\mathbf{Y}}\mathbf{W}_{Y}^{-1}\widetilde{\mathbf{q}}) = \\ (\mathrm{M}_{X}^{-\frac{1}{2}}\mathbf{Z}_{\mathbf{X}}\mathbf{W}_{X}^{-\frac{1}{2}}\mathbf{W}_{X}^{-\frac{1}{2}}\widetilde{\mathbf{p}})^{T}(\mathrm{M}_{Y}^{-\frac{1}{2}}\mathbf{Z}_{\mathbf{Y}}\mathbf{W}_{Y}^{-\frac{1}{2}}\mathbf{W}_{Y}^{-\frac{1}{2}}\widetilde{\mathbf{q}}) = \\ (\widetilde{\mathbf{Z}}_{\mathbf{X}}\mathbf{W}_{X}^{-\frac{1}{2}}\widetilde{\mathbf{p}})^{T}(\widetilde{\mathbf{Z}}_{\mathbf{Y}}\mathbf{W}_{Y}^{-\frac{1}{2}}\widetilde{\mathbf{q}}) = (\widetilde{\mathbf{Z}}_{\mathbf{X}}\mathbf{W}_{X}^{-\frac{1}{2}}\mathbf{W}_{X}^{\frac{1}{2}}\widetilde{\mathbf{u}})^{T}(\widetilde{\mathbf{Z}}_{\mathbf{Y}}\mathbf{W}_{Y}^{-\frac{1}{2}}\mathbf{W}_{Y}^{\frac{1}{2}}\widetilde{\mathbf{v}}) = \\ (\widetilde{\mathbf{Z}}_{\mathbf{X}}\widetilde{\mathbf{u}})^{T}(\widetilde{\mathbf{Z}}_{\mathbf{Y}}\widetilde{\mathbf{v}}) = \widetilde{\mathbf{u}}\widetilde{\mathbf{Z}}_{\mathbf{X}}^{T}\widetilde{\mathbf{Z}}_{\mathbf{Y}}\widetilde{\mathbf{v}} = \widetilde{\mathbf{u}}\mathbf{Z}_{\mathbf{R}}\widetilde{\mathbf{v}} = \widetilde{\mathbf{u}}\mathbf{U}\boldsymbol{\Delta}\mathbf{V}\widetilde{\mathbf{v}} = \widetilde{\delta},\end{aligned} \tag{9}$$

where $\widetilde{\delta}$ is the first singular value from $\boldsymbol{\Delta}$ for each c step. PLS-CA-R maximization is subject to the orthogonality constraint that $\boldsymbol{\ell}_{\mathbf{X},c}^{T}\boldsymbol{\ell}_{\mathbf{X},c'} = 0$ when $c \neq c'$. This orthogonality constraint propagates through to many of the vectors and matrices associated with $\mathbf{Z}_{\mathbf{X}}$ where $\mathbf{T}_{\mathbf{X}}^{T}\mathbf{T}_{\mathbf{X}} = \widetilde{\mathbf{P}}^{T}\mathbf{W}_{J}^{-1}\widetilde{\mathbf{P}} = \widetilde{\mathbf{U}}^{T}\widetilde{\mathbf{U}} = \mathbf{I}$; these orthogonality constraints do not apply to the various vectors and matrices associated with $\mathbf{Y}$.

2.4 Decomposition and reconstitution

PLS-CA-R is a "double decomposition" where

$$\begin{aligned}\mathbf{Z}_{\mathbf{X}} = \mathrm{M}_{\mathbf{X}}^{\frac{1}{2}}\mathbf{T}\widehat{\mathbf{U}}^{T}\mathbf{W}_{\mathbf{X}}^{\frac{1}{2}} \text{ and} \\ \widehat{\mathbf{Z}}_{\mathbf{Y}} = \mathrm{M}_{\mathbf{Y}}^{\frac{1}{2}}\mathbf{T}_{\mathbf{X}}\mathbf{B}\widetilde{\mathbf{V}}^{T}\mathbf{W}_{\mathbf{Y}}^{\frac{1}{2}} = \mathrm{M}_{\mathbf{Y}}^{\frac{1}{2}}\widetilde{\mathbf{Z}}_{\mathbf{X}}\widehat{\mathbf{U}}^{T+}\mathbf{B}\widetilde{\mathbf{V}}^{T}\mathbf{W}_{\mathbf{Y}}^{\frac{1}{2}},\end{aligned} \tag{10}$$

where $\widetilde{\mathbf{Z}}_{\mathbf{X}} = \mathbf{T}\widehat{\mathbf{U}}^{T}$ and $\widehat{\widetilde{\mathbf{Z}}}_{\mathbf{Y}} = \mathbf{T}_{\mathbf{X}}\mathbf{B}\widetilde{\mathbf{V}}^{T} = \widetilde{\mathbf{Z}}_{\mathbf{X}}\widehat{\mathbf{U}}^{T+}\mathbf{B}\widetilde{\mathbf{V}}^{T}$. PLS-CA-R, like PLS-R, provides the same estimated predicted values as OLS under the conditions that $\mathbf{Z}_{\mathbf{X}}$ is (1) full rank, (2) non-singular, (3) not excessively multicollinear:

$$\widehat{\mathbf{Z}}_{\mathbf{Y}} = \mathrm{M}_{\mathbf{Y}}^{\frac{1}{2}}\mathbf{T}_{\mathbf{X}}\mathbf{B}\widetilde{\mathbf{V}}^{T}\mathbf{W}_{\mathbf{Y}}^{\frac{1}{2}} = \widetilde{\mathbf{Z}}_{\mathbf{X}}(\widetilde{\mathbf{Z}}_{\mathbf{X}}^{T}\widetilde{\mathbf{Z}}_{\mathbf{X}})^{+}\widetilde{\mathbf{Z}}_{\mathbf{X}}^{T}\mathbf{Z}_{\mathbf{Y}}, \tag{11}$$

where $\mathbf{T}_{\mathbf{X}}\mathbf{B}\widetilde{\mathbf{V}}^{T} = \widetilde{\mathbf{Z}}_{\mathbf{X}}(\widetilde{\mathbf{Z}}_{\mathbf{X}}^{T}\widetilde{\mathbf{Z}}_{\mathbf{X}})^{+}\widetilde{\mathbf{Z}}_{\mathbf{X}}^{T}\widetilde{\mathbf{Z}}_{\mathbf{Y}}$. This connection to OLS shows how to residualize (i.e., "regress out" or "correct") for known confounding effects, akin to how residuals are computed in OLS. We do so with the original $\mathbf{Z}_{\mathbf{Y}}$ as $\mathbf{Z}_{\mathbf{Y}} - \widehat{\mathbf{Z}}_{\mathbf{Y}}$. PLS-CA-R produces a both predicted and residualized version of $\mathbf{Y}$. Recall that $\widehat{\mathbf{Z}}_{\mathbf{Y}} = \mathrm{M}_{\mathbf{Y}}^{\frac{1}{2}}\mathbf{T}_{\mathbf{X}}\mathbf{B}\widetilde{\mathbf{V}}^{T}\mathbf{W}_{\mathbf{Y}}^{\frac{1}{2}}$. We compute a reconstituted form of $\mathbf{Y}$ as

$$\widehat{\mathbf{Y}} = (\widehat{\mathbf{Z}}_{\mathbf{Y}} + \mathbf{E}_{\mathbf{Y}}) \times (\mathbf{1}^{T}\mathbf{Y}\mathbf{1}), \tag{12}$$

which is the opposite steps of computing the deviations matrix. We add back in the expected values and then scale the data by the total sum of the original matrix. The same can be done for residualized values (i.e., “error”) as

$$\mathbf{Y}_{\epsilon} = [(\mathbf{Z_Y} - \widehat{\mathbf{Z}}_\mathbf{Y}) + \mathbf{E_Y}] \times (\mathbf{1}^T\mathbf{Y}\mathbf{1}). \tag{13}$$

Typically, $\mathbf{E_Y}$ is derived from the model of the data (as noted in Eq. 4). However, the reconstituted space could come from any model by way of generalized correspondence analysis (Escofier 1983; Escofier 1984; for more details and background see also Beaton et al. 2018). With GCA we could use any reasonable model with alternates to $\mathbf{E_Y}$ that could be obtained, for examples, from known priors, a theoretical model, out of sample data, or even population estimates. The same procedures can be applied to obtain a reconstituted $\mathbf{X}$. Finally, CA can then be applied directly to either $\widehat{\mathbf{Y}}$ or $\mathbf{Y}_{\epsilon}$.

2.5 Concluding remarks

Now that we have formalized PLS-CA-R, we want to point out small variations, some caveats, and some additional features here. We go into much more detail in later sections on larger variations and broader generalizations through our formulation.

In PLS-CA-R (and PLS-R) each subsequent $\widetilde{\delta}$ is not guaranteed to be smaller than the previous, with the exception of all $\widetilde{\delta}$ are smaller than the first. This is a by-product of the iterative process and the deflation steps, and does not occur with just a single pass of the SVD or GSVD (i.e., PLS-correlation decomposition). This poses two issues: (1) visualization of component scores and (2) explained variance. For visualization of the component scores—which use $\widetilde{\delta}$—there is an alternative computation: $\mathbf{F}^{'}_{J} = \mathbf{W}_J\widetilde{\mathbf{P}}$ and $\mathbf{F}^{'}_{K} = \mathbf{W}_K\widetilde{\mathbf{Q}}$. This alternative is referred to as “asymmetric component scores” in the correspondence analysis literature (Abdi & Béra 2014, Greenacre 1993). Additionally, instead of computing the variance per component or latent variable, we can instead compute the amount of variance explained by each component in $\mathbf{X}$ and $\mathbf{Y}$. To do so we require the sum of the eigenvalues of each of the respective matrices per iteration. The trace for each is computed via CA (with the GSVD). Before the first iteration of PLS-CA-R we can obtain the the full variance (sum of the eigenvalues) of each matrix from GSVD($\mathbf{M}^{-1}_\mathbf{X}, \mathbf{Z_X}, \mathbf{W}^{-1}_\mathbf{X}$) and GSVD($\mathbf{M}^{-1}_\mathbf{Y}, \mathbf{Z_Y}, \mathbf{W}^{-1}_\mathbf{Y}$), which we refer to as $\phi_\mathbf{X}$ and $\phi_\mathbf{Y}$, respectively. We can compute the sum of the eigenvalues for each deflated version of $\mathbf{Z_X}$ and $\mathbf{Z_Y}$ through the GSVD just as above, referred to as $\phi_{\mathbf{X},c}$ and $\phi_{\mathbf{Y},c}$. For each c component the proportion of explained variance for each matrix is $\frac{\phi_{\mathbf{X}}-\phi_{\mathbf{X},c}}{\phi_{\mathbf{X}}}$ and $\frac{\phi_{\mathbf{Y}}-\phi_{\mathbf{Y},c}}{\phi_{\mathbf{Y}}}$.

In our formulation, the weights we use are derived from the χ^2 assumption of independence. However nearly any choices of weights could be used, so long as the weight matrices are positive semi-definite. However, if alternate row weights (i.e., $\mathbf{M_X}$ or $\mathbf{M_Y}$) were chosen, then the fitted values and residuals are no longer guaranteed to be orthogonal (the same condition is true in weighted OLS). Likewise, alternate column weights (i.e., $\mathbf{W_X}$ or $\mathbf{W_Y}$) may deviate from the assumptions of χ^2 and may no longer reflect independence, or is an alternate metric altogether. There are some well-established alternates that we mention further in later sections, and through the concept of generalized correspondence analysis (Escofier 1983, 1984) almost any weights or model could be used. However, those changes should be informed by sound statistical and theoretical assumptions.

Finally, though we formalized PLS-CA-R as a method for categorical (nominal) data coded in complete disjunctive format (as seen in Table 1—see SEX columns—or Table 2), PLS-CA-R can easily accomodate various data types without loss of information. Specifically, both continuous and ordinal data can be handled with relative ease and in a "pseudo-disjunctive" format, also referred to as "fuzzy coding" where complete disjunctive would be a "crisp coding" (Greenacre 2014). We explain exactly to handle various data types as Section 3 progresses, which reflects more "real world" problems: complex, mixed data types, and multi-source data.

3 Applications & Examples

The goal of this section is to provide a several illustrative examples with real data from the Alzheimer's Disease Neuroimaging Initiative (ADNI). These examples highlight how to approach mixed data with PLS-CA-R as well as the multiple uses of PLS-CA-R (e.g., for analyses, as a residualization procedure). We present three sets of analyses to illustrate multiple uses of PLS-CA-R. First we introduce PLS-CA-R through a typical and relatively straightforward example: predict genotypes (categorical) from groups (categorical) where we can highlight multiple features of PLS-CA-R. Next we present analyses with the goal to predict genotypes from a small set of behavioral and brain variables. This second example serves multiple purposes: (1) how to recode and analyze mixed data (categorical, ordinal, and continuous), (2) how to use PLS-CA-R as an analysis technique, and (3) how to use PLS-CA-R as residualization technique (i.e., adjust for confounds) prior to subsequent analyses. Finally, we present a larger analysis with the goal to predict genotypes from cortical uptake of AV45 (i.e., a radiotracer) PET scan for beta-amyloid ("Aβ") deposition. This final example also makes use of residualization as illustrated in the second example.

3.1 ADNI Data

Data used in the preparation of this article come from 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).

The data we use in the following examples come from several modalities of the ADNI data entirely from the ADNI-GO/2 cohort. Generally, the data come from two sources available from the ADNI download site (`http://adni.loni.usc.edu/`): genome-wide data and the TADPOLE challenge data (`https://tadpole.grand-challenge.org/`) which contains a wide variety of data (e.g., demographics, diagnosis, cognitive and behavioral data, and some neuroimaging data). Because the genetics data are used in every example, we provide all genetics preprocessing details here, and then describe any preprocessing for other data as we discuss specific examples.

For all examples in this paper we use a candidate set of single nucleotide polymorphisms (SNPs) extracted from the genome-wide data. We extracted only SNPs associated with the *MAPT*, *APP*, *ApoE*, and *TOMM40* genes because they are considered as candidate contributors to various AD pathologies: *MAPT* because of its association with tau proteins, AD pathology, or cognitive decline (Myers et al. 2005, Trabzuni et al. 2012, Desikan et al. 2015, Cruchaga et al. 2012, Peterson et al. 2014), *APP* because of its association with β-amyloid proteins (Cruchaga et al. 2012, Huang et al. 2017, Jonsson et al. 2012), as well as *ApoE* and *TOMM40* because of their strong association with the diagnosis of AD and presence of various AD pathologies (Linnertz et al. 2014, Roses et al. 2010, Bennet et al. 2010, Huang et al. 2017). SNPs were processed as follows via Purcell et al. (2007) with additional `R` code as necessary: minor allele frequency (MAF) $>$ 5% and missingness for individuals and genotypes $\leq$ 10%. Because the SNPs are coded as categorical variables (i.e., for each genotype) we performed an additional level of preprocessing: genotypes $>$ 5% because even with MAF $>$ 5%, it was possible that some genotypes (e.g., the heterozygote or minor homozygote) could still have very few occurrences. Therefore if any genotypes were $\leq$ 5% they were combined with another genotype. In all cases the minor homozygote ('aa') fell below that threshold and was then combined with its respective heterozygote ('Aa'); thus some SNPs were effectively coded as the dominant model (i.e., the major homozygote vs. the presence of a minor allele). See Table for an example of SNP data coding examples. From the ADNI-GO/2 cohort there were 791 available participants. After preprocessing there were 791 participants with 134 total SNPs across the four candidate genes. The 134 SNPs span 349 columns in disjunctive coding (see Table 2). Other data include diagnosis and demographics, some behavioral and cognitive instruments, and several types of brain-based measures. We discuss these additional data in further detail when we introduce these data.

3.2 Diagnosis and genotypes

Our first example asks and answers the question: "which genotypes are associated with which diagnostic category?". We do so through the prediction of genotypes from diagnosis. Diagnosis at baseline in the ADNI study is a mutually exclusive categorical variable that denotes which group each participant belongs to (at the first visit): control (CN; $N = 155$), subjective memory complaints (SMC; $N = 99$), early mild cognitive impairment (EMCI; $N = 277$), late mild cognitive impairment (LMCI; $N = 134$), and Alzheimer's disease (AD; $N = 126$). We present this first example analysis in two ways: akin to a standard regression

Table 2: An example of a SNP with its genotypes for respective individuals and disjunctive coding for three types of genetic models: genotypic (three levels), dominant (two levels: major homozygote vs. presence of minor allele), and recessive (two levels: presence of major allele vs. minor homozygote).

	SNP	Genotypic			Dominant		Recessive	
	Genotype	AA	Aa	aa	AA	Aa+aa	AA+Aa	aa
SUBJ 1	Aa	0	1	0	0	1	1	0
SUBJ 2	aa	0	0	1	0	1	1	0
SUBJ 3	aa	0	0	1	0	1	0	1
SUBJ 4	AA	1	0	0	1	0	1	0
SUBJ 5	Aa	0	1	0	0	1	1	0
SUBJ 6	AA	1	0	0	1	0	1	0

progblem a la Wold [Wold (1975); Wold et al. (1984); Wold et al. (1987); cf. Eq. 11) and then again in the more recent, and now typical, multivariate perspective of "projection onto latent structures" (Abdi 2010).

Table 3: Descriptives and demographics for the sample.

	N	AGE mean (sd)	EDU mean (sd)	Males (Females)
AD	126	74.53 (8.42)	15.78 (2.68)	75 (51)
CN	155	74 (6.02)	16.41 (2.52)	80 (75)
EMCI	277	71.14 (7.39)	15.93 (2.64)	156 (121)
LMCI	134	72.24 (7.68)	16.43 (2.57)	73 (61)
SMC	99	72.19 (5.71)	16.81 (2.51)	40 (59)

For this example we refer to diagnosis groups as the predictors ($\mathbf{X}$) and the genotypic data as the responses ($\mathbf{Y}$). Both data types are coded in disjunctive format (see Tables 1 and 2). Because there are five columns (groups) in $\mathbf{X}$, PLS-CA-R produces only four latent variables (a.k.a. components). Table 4 presents the cumulative explained variance for both

$\mathbf{X}$ and $\mathbf{Y}$ and shows that groups explain only a small amount of genotypic variance: $R^2 = 0.0065$.

Table 4: The R-squared values over the four latent variables for both groups and genotypes. The full variance of groups is explained over the four latent variables. The groups explained 0.65% of the genotypes.

	X (groups) R-squared cumulative	Y (genotypes) R-squared cumulative
Latent variable 1	0.25	0.0026
Latent variable 2	0.50	0.0042
Latent variable 3	0.75	0.0055
Latent variable 4	1.00	0.0065

In a simple regression-like framework we can compute the variance contributed by genotypes or group (i.e., levels of variables) or variance contributed by entire variables (in this example: SNPs). First we compute the contributions to the variance of the genotypes as the sum of the squared loadings for each item: $[(\mathbf{V} \odot \mathbf{V})\mathbf{1}] \times C^{-1}$, where $\mathbf{1}$ is a conformable vector of ones. Total contribution values exist between 0 and 1 and describe the proportional variance each genotype contributes. These contributions to components can be computed as $\mathbf{v}_c \odot \mathbf{v}_c$. Because the contributions are additive we can compute the contributions for a SNP through all of its all. A simple criterion to identify genotypes or SNPs that contribute to the model is to identify which genotype or SNP contributes more variance than expected, which is one divided by the total number of original variables (i.e., SNPs). In this case that would be $1/134 = 0.0075$. This criterion can be applied on the whole or component-wise. We show the genotypes and SNPs with above exepcted variance for the whole model (i.e., high contributing variables a regression framework) in Figure 1.

Though PLS-R was initally developed as a regression approach—especially to handle highly collinear predictors or a set of predictors that are not full rank (see explanations in Wold et al. 1984)—it is far more common to use PLS to find latent structures (i.e., components or latent variables) (Abdi 2010). From this point forward we show only the more common "projection onto latent structures" perspectives. We show the latent vari-

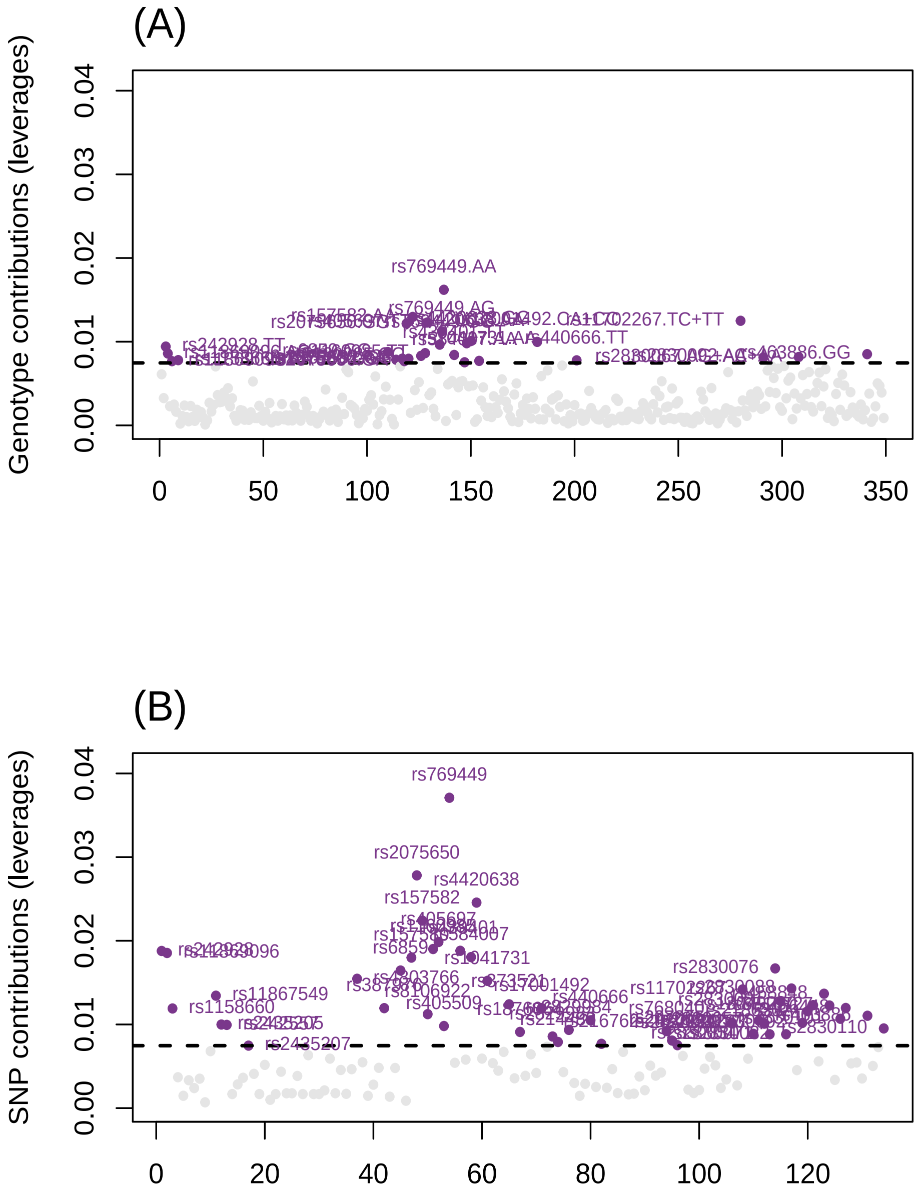

Figure 1: Regression approach to prediction of genotypes from groups. Contributions across all components for genotypes (A; top) and the SNPs (B; bottom) computed as the summation of genotypes within a SNP. The horizontal line shows the expected variance and we only highlight genotypes (A; top) or SNPs (B; bottom) greater than the expected variance. Some of the highest contributing genotypes (e.g., AA and AG genotypes for rs769449) or SNPs (e.g., rs769449 and rs20756560) come from the APOE and TOMM40 genes.

able scores (observations) and component scores (variables) for the first two latent variables/components in Figure 2. The first latent variable scores (Fig. 2a) shows a gradient from the control (CN) group through to the Alzheimer's Disease (AD) groups (CN to SMC to EMCI to LMCI to AD). The second latent variable shows a dissociation of the EMCI group from all other groups (Fig. 2b). Figure 2c and d show the component scores for the variables. Genotypes on the left side of first latent variable (a.k.a., component; horizontal axis in Figs. 2c and d) are more associated with CN and SMC than the other groups, where as genotypes on the right side are more associated with AD and LMCI than the other groups. Genotypes highlighted in purple are those that contribute more than expected variance to the first component. Through the latent structures approach we can more clearly see the relationships between groups and genotypes. Because we treat the data categorically and code for genotypes, we can identify the specific genotypes that contribute to these effects. For example the 'AA' genotype of rs769449 and the 'GG' genotype of rs2075650 are more associated with AD and LMCI than the other groups. In conrast, the 'TT' genotype of rs405697 and the 'TT' genotype rs439401 are more associated with the CN group than other groups (and thus could suggest potential protective effects).

This group-based analysis is also a discriminant analysis because it maximally separates groups. Thus we can classify observations by assigning them to the closest group. To correctly project observations onto the latent variables we compute $\mathbf{L}_{\mathbf{Y}} \times I^{\frac{1}{2}} = [\mathbf{O}_{\mathbf{Y}} \oslash (\mathbf{m}_{\mathbf{Y}}\mathbf{1}^{T})]\mathbf{F}_{K}\mathbf{\Delta}^{-1}$ where 1 is a $1 \times K$ vector of ones where $\mathbf{O}_{\mathbf{Y}} \oslash (\mathbf{m}_{\mathbf{Y}}\mathbf{1}^{T})$ are "row profiles" of $\mathbf{Y}$ (i.e., each element of $\mathbf{Y}$ divided by its respective row sum). Observations from $\mathbf{L}_{\mathbf{Y}} \times I^{\frac{1}{2}}$ are then assigned to the closest group in $\mathbf{F}_{J}$, either for per component, across a subset of components, or all components. For this example we use the full set (four) of components. The assigned groups can then be compared to the *a priori* groups to compute a classification accuracy. Figure 3 shows the results of the discriminant analysis but only visualized on the first two components. Figures 3a and b show the scores for $\mathbf{F}_{J}$ and $\mathbf{L}_{\mathbf{Y}} \times I^{\frac{1}{2}}$, respectively. Figure 3c shows the assignment of observations to their closest group. Figure 3d visualizes the accuracy of the assignment, where observations in black are correct assignments (gray are incorrect assignments). The total classification accuracy 38.69% (where chance accuracy was 23.08%). Finally, typical PLS-R discriminant analyses are applied in scenarios where

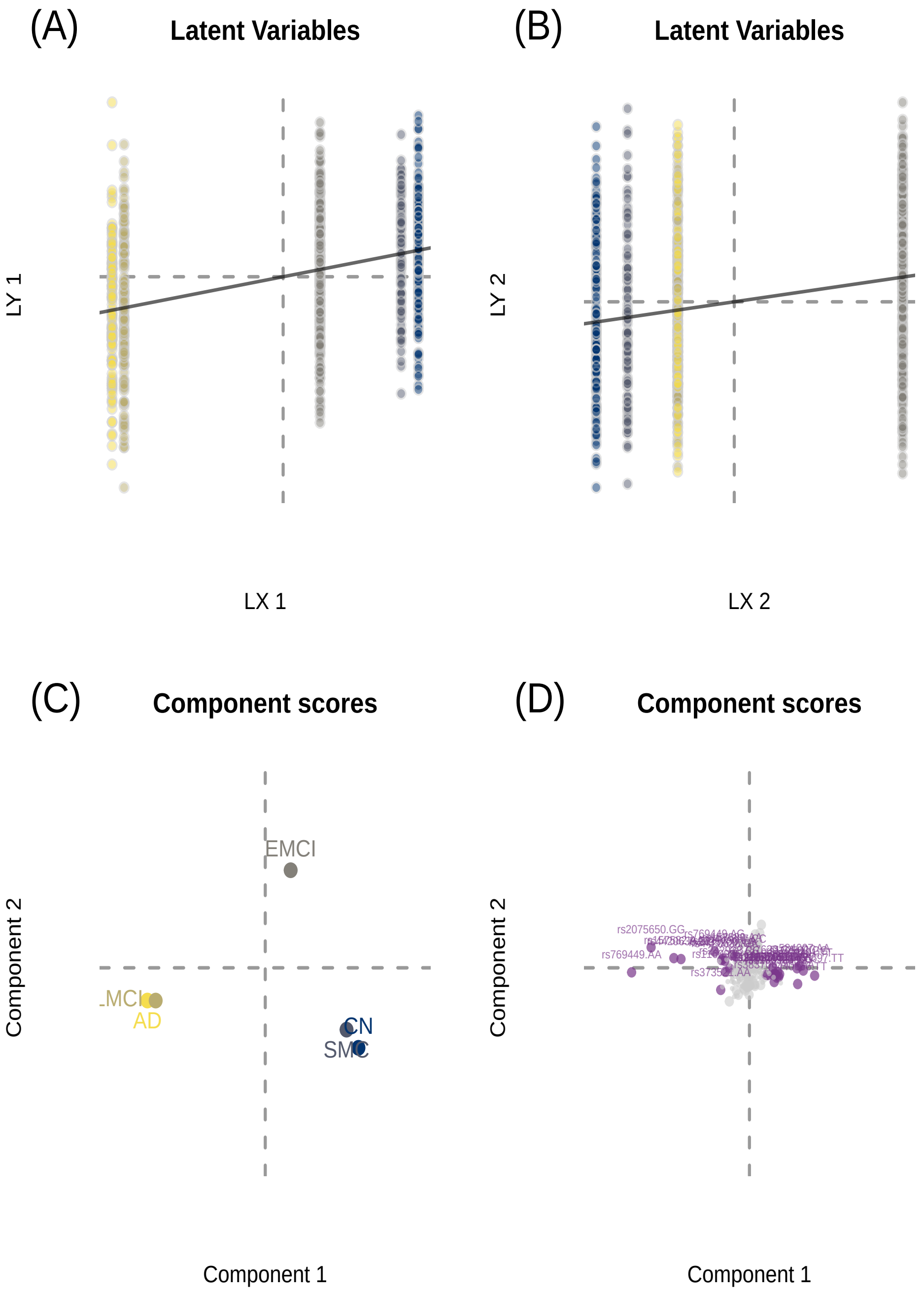

Figure 2: Latent variable projection approach to prediction of genotypes from groups. (A) and (B) show the latent variable scores for latent variables (LVs; components) one and two, respectively; (C) shows the component scores of the groups, and (D) shows the component scores of the genotypes. In (D) we highlight genotypes with above expected contribution to Latent Variable (Component) 1 in purple and make all other genotypes gray.

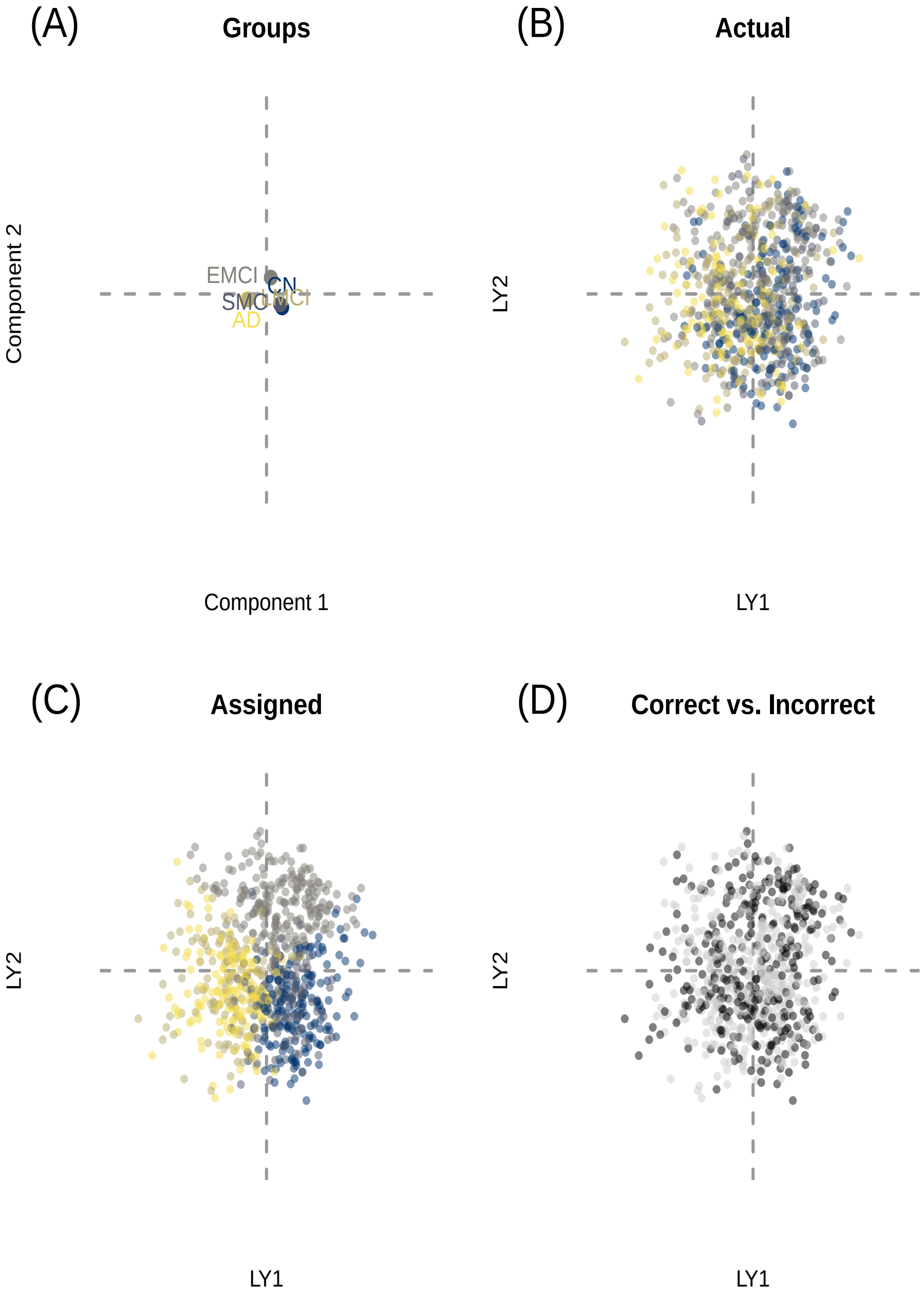

Figure 3: Discriminant PLS-CA-R. (A) shows the component scores for the group on Latent Variables (LV) 1 and 2 (horizontal and vertical respectively), (B) shows the latent variable scores for the genotype ('LY') LV scores for LVs 1 and 2, colored by *a priori* group association, (C) shows the latent variable scores for the genotype ('LY') LV scores for LVs 1 and 2, colored by *assigned* group association (i.e., nearest group assignment across all LVs), and (D) shows correct vs. incorrect assignment in black and gray, respectively.

a small set of, or even a single, (typically) categorical responses are predicted from many predictors (Pérez-Enciso & Tenenhaus 2003). However, such an approach appears to be "over optimistic" in its prediction and classification (Rodríguez-Pérez et al. 2018), which is why we present discriminant PLS-CA-R more akin to a typical regression problem (i.e., here a single predictor with multiple responses).

Table 5: The *a priori* (rows) vs. assigned (columns) accuracies for the discriminant analysis.

	CN	SMC	EMCI	LMCI	AD
CN	62	15	54	16	17
SMC	20	40	33	18	10
EMCI	34	20	110	23	29
LMCI	14	12	39	44	20
AD	25	12	41	33	50

3.3 Mixed data and residualization

Our second example illustrates the prediction of genotypes from multiple brain and behavioral variables: (1) three behavioral/clinical scales: Montreal Cognitive Assessment (MoCA) (Nasreddine et al. 2005), Clinical Dementia Rating-Sum of Boxes (CDRSB) (Morris 1993), and Alzheimer's Disease Assessment Scale (ADAS13) (Skinner et al. 2012), (2) volumetric brain measures in mm^3: hippocampus (HIPPO), ventricles (VENT), and whole brain (WB), and (3) global estimates of brain function via PET scans: average FDG (for cerebral blood flow; metabolism) in angular, temporal, and posterior cingulate and average AV45 (Aβ tracer) standard uptake value ratio (SUVR) in frontal, anterior cingulate, precuneus, and parietal cortex relative to the cerebellum. This example higlights two features of PLS-CA-R: (1) the ability to accomodate mixed data types (continuous, ordinal, and categorical) and (2) as a way to residualize (orthogonalize; cf. Eq. 13) with respect to known or assumed confounds.

Here, the predictors encompass a variety of data types: all of the brain markers (volumetric MRI estimates, functional PET estimates) and the ADAS13 appear as generally continuous data, whereas the MoCA and especially the CDRSB are generally ordinal because they have limited values constrained by a minimum and maximum score: the CDRSB exists between 0 and 9 generally by steps of 1, and the MoCA exists between 0 and 30, though values below 20 are exceedingly rare. Furthermore, the assumed differences between each level are not considered the same, for example, MoCA scores of 29 and 30 are regarded as preserved and normal (high) levels of cognition, where as 26 and 27 is the (clinical) line between impaired and unimpaired. There are many properties of PLS-CA-R by way of CA that allow for easy inclusion of mixed data types. In particular, continuous and ordinal data types can be coded into what is called thermometer (Beaton et al. 2018), fuzzy, or "bipolar" coding (because it has two poles) (Greenacre 2014); an idea initially propsosed by Escofier for continuous data (Escofier 1979). The "Escofier transform" allows continuous data to be analyzed by CA and produces the exact same results as PCA (Escofier 1979). The same principles can be applied to ordinal data as well (Beaton et al. 2018). Continuous and ordinal data can be transformed into a "pseudo-disjunctive" format that behaves exactly like complete disjunctive data (see Table 1) but preserves the values (as opposed to binning, or dichotomizing). Here, we refer to the transform for continuous data as the "Escofier transform" or "Escofier coding" (Beaton et al. 2016) and the transform for ordinal data as the "thermometer transform" or "thermometer coding". Because continuous, ordinal, and categorical data can all be trasnformed into a disjunctive-like format, they can all be analyzed with PLS-CA-R.

While the overall objective of this example is to understand the relationship between routine markers of AD and genetics, confounds exist for both the predictors (behavioral and brain data) and the responses (genotype data): age, sex, and education influence the behavioral and brain variables, whereas sex, race, and ethnicity influence the genotypic variables. To note, these confounds are also of mixed types (e.g., sex is categorical, age is generally continuous). Thus in this example we illustrate the mixed analysis in two ways—unadjusted and then adjusted for these confounds. First we show the effects of the confounds on the separate data sets, and then compare and contrast adjusted vs. unadjusted analyes. For the "mixed" data analyses, the volumetric data were also normalized (divided by) by intracranial volume prior to these analyses; effectively transformed into proportional volumes within each participant. Any other adjustments are described when needed.

First we show the PLS-CA-R between each data set and their respective confounds. The main effects of age, sex, and education explained 11.17% of the variance of the behavioral and brain data, where the main effects of sex, race, and ethnicity explained 2.1% of the variance of the genotypic data. The first two components of each analysis are shown in Figure 4. In the brain and behavioral data, age explains a substantial amount of variance and effectively explains Component 1. In the genotypic analysis, race is the primary explanatory effect; more specifically, the first two components are explained by those that identify as black or African-American (Component 1) vs. those that identify as Asian, Native, Hawaiian, or Latino/Hispanic (Component 2). Both data sets were reconstituted (i.e., $\mathbf{Y}_\epsilon$ from Eq. 13) from their residuals.

Next we performed two analyses with the same goal: understand the relationship between genetics and the behavioral and brain markers. In the unadjusted analysis, the brain and behavioral data explained 1.6% of variance in the genotypic data, whereas in the adjusted analysis, the brain and behavioral data explained 1.54% of variance in the genotypic data. The first two components of the PLS-CA-R results can be seen in Figure 4.

In the unadjusted analysis (Figure 4a and c) vs. the adjusted analysis (Figure 4b and d), we can some similarities and differences, especially with respect to the behavioral and brain data. AV45 shows little change after the residualization, and generally explains a substantial amount of variance as it contributes highly to the first two components in both analyses. The effects of the structural data—especially the hippocampus—are dampened after adjustment (see Figure 4a vs b), where the effects of FDG and CDRSB are now (relatively) increased (see Figure 4a vs b). On the subject level, the differences are not substantial, but there are noticeable effects especially with the ability to distinguish between groups (see Figure 6). One important effect is that on a spectrum from CON to AD, we can see that the residualization has a larger impact on the CON side, where the AD side remains somewhat homgeneous (see Figure 6c) for the brain and behavioral variables. With respect to the genotypic LV, there is much less of an effect (see Figure 6d), wherein the observations appear relatively unchanged. However, both pre- (horizontal axis; Figure 6d)

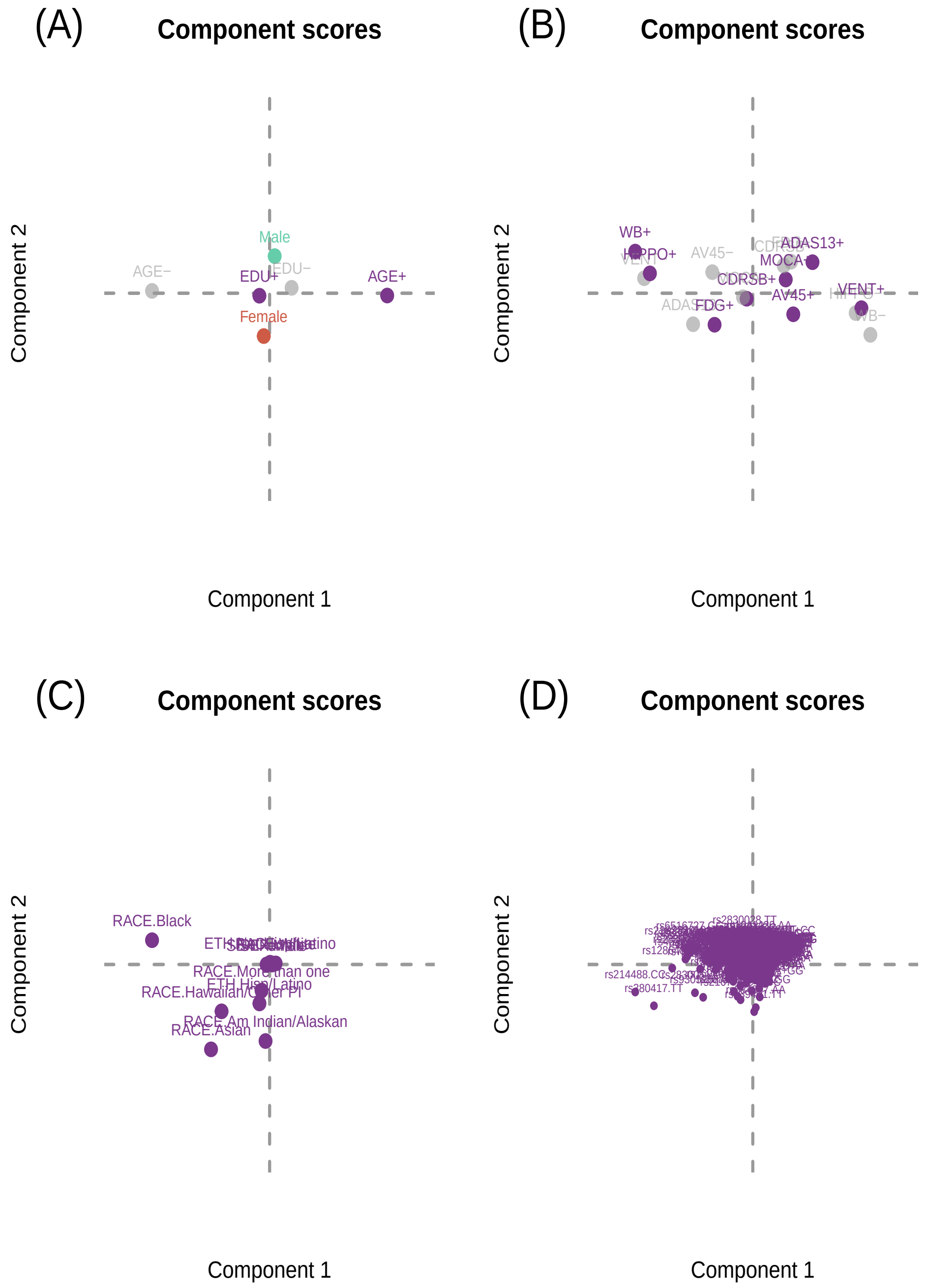

Figure 4: PLS-CA-R used as a way to residualize (orthogonalize) data. The top figures (A) and (B) show prediction of the brain and behavior markers from age, sex, and education. Gray items are one side (lower end) of the "bipolar" or pseudo-disjunctive variables. The bottom figures (C) and (D) show the prediction of genotypes from sex, race, and ethnicity.

and post- (vertical axis; Figure 6d) residualization shows that there are individuals with unique genotypic patterns that remain unaffected by the residualization process (i.e., those at the tails).

From this point forward we emphasize the results from the adjusted analyses because they are more realistic in terms of how analyses are performed. For this we refer to Figure 6b—which shows the latent variable scores for the observations and the averages of those scores for the groups—and Figures 5b and 5d—which show the component scores for the brain and behavioral markers and the genotypes, respectively. The first latent variable (Fig. 6b) shows a gradient from control (CON) on the left to Alzheimer's Disease (AD) on the right. Brain and behavioral variables on the right side of the first component (horizontal axis in Fig. 5b) are more associated with genotypes on the right side (Fig. 5d), where brain and behavioral variables on the left side of are more associated with genotypes on the left side. In particular, the AA genotype of rs769449, GG genotype of rs2075650, GG genotype of rs4420638, and AA genotype of rs157582 (amongst others) are related to increased AV45 (AV45+), decreased FDG (FDG-), and increased ADAS13 scores (ADAS13+), where as the TT genotype of rs405697, GG genotype of rs157580, and TC+TT genotypes of rs7412 (amongst others) are more associated with control or possibly protective effects (i.e., decreased AV4, increased FDG, and decreased ADAS13 scores).

3.4 SUVR and genotypes

In this final example we make use of all the features of PLS-CA-R: an example with mixed data types within and between data sets, each with confounds (and thus require residualization). This example serves as something more akin to the typical analysis pipeline with similar objectives. The goal of this example is to predict genotypes from $\beta-$amyloid burden ("AV45 uptake") across regions of the cortex. In this case, we also assume that the distribution of AV45 uptake across cortical regions approximately follows that of χ^2 in that we compute the deviations from independence (produced from the product between the row and column probabilities). However we want to note that this is only one possible way to handle such data. It is possible to treat these data as row-wise proportions (i.e., percentage of total uptake per region within each subject) or even as continuous data; though these

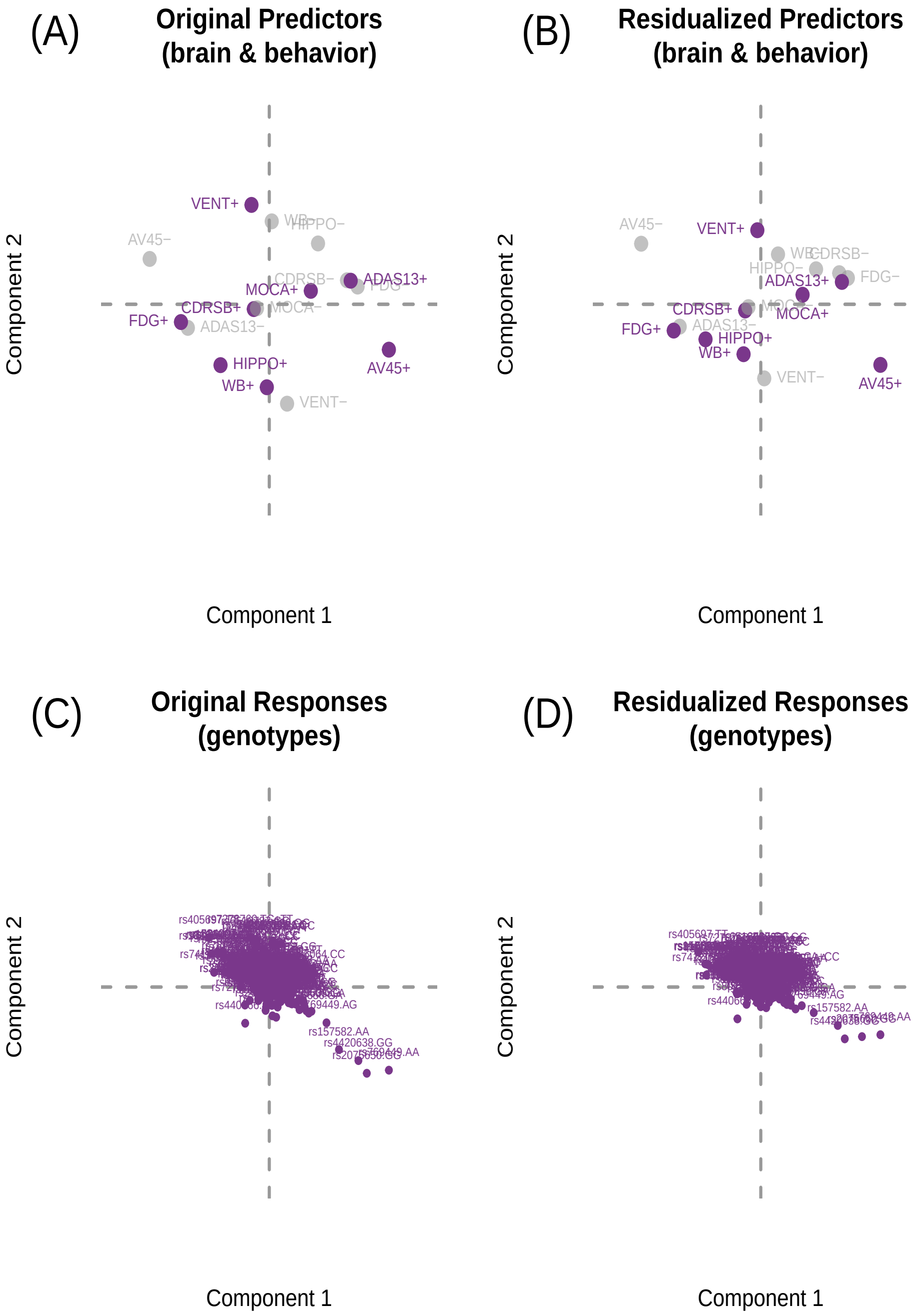

Figure 5: PLS-CA-R to predict genotypes from brain and behavioral markers on the original and residualized data shown on the first two latent variables (components). The top figures (A) and (B) show the component scores for the brain and behavioral markers for the original and residualized data, respectively, and the bottom figures (C) and (D) show the component scores for the genotypes for the original and residualized data, respectively.

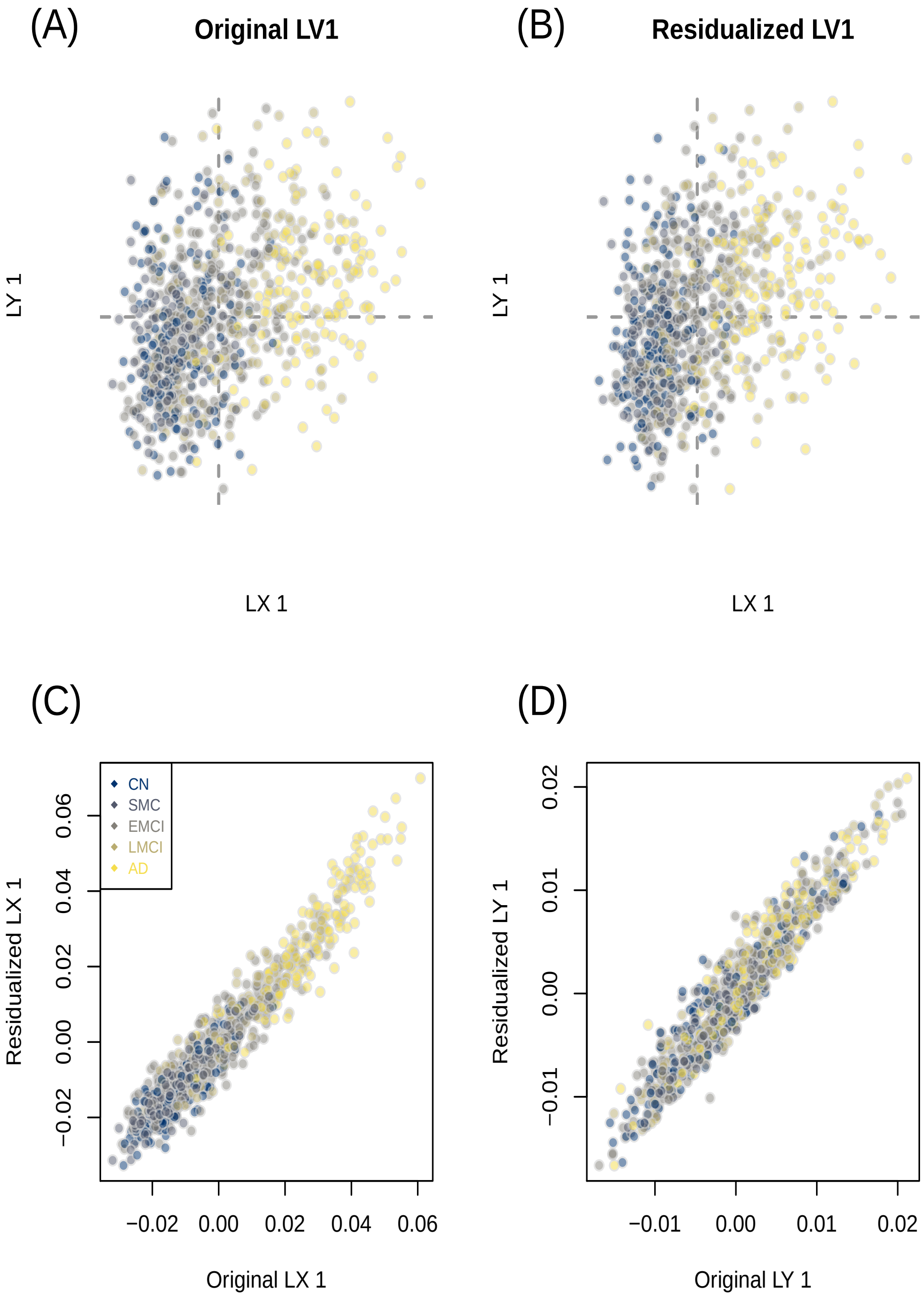

Figure 6: Latent variable scores (observations) for the first latent variable. The top figures (A) and (B) show the projection of the latent variable scores from each set: LX are the brain and behavioral markers, where as LY are the genotypes, for the original and residualized, respectively. The bottom figures (C) and (D) show the the original and residualized scores for the first latent variable compared to one another for each set: the brain and behavioral markers (LX) and the genotypes (LY), respectively.

data are strictly non-negative. Ultimately, it is up to the analyst to decide how to treat such data and how it fits into the analysis framework.

Because not all subjects have complete AV45 and genotypic data, the sample for this example is slightly smaller: $N = 778$. Ethnicity, race, and sex (all categorical) explains 2.07% of the variance in the genotypic data where age (numeric), education (ordinal), and sex (categorical) explains 2.22% of the variance in the in the AV45 uptake data. Overall, AV45 brain data explains 9.08% of the variance in the genotypic data. With the adjusted data we can now perform our intended analyses. Although this analysis produced 67 components (latent variables), we focus on just the first (0.57% of genotypic variance explained by AV45 brain data).

The first latent variable in Figure 7a is associated with only the horizontal axes (Component 1) in Figure 7b and c. The horizontal axis in Fig. 7a is associated with the horizontal axis in Fig. 7b whereas the vertical axis in Fig. 7a is associated with the horizontal axis in Fig. 7c. The first latent variable (Figure 7a) shows a gradient: from left to right we see the groups configured from CN to AD. On the first latent variable we do also see a group-level dissociation where AD+LMCI are entirely on one side whereas EMCI+SMC+CN are on the opposite side for both $\mathbf{L_X}$ (AV45 uptake, horizontal) and $\mathbf{L_Y}$ (genotypes, vertical); effectively the means of AD and LMCI exist in the upper right quadrant and the means of the EMCI, SMC, and CN groups exist in the lower left quadrant. Higher relative AV45 uptake for the regions on the left side of Component 1 are more associated with EMCI, SMC, and CN than with the other groups, whereas higher relative AV45 uptake for the regions on the right side of Component 1 are more associated with AD and LMCI (Fig. 7b). The genotypes on the left side are associated with the uptake in regions on the left side and the genotypes on the right side are associated with the uptake in regions on the right side (Fig. 7c). For example, LV/Component 1 shows relative uptake in right and left frontal pole, rostral middle frontal, and medial orbitofrontal regions are more associated with the following genotypes: AA and AG from rs769449, GG from rs2075650, GG from rs4420638, and AA from rs157582, than with other genotypes; these effects are generally driven by the AD and LMCI groups. Conversely, LV/Component 1 shows higher relative uptake in right and left lingual, cuneus, as well left parahippocampal and left entorhinal are more

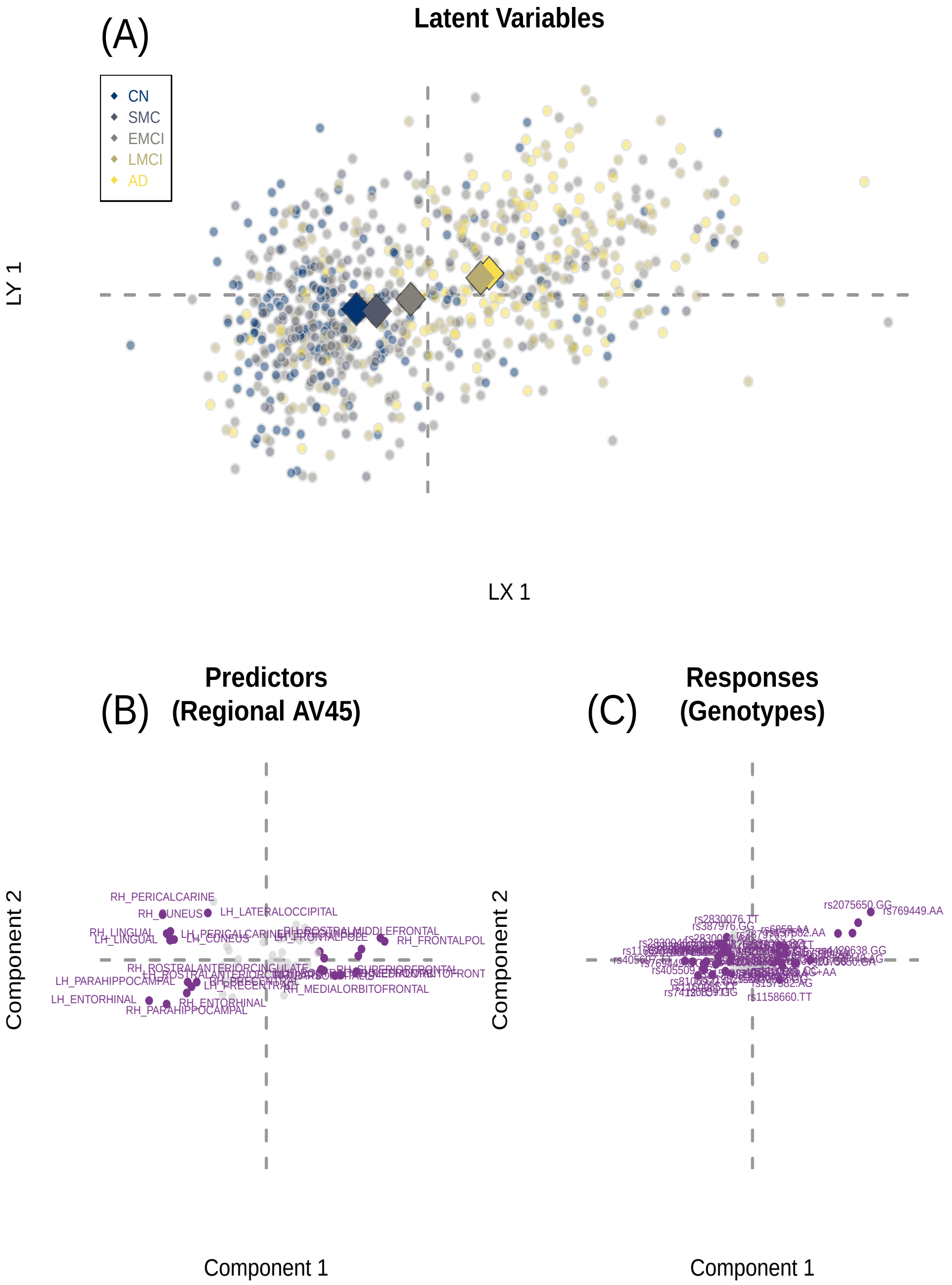

Figure 7: PLS-CA-R to predict genotypes from amyloid burden ("AV45 uptake"). The top figure (A) shows the latent variable scores for the observations on the first latent variable with group averages. The bottom figures (B) and (C) show the amyloid burden in cortical regions and the genotypes, respecively. In (A) we see a gradient from the Alzheimer's Disease (AD) group to the control (CON) group. Only items with above expected contribution to variance on the first LV are highlighed in purple.

associated with the following genotypes: TT from rs405697, GG from rs6859, TC+TT from rs7412, TT from rs2830076, GG from rs157580, and AA from rs4420638 genotypes than with other genotypes; these effects are generally driven by the CN, SMC, and EMCI cohorts. In summary, from the PLS-CA-R results we see that particular patterns of regional AV45 uptake predict particular genotypic patterns across many SNPs, and that the sources these effects are generally driven by the groups. Furthermore the underlying brain and genotypic effects of the groups exist along a spectrum of severity.

4 Discussion

Many modern studies, like ADNI, aim to measure individuals at a variety of scales: genetics and genomics, brain structure and function, many aspects of cognition and behavior, batteries of clinical measures, and almost anything in between all of these levels. These data are extremely complex: they are heterogeneous and more often than not "wide" (many more variables than subjects). But many current strategies and approaches to handle such multivariate heterogeneous data often requires compromises or sacrifices (e.g., the presumption of single numeric model for categorical data such as the additive model for SNPs; Z-scores of ordinal values; or "dichotomania" (https://www.fharrell.com/post/errmed/#catg): the binning of continuous values into categories). Many of those strategies and approaches presume that data are interval scale, or completely misrepresent data, and therefore the properties of those data types are Effectively ignored. Because of the many features and flexibility of PLS-CA-R—e.g., best fit to predictors, orthogonal latent variables, accommodation for virtually any data type—we are able to identify distinct variables and levels (e.g., genotypes) that define or contribute to control (CON) vs. disease (AD) effects (e.g., Fig. 2) or reveal particular patterns anchored by the polar control and disease effects (CON → SMC → EMCI → LMCI → AD; see, e.g., Fig. 7).

While we focused on particular ways of coding and transforming data, there are many alternatives that could be used with PLS-CA-R. For example, we used a disjunctive approach for SNPs because they are categorical, which matches the genotypic model. However, through various disjunctive schemes, or other forms of Escofier or fuzzy coding, we could have used any genetic model: if all SNPs were coded as the major vs. the minor allele

('AA' vs. {'Aa+aa'}), this would be the dominant model, or we could have assumed the additive model —i.e., 0, 1, 2 for 'AA', 'Aa', and 'aa', respectively—and transformed the data with the ordinal approach (but we strongly emphasize *not* the continuous approach). We previously provided a comprehensive guide on how transform various SNP genetic models for use in PLS-CA and CA elsewhere (see Appendix of Beaton et al. 2016). Furthermore, we only highlighted one of many possible methods to transform ordinal data. The term "fuzzy coding" applies more generally to the recoding of ordinal, ranked, preference, and even continuous data across a number of schemes, all of which conform to the same properties as disjunctive data. The many "fuzzy" and "double" coding schemes are generally found in Escofier (1979), Lebart et al. (1984), or Greenacre (2014). However, for ordinal data—especially with fewer than or equal to 10 levels, and without excessively rare ($\leq 1\%$) occurences—we recommend to treat ordinal values as categorical levels. When ordinal data are treated as categorial (and disjunctively coded), greater detail about the levels emerges and in most cases reveal non-linear patterns of the ordinal levels.

Though we have presented PLS-CA-R as a generalization of PLS-R that accomodates virutally any data type (by way of CA), the way we formalized PLS-CA-R—in Section 2.2 and describe its algorithm in Algorithm 2—leads to further variants and broader generalizations, that span various PLS, CA, and related approaches, several typical PLS algorithms, a variety of optimizations (e.g., canonical correlation), and ridge-like regularization.

4.1 GPLS algorithms

In general there exist three primary PLS algorithms: PLS correlation decomposition (Bookstein 1994, Ketterlinus et al. 1989) generally more known in neuroimaging (McIntosh et al. 1996, McIntosh & Lobaugh 2004, Krishnan et al. 2011) which has numerous alternate names such as PLS-SVD and Tucker's interbattery factor analysis (Tucker 1958) amongst others (see also Beaton et al. 2016), PLS regression decomposition (cf. Section 2.2 and also Algorithm 2) and the PLS canonical decomposition (Tenenhaus 1998, Wegelin et al. 2000), which is a symmetric method with iterative deflation (i.e., it has features of both PLS-C and PLS-R). Given the way in which we formalize PLS-CA-R—as a generalized PLS-R—here we show how PLS-CA-R provides the basis of generalizations of these three algorithms, as well as further optimizations, similar to Borga et al. (1992), Indahl et al. (2009), and de Micheaux et al. (2019) but we do so in a more comprehensive way that incorporates more methods than other unification strategies, and we also do so in a way that accomodates multiple data types. We refer to the three techniques under the umbrella of generalized partial least squares (GPLS) as GPLS-COR, GPLS-REG, and GPLS-CAN, for the "correlation", "regression", and "canonical" decompositions respectively. GPLS-COR and GPLS-CAN are symmetric decomposition approaches where neither $\mathbf{Z}_\mathbf{X}$ nor $\mathbf{Z}_\mathbf{Y}$ are privileged. GPLS-REG is an asymmetric decomposition approach where $\mathbf{Z}_\mathbf{X}$ is privileged. We present the GPLS-COR, GPLS-REG, and then GPLS-CAN algorithms with their respective optimizations. We do so in the previously mentioned order because GPLS-COR is used as the basis of all three algorithms and GPLS-CAN shares features and concepts with both GPLS-COR and GPLS-REG. For all of these we rely on the basis of PLS-CA-R we established in Section 2.2—specifically for various mixed data types under the χ^2 model (as used in CA).

The GPLS-COR decomposition is the simplest GPLS technique. It requires only a single pass of the SVD—or in our case the GPLSSVD. There are no explicit iterative steps in GPLS-COR. GPLS-COR takes as input the two preprocessed matrices—$\mathbf{Z}_\mathbf{X}$ and $\mathbf{Z}_\mathbf{Y}$—and their respective row and column weights: $\mathbf{M}_\mathbf{X}$ and $\mathbf{W}_\mathbf{X}$ for $\mathbf{Z}_\mathbf{X}$, and $\mathbf{M}_\mathbf{Y}$ and $\mathbf{W}_\mathbf{Y}$ for $\mathbf{Z}_\mathbf{Y}$, where C is the desired number of components to return. GPLS-COR is shown in Algorithm 1.

```
Result: Generalized PLS-correlation between $\mathbf{Z}_\mathbf{X}$ and $\mathbf{Z}_\mathbf{Y}$
Input  : $\mathbf{M}_\mathbf{X}$, $\mathbf{Z}_\mathbf{X}$, $\mathbf{W}_\mathbf{X}$, $\mathbf{M}_\mathbf{Y}$, $\mathbf{Z}_\mathbf{Y}$, $\mathbf{W}_\mathbf{Y}$, $C$
Output: $\mathbf{U}$, $\mathbf{V}$, $\mathbf{P}$, $\mathbf{Q}$, $\mathbf{F}_J$, $\mathbf{F}_K$, $\mathbf{L}_\mathbf{X}$, $\mathbf{L}_\mathbf{Y}$, $\boldsymbol{\Delta}$

GPLSSVD($\mathbf{M}_\mathbf{X}, \mathbf{Z}_\mathbf{X}, \mathbf{W}_\mathbf{X}, \mathbf{M}_\mathbf{Y}, \mathbf{Z}_\mathbf{Y}, \mathbf{W}_\mathbf{Y}, C$)
```

Algorithm 1: Generalized PLS-correlation algorithm. GPLS-COR is the GPLSSVD and provides the basis of other GPLS techniques. Furthermore, GPLS-COR easily allows for a variety of optmizations for examples canonical correlation, reduced rank regression (redundancy analysis), and even ridge-like regularization.

GPLS-COR maximizes the relationship between $\mathbf{L}_\mathbf{X}$ and $\mathbf{L}_\mathbf{Y}$ with the orthogonality constraint $\boldsymbol{\ell}_{\mathbf{X},c}^T\boldsymbol{\ell}_{\mathbf{Y},c'} = 0$ when $c \neq c'$ where $\boldsymbol{\ell}_{\mathbf{X},c}^T\boldsymbol{\ell}_{\mathbf{Y},c} = \delta_c$ and thus $\mathbf{L}_\mathbf{X}^T\mathbf{L}_\mathbf{Y} = \mathbf{U}^T\widetilde{\mathbf{Z}}_\mathbf{X}^T\widetilde{\mathbf{Z}}_\mathbf{Y}\mathbf{V}^T =$

$\mathbf{U}^T\widetilde{\mathbf{Z}}_\mathbf{R}\mathbf{V}^T = \mathbf{U}^T\mathbf{U}\boldsymbol{\Delta}\mathbf{V}^T\mathbf{V}^T = \boldsymbol{\Delta}$. We can show this with the generalized vectors and weight as $\mathbf{L}_\mathbf{X}^T\mathbf{L}_\mathbf{Y} = \mathbf{P}^T\mathbf{W}_\mathbf{X}\mathbf{Z}_\mathbf{X}^T\mathbf{M}_\mathbf{X}^{\frac{1}{2}}\mathbf{M}_\mathbf{Y}^{\frac{1}{2}}\mathbf{Z}_\mathbf{Y}\mathbf{W}_\mathbf{Y}\mathbf{Q}^T = \mathbf{P}^T\mathbf{W}_\mathbf{X}\mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^T\mathbf{W}_\mathbf{Y}\mathbf{Q} = \boldsymbol{\Delta}$. Furthermore, GPLS-COR (via GPLSSVD) provides all of the other outputs as previously described in Section 2.1. GPLS-COR—which is the GPLSSVD—provides the basis for the other two algorithms: both GPLS-REG and GPLS-CAN make use of GPLS-COR (i.e., the GPLSSVD) with rank 1 solutions iteratively.

The GPLS-REG decomposition builds off of the GPLS-COR algorithm, but does so by way of the GPLSSVD septuplet iteratively for C iterations, with only a rank 1 solution is provided for each use of the GPLSSVD. Then the two data matrices—$\mathbf{Z_X}$ and $\mathbf{Z_Y}$—are deflated for each step asymmetrically, with a privileged $\mathbf{Z_X}$. GPLS-REG is shown in Algorithm 2.

```
Result: Generalized PLS-regression between $\mathbf{Z_X}$ and $\mathbf{Z_Y}$
Input  : $\mathbf{M_X}$, $\mathbf{Z_X}$, $\mathbf{W_X}$, $\mathbf{M_Y}$, $\mathbf{Z_Y}$, $\mathbf{W_Y}$, $C$
Output: $\widetilde{\mathbf{U}}$, $\widetilde{\mathbf{V}}$, $\widetilde{\mathbf{P}}$, $\widetilde{\mathbf{Q}}$, $\widetilde{\mathbf{F}}_J$, $\widetilde{\mathbf{F}}_K$, $\mathbf{L_X}$, $\mathbf{L_Y}$, $\widetilde{\boldsymbol{\Delta}}$, $\mathbf{T_X}$, $\widehat{\mathbf{U}}$, $\mathbf{B}$

for $c = 1, \ldots, C$ do
    GPLSSVD$(\mathbf{M_X}, \mathbf{Z_X}, \mathbf{W_X}, \mathbf{M_Y}, \mathbf{Z_Y}, \mathbf{W_Y}, 1)$
    $\mathbf{t_X} \leftarrow \boldsymbol{\ell}_\mathbf{X} \times ||\boldsymbol{\ell}_\mathbf{X}||^{-1}$
    $b \leftarrow \boldsymbol{\ell}_\mathbf{Y}^T\mathbf{t_X}$
    $\widehat{\mathbf{u}} \leftarrow (\mathbf{M}_\mathbf{X}^{\frac{1}{2}}\mathbf{Z_X}\mathbf{W}_\mathbf{X}^{\frac{1}{2}})^T\mathbf{t_X}$
    $\mathbf{Z_X} \leftarrow \mathbf{Z_X} - [\mathbf{M}_\mathbf{X}^{-\frac{1}{2}}(\mathbf{t_X}\widehat{\mathbf{u}}^T)\mathbf{W}_\mathbf{X}^{-\frac{1}{2}}]$
    $\mathbf{Z_Y} \leftarrow \mathbf{Z_Y} - [\mathbf{M}_\mathbf{Y}^{-\frac{1}{2}}(b\mathbf{t_X}\widetilde{\mathbf{v}}^T)\mathbf{W}_\mathbf{Y}^{-\frac{1}{2}}]$
end
```

Algorithm 2: Generalized PLS-regression algorithm. The results of a rank 1 GPLSSVD are used to compute the latent variables and values necessary for deflation of $\mathbf{Z_X}$ and $\mathbf{Z_Y}$. PLS-CA-R is a specific instance of GPLS-REG, which we defined in Section 2.2.

GPLS-REG maximizes the relationship between $\mathbf{L_X}$ and $\mathbf{L_Y}$ with the orthogonality constraint $\boldsymbol{\ell}_{\mathbf{X},c}^T\boldsymbol{\ell}_{\mathbf{X},c'} = 0$ when $c \neq c'$ where $\boldsymbol{\ell}_{\mathbf{X},c}^T\boldsymbol{\ell}_{\mathbf{Y},c} = \delta_c$ which is also $\text{diag}\{\mathbf{L}_\mathbf{X}^T\mathbf{L}_\mathbf{Y}\} = \text{diag}\{\widetilde{\boldsymbol{\Delta}}\}$.

The GPLS-CAN decomposition builds off of the GPLS-COR algorithm, but does so by way of the GPLSSVD septuplet iteratively for C iterations, with only a rank 1 solution is provided for each use of the GPLSSVD. Then the two data matrices—$\mathbf{Z_X}$ and $\mathbf{Z_Y}$—are deflated for each step symmetrically. GPLS-CAN is shown in Algorithm 3

```
Result: Generalized PLS-canonical between $\mathbf{Z_X}$ and $\mathbf{Z_Y}$
Input  : $\mathbf{M_X}$, $\mathbf{Z_X}$, $\mathbf{W_X}$, $\mathbf{M_Y}$, $\mathbf{Z_Y}$, $\mathbf{W_Y}$, $C$
Output: $\widetilde{\mathbf{U}}$, $\widetilde{\mathbf{V}}$, $\widetilde{\mathbf{P}}$, $\widetilde{\mathbf{Q}}$, $\widetilde{\mathbf{F}}_J$, $\widetilde{\mathbf{F}}_K$, $\mathbf{L_X}$, $\mathbf{L_Y}$, $\widetilde{\boldsymbol{\Delta}}$, $\mathbf{T_X}$, $\mathbf{T_Y}$, $\widehat{\mathbf{U}}$, $\widehat{\mathbf{V}}$

for $c = 1, \dots, C$ do
    GPLSSVD$(\mathbf{M_X}, \mathbf{Z_X}, \mathbf{W_X}, \mathbf{M_Y}, \mathbf{Z_Y}, \mathbf{W_Y}, 1)$
    $\mathbf{t_X} \leftarrow \boldsymbol{\ell}_\mathbf{X} \times ||\boldsymbol{\ell}_\mathbf{X}||^{-1}$
    $\mathbf{t_Y} \leftarrow \boldsymbol{\ell}_\mathbf{Y} \times ||\boldsymbol{\ell}_\mathbf{Y}||^{-1}$
    $\widehat{\mathbf{u}} \leftarrow (\mathbf{M}_\mathbf{X}^{\frac{1}{2}}\mathbf{Z_X}\mathbf{W}_\mathbf{X}^{\frac{1}{2}})^T\mathbf{t_X}$
    $\widehat{\mathbf{v}} \leftarrow (\mathbf{M}_\mathbf{Y}^{\frac{1}{2}}\mathbf{Z_Y}\mathbf{W}_\mathbf{Y}^{\frac{1}{2}})^T\mathbf{t_Y}$
    $\mathbf{Z_X} \leftarrow \mathbf{Z_X} - [\mathbf{M}_\mathbf{X}^{-\frac{1}{2}}(\mathbf{t_X}\widehat{\mathbf{u}}^T)\mathbf{W}_\mathbf{X}^{-\frac{1}{2}}]$
    $\mathbf{Z_Y} \leftarrow \mathbf{Z_Y} - [\mathbf{M}_\mathbf{Y}^{-\frac{1}{2}}(\mathbf{t_Y}\widehat{\mathbf{v}}^T)\mathbf{W}_\mathbf{Y}^{-\frac{1}{2}}]$
end
```

Algorithm 3: Generalized PLS-canonical algorithm. The results of a rank 1 GPLSSVD are used to compute the latent variables and values necessary for deflation of $\mathbf{Z_X}$ and $\mathbf{Z_Y}$. Note that the deflation in GPLS-CAN differs from GPLS-REG in Algorithm 2.

GPLS-CAN maximizes the relationship between $\mathbf{L_X}$ and $\mathbf{L_Y}$ with the orthogonality constraints $\boldsymbol{\ell}_{\mathbf{X},c}^T\boldsymbol{\ell}_{\mathbf{X},c'} = 0$ and $\boldsymbol{\ell}_{\mathbf{Y},c}^T\boldsymbol{\ell}_{\mathbf{Y},c'} = 0$ when $c \neq c'$ where $\boldsymbol{\ell}_{\mathbf{X},c}^T\boldsymbol{\ell}_{\mathbf{Y},c} = \delta_c$ which is also $\text{diag}\{\mathbf{L}_\mathbf{X}^T\mathbf{L_Y}\} = \text{diag}\{\widetilde{\boldsymbol{\Delta}}\}$.

Note that across all three algorithms defined here, that the first component is identical when the same preprocessed data and constraints are provided to the GPLSSVD. In nearly all cases, subsequent components across the three algorithms differ, but also generally they do not differ substantially. The similarities can be traced back to the common maximization of $\boldsymbol{\ell}_{\mathbf{X},c}^T\boldsymbol{\ell}_{\mathbf{Y},c} = \delta_c$, where the differences can be traced back to the specific orthogonality optimizations when $c \neq c'$ where: (1) GPLS-COR in Algorithm 1 is $\boldsymbol{\ell}_{\mathbf{X},c}^T\boldsymbol{\ell}_{\mathbf{Y},c'} = 0$, (2) GPLS-REG in Algorithm 2 is $\boldsymbol{\ell}_{\mathbf{X},c}^T\boldsymbol{\ell}_{\mathbf{X},c'} = 0$, and (3) GPLS-CAN in Algorithm 3 is both $\boldsymbol{\ell}_{\mathbf{X},c}^T\boldsymbol{\ell}_{\mathbf{X},c'} = 0$ and $\boldsymbol{\ell}_{\mathbf{Y},c}^T\boldsymbol{\ell}_{\mathbf{Y},c'} = 0$.

4.2 GPLS optimizations and further generalizations

From the GPLS perspective, we can better unify the wide variety of approaches with similar goals but variations of metric, transformations, and optimizations that often appear under a wide variety of names (e.g., PLS, CCA, interbattery factor analysis, co-inertia analysis, canonical variates, PLS-CA, and so on; see Abdi et al. (2017)). The way we defined the GPLS algorithms—in particular with the constraints applied to the rows and columns of each data matrix—leads to numerous further generalizations.

For simplicity, let us first focus on Algorithm 1, and assume that $\mathbf{X}$ and $\mathbf{Y}$ are continuous data, where $\mathbf{Z_X}$ and $\mathbf{Z_Y}$ are column-wise centered and/or scaled versions of $\mathbf{X}$ and $\mathbf{Y}$. Though we have established Algorithm 1 as GPLS-COR—and more generally as the GPLSSVD—we can obtain the results of three of the most common "two-table" techniques: PLS correlation (PLSC), canonical correlation analysis (CCA), and redundancy analysis (RDA, a.k.a., reduced rank regression [RRR]). Standard PLSC is performed as GPLSSVD$(\mathbf{I}, \mathbf{Z_X}, \mathbf{I}, \mathbf{I}, \mathbf{Z_Y}, \mathbf{I})$, CCA is performed as GPLSSVD$(\mathbf{I}, \mathbf{Z_X}, (\mathbf{Z}_\mathbf{X}^T\mathbf{Z_X})^{-1}, \mathbf{I}, \mathbf{Z_Y}, (\mathbf{Z}_\mathbf{Y}^T\mathbf{Z_Y})^{-1})$, and RDA—where $\mathbf{X}$ is privileged—is performed as GPLSSVD$(\mathbf{I}, \mathbf{Z_X}, (\mathbf{Z}_\mathbf{X}^T\mathbf{Z_X})^{-1}, \mathbf{I}, \mathbf{Z_Y}, \mathbf{I})$. Furthermore, these three variants—PLSC, CCA, and RDA/RRR—also generalize discriminant analyses under different optimizations so long as $\mathbf{X}$ is a dummy-coded or complete disjunctive matrix to assign each observation (row) to a specific group or category (columns).

Most importantly, because of the ways we formalized the GPLS algorithms—see also Section 2.2—and the variety of ways to suitably transform data (e.g., the various coding schemes we have shown) allow application of PLS-CA-R and GPLS algorithms on a variety of different problems or models such as log or power transformations and alternate choices for weights (see Eq. 3) or models (see Eq. 4). That means that the GPLS algorithms further generalize many approaches, especially the numerous variants of CA. Generally in the cases of strictly positive data, there may be a need to preprocess data within the family of power transformations for CA (Greenacre 2009) or alternate distance metrics such as Hellinger distance (Rao 1995, Escofier 1978). Finally, with the choices of weights can change, as they do for Hellinger CA, and for the variations of "non-symmetrical CA" (D'Ambra & Lauro 1992, Kroonenberg & Lombardo 1999, Takane et al. 1991), where both types of variants require one set of weights as $\mathbf{I}$ (akin to RDA/RRR-type optimizations with CA/χ^2 models across any of the GPLS algorithms).

4.3 Ridge-like regularization

It is also possible to apply ridge-like regularization to PLS-CA regression, correlation, and canonical decompositions. We show two possible strategies for ridge-like regularization under the data/model assumptions and the preprocessing we established in Section 2.2.

The first approach is based on Takane's regularized multiple CA (Takane & Hwang 2006) and regularized nonsymmetric CA (Takane & Jung 2009). To do so, it is convenient to slightly reformulate PLS-CA-R, but still require $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{O_X}$, $\mathbf{O_Y}$, $\mathbf{E_X}$, and $\mathbf{E_Y}$ as defined in Section 2.2. First we re-define $\mathbf{Z_X} = (\mathbf{O_X} - \mathbf{E_X}) \times (\mathbf{1}^T\mathbf{X}1)$ and $\mathbf{Z_Y} = (\mathbf{O_Y} - \mathbf{E_Y}) \times (\mathbf{1}^T\mathbf{Y}1)$; which are the same as in Eq. 5 except scaled by the grand sum of its respective source data matrix. Next we define the following additional matrices: $\mathbf{D}_{\mathbf{X},I} = \text{diag}\{\mathbf{X1}\}$, and $\mathbf{D}_{\mathbf{Y},I} = \text{diag}\{\mathbf{Y1}\}$ which are diagonal matrices of the row sums of $\mathbf{X}$ and $\mathbf{Y}$, respectively and $\mathbf{D}_{\mathbf{X},J} = \text{diag}\{\mathbf{1}^{\mathrm{T}}\mathbf{X}\}$, and $\mathbf{D}_{\mathbf{Y},K} = \text{diag}\{\mathbf{1}^{\mathrm{T}}\mathbf{Y}\}$ which are the column sums of $\mathbf{X}$ and $\mathbf{Y}$. Then PLS-CA correlation, regression, and canonical decompositions replace the GPLSSVD step in Algorithms 1, 2, 3 with GPLSSVD$(\mathbf{D}_{\mathbf{X},I}^{-1}, \mathbf{Z}_{\mathbf{X}}^T, \mathbf{D}_{\mathbf{X},J}^{-1}, \mathbf{D}_{\mathbf{Y},I}^{-1}, \mathbf{Z}_{\mathbf{Y}}^T, \mathbf{D}_{\mathbf{Y},K}^{-1})$. The only differences between this Takane-ian reformulation and what we originally established is that the generalized singular vectors ($\mathbf{P}$ and $\mathbf{Q}$) and the component scores ($\mathbf{F_J}$ and $\mathbf{F_K}$) differ by constant scaling factors (which come from the sums of $\mathbf{X}$ and $\mathbf{Y}$).

We can regularize PLS-CA-R in the same way as Takane's RMCA. To do so we require (1) a ridge parameter which we refer to as λ and (2) variants of $\mathbf{D}_{\mathbf{X},I}$, $\mathbf{D}_{\mathbf{X},J}$, $\mathbf{D}_{\mathbf{Y},I}$, and $\mathbf{D}_{\mathbf{Y},K}$ that we refer to as $\mathbb{D}_{\mathbf{X},I} = \mathbf{D}_{\mathbf{X},I} + [\lambda \times (\mathbf{Z_X}\mathbf{Z}_{\mathbf{X}}^T)^+]$, $\mathbb{D}_{\mathbf{Y},I} = \mathbf{D}_{\mathbf{Y},I} + [\lambda \times (\mathbf{Z_Y}\mathbf{Z}_{\mathbf{Y}}^T)^+]$, $\mathbb{D}_{\mathbf{X},J} = \mathbf{D}_{\mathbf{X},J} + [\lambda \times \mathbf{Z}_{\mathbf{X}}^T(\mathbf{Z_X}\mathbf{Z}_{\mathbf{X}}^T)^+\mathbf{Z_X}]$, and $\mathbb{D}_{\mathbf{Y},K} = \mathbf{D}_{\mathbf{Y},K} + [\lambda \times \mathbf{Z}_{\mathbf{Y}}^T(\mathbf{Z_Y}\mathbf{Z}_{\mathbf{Y}}^T)^+\mathbf{Z_Y}]$. When $\lambda = 0$ then $\mathbb{D}_{\mathbf{X},I} = \mathbf{D}_{\mathbf{X},I}$, $\mathbb{D}_{\mathbf{Y},I} = \mathbf{D}_{\mathbf{Y},I}$, $\mathbb{D}_{\mathbf{X},J} = \mathbf{D}_{\mathbf{X},J}$, $\mathbb{D}_{\mathbf{Y},K} = \mathbf{D}_{\mathbf{Y},K}$. We obtain regularized forms of PLS-CA for the correlation, regression, and canonical decompositions if we simply replace the GPLSSVD step in each algorithm with GPLSSVD$(\mathbb{D}_{\mathbf{X},I}^{-1}, \mathbf{Z}_{\mathbf{X}}^T, \mathbb{D}_{\mathbf{X},J}^{-1}, \mathbb{D}_{\mathbf{Y},I}^{-1}, \mathbf{Z}_{\mathbf{Y}}^T, \mathbb{D}_{\mathbf{Y},K}^{-1})$. As per Takane's recommendation (Takane & Hwang 2006), λ could be any positive value, though integers in the range from 1 to 20 provide sufficient regularization, especially as λ increases.

However, the Takane-ian approach may not be feasible when I, J, and/or K are particularly large because the various crossproduct and projection matrices require a large amount of memory and/or computational expense. So we now introduce a "truncated" version of the Takane regularization which is far more computationally efficient, and analogous to the regularization procedure of Allen (Allen 2013, Allen et al. 2014). We re-define $\mathbb{D}_{\mathbf{X},I} = \mathbf{D}_{\mathbf{X},I} + (\lambda \times \mathbf{I})$ and $\mathbb{D}_{\mathbf{Y},I} = \mathbf{D}_{\mathbf{Y},I} + (\lambda \times \mathbf{I})$ and then also $\mathbb{D}_{\mathbf{X},J} = \mathbf{D}_{\mathbf{X},J} + (\lambda \times \mathbf{I})$ and $\mathbb{D}_{\mathbf{Y},K} = \mathbf{D}_{\mathbf{Y},K} + (\lambda \times \mathbf{I})$ where $\mathbf{I}$ are identity matrices (1s on the diagonal) of appropriate size. Like in the previous formulation, we replace the values we have in the GPLSSVD step where GPLSSVD($\mathbb{D}^{-1}_{\mathbf{X},I}, \mathbf{Z}^{T}_{\mathbf{X}}, \mathbb{D}^{-1}_{\mathbf{X},J}, \mathbb{D}^{-1}_{\mathbf{Y},I}, \mathbf{Z}^{T}_{\mathbf{Y}}, \mathbb{D}^{-1}_{\mathbf{Y},K}$); and in this particular case, the constraint matrices are all diagonal matrices, which allows for a lower memory footprint and less computational burden.

Finally, we have two concluding remarks on ridge-like regularization. The first point is that the more simplified Takane/Allen hybrid approach to ridge-like regularization also applies much more generally to virtually any technique for the SVD or GPLSSVD. For any approach, we only require some inflation factor (λ) for the constraints so long as those constraints are diagonal matrices. The second point is that while we have presented ridge-like regularization with a single λ it is entirely possible to use different λs for each set of constraints. Though it is possible, we do not necessarily recommend this approach, as it would require a complex grid search over all the various λ parameters; or one could minimize the number of parameters to search and set some of the λs to 0 and, for example, use only one or two λ values instead of four possible λ values.

4.4 Conclusions

The primary motivation to develop PLS-CA-R was to address the need of many fields that require *data type general* methods. We introduced PLS-CA-R in a way that emphasizes various recoding schemes to accomodate different data types all with respect to CA and the χ^2 model. While that was the bulk of this work, our secondary goal was to further generalize the PLS-CA approach and to better unify many methods under a simpler framework, specifically by way of the GPLSSVD and our three GPLS algorithms. Thus our generalizations—first established in Section 2.2, and expanded upon in Discussion—accomodate: almost any data type, various metrics (e.g., Hellinger distance), various optimizations (e.g., PLS, CCA, or RDA type optmizations), and even two strategies for ridge-like regularization. We have foregone any discussions of inference, stability, and resampling for PLS-CA-R because, as a generalization of PLS-R, many inference and stability approaches still apply—such as feature selection or sparsification (Sutton et al. 2018), additional regularization or sparsification approaches (Le Floch et al. 2012, Guillemot et al. 2019, Tenenhaus et al. 2014, Tenenhaus & Tenenhaus 2011), cross-validation (Wold et al. 1987, Rodríguez-Pérez et al. 2018, Kvalheim et al. 2019, Abdi 2010), permutation (Berry et al. 2011), various bootstrap (Efron 1979, Chernick 2008) approaches (Abdi 2010, Takane & Jung 2009) or tests (McIntosh & Lobaugh 2004, Krishnan et al. 2011), and other frameworks such as split-half resampling (Strother et al. 2002, Kovacevic et al. 2013, Strother et al. 2004)—and are easily adapted for the PLS-CA-R and GPLS frameworks.

PLS-CA-R was designed primarily as the mixed-data generalization of PLSR that provides for us a technique that both produces latent variables and performs regression when standard assumptions are not met (e.g., HDLSS or high collinearity). PLS-CA-R—and GPLS—addresses the need of many fields that require *data type general* methods across multi-source and multi-domain data sets where we require careful considerations about how we prepare and understand our data (Nguyen & Holmes 2019). We introduced PLS-CA-R in a way that emphasizes various recoding schemes to accomodate different data types all with respect to CA and the χ^2 model. PLS-CA-R provides key features necessary for data analyses as data-rich and data-heavy disciplines and fields rapidly move towards and depend on fundamental techniques in machine and statistical learning (e.g., PLSR, CCA). Finally, with techniques such as mixed-data MFA (Bécue-Bertaut & Pagès 2008), PLS-CA-R provides a much needed basis for development of future methods designed forsuch complex data sets.

5 Acknowledgements

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 DOD 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: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; 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.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. 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 Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

Abdi, H. (2010), 'Partial least squares regression and projection on latent structure regression (PLS Regression)', *Wiley Interdisciplinary Reviews: Computational Statistics* **2**(1), 97–106.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1002/wics.51*

Abdi, H. & Béra, M. (2014), 'Correspondence analysis', *Encyclopedia of social network analysis and mining* pp. 275–284.

Abdi, H., Guillemot, V., Eslami, A. & Beaton, D. (2017), 'Canonical correlation analysis', *Encyclopedia of Social Network Analysis and Mining* pp. 1–16. Springer.

Allen, G. I. (2013), 'Sparse and Functional Principal Components Analysis',

arXiv:1309.2895 [stat] . arXiv: 1309.2895.
URL: *http://arxiv.org/abs/1309.2895*

Allen, G. I., Grosenick, L. & Taylor, J. (2014), 'A Generalized Least-Square Matrix Decomposition', *Journal of the American Statistical Association* **109**(505), 145–159.
URL: *http://dx.doi.org/10.1080/01621459.2013.852978*

Bécue-Bertaut, M. & Pagès, J. (2008), 'Multiple factor analysis and clustering of a mixture of quantitative, categorical and frequency data', *Computational Statistics & Data Analysis* **52**(6), 3255–3268.
URL: *http://www.sciencedirect.com/science/article/pii/S016794730700360X*

Beaton, D., Dunlop, J., Abdi, H. & Alzheimer's Disease Neuroimaging Initiative (2016), 'Partial Least Squares Correspondence Analysis: A Framework to Simultaneously Analyze Behavioral and Genetic Data', *Psychological Methods* **21**(4), 621–651.

Beaton, D., Sunderland, K. M., Levine, B., Mandzia, J., Masellis, M., Swartz, R. H., Troyer, A. K., Binns, M. A., Abdi, H., Strother, S. C. & others (2018), 'Generalization of the minimum covariance determinant algorithm for categorical and mixed data types', *bioRxiv* p. 333005.

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: JAD* **22**(1), 129–134.

Benzécri, J. P. (1973), *L'analyse des données: L'analyse des correspondances*, Dunod. Google-Books-ID: sDTwAAAAMAAJ.

Berry, K. J., Johnston, J. E. & Mielke, P. W. (2011), 'Permutation methods', *Wiley Interdisciplinary Reviews: Computational Statistics* **3**, 527–542.
URL: *http://wires.wiley.com/WileyCDA/WiresArticle/wisId-WICS177.html*

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

Borga, M., Landelius, T. & Knutsson, H. (1992), *A Unified Approach to PCA, PLS, MLR and CCA*.

Bürkner, P. C. & Vuorre, M. (n.d.), 'Ordinal Regression Models in Psychology: A Tutorial'. **URL:** *https://osf.io/x8swp*

Chernick, M. (2008), *Bootstrap methods: A guide for practitioners and researchers*, Vol. 619, Wiley-Interscience.

Cruchaga, C., Chakraverty, S., Mayo, K., Vallania, F. L. M., Mitra, R. D., Faber, K., Williamson, J., Bird, T., Diaz-Arrastia, R., Foroud, T. M., Boeve, B. F., Graff-Radford, N. R., St. Jean, P., Lawson, M., Ehm, M. G., Mayeux, R., Goate, A. M. & for the NIA-LOAD/NCRAD Family Study Consortium (2012), 'Rare Variants in APP, PSEN1 and PSEN2 Increase Risk for AD in Late-Onset Alzheimer's Disease Families', *PLoS ONE* **7**(2), e31039.
URL: *http://dx.doi.org/10.1371/journal.pone.0031039*

D'Ambra, L. & Lauro, N. C. (1992), 'Non symmetrical exploratory data analysis', *Statistica Applicata* **4**(4), 511–529.

de Micheaux, P. L., Liquet, B., Sutton, M. et al. (2019), 'Pls for big data: A unified parallel algorithm for regularised group pls', *Statistics Surveys* **13**, 119–149.

Desikan, R. S., Schork, A. J., Wang, Y., Witoelar, A., Sharma, M., McEvoy, L. K., Holland, D., Brewer, J. B., Chen, C.-H., Thompson, W. K., Harold, D., Williams, J., Owen, M. J., O'Donovan, M. C., Pericak-Vance, M. A., Mayeux, R., Haines, J. L., Farrer, L. A., Schellenberg, G. D., Heutink, P., Singleton, A. B., Brice, A., Wood, N. W., Hardy, J., Martinez, M., Choi, S. H., DeStefano, A., Ikram, M. A., Bis, J. C., Smith, A., Fitzpatrick, A. L., Launer, L., van Duijn, C., Seshadri, S., Ulstein, I. D., Aarsland, D., Fladby, T., Djurovic, S., Hyman, B. T., Snaedal, J., Stefansson, H., Stefansson, K., Gasser, T., Andreassen, O. A. & Dale, A. M. (2015), 'Genetic overlap between Alzheimer's disease and Parkinson's disease at the MAPT locus', *Molecular Psychiatry* .
URL: *http://www.nature.com/mp/journal/vaop/ncurrent/full/mp20156a.html*

Duchesne, S., Chouinard, I., Potvin, O., Fonov, V. S., Khademi, A., Bartha, R., Bellec, P., Collins, D. L., Descoteaux, M., Hoge, R., McCreary, C. R., Ramirez, J., Scott, C. J. M., Smith, E. E., Strother, S. C. & Black, S. E. (2019), 'The Canadian Dementia Imaging Protocol: Harmonizing National Cohorts', *Journal of Magnetic Resonance Imaging* **49**(2), 456–465.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1002/jmri.26197*

Efron, B. (1979), 'Bootstrap Methods: Another Look at the Jackknife', *The Annals of Statistics* **7**(1), 1–26. ArticleType: primary_article / Full publication date: Jan., 1979 / Copyright © 1979 Institute of Mathematical Statistics.
URL: *http://www.jstor.org.libproxy.utdallas.edu/stable/2958830*

Escofier, B. (1978), 'Analyse factorielle et distances répondant au principe d'équivalence distributionnelle', *Revue de statistique appliquée* **26**(4), 29–37.

Escofier, B. (1979), 'Traitement simultané de variables qualitatives et quantitatives en analyse factorielle', *Cahiers de l'Analyse des Données* **4**(2), 137–146.

Escofier, B. (1983), 'Analyse de la différence entre deux mesures définies sur le produit de deux mêmes ensembles', *Cahiers de l'Analyse des Données* **8**(3), 325–329.

Escofier, B. (1984), 'Analyse factorielle en reférence à un modéle. Application à lanalyse de tableaux dechanges', *Revue de Statistique Appliquée* **32**(4), 25–36.

Escofier-Cordier, B. (1965), L'Analyse des Correspondences., Thèse, Faculté des Sciences de Rennes, Université de Rennes.

Farhan, S. M. K., Bartha, R., Black, S. E., Corbett, D., Finger, E., Freedman, M., Greenberg, B., Grimes, D. A., Hegele, R. A., Hudson, C., Kleinstiver, P. W., Lang, A. E., Masellis, M., McIlroy, W. E., McLaughlin, P. M., Montero-Odasso, M., Munoz, D. G., Munoz, D. P., Strother, S., Swartz, R. H., Symons, S., Tartaglia, M. C., Zinman, L. & Strong, M. J. (2016), 'The Ontario Neurodegenerative Disease Research Initiative (ONDRI)', *Canadian Journal of Neurological Sciences* pp. 1–7.

URL: *https://www.cambridge.org/core/journals/canadian-journal-of-neurological-sciences/article/div-classtitlethe-ontario-neurodegenerative-disease-research-initiative-ondridiv/3D9558108D69BBF1A4B158DCF2EF6329/core-reader*

Genin, E., Hannequin, D., Wallon, D., Sleegers, K., Hiltunen, M., Combarros, O., Bullido, M. J., Engelborghs, S., De Deyn, P., Berr, C., Pasquier, F., Dubois, B., Tognoni, G., Fiévet, N., Brouwers, N., Bettens, K., Arosio, B., Coto, E., Del Zompo, M., Mateo, I., Epelbaum, J., Frank-Garcia, A., Helisalmi, S., Porcellini, E., Pilotto, A., Forti, P., Ferri, R., Scarpini, E., Siciliano, G., Solfrizzi, V., Sorbi, S., Spalletta, G., Valdivieso, F., Vepsäläinen, S., Alvarez, V., Bosco, P., Mancuso, M., Panza, F., Nacmias, B., Bossù, P., Hanon, O., Piccardi, P., Annoni, G., Seripa, D., Galimberti, D., Licastro, F., Soininen, H., Dartigues, J.-F., Kamboh, M. I., Van Broeckhoven, C., Lambert, J. C., Amouyel, P. & Campion, D. (2011), 'APOE and Alzheimer disease: a major gene with semi-dominant inheritance', *Molecular psychiatry* **16**(9), 903–907.

Greenacre, M. (2009), 'Power transformations in correspondence analysis', *Computational Statistics & Data Analysis* **53**(8), 3107–3116.

Greenacre, M. (2014), Data Doubling and Fuzzy Coding, *in* J. Blasius & M. Greenacre, eds, 'Visualization and Verbalization of Data', CRC Press, Philadelphia, PA, USA, pp. 239–253.

Greenacre, M. J. (1984), *Theory and Applications of Correspondence Analysis*, Academic Press.
URL: *http://books.google.com/books?id=LsPaAAAAMAAJ*

Greenacre, M. J. (1993), 'Biplots in correspondence analysis', *Journal of Applied Statistics* **20**(2), 251–269.

Greenacre, M. J. (2010), 'Correspondence analysis', *Wiley Interdisciplinary Reviews: Computational Statistics* **2**(5), 613–619.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1002/wics.114*

Guillemot, V., Beaton, D., Gloaguen, A., Löfstedt, T., Levine, B., Raymond, N., Tenenhaus, A. & Abdi, H. (2019), 'A constrained singular value decomposition method that integrates sparsity and orthogonality', *PloS one* **14**(3), e0211463.

Holmes, S. (2008), 'Multivariate data analysis: The French way', *arXiv:0805.2879 [stat]* pp. 219–233. arXiv: 0805.2879.
URL: *http://arxiv.org/abs/0805.2879*

Huang, Y.-W. A., Zhou, B., Wernig, M. & Südhof, T. C. (2017), 'ApoE2, ApoE3, and ApoE4 Differentially Stimulate APP Transcription and Aβ Secretion', *Cell* .
URL: *//www.sciencedirect.com/science/article/pii/S0092867416317603*

Indahl, U. G., Liland, K. H. & Næs, T. (2009), 'Canonical partial least squares—a unified pls approach to classification and regression problems', *Journal of Chemometrics: A Journal of the Chemometrics Society* **23**(9), 495–504.

Jonsson, T., Atwal, J. K., Steinberg, S., Snaedal, J., Jonsson, P. V., Bjornsson, S., Stefansson, H., Sulem, P., Gudbjartsson, D., Maloney, J., Hoyte, K., Gustafson, A., Liu, Y., Lu, Y., Bhangale, T., Graham, R. R., Huttenlocher, J., Bjornsdottir, G., Andreassen, O. A., Jönsson, E. G., Palotie, A., Behrens, T. W., Magnusson, O. T., Kong, A., Thorsteinsdottir, U., Watts, R. J. & Stefansson, K. (2012), 'A mutation in *APP* protects against Alzheimer's disease and age-related cognitive decline', *Nature* **488**(7409), 96–99.
URL: *https://www.nature.com/articles/nature11283*

Ketterlinus, R. D., Bookstein, F. L., Sampson, P. D. & Lamb, M. E. (1989), 'Partial least squares analysis in developmental psychopathology', *Development and Psychopathology* **1**(4), 351–371.

Kovacevic, N., Abdi, H., Beaton, D. & McIntosh, A. R. (2013), Revisiting pls resampling: comparing significance versus reliability across range of simulations, *in* 'New Perspectives in Partial Least Squares and Related Methods', Springer, pp. 159–170.

Krishnan, A., Williams, L. J., McIntosh, A. R. & Abdi, H. (2011), 'Partial Least Squares (PLS) methods for neuroimaging: A tutorial and review', *NeuroImage* **56**(2), 455 – 475. Multivariate Decoding and Brain Reading.
URL: *http://www.sciencedirect.com/science/article/pii/S1053811910010074*

Kroonenberg, P. M. & Lombardo, R. (1999), 'Nonsymmetric correspondence analysis: A tool for analysing contingency tableswith a dependence structure', *Multivariate Behavioral Research* **34**(3), 367–396.

Kvalheim, O. M., Grung, B. & Rajalahti, T. (2019), 'Number of components and prediction error in partial least squares regression determined by Monte Carlo resampling strategies', *Chemometrics and Intelligent Laboratory Systems* .
URL: *http://www.sciencedirect.com/science/article/pii/S0169743918307056*

Le Floch, d., Guillemot, V., Frouin, V., Pinel, P., Lalanne, C., Trinchera, L., Tenenhaus, A., Moreno, A., Zilbovicius, M., Bourgeron, T., Dehaene, S., Thirion, B., Poline, J.-B. & Duchesnay, d. (2012), 'Significant correlation between a set of genetic polymorphisms and a functional brain network revealed by feature selection and sparse Partial Least Squares', *NeuroImage* **63**(1), 11–24.
URL: *http://www.sciencedirect.com/science/article/pii/S1053811912006775*

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

Lettre, G., Lange, C. & Hirschhorn, J. N. (2007), 'Genetic model testing and statistical power in population-based association studies of quantitative traits', *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society* **31**(4), 358–362.

Linnertz, C., Anderson, L., Gottschalk, W., Crenshaw, D., Lutz, M. W., Allen, J., Saith, S., Mihovilovic, M., Burke, J. R., Welsh-Bohmer, K. A., Roses, A. D. & Chiba-Falek, O. (2014), 'The cis-regulatory effect of an Alzheimer's disease-associated poly-T locus on expression of TOMM40 and apolipoprotein E genes', *Alzheimer's & Dementia* **10**(5), 541–551.
URL: *http://www.sciencedirect.com/science/article/pii/S155252601302801X*

McIntosh, A., Bookstein, F., Haxby, J. & Grady, C. (1996), 'Spatial Pattern Analysis of Functional Brain Images Using Partial Least Squares', *NeuroImage* **3**(3), 143–157.
URL: *http://www.sciencedirect.com/science/article/pii/S1053811996900166*

McIntosh, A. R. & Lobaugh, N. J. (2004), 'Partial least squares analysis of neuroimaging data: applications and advances', *Neuroimage* **23**, S250–S263.

Montero-Odasso, M., Pieruccini-Faria, F., Bartha, R., Black, S. E., Finger, E., Freedman, M., Greenberg, B., Grimes, D. A., Hegele, R. A., Hudson, C., Kleinstiver, P. W., Lang, A. E., Masellis, M., McLaughlin, P. M., Munoz, D. P., Strother, S., Swartz, R. H., Symons, S., Tartaglia, M. C., Zinman, L., Strong, M. J., ONDRI Investigators & McIlroy, W. (2017), 'Motor Phenotype in Neurodegenerative Disorders: Gait and Balance Platform Study Design Protocol for the Ontario Neurodegenerative Research Initiative (ONDRI)', *Journal of Alzheimer's disease: JAD* **59**(2), 707–721.

Morris, J. C. (1993), 'The Clinical Dementia Rating (CDR): current version and scoring rules.', *Neurology* .

Myers, A. J., Kaleem, M., Marlowe, L., Pittman, A. M., Lees, A. J., Fung, H. C., Duckworth, J., Leung, D., Gibson, A., Morris, C. M., Silva, R. d. & Hardy, J. (2005), 'The H1c haplotype at the MAPT locus is associated with Alzheimer's disease', *Human Molecular Genetics* **14**(16), 2399–2404.
URL: *http://hmg.oxfordjournals.org/content/14/16/2399*

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J. L. & Chertkow, H. (2005), 'The Montreal Cognitive Assessment, MoCA: A Brief Screening Tool For Mild Cognitive Impairment', *Journal of the American Geriatrics Society* **53**(4), 695–699.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1532-5415.2005.53221.x*

Nguyen, L. H. & Holmes, S. (2019), 'Ten quick tips for effective dimensionality reduction', *PLOS Computational Biology* **15**(6), e1006907.

Peterson, D., Munger, C., Crowley, J., Corcoran, C., Cruchaga, C., Goate, A. M., Norton, M. C., Green, R. C., Munger, R. G., Breitner, J. C. S., Welsh-Bohmer, K. A., Lyketsos, C., Tschanz, J. & Kauwe, J. S. K. (2014), 'Variants in PPP3r1 and MAPT are associated with more rapid functional decline in Alzheimer's disease: The Cache County Dementia

Progression Study', *Alzheimer's & Dementia* **10**(3), 366–371.
URL: *http://www.sciencedirect.com/science/article/pii/S1552526013000861*

Pérez-Enciso, M. & Tenenhaus, M. (2003), 'Prediction of clinical outcome with microarray data: a partial least squares discriminant analysis (PLS-DA) approach', *Human Genetics* **112**(5), 581–592.
URL: *https://doi.org/10.1007/s00439-003-0921-9*

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J. et al. (2007), 'Plink: a tool set for whole-genome association and population-based linkage analyses', *The American journal of human genetics* **81**(3), 559–575.

Rao, C. R. (1995), 'A review of canonical coordinates and an alternative to correspondence analysis using Hellinger distance', *Qüestiió: quaderns d'estadística i investigació operativa* **19**(1).

Rodríguez-Pérez, R., Fernández, L. & Marco, S. (2018), 'Overoptimism in cross-validation when using partial least squares-discriminant analysis for omics data: a systematic study', *Analytical and Bioanalytical Chemistry* **410**(23), 5981–5992.
URL: *https://doi.org/10.1007/s00216-018-1217-1*

Roses, A. D., Lutz, M. W., Amrine-Madsen, H., Saunders, A. M., Crenshaw, D. G., Sundseth, S. S., Huentelman, M. J., Welsh-Bohmer, K. A. & Reiman, E. M. (2010), 'A TOMM40 variable-length polymorphism predicts the age of late-onset Alzheimer's disease', *The Pharmacogenomics Journal* **10**(5), 375–384.

Skinner, J., Carvalho, J. O., Potter, G. G., Thames, A., Zelinski, E., Crane, P. K. & Gibbons, L. E. (2012), 'The Alzheimer's Disease Assessment Scale-Cognitive-Plus (ADAS-Cog-Plus): an expansion of the ADAS-Cog to improve responsiveness in MCI', *Brain imaging and behavior* **6**(4).
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3873823/*

Strother, S. C., Anderson, J., Hansen, L. K., Kjems, U., Kustra, R., Sidtis, J., Frutiger, S., Muley, S., LaConte, S. & Rottenberg, D. (2002), 'The Quantitative Evaluation of

Functional Neuroimaging Experiments: The NPAIRS Data Analysis Framework', *NeuroImage* **15**(4), 747–771.
URL: *http://www.sciencedirect.com/science/article/pii/S1053811901910341*

Strother, S., La Conte, S., Hansen, L. K., Anderson, J., Zhang, J., Pulapura, S. & Rottenberg, D. (2004), 'Optimizing the fmri data-processing pipeline using prediction and reproducibility performance metrics: I. a preliminary group analysis', *Neuroimage* **23**, S196–S207.

Sutton, M., Thiébaut, R. & Liquet, B. (2018), 'Sparse partial least squares with group and subgroup structure', *Statistics in Medicine* **37**(23), 3338–3356.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.7821*

Takane, Y. (2003), Relationships among Various Kinds of Eigenvalue and Singular Value Decompositions, *in* H. Yanai, A. Okada, K. Shigemasu, Y. Kano & J. J. Meulman, eds, 'New Developments in Psychometrics', Springer Japan, pp. 45–56.

Takane, Y. & Hwang, H. (2006), 'Regularized multiple correspondence analysis', *Multiple correspondence analysis and related methods* pp. 259–279.

Takane, Y. & Jung, S. (2009), 'Regularized nonsymmetric correspondence analysis', *Computational Statistics & Data Analysis* **53**(8), 3159–3170.
URL: *http://www.sciencedirect.com/science/article/pii/S0167947308004313*

Takane, Y., Yanai, H. & Mayekawa, S. (1991), 'Relationships among several methods of linearly constrained correspondence analysis', *Psychometrika* **56**(4), 667–684.

Tenenhaus, A., Philippe, C., Guillemot, V., Cao, K.-A. L., Grill, J. & Frouin, V. (2014), 'Variable selection for generalized canonical correlation analysis', *Biostatistics* p. kxu001.
URL: *http://biostatistics.oxfordjournals.org/content/early/2014/02/17/biostatistics.kxu001*

Tenenhaus, A. & Tenenhaus, M. (2011), 'Regularized Generalized Canonical Correlation Analysis', *Psychometrika* **76**(2), 257–284.
URL: *http://www.springerlink.com/content/8r17621w56k3025w/abstract/*

Tenenhaus, M. (1998), *La régression PLS: théorie et pratique*, Editions TECHNIP. Google-Books-ID: OesjK2KZhsAC.

Trabzuni, D., Wray, S., Vandrovcova, J., Ramasamy, A., Walker, R., Smith, C., Luk, C., Gibbs, J. R., Dillman, A., Hernandez, D. G., Arepalli, S., Singleton, A. B., Cookson, M. R., Pittman, A. M., Silva, R. d., Weale, M. E., Hardy, J. & Ryten, M. (2012), 'MAPT expression and splicing is differentially regulated by brain region: relation to genotype and implication for tauopathies', *Human Molecular Genetics* **21**(18), 4094–4103.
URL: *http://hmg.oxfordjournals.org/content/21/18/4094*

Tucker, L. R. (1958), 'An inter-battery method of factor analysis', *Psychometrika* **23**(2), 111–136.
URL: *http://www.springerlink.com/content/u74149nr2h0124l0/*

Vormfelde, S. V. & Brockmöller, J. (2007), 'On the value of haplotype-based genotype–phenotype analysis and on data transformation in pharmacogenetics and -genomics', *Nature Reviews Genetics* **8**(12).
URL: *http://www.nature.com.libproxy.utdallas.edu/nrg/journal/v8/n12/full/nrg1916-c1.html*

Wegelin, J. A. et al. (2000), 'A survey of partial least squares (pls) methods, with emphasis on the two-block case', *University of Washington, Tech. Rep* .

Wold, H. (1975), 'Soft modelling by latent variables: the non-linear iterative partial least squares (NIPALS) approach', *Perspectives in Probability and Statistics, In Honor of MS Bartlett* pp. 117–144.

Wold, S., Esbensen, K. & Geladi, P. (1987), 'Principal component analysis', *Chemometrics and Intelligent Laboratory Systems* **2**(1), 37–52.
URL: *http://www.sciencedirect.com/science/article/pii/0169743987800849*

Wold, S., Ruhe, A., Wold, H. & Dunn, III, W. (1984), 'The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses', *SIAM Journal on Scientific and Statistical Computing* **5**(3), 735–743.

Wold, S., Sjöström, M. & Eriksson, L. (2001), 'PLS-regression: a basic tool of chemometrics', *Chemometrics and Intelligent Laboratory Systems* **58**(2), 109–130.
URL: *http://www.sciencedirect.com/science/article/pii/S0169743901001551*