Jendoubi and Strimmer *BMC Bioinformatics* (2019) 20:15
https://doi.org/10.1186/s12859-018-2572-9

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

# A whitening approach to probabilistic canonical correlation analysis for omics data integration

Takoua Jendoubi[1,2]* and Korbinian Strimmer[3]

## Abstract

**Background:** Canonical correlation analysis (CCA) is a classic statistical tool for investigating complex multivariate data. Correspondingly, it has found many diverse applications, ranging from molecular biology and medicine to social science and finance. Intriguingly, despite the importance and pervasiveness of CCA, only recently a probabilistic understanding of CCA is developing, moving from an algorithmic to a model-based perspective and enabling its application to large-scale settings.

**Results:** Here, we revisit CCA from the perspective of statistical whitening of random variables and propose a simple yet flexible probabilistic model for CCA in the form of a two-layer latent variable generative model. The advantages of this variant of probabilistic CCA include non-ambiguity of the latent variables, provisions for negative canonical correlations, possibility of non-normal generative variables, as well as ease of interpretation on all levels of the model. In addition, we show that it lends itself to computationally efficient estimation in high-dimensional settings using regularized inference. We test our approach to CCA analysis in simulations and apply it to two omics data sets illustrating the integration of gene expression data, lipid concentrations and methylation levels.

**Conclusions:** Our whitening approach to CCA provides a unifying perspective on CCA, linking together sphering procedures, multivariate regression and corresponding probabilistic generative models. Furthermore, we offer an efficient computer implementation in the "whitening" R package available at https://CRAN.R-project.org/package=whitening.

**Keywords:** Multivariate analysis, Probabilistic canonical correlation analysis, Data integration

## Background

Canonical correlation analysis (CCA) is a classic and highly versatile statistical approach to investigate the linear relationship between two sets of variables [1, 2]. CCA helps to decode complex dependency structures in multivariate data and to identify groups of interacting variables. Consequently, it has numerous practical applications in molecular biology, for example omics data integration [3] and network analysis [4], but also in many other areas such as econometrics or social science.

In its original formulation CCA is viewed as an algorithmic procedure optimizing a set of objective functions, rather than as a probablistic model for the data. Only relatively recently this perspective has changed. Bach and Jordan [5] proposed a latent variable model for CCA building on earlier work on probabilistic principal component analysis (PCA) by [6]. The probabilistic approach to CCA not only allows to derive the classic CCA algorithm but also provide an avenue for Bayesian variants [7, 8].

In parallel to establishing probabilistic CCA the classic CCA approach has also been further developed in the last decade by introducing variants of the CCA algorithm that are more pertinent for high-dimensional data sets now routinely collected in the life and physical sciences. In particular, the problem of singularity in the original CCA

*Correspondence: t.jendoubi14@imperial.ac.uk
[1] Epidemiology and Biostatistics, School of Public Health, Imperial College London, Norfolk Place, W2 1PG London, UK
[2] Statistics Section, Department of Mathematics, Imperial College London, South Kensington Campus, SW7 2AZ London, UK
Full list of author information is available at the end of the article

algorithm is resolved by introducing sparsity and regularization [9–13] and, similarly, large-scale computation is addressed by new algorithms [14, 15].

In this note, we revisit both classic and probabilistic CCA from the perspective of whitening of random variables [16]. As a result, we propose a simple yet flexible probabilistic model for CCA linking together multivariate regression, latent variable models, and high-dimensional estimation. Crucially, this model for CCA not only facilitates comprehensive understanding of both classic and probabilistic CCA via the process of whitening but also extends CCA by allowing for negative canonical correlations and providing the flexibility to include non-normal latent variables.

The remainder of this paper is as follows. First, we present our main results. After reviewing classical CCA we demonstrate that the classic CCA algorithm is special form of whitening. Next, we show that the link of CCA with multivariate regression leads to a probabilistic two-level latent variable model for CCA that directly reproduces classic CCA without any rotational ambiguity. Subsequently, we discuss our approach by applying it to both synthetic data as well as to multiple integrated omics data sets. Finally, we describe our implementation in R and highlight computational and algorithmic aspects.

Much of our discussion is framed in terms of random vectors and their properties rather than in terms of data matrices. This allows us to study the probabilistic model underlying CCA separate from associated statistical procedures for estimation.

### Multivariate notation

We consider two random vectors $\boldsymbol{X} = \left(X_1, \ldots, X_p\right)^T$ and $\boldsymbol{Y} = \left(Y_1, \ldots, Y_q\right)^T$ of dimension $p$ and $q$. Their respective multivariate distributions $F_{\boldsymbol{X}}$ and $F_{\boldsymbol{Y}}$ have expectation $\text{E}(\boldsymbol{X}) = \boldsymbol{\mu}_{\boldsymbol{X}}$ and $\text{E}(\boldsymbol{Y}) = \boldsymbol{\mu}_{\boldsymbol{Y}}$ and covariance $\text{var}(\boldsymbol{X}) = \boldsymbol{\Sigma}_{\boldsymbol{X}}$ and $\text{var}(\boldsymbol{Y}) = \boldsymbol{\Sigma}_{\boldsymbol{Y}}$. The cross-covariance between $\boldsymbol{X}$ and $\boldsymbol{Y}$ is given by $\text{cov}(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{\Sigma}_{\boldsymbol{XY}}$. The corresponding correlation matrices are denoted by $\boldsymbol{P}_{\boldsymbol{X}}$, $\boldsymbol{P}_{\boldsymbol{Y}}$, and $\boldsymbol{P}_{\boldsymbol{XY}}$. By $\boldsymbol{V}_{\boldsymbol{X}} = \text{diag}(\boldsymbol{\Sigma}_{\boldsymbol{X}})$ and $\boldsymbol{V}_{\boldsymbol{Y}} = \text{diag}(\boldsymbol{\Sigma}_{\boldsymbol{Y}})$ we refer to the diagonal matrices containing the variances only, allowing to decompose covariances as $\boldsymbol{\Sigma} = \boldsymbol{V}^{1/2}\boldsymbol{P}\boldsymbol{V}^{1/2}$. The composite vector $\left(\boldsymbol{X}^T, \boldsymbol{Y}^T\right)^T$ has therefore mean $\left(\boldsymbol{\mu}_{\boldsymbol{X}}^T, \boldsymbol{\mu}_{\boldsymbol{Y}}^T\right)^T$ and covariance $\begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{X}} & \boldsymbol{\Sigma}_{\boldsymbol{XY}} \\ \boldsymbol{\Sigma}_{\boldsymbol{XY}}^T & \boldsymbol{\Sigma}_{\boldsymbol{Y}} \end{pmatrix}$.

Vector-valued samples of the random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ are denoted by $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ so that $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_i, \ldots, \boldsymbol{x}_n)^T$ is the $n \times p$ data matrix for $\boldsymbol{X}$ containing $n$ observed samples (one in each row). Correspondingly, the empirical mean for $\boldsymbol{X}$ is given by $\hat{\boldsymbol{\mu}}_{\boldsymbol{X}} = \bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i$, the unbiased covariance estimate is $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{X}} = \boldsymbol{S}_{\boldsymbol{X}} = \frac{1}{n-1}\sum_{i=1}^n (\boldsymbol{x}_i - \bar{\boldsymbol{x}})\,(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T$, and the corresponding correlation estimate is denoted by $\widehat{\boldsymbol{P}}_{\boldsymbol{X}} = \boldsymbol{R}_{\boldsymbol{X}}$.

## Results

We first introduce CCA from a classical perspective, then we demonstrate that CCA is best understood as a special and uniquely defined type of whitening transformation. Next, we investigate the close link of CCA with multivariate regression. This not only allows to interpret CCA as regression model and to better understand canonical correlations, but also provides the basis for a probabilistic generative latent variable model of CCA based on whitening. This model is introduced in the last subsection.

### Classical CCA

In canonical correlation analysis the aim is to find mutually orthogonal pairs of maximally correlated linear combinations of the components of $\boldsymbol{X}$ and of $\boldsymbol{Y}$. Specifically, we seek canonical directions $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}_j$ (i.e. vectors of dimension $p$ and $q$, respectively) for which

$$\text{cor}\left(\boldsymbol{\alpha}_i^T\boldsymbol{X}, \boldsymbol{\beta}_j^T\boldsymbol{Y}\right) = \begin{cases} \lambda_i & \text{maximal for } i = j \\ 0 & \text{otherwise,} \end{cases} \quad (1)$$

where $\lambda_i$ are the canonical correlations, and simultaneously

$$\text{cor}\left(\boldsymbol{\alpha}_i^T\boldsymbol{X}, \boldsymbol{\alpha}_j^T\boldsymbol{X}\right) = \begin{cases} 1 & \text{for } i = j \\ 0 & \text{otherwise,} \end{cases} \quad (2)$$

and

$$\text{cor}\left(\boldsymbol{\beta}_i^T\boldsymbol{Y}, \boldsymbol{\beta}_j^T\boldsymbol{Y}\right) = \begin{cases} 1 & \text{for } i = j \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

In matrix notation, with $\boldsymbol{A} = \left(\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_p\right)^T$, $\boldsymbol{B} = \left(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_q\right)^T$, and $\boldsymbol{\Lambda} = \text{diag}(\lambda_i)$, the above can be written as $\text{cor}(\boldsymbol{AX}, \boldsymbol{BY}) = \boldsymbol{\Lambda}$ as well as $\text{cor}(\boldsymbol{AX}) = \boldsymbol{I}$ and $\text{cor}(\boldsymbol{BY}) = \boldsymbol{I}$. The projected vectors $\boldsymbol{AX}$ and $\boldsymbol{BY}$ are also called the CCA scores or the canonical variables.

Hotelling (1936) [1] showed that there are, assuming full rank covariance matrices $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{Y}}$, exactly $m = \min(p, q)$ canonical correlations and pairs of canonical directions $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}_i$, and that these can be computed analytically from a generalized eigenvalue problem (e.g., [2]). Further below we will see how canonical directions and correlations follow almost effortlessly from a whitening perspective of CCA.

Since correlations are invariant against rescaling, optimizing Eq. 1 determines the canonical directions $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}_i$ only up to their respective lengths, and we can thus arbitrarily fix the magnitude of the vectors $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}_i$. A common choice is to simply normalize them to unit length so that $\boldsymbol{\alpha}_i^T\boldsymbol{\alpha}_i = 1$ and $\boldsymbol{\beta}_i^T\boldsymbol{\beta}_i = 1$.

Similarly, the overall sign of the canonical directions $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}_j$ is also undetermined. As a result, different implementations of CCA may yield canonical directions with different signs, and depending on the adopted convention this can be used either to enforce positive or to allow negative canonical correlations, see below for further discussion in the light of CCA as a regression model.

Because it optimizes correlation, CCA is invariant against location translation of the original vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$, yielding identical canonical directions and correlations in this case. However, under scale transformation of $\boldsymbol{X}$ and $\boldsymbol{Y}$ only the canonical correlations $\lambda_i$ remain invariant whereas the directions will differ as they depend on the variances $\boldsymbol{V}_X$ and $\boldsymbol{V}_Y$. Therefore, to facilitate comparative analysis and interpretation the canonical directions the random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ (and associated data) are often standardized.

Classical CCA uses the empirical covariance matrix $\boldsymbol{S}$ to obtain canonical correlations and directions. However, $\boldsymbol{S}$ can only be safely employed if the number of observations is much larger than the dimensions of either of the two random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$, since otherwise $\boldsymbol{S}$ constitutes only a poor estimate of the underlying covariance structure and in addition may also become singular. Therefore, to render CCA applicable to small sample high-dimensional data two main strategies are common: one is to directly employ regularization on the level of the covariance and correlation matrices to stabilize and improve their estimation; the other is to devise probabilistic models for CCA to facilitate application of Bayesian inference and other regularized statistical procedures.

## Whitening transformations and CCA

### *Background on whitening*

Whitening, or sphering, is a linear statistical transformation that converts a random vector $\boldsymbol{X}$ with covariance matrix $\boldsymbol{\Sigma}_X$ into a random vector

$$\widetilde{\boldsymbol{X}} = \boldsymbol{W}_X \boldsymbol{X} \tag{4}$$

with unit diagonal covariance $\text{var}\left(\widetilde{\boldsymbol{X}}\right) = \boldsymbol{\Sigma}_{\widetilde{X}} = \boldsymbol{I}_p$. The matrix $\boldsymbol{W}_X$ is called the *whitening matrix* or *sphering matrix* for $\boldsymbol{X}$, also known as the *unmixing matrix*. In order to achieve whitening the matrix $\boldsymbol{W}_X$ has to satisfy the condition $\boldsymbol{W}_X\boldsymbol{\Sigma}_X\boldsymbol{W}_X^T = \boldsymbol{I}_p$, but this by itself is not sufficient to completely identify $\boldsymbol{W}_X$. There are still infinitely many possible whitening transformations, and the family of whitening matrices for $\boldsymbol{X}$ can be written as

$$\boldsymbol{W}_X = \boldsymbol{Q}_X\boldsymbol{P}_X^{-1/2}\boldsymbol{V}_X^{-1/2}. \tag{5}$$

Here, $\boldsymbol{Q}_X$ is an orthogonal matrix; therefore the whitening matrix $\boldsymbol{W}_X$ itself is not orthogonal unless $\boldsymbol{P}_X = \boldsymbol{V}_X = \boldsymbol{I}_p$. The choice of $\boldsymbol{Q}_X$ determines the type of whitening [16]. For example, using $\boldsymbol{Q}_X = \boldsymbol{I}_p$ leads to ZCA-cor whitening, also known as Mahalanobis whitening based on the correlation matrix. PCA-cor whitening, another widely used sphering technique, is obtained by setting $\boldsymbol{Q}_X = \boldsymbol{G}^T$, where $\boldsymbol{G}$ is the eigensystem resulting from the spectral decomposition of the correlation matrix $\boldsymbol{P}_X = \boldsymbol{G}\boldsymbol{\Theta}\boldsymbol{G}^T$. Since there is a sign ambiguity in the eigenvectors $\boldsymbol{G}$ we adopt the convention of [16] to adjust columns signs of $\boldsymbol{G}$, or equivalently row signs of $\boldsymbol{Q}_x$, so that the rotation matrix $\boldsymbol{Q}_X$ has a positive diagonal.

The corresponding inverse relation $\boldsymbol{X} = \boldsymbol{W}_X^{-1}\widetilde{\boldsymbol{X}} = \boldsymbol{\Phi}_X^T\widetilde{\boldsymbol{X}}$ is called a *coloring* transformation, where the matrix $\boldsymbol{W}_X^{-1} = \boldsymbol{\Phi}_X^T$ is the *mixing matrix*, or *coloring matrix* that we can write in terms of rotation matrix $\boldsymbol{Q}_X$ as

$$\boldsymbol{\Phi}_X = \boldsymbol{Q}_X\boldsymbol{P}_X^{1/2}\boldsymbol{V}_X^{1/2} \tag{6}$$

Like $\boldsymbol{W}_X$ the mixing matrix $\boldsymbol{\Phi}_X$ is not orthogonal. The entries of the matrix $\boldsymbol{\Phi}_X$ are called the *loadings*, i.e. the coefficients linking the whitened variable $\widetilde{\boldsymbol{X}}$ with the original $\boldsymbol{x}$. Since $\widetilde{\boldsymbol{X}}$ is a white random vector with $\text{cov}\left(\widetilde{\boldsymbol{X}}\right) = \boldsymbol{I}_p$ the loadings are equivalent to the covariance $\text{cov}\left(\widetilde{\boldsymbol{X}}, \boldsymbol{X}\right) = \boldsymbol{\Phi}_X$. The corresponding correlations, also known as *correlation-loadings*, are

$$\text{cor}\left(\widetilde{\boldsymbol{X}}, \boldsymbol{X}\right) = \boldsymbol{\Psi}_X = \boldsymbol{\Phi}_X\boldsymbol{V}_X^{-1/2} = \boldsymbol{Q}_X\boldsymbol{P}_X^{1/2}. \tag{7}$$

Note that the sum of squared correlations in each column of $\boldsymbol{\Psi}_X$ sum up to 1, as $\text{diag}\left(\boldsymbol{\Psi}_X^T\boldsymbol{\Psi}_X\right) = \text{diag}(\boldsymbol{P}_X) = \boldsymbol{I}_p$.

### *CCA whitening*

We will show now that CCA has a very close relationship to whitening. In particular, the objective of CCA can be seen to be equivalent to simultaneous whitening of both $\boldsymbol{X}$ and $\boldsymbol{Y}$, with a diagonality constraint on the cross-correlation matrix between the whitened $\widetilde{\boldsymbol{X}}$ and $\widetilde{\boldsymbol{Y}}$.

First, we make the choice to standardize the canonical directions $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}_i$ according to $\text{var}\left(\boldsymbol{\alpha}_i^T\boldsymbol{X}\right) = \boldsymbol{\alpha}_i^T\boldsymbol{\Sigma}_X\boldsymbol{\alpha}_i = 1$ and $\text{var}\left(\boldsymbol{\beta}_i^T\boldsymbol{Y}\right) = \boldsymbol{\beta}_i^T\boldsymbol{\Sigma}_Y\boldsymbol{\beta}_i = 1$. As a result $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}_i$ form the basis of two whitening matrices, $\boldsymbol{W}_X = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_p)^T = \boldsymbol{A}$ and $\boldsymbol{W}_Y = \left(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_q\right)^T = \boldsymbol{B}$, with *rows* containing the canonical directions. The length constraint $\boldsymbol{\alpha}_i^T\boldsymbol{\Sigma}_X\boldsymbol{\alpha}_i = 1$ thus becomes $\boldsymbol{W}_X\boldsymbol{\Sigma}_X\boldsymbol{W}_X^T = \boldsymbol{I}_p$ meaning that $\boldsymbol{W}_X$ (and $\boldsymbol{W}_Y$) is indeed a valid whitening matrix.

Second, after whitening $\boldsymbol{X}$ and $\boldsymbol{Y}$ individually to $\widetilde{\boldsymbol{X}}$ and $\widetilde{\boldsymbol{Y}}$ using $\boldsymbol{W}_X$ and $\boldsymbol{W}_Y$, respectively, the joint covariance of $\left(\widetilde{\boldsymbol{X}}^T, \widetilde{\boldsymbol{Y}}^T\right)^T$ is $\begin{pmatrix} \boldsymbol{I}_p & \boldsymbol{P}_{\widetilde{X}\widetilde{Y}} \\ \boldsymbol{P}_{\widetilde{X}\widetilde{Y}}^T & \boldsymbol{I}_q \end{pmatrix}$. Note that whitening of $\left(\boldsymbol{X}^T, \boldsymbol{Y}^T\right)^T$ simultaneously would in contrast lead to a fully diagonal covariance matrix. In the above $\boldsymbol{P}_{\widetilde{X}\widetilde{Y}} = \text{cor}\left(\widetilde{\boldsymbol{X}}, \widetilde{\boldsymbol{Y}}\right) = \text{cov}\left(\widetilde{\boldsymbol{X}}, \widetilde{\boldsymbol{Y}}\right)$ is the cross-correlation matrix between the two whitened vectors and can be expressed as

$$\boldsymbol{P}_{\widetilde{X}\widetilde{Y}} = \boldsymbol{W}_X\boldsymbol{\Sigma}_{XY}\boldsymbol{W}_Y^T = \boldsymbol{Q}_X\boldsymbol{K}\boldsymbol{Q}_Y^T = (\widetilde{\rho}_{ij}) \tag{8}$$

and

$$\boldsymbol{K} = \boldsymbol{P}_X^{-1/2}\boldsymbol{P}_{XY}\boldsymbol{P}_Y^{-1/2} = (k_{ij}). \tag{9}$$

Following the terminology in [17] we may call $\boldsymbol{K}$ the correlation-adjusted cross-correlation matrix between $\boldsymbol{X}$ and $\boldsymbol{Y}$.

With this setup the CCA objective can be framed simply as the demand that $\text{cor}\left(\widetilde{\boldsymbol{X}}, \widetilde{\boldsymbol{Y}}\right) = \boldsymbol{P}_{\widetilde{X}\widetilde{Y}}$ must be diagonal. Since in whitening the orthogonal matrices $\boldsymbol{Q}_{\boldsymbol{X}}$ and $\boldsymbol{Q}_{\boldsymbol{Y}}$ can be freely selected we can achieve diagonality of $\boldsymbol{P}_{\widetilde{X}\widetilde{Y}}$ and hence pinpoint the CCA whitening matrices by applying singular value decomposition to

$$\boldsymbol{K} = \left(\boldsymbol{Q}_{\boldsymbol{X}}^{\text{CCA}}\right)^T \boldsymbol{\Lambda} \boldsymbol{Q}_{\boldsymbol{Y}}^{\text{CCA}} . \tag{10}$$

This provides the rotation matrices $\boldsymbol{Q}_{\boldsymbol{X}}^{\text{CCA}}$ and the $\boldsymbol{Q}_{\boldsymbol{Y}}^{\text{CCA}}$ of dimensions $m \times p$ and $m \times q$, respectively, and the $m \times m$ matrix $\boldsymbol{\Lambda} = \text{diag}(\lambda_i)$ containing the singular values of $\boldsymbol{K}$, which are also the singular values of $\boldsymbol{P}_{\widetilde{X}\widetilde{Y}}$. Since $m = \min(p, q)$ the larger of the two rotation matrices will not be a square matrix but it can nonetheless be used for whitening via Eqs. 4 and 5 since it still is semi-orthogonal with $\boldsymbol{Q}_{\boldsymbol{X}}^{\text{CCA}} \left(\boldsymbol{Q}_{\boldsymbol{X}}^{\text{CCA}}\right)^T = \boldsymbol{Q}_{\boldsymbol{Y}}^{\text{CCA}} \left(\boldsymbol{Q}_{\boldsymbol{Y}}^{\text{CCA}}\right)^T = \boldsymbol{I}_m$. As a result, we obtain $\text{cor}\left(\widetilde{X}_i^{\text{CCA}}, \widetilde{Y}_i^{\text{CCA}}\right) = \lambda_i$ for $i = 1 \ldots m$, i.e. the canonical correlations are identical to the singular values of $\boldsymbol{K}$.

Hence, CCA may be viewed as the outcome of a uniquely determined whitening transformation with underlying sphering matrices $\boldsymbol{W}_{\boldsymbol{X}}^{\text{CCA}}$ and $\boldsymbol{W}_{\boldsymbol{Y}}^{\text{CCA}}$ induced by the rotation matrices $\boldsymbol{Q}_{\boldsymbol{X}}^{\text{CCA}}$ and $\boldsymbol{Q}_{\boldsymbol{Y}}^{\text{CCA}}$. Thus, the distinctive feature of CCA whitening, in contrast to other common forms of whitening described in [16], is that by construction it is not only informed by $\boldsymbol{P}_{\boldsymbol{X}}$ and $\boldsymbol{P}_{\boldsymbol{Y}}$ but also by $\boldsymbol{P}_{\boldsymbol{XY}}$, which fixes all remaining rotational freedom.

## CCA and multivariate regression

### *Optimal linear multivariate predictor*

In multivariate regression the aim is to build a model that, given an input vector $\boldsymbol{X}$, predicts a vector $\boldsymbol{Y}$ as well as possible according to a specific measure such as squared error. Assuming a linear relationship $\boldsymbol{Y}^{\star} = \boldsymbol{a} + \boldsymbol{b}^T\boldsymbol{X}$ is the predictor random variable, with mean $\text{E}(\boldsymbol{Y}^{\star}) = \boldsymbol{\mu}_{\boldsymbol{Y}^{\star}} = \boldsymbol{a} + \boldsymbol{b}^T\boldsymbol{\mu}_{\boldsymbol{X}}$. The expected squared difference between $\boldsymbol{Y}$ and $\boldsymbol{Y}^{\star}$, i.e. the mean squared prediction error

$$\begin{aligned} \text{MSE} &= \text{Tr}\left(\text{E}\left(\left(\boldsymbol{Y} - \boldsymbol{Y}^{\star}\right)\left(\boldsymbol{Y} - \boldsymbol{Y}^{\star}\right)^T\right)\right) \\ &= \sum_{i=1}^{q} \text{E}\left(\left(Y_i - Y_i^{\star}\right)^2\right), \end{aligned} \tag{11}$$

is a natural measure of how well $\boldsymbol{Y}^{\star}$ predicts $\boldsymbol{Y}$. As a function of the model parameters $\boldsymbol{a}$ and $\boldsymbol{b}$ the predictive MSE becomes

$$\begin{aligned} \text{MSE}(\boldsymbol{a}, \boldsymbol{b}) = &\text{Tr}\Big(\left(\boldsymbol{\mu}_{\boldsymbol{Y}} - \boldsymbol{\mu}_{\boldsymbol{Y}^{\star}}\right)\left(\boldsymbol{\mu}_{\boldsymbol{Y}} - \boldsymbol{\mu}_{\boldsymbol{Y}^{\star}}\right)^T + \\ &\boldsymbol{\Sigma}_{\boldsymbol{Y}} + \boldsymbol{b}^T\boldsymbol{\Sigma}_{\boldsymbol{X}}\boldsymbol{b} - 2\boldsymbol{b}^T\boldsymbol{\Sigma}_{\boldsymbol{XY}}\Big) . \end{aligned} \tag{12}$$

Optimal parameters for best linear predictor are found by minimizing this MSE function. For the offset $\boldsymbol{a}$ this yields

$$\boldsymbol{a}^{\cdot} = \boldsymbol{\mu}_{\boldsymbol{Y}} - \left(\boldsymbol{b}^{\cdot}\right)^T\boldsymbol{\mu}_{\boldsymbol{X}} \tag{13}$$

which regardless of the value of $\boldsymbol{b}^{\cdot}$ ensures $\boldsymbol{\mu}_{\boldsymbol{Y}^{\star}} - \boldsymbol{\mu}_{\boldsymbol{Y}} = 0$. Likewise, for the matrix of regression coefficients minimization results in

$$\boldsymbol{b}^{\text{all}} = \boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{XY}} \tag{14}$$

with minimum achieved $\text{MSE}\left(\boldsymbol{a}^{\text{all}}, \boldsymbol{b}^{\text{all}}\right) = \text{Tr}\left(\boldsymbol{\Sigma}_{\boldsymbol{Y}}\right) - \text{Tr}\left(\boldsymbol{\Sigma}_{\boldsymbol{YX}}\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{XY}}\right)$.

If we exclude predictors from the model by setting regression coefficients $\boldsymbol{b}^{\text{zero}} = 0$ then the corresponding optimal intercept is $\boldsymbol{a}^{\text{zero}} = \boldsymbol{\mu}_{\boldsymbol{Y}}$ and the minimum achieved $\text{MSE}\left(\boldsymbol{a}^{\text{zero}}, \boldsymbol{b}^{\text{zero}}\right) = \text{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{Y}})$. Thus, by adding predictors $\boldsymbol{X}$ to the model the predictive MSE is reduced, and hence the fit of the model correspondingly improved, by the amount

$$\begin{aligned} \Delta &= \text{MSE}\left(\boldsymbol{a}^{\text{zero}}, \boldsymbol{b}^{\text{zero}}\right) - \text{MSE}\left(\boldsymbol{a}^{\text{all}}, \boldsymbol{b}^{\text{all}}\right) \\ &= \text{Tr}\left(\boldsymbol{\Sigma}_{\boldsymbol{YX}}\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{XY}}\right) \\ &= \text{Tr}\left(\text{cov}\left(\boldsymbol{Y}, \boldsymbol{Y}^{\text{all}\star}\right)\right) . \end{aligned} \tag{15}$$

If the response $Y$ is univariate ($q = 1$) then $\Delta$ reduces to the variance-scaled coefficient of determination $\sigma_Y^2\boldsymbol{P}_{Y\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{X}}^{-1}\boldsymbol{P}_{\boldsymbol{X}Y}$. Note that in the above no distributional assumptions are made other than specification of means and covariances.

### *Regression view of CCA*

The first step to understand CCA as a regression model is to consider multivariate regression between two whitened vectors $\widetilde{\boldsymbol{X}}$ and $\widetilde{\boldsymbol{Y}}$ (considering whitening of any type, including but not limited to CCA-whitening). Since $\boldsymbol{\Sigma}_{\widetilde{\boldsymbol{X}}} = \boldsymbol{I}_p$ and $\boldsymbol{\Sigma}_{\widetilde{X}\widetilde{Y}} = \boldsymbol{P}_{\widetilde{X}\widetilde{Y}}$ the optimal regression coefficients to predict $\widetilde{\boldsymbol{Y}}$ from $\widetilde{\boldsymbol{X}}$ are given by

$$\boldsymbol{b}^{\text{all}} = \boldsymbol{P}_{\widetilde{X}\widetilde{Y}} , \tag{16}$$

i.e. the pairwise correlations between the elements of the two vectors $\widetilde{\boldsymbol{X}}$ and $\widetilde{\boldsymbol{Y}}$. Correspondingly, the decrease in predictive MSE due to including the predictors $\widetilde{\boldsymbol{X}}$ is

$$\begin{aligned} \Delta &= \text{Tr}\left(\boldsymbol{P}_{\widetilde{X}\widetilde{Y}}^T\boldsymbol{P}_{\widetilde{X}\widetilde{Y}}\right) = \sum_{i,j} \widetilde{\rho}_{ij}^2 \\ &= \text{Tr}\left(\boldsymbol{K}^T\boldsymbol{K}\right) = \sum_{i,j} k_{ij}^2 \\ &= \text{Tr}\left(\boldsymbol{\Lambda}^2\right) = \sum_{i} \lambda_i^2 . \end{aligned} \tag{17}$$

In the special case of CCA-whitening the regression coefficients further simplify to $b_{ii}^{\text{all}} = \lambda_i$, i.e. the canonical correlations $\lambda_i$ act as the regression coefficients linking CCA-whitened $\widetilde{\boldsymbol{Y}}$ and $\widetilde{\boldsymbol{X}}$. Furthermore, as the decrease in predictive MSE $\Delta$ is the sum of the squared canonical correlations (cf. Eq. 17), each $\lambda_i^2$ can be interpreted as the variable importance of the corresponding variable in $\widetilde{\boldsymbol{X}}^{\text{CCA}}$ to predict the outcome $\widetilde{\boldsymbol{Y}}^{\text{CCA}}$. Thus, CCA directly results from multivariate regression between CCA-whitened random vectors, where the canonical correlations $\lambda_i$ assume the role of regression coefficients and $\lambda_i^2$ provides a natural measure to rank the canonical components in order of their respective predictive capability.

A key difference between classical CCA and regression is that in the latter both positive and negative coefficients are allowed to account for the directionality of the influence of the predictors. In contrast, in classical CCA only positive canonical correlations are permitted by convention. To reflect that CCA analysis is inherently a regression model we advocate here that canonical correlations should indeed be allowed to assume both positive and negative values, as fundamentally they are regression coefficients. This can be implemented by exploiting the sign ambiguity in the singular value decomposition of $\boldsymbol{K}$ (Eq. 10). In particular, the rows signs of $\boldsymbol{Q}_{\boldsymbol{X}}^{\text{CCA}}$ and $\boldsymbol{Q}_{\boldsymbol{Y}}^{\text{CCA}}$ and the signs of $\lambda_i$ can be revised simultaneously without affecting $\boldsymbol{K}$. We propose to choose $\boldsymbol{Q}_{\boldsymbol{X}}^{\text{CCA}}$ and $\boldsymbol{Q}_{\boldsymbol{Y}}^{\text{CCA}}$ such that both rotation matrices have a positive diagonal, and then to adjust the signs of the $\lambda_i$ accordingly. Note that orthogonal matrices with positive diagonals are closest to the identity matrix (e.g. in terms of the Frobenius norm) and thus constitute minimal rotations.

### Generative latent variable model for CCA

With the link of CCA to whitening and multivariate regression established it is straightforward to arrive at simple and easily interpretable generative probabilistic latent variable model for CCA. This model has two levels of hidden variables: it uses uncorrelated latent variables $\boldsymbol{Z}^{\boldsymbol{X}}, \boldsymbol{Z}^{\boldsymbol{Y}}, \boldsymbol{Z}^{\text{shared}}$ (level 1) with zero mean and unit variance to generate the CCA-whitened variables $\widetilde{\boldsymbol{X}}^{\text{CCA}}$ and $\widetilde{\boldsymbol{Y}}^{\text{CCA}}$ (level 2) which in turn produce the observed vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ – see Fig. 1

Specifically, on the first level we have latent variables

$$\begin{aligned} \boldsymbol{Z}^{\boldsymbol{X}} &\sim F_{\boldsymbol{Z}_{\boldsymbol{X}}}, \\ \boldsymbol{Z}^{\boldsymbol{Y}} &\sim F_{\boldsymbol{Z}_{\boldsymbol{Y}}}, \text{ and} \\ \boldsymbol{Z}^{\text{shared}} &\sim F_{\boldsymbol{Z}_{\text{shared}}}, \end{aligned} \qquad (18)$$

with $\text{E}\left(\boldsymbol{Z}^{\boldsymbol{X}}\right) = \text{E}\left(\boldsymbol{Z}^{\boldsymbol{Y}}\right) = \text{E}\left(\boldsymbol{Z}^{\text{shared}}\right) = 0$ and $\text{var}\left(\boldsymbol{Z}^{\boldsymbol{X}}\right) = \boldsymbol{I}_p$, $\text{var}\left(\boldsymbol{Z}^{\boldsymbol{Y}}\right) = \boldsymbol{I}_q$, and $\text{var}\left(\boldsymbol{Z}^{\text{shared}}\right) = \boldsymbol{I}_m$ and no mutual correlation among the components of $\boldsymbol{Z}^{\boldsymbol{X}}$, $\boldsymbol{Z}^{\boldsymbol{Y}}$, and $\boldsymbol{Z}^{\text{shared}}$. The second level latent variables are then generated by mixing shared and non-shared variables according to

$$\begin{aligned} \widetilde{X}_i^{\text{CCA}} &= \sqrt{1-|\lambda_i|}\, Z_i^{\boldsymbol{X}} + \sqrt{|\lambda_i|} Z_i^{\text{shared}} \\ \widetilde{Y}_i^{\text{CCA}} &= \sqrt{1-|\lambda_i|}\, Z_i^{\boldsymbol{Y}} + \sqrt{|\lambda_i|} Z_i^{\text{shared}}\, \text{sign}(\lambda_i) \end{aligned} \qquad (19)$$

where the parameters $\lambda_1, \ldots, \lambda_m$ can be positive as well as negative and range from -1 to 1. The components $i > m$ are always non-shared and taken from $\boldsymbol{Z}^{\boldsymbol{X}}$ or $\boldsymbol{Z}^{\boldsymbol{Y}}$ as appropriate, i.e. as above but with $\lambda_{i>m} = 0$. By construction, this results in $\text{var}\left(\widetilde{\boldsymbol{X}}^{\text{CCA}}\right) = \boldsymbol{I}_p$, $\text{var}\left(\widetilde{\boldsymbol{Y}}^{\text{CCA}}\right) = \boldsymbol{I}_q$ and $\text{cov}\left(\widetilde{X}_i^{\text{CCA}}, \widetilde{Y}_i^{\text{CCA}}\right) = \lambda_i$. Finally, the observed variables are produced by a coloring transformation and subsequent translation

$$\begin{aligned} \boldsymbol{X} &= \boldsymbol{\Phi}_{\boldsymbol{X}}^T \widetilde{\boldsymbol{X}}^{\text{CCA}} + \boldsymbol{\mu}_{\boldsymbol{X}} \\ \boldsymbol{Y} &= \boldsymbol{\Phi}_{\boldsymbol{Y}}^T \widetilde{\boldsymbol{Y}}^{\text{CCA}} + \boldsymbol{\mu}_{\boldsymbol{Y}} \end{aligned} \qquad (20)$$

To clarify the workings behind Eq. 19 assume there are three uncorrelated random variables $Z_1$, $Z_2$, and $Z_3$ with mean 0 and variance 1. We construct $X_1$ as a mixture of $Z_1$ and $Z_3$ according to $X_1 = \sqrt{1-\alpha} Z_1 + \sqrt{\alpha} Z_3$ where $\alpha \in [0,1]$, and, correspondingly, $X_2$ as a mixture of $Z_2$ and $Z_3$ via $X_2 = \sqrt{1-\alpha} Z_2 + \sqrt{\alpha} Z_3$. If $\alpha = 0$ then $X_1 = Z_1$ and $X_2 = Z_2$, and if $\alpha = 1$ then $X_1 = X_2 = Z_3$. By design, the new variables have mean zero ($\text{E}(X_1) = \text{E}(X_2) = 0$) and unit variance ($\text{var}(X_1) = \text{var}(X_2) = 1$). Crucially, the weight $\alpha$ of the latent variable $Z_3$ common to both mixtures induces a correlation between $X_1$ and $X_2$. The covariance between $X_1$ and $X_2$ is $\text{cov}(X_1, X_2) = \text{cov}\left(\sqrt{\alpha} Z_3, \sqrt{\alpha} Z_3\right) = \alpha$, and since $X_1$ and

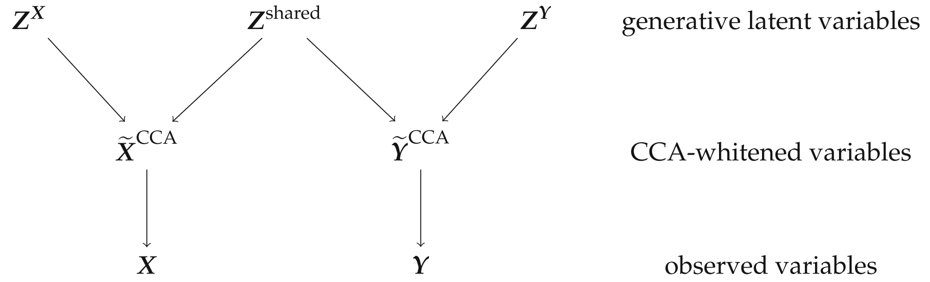
$\boldsymbol{Z}^{\boldsymbol{X}}$
$\boldsymbol{Z}^{\text{shared}}$
$\boldsymbol{Z}^{\boldsymbol{Y}}$
generative latent variables
$\widetilde{\boldsymbol{X}}^{\text{CCA}}$
$\widetilde{\boldsymbol{Y}}^{\text{CCA}}$
CCA-whitened variables
$\boldsymbol{X}$
$\boldsymbol{Y}$
observed variables

**Fig. 1** Probabilistic CCA as a two layer latent variable generative model. The middle layer contains the CCA-whitened variables $\widetilde{\boldsymbol{X}}^{\text{CCA}}$ and $\widetilde{\boldsymbol{Y}}^{\text{CCA}}$, and the top layer the uncorrelated generative latent variables $\boldsymbol{Z}^{\boldsymbol{X}}$, $\boldsymbol{Z}^{\boldsymbol{Y}}$, and $\boldsymbol{Z}^{\text{shared}}$

$X_2$ have variance 1 we have $\text{cor}(X_1, X_2) = \alpha$. In Eq. 19 this is further extended to allow a signed $\alpha$ and hence negative correlations.

Note that the above probabilistic model for CCA is in fact not a single model but a family of models, since we do not completely specify the underlying distributions, only their means and (co)variances. While in practice we will typically assume normally distributed generative latent variables, and hence normally distributed observations, it is equally possible to employ other distributions for the first level latent variables. For example, a rescaled $t$-distribution with a wider tail than the normal distribution may be employed to obtain a robustified version of CCA [18].

## Discussion

### Synthetic data

In order to test whether our algorithm allows to correctly identify negative canonical correlations we conducted simulations using simulated data. Specifically, we generated data $\boldsymbol{X}_i$ and $\boldsymbol{y}_i$ from a $p+q$ dimensional multivariate normal distribution with zero mean and covariance matrix $\begin{pmatrix} \boldsymbol{\Sigma}_X & \boldsymbol{\Sigma}_{XY} \\ \boldsymbol{\Sigma}_{XY}^T & \boldsymbol{\Sigma}_Y \end{pmatrix}$ where $\boldsymbol{\Sigma}_X = \boldsymbol{I}_p$, $\boldsymbol{\Sigma}_Y = \boldsymbol{I}_q$ and $\boldsymbol{\Sigma}_{XY} = \text{diag}(\lambda_i)$. The canonical correlations where set to have alternating positive and negative signs $\lambda_1 = \lambda_3 = \lambda_5 = \lambda_7 = \lambda_9 = \lambda$ and $\lambda_2 = \lambda_4 = \lambda_6 = \lambda_8 = \lambda_{10} = -\lambda$ with varying strength $\lambda \in \{0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$. A similar setup was used in [14]. The dimensions were fixed at $p = 60$ and $q = 10$ and the sample size was $n \in \{20, 30, 50, 100, 200, 500\}$ so that both the small and large sample regime was covered. For each combination of $n$ and $\lambda$ the simulations were repeated 500 times, and our algorithm using shrinkage estimation of the underlying covariance matrices was employed to each of the 500 data sets to fit the CCA model. The resulting estimated canonical correlations were then compared with the corresponding true canonical correlations, and the proportion of correctly estimated signs was recorded.

The outcome from this simulation study is summarized graphically in Fig. 2. The key finding is that, depending on the strength of correlation $\lambda$ and sample size $n$, our algorithm correctly determines the sign of both negative and positive canonical correlations. As expected, the proportion of correctly classified canonical correlations increases with sample size and with the strength of correlation. Remarkably, even for comparatively weak correlation such as $\lambda = 0.5$ and low sample size still the majority of canonical correction were estimated with the true sign. In short, this simulation demonstrates that if there are negative canonical correlations between pairs of canonical variables these will be detected by our approach.

### Nutrimouse data

We now analyze two experimental omics data sets to illustrate our approach. Specifically, we demonstrate the capability of our variant of CCA to identify negative canonical correlations among canonical variates as well its application to high-dimensional data where the number of samples $n$ is smaller than the number of variables $p$ and $q$.

The first data set is due to [19] and results from a nutrigenomic study in the mouse studying $n = 40$ animals. The $\boldsymbol{X}$ variable collects the measurements of the gene expression of $p = 120$ genes in liver cells. These were selected a priori considering the biological relevance for the study. The $\boldsymbol{Y}$ variable contains lipid concentrations of $q = 21$ hepatic fatty acids, measured on the same animals. Before further analysis we standardized both $\boldsymbol{X}$ and $\boldsymbol{Y}$.

Since the number of available samples $n$ is smaller than the number of genes $p$ we used shrinkage estimation to obtain the joint correlation matrix which resulted in a shrinkage intensity of $\lambda_{\text{cor}} = 0.16$. Subsequently, we computed canonical directions and associated canonical correlations $\lambda_1, \ldots, \lambda_{21}$. The canonical correlations are shown in Fig. 3, and range in value between -0.96 and 0.87. As can be seen, 16 of the 21 canonical correlations are negative, including the first three top ranking correlations. In Fig. 4 we depict the squared correlation loadings between the first 5 components of the canonical covariates $\widetilde{\boldsymbol{X}}^{\text{CCA}}$ and $\widetilde{\boldsymbol{Y}}^{\text{CCA}}$ and the corresponding observed variables $\boldsymbol{X}$ and $\boldsymbol{Y}$. This visualization shows that most information about the correlation structure within and between the two data sets (gene expression and lipid concentrations) is concentrated in the first few latent components.

This is confirmed by further investigation of the scatter plots both between corresponding pairs of $\widetilde{\boldsymbol{X}}^{\text{CCA}}$ and $\widetilde{\boldsymbol{Y}}^{\text{CCA}}$ canonical variates (Fig. 5) as well as within each variate (Fig. 6). Specifically, the first CCA component allow to identify the genotype of the mice (wt: wild type; ppar: PPAR-$\alpha$ deficient) whereas the subsequent few components reveal the imprint of the effect of the various diets (COC: coconut oil; FISH: fish oils; LIN: linseed oils; REF: reference diet; SUN: sunflower oil) on gene expression and lipid concentrations.

### The Cancer Genome Atlas LUSC data

As a further illustrative example we studied genomic data from The Cancer Genome Atlas (TCGA), a public resource that catalogues clinical data and molecular characterizations of many cancer types [20]. We used the TCGA2STAT tool to access the TCGA database from within R [21].

Specifically, we retrieved gene expression (RNASeq2) and methylation data for lung squamous cell carcinoma (LUSC) which is one of the most common types of lung cancer. After download, calibration and filtering as well

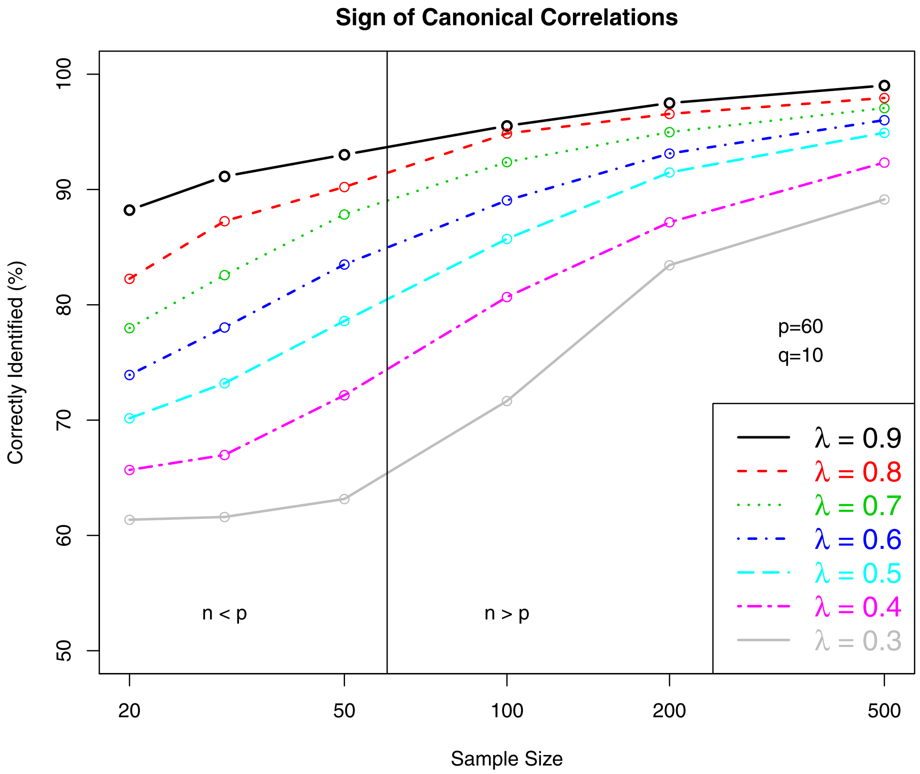

**Fig. 2** Percentage of estimated canonical correlations with correctly identified signs in dependence of the sample size and the strength of the true canonical correlation

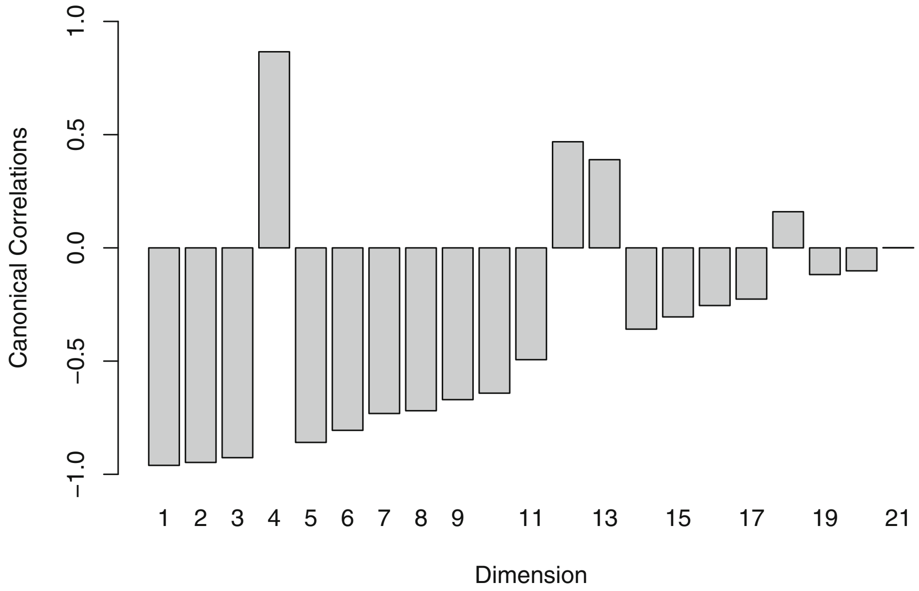

**Fig. 3** Plot of the estimated canonical correlations for the Nutrimouse data. The majority of the correlations indicate a negative assocation between the corresponding canonical variables

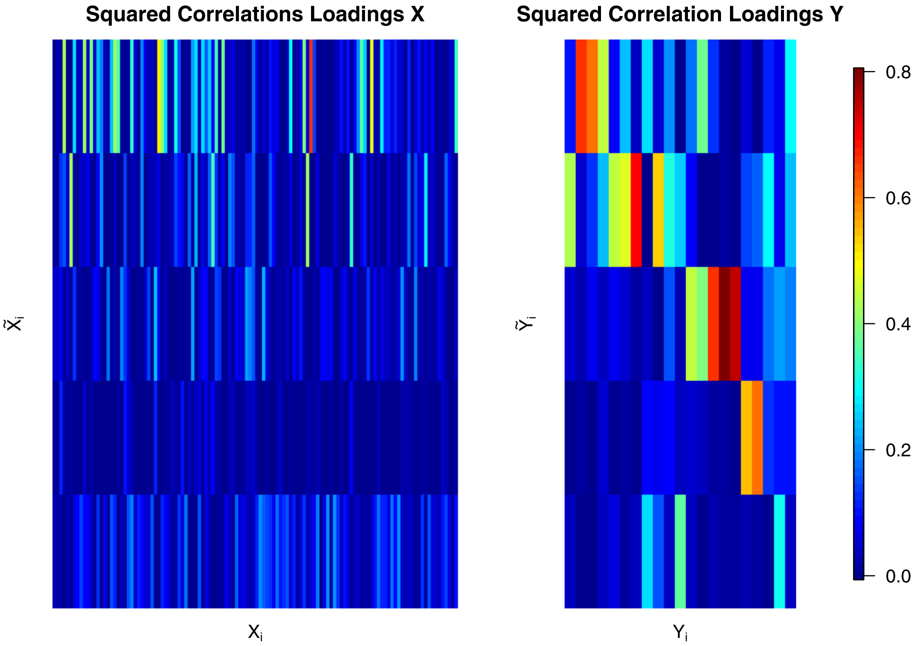

**Fig. 4** Squared correlations loadings between the first 5 components of the canonical covariates $\widetilde{\boldsymbol{X}}^{\text{CCA}}$ and $\widetilde{\boldsymbol{Y}}^{\text{CCA}}$ and the corresponding observed variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ for the Nutrimouse data

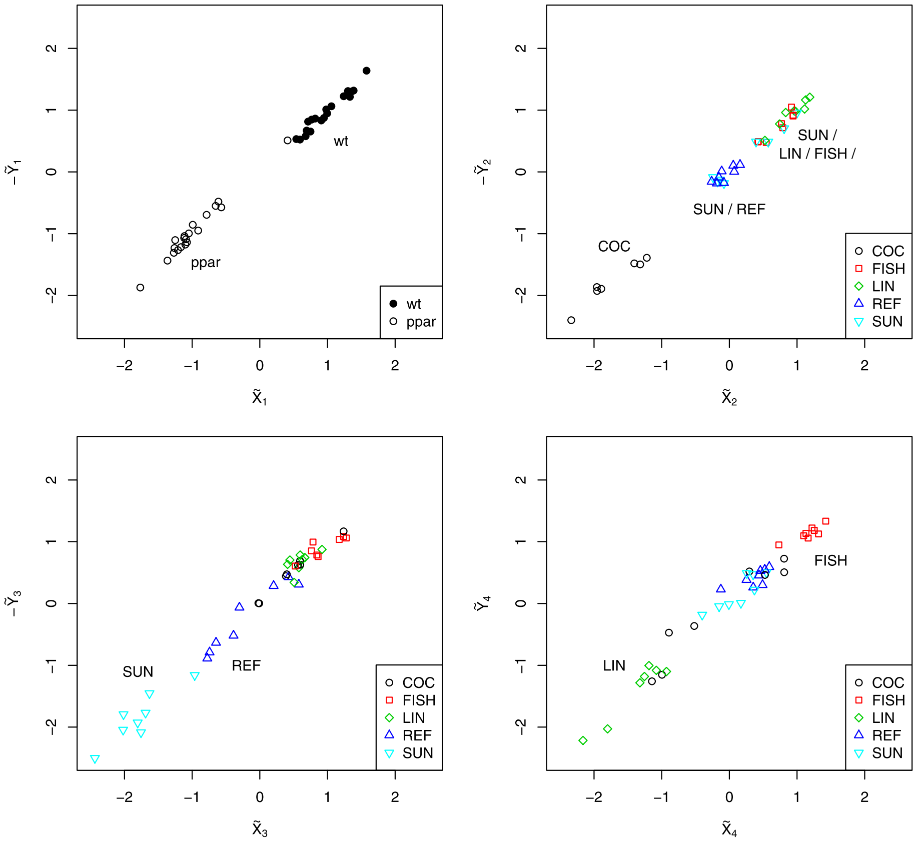

**Fig. 5** Scatter plots between corresponding pairs of canonical covariates for the Nutrimouse data

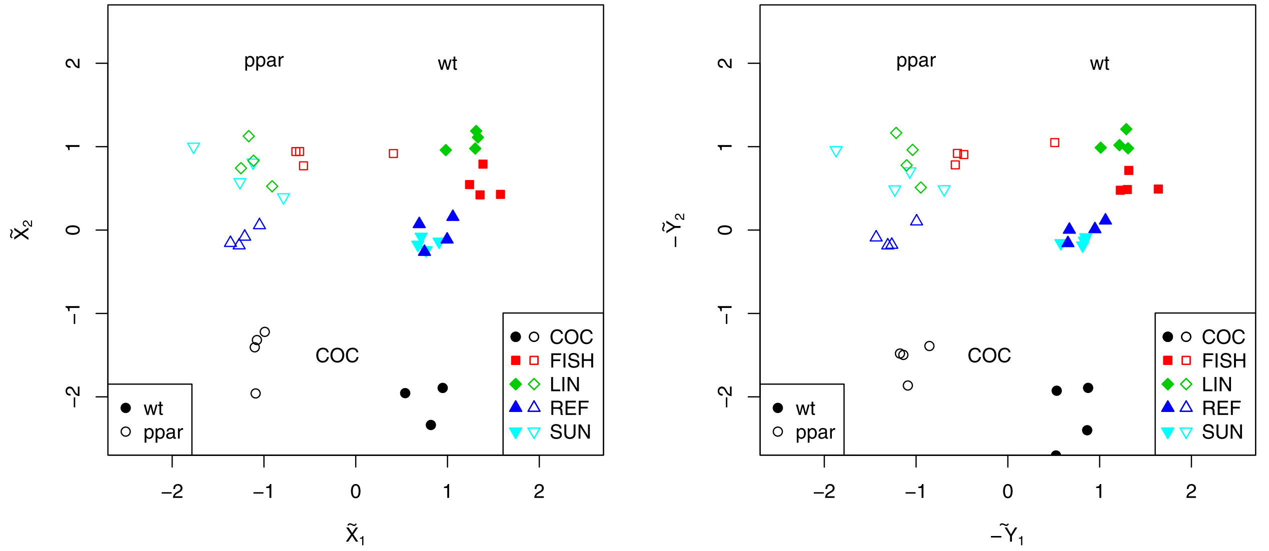

**Fig. 6** Scatter plots between first and second components within each canonical covariate for the Nutrimouse data

as matching the two data types to 130 common patients following the guidelines in [21] we obtained two data matrices, one ($\boldsymbol{X}$) measuring gene expression of $p = 206$ genes and one ($\boldsymbol{Y}$) containing methylation levels corresponding to $q = 234$ probes. As clinical covariates the sex of each of the 130 patients (97 males, 33 females) was downloaded as well as the vital status (46 events in males, and 11 in females) and cancer end points, i.e. the number of days to last follow-up or the days to death. In addition, since smoking cigarettes is a key risk factor for lung cancer, the number of packs per year smoked was also recorded. The number of packs ranged from 7 to 240, so all of the patients for which this information was available were smokers.

As above we applied the shrinkage CCA approach to the LUSC data which resulted in a correlation shrinkage intensity of $\lambda_{\text{cor}} = 0.19$. Subsequently, we computed canonical directions and associated canonical correlations $\lambda_1, \ldots, \lambda_{21}$. The canonical correlations are shown in Fig. 7, and range in value between -0.92 and 0.98. Among the top 10 strongest correlated pairs of canonical covariates only one has a negative coefficient. The plot of the squared correlation loadings (Fig. 8) for these 10 components already indicates that the data can be sufficiently summarized by a few canonical covariates.

Scatter plots between the first pair of canonical components and between the first two components of $\widetilde{\boldsymbol{X}}^{\text{CCA}}$ are presented in Fig. 9. These plots show that the first canonical component corresponds to the sex of the patients, with males and females being clearly separated by underlying patterns in gene expression and methylation. The survival probabilities computed for both groups show that there is a statistically significant different risk pattern between males and females (Fig. 10). However, inspection of the second order canonical variates reveals that the difference in risk is likely due to overrepresentation of strong smokers in male patients rather than being directly attributed to the sex of the patient (Fig. 9 right).

## Conclusions

CCA is crucially important procedure for integration of multivariate data. Here, we have revisited CCA from the perspective of whitening that allows a better understanding of both classical CCA and its probabilistic variant. In particular, our main contributions in this paper are:

- first, we show that CCA is procedurally equivalent to a special whitening transformation, that unlike other general whitening procedures, is uniquely defined and without any rotational ambiguity;
- second, we demonstrate the direct connection of CCA with multivariate regression and demonstrate that CCA is effectively a linear model between whitened variables, and that correspondingly canonical correlations are best understood as regression coefficients;
- third, the regression perspective advocates for permitting both positive and negative canonical correlations and we show that this also allows to resolve the sign ambiguity present in the canonical directions;
- fourth, we propose an easily interpretable probabilistic generative model for CCA as a two-layer latent variable framework that not only admits canonical correlations of both signs but also allows non-normal latent variables;

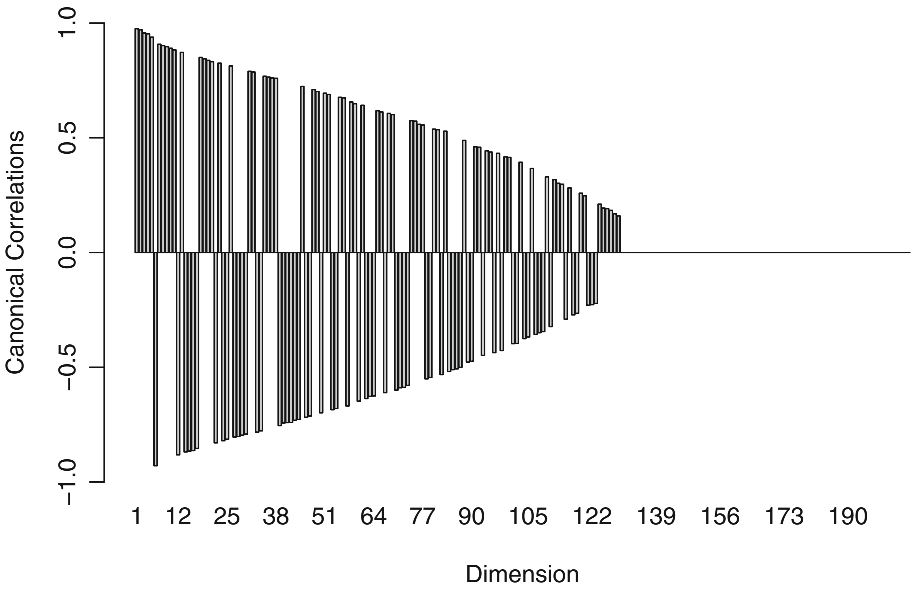

**Fig. 7** Plot of the estimated canonical correlations for the TCGA LUSC data

- and fifth, we provide a computationally effective computer implementation in the “whitening” R package based on high-dimensional shrinkage estimation of the underlying covariance and correlation matrices and show that this approach performs well both for simulated data as well as in application to the analysis of various types of omics data.

In short, this work provides a unifying perspective on CCA, linking together sphering procedures, multivariate regression and corresponding probabilistic generative

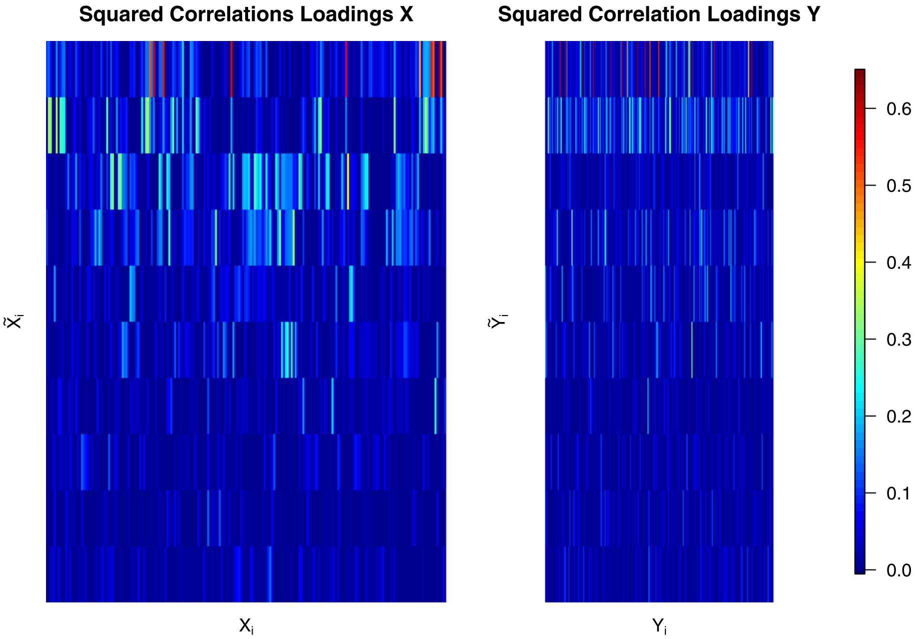

**Fig. 8** Squared correlations loadings between the first 10 components of the canonical covariates $\widetilde{\boldsymbol{X}}^{\text{CCA}}$ and $\widetilde{\boldsymbol{Y}}^{\text{CCA}}$ and the corresponding observed variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ for the TCGA LUSC data

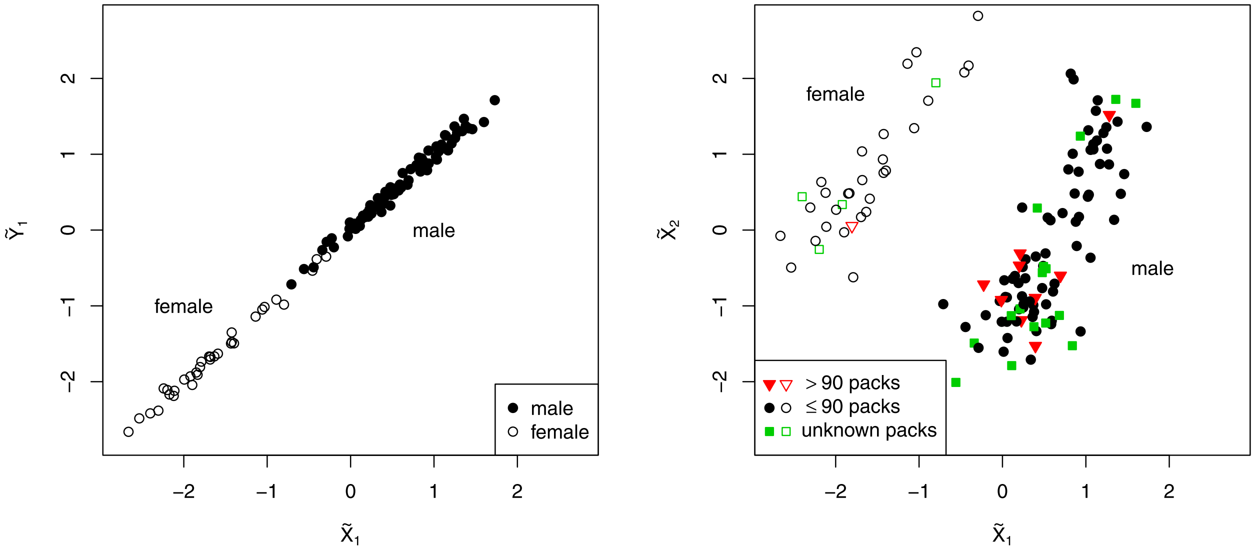

**Fig. 9** Scatter plots between first component of $\widetilde{X}^{\text{CCA}}$ and $\widetilde{Y}^{\text{CCA}}$ (left) and within the first two components of $\widetilde{X}^{\text{CCA}}$ (right) for the TCGA LUSC data

models, and also offers a practical tool for high-dimensional CCA for practitioners in applied statistical data analysis.

## Methods

### Implementation in R

We have implemented our method for high-dimensional CCA allowing for potential negative canonical correlations in the R package "whitening" that is freely available from https://CRAN.R-project.org/package=whitening. The functions provided in this package incorporate the computational efficiencies described below. The R package also includes example scripts. The "whitening" package has been used to conduct the data analysis described in this paper. Further information and R code to reproduce the analyses in this paper is available at http://strimmerlab.org/software/whitening/.

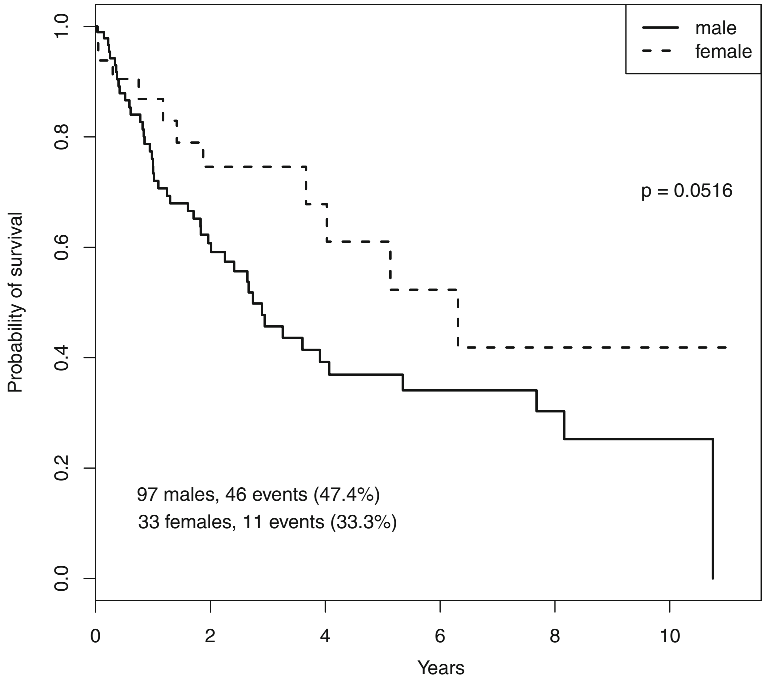

**Fig. 10** Plot of the survival probabilities for male and female patients for the TCGA LUSC data

### High-dimensional estimation

Practical application of CCA, in both the classical and probabilistic variants, requires estimation of the joint covariance of $\boldsymbol{X}$ and $\boldsymbol{Y}$ from data, as well as the computation of the corresponding underlying whitening matrices $\boldsymbol{W}_{\boldsymbol{X}}^{\text{CCA}}$ and $\boldsymbol{W}_{\boldsymbol{Y}}^{\text{CCA}}$ (i.e. canonical directions) and canonical correlations $\lambda_i$.

In moderate dimensions and large sample size $n$, i.e. when both $p$ and $q$ are not excessively big and $n$ is larger than both $p$ and $q$ the classic CCA algorithm is applicable and empirical or maximum likelihood estimates may be used. Conversely, if the sample size $n$ is small compared to $p$ and $q$ then there exist numerous effective Bayesian, penalized likelihood and other related regularized estimators to obtain *statistically efficient estimates* of the required covariance matrices (e.g., [22–25]). In our implementation in R and in the analysis below we use the shrinkage covariance estimation approach developed in [22] and also employed for CCA analysis in [14]. However, in principle any other preferred covariance estimator may be applied.

### Algorithmic efficiencies

In addition to statistical issues concerning accurate estimation, high dimensionality also poses substantial challenges in *algorithmic* terms, with regard both to memory requirements as well as to computing time. Specifically, for large values of $p$ and $q$ directly performing the matrix operations necessary for CCA, such as computing the matrix square root or even simple matrix multiplication, will be prohibitive since these procedures typically scale in cubic order of $p$ and $q$.

In particular, in a CCA analysis this affects i) the computation and estimation of the matrix $\boldsymbol{K}$ (Eq. 9) containing the adjusted cross-correlations, and ii) the calculation of the whitening matrices $\boldsymbol{W}_{\boldsymbol{X}}^{\text{CCA}}$ and $\boldsymbol{W}_{\boldsymbol{Y}}^{\text{CCA}}$ with the canonical directions $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}_i$ from the rotation matrices $\boldsymbol{Q}_{\boldsymbol{X}}^{\text{CCA}}$ and $\boldsymbol{Q}_{\boldsymbol{Y}}^{\text{CCA}}$ (Eq. 5). These computational steps involve multiplication and square-root calculations involving possibly very large matrices of dimension $p \times p$ and $q \times q$.

Fortunately, in the small sample domain with $n \leq p, q$ there exist computational tricks to perform these matrix operations in a very effective and both time- and memory-saving manner that avoids to directly compute and handle the large-scale covariance matrices and their derived quantities [e.g. [26]]. Note this requires the use of regularized estimators, e.g. shrinkage or ridge-type estimation. Specifically, in our implementation of CCA we capitalize on an algorithm described in [27] (see "Zuber et al. algorithm" section for details) that allows to compute the matrix product of the inverse matrix square root of the shrinkage estimate of the correlation matrix $\boldsymbol{R}$ with a matrix $\boldsymbol{M}$ without the need to store or compute the full estimated correlation matrices. The computationals savings due to effective matrix operations for $n < p$ and $n < q$ can be substantial, going from $O\left(p^3\right)$ and $O\left(q^3\right)$ down to $O\left(n^3\right)$ in terms of algorithmic complexity. Correspondingly, for example for $p/n = 3$ this implies time savings of factor 27 compared to "naive" direct computation.

### Zuber et al. algorithm

Zuber et al. (2012) [27] describe an algorithm that allows to compute the matrix product of the inverse matrix square root of the shrinkage estimate of the correlation matrix $\boldsymbol{R}$ with a matrix $\boldsymbol{M}$ without the need to store or compute the full estimated correlation matrices. Specifically, writing the correlation estimator in the form

$$\underbrace{\boldsymbol{R}}_{p\times p} = \lambda \left( \boldsymbol{I}_p + \underbrace{\boldsymbol{U}}_{p\times n} \underbrace{\boldsymbol{N}}_{n\times n} \boldsymbol{U}^T \right) \tag{21}$$

allows for algorithmically effective matrix multiplication of

$$\underbrace{\boldsymbol{R}^{-1/2}}_{p\times p} \underbrace{\boldsymbol{M}}_{p\times d} = \lambda^{-1/2} \left( \boldsymbol{M} - \boldsymbol{U} \underbrace{\left(\boldsymbol{I}_n - (\boldsymbol{I}_n + \boldsymbol{N})^{-1/2}\right)}_{n\times n} \left( \underbrace{\boldsymbol{U}^T \boldsymbol{M}}_{n\times d} \right) \right) . \tag{22}$$

Note that on the right-hand side of these two equations no matrix of dimension $p \times p$ appears; instead all matrices are of much smaller size.

In the CCA context we apply this procedure to Eq. 5 in order to obtain the whitening matrix and the canonical directions and also to Eq. 9 to efficiently compute the matrix $\boldsymbol{K}$.

**Acknowledgements**

KS thanks Martin Lotz for discussion. We also thank the anonymous referees for their many useful suggestions.

**Funding**

TJ was funded by a Wellcome Trust ISSF Ph.D. studentship. The funding body did not play any role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript

**Availability of data and materials**

The proposed method has been implemented in the R package "whitening" that is distributed from https://CRAN.R-project.org/package=whitening. R code to reproduce the analysis is available from http://strimmerlab.org/software/whitening/.

**Authors' contributions**

TJ and KS jointly conducted the research and wrote the paper. Both authors read and approved the manuscript.

**Ethics approval and consent to participate**

Not applicable.

**Consent for publication**

Not applicable.

**Competing interests**

The authors declare that they have no competing interests.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

### Author details

[1]Epidemiology and Biostatistics, School of Public Health, Imperial College London, Norfolk Place, W2 1PG London, UK. [2]Statistics Section, Department of Mathematics, Imperial College London, South Kensington Campus, SW7 2AZ London, UK. [3]School of Mathematics, University of Manchester, Alan Turing Building, Oxford Road, M13 9PL Manchester, UK.

Received: 13 July 2018 Accepted: 10 December 2018
Published online: 09 January 2019

### References

1. Hotelling H. Relations between two sets of variates. Biometrika. 1936;28: 321–77.
2. Härdle WK, Simar L. Canonical correlation analysis. In: Applied Multivariate Statistical Analysis. Chap. 16. Berlin: Springer; 2015. p. 443–54.
3. Cao D-S, Liu S, Zeng W-B, Liang Y-Z. Sparse canonical correlation analysis applied to -omics studies for integrative analysis and biomarker discovery. J Chemometrics. 2015;29:371–8.
4. Hong S, Chen X, Jin L, Xiong M. Canonical correlation analysis for RNA-seq co-expression networks. Nucleic Acids Res. 2013;41:95.
5. Bach FR, Jordan MI. A probabilistic interpretation of canonical correlation analysis. Technical Report No. 688, Department of Statistics. Berkeley: University of California; 2005.
6. Tipping ME, Bishop CM. Probabilistic principal component analysis. J R Statist Soc B. 1999;61(3):611–22. https://doi.org/10.1111/1467-9868.00196.
7. Wang C. Variational Bayesian approach to canonical correlation analysis. IEEE T Neural Net. 2007;18:905–10.
8. Klami A, Kaski S. Local dependent components. Proceedings of the 24th International Conference on Machine Learning (ICML 2007). 2007;24: 425–32.
9. Waaijenborg S, de Witt Hamer PCV, Zwinderman AH. Quantifying the association between gene expressions and DNA-markers by penalized canonical correlation analysis. Stat Appl Genet Molec Biol. 2008;7(1). Article 3. https://doi.org/10.2202/1544-6115.1329.
10. Parkhomenko E, Tritchler D, Beyene J. Sparse canonical correlation analysis with application to genomic data integration. Stat Appl Genet Molec Biol. 2009;8:1.
11. Witten D, Tibshirani R, Hastie T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics. 2009;10(3):515–34. https://doi.org/10.1093/biostatistics/kxp008.
12. Hardoon DR, Shawe-Taylor J. Sparse canonical correlation analysis. Mach Learn. 2011;83:331–53.
13. Wilms I, Croux C. Sparse canonical correlation analysis from a predictive point of view. Biomet J. 2015;57:834–51.
14. Cruz-Cano R, Lee M-LT. Fast regularized canonical correlation analysis. Comp Stat Data Anal. 2014;70:88–100.
15. Ma Z, Lu Y, Foster D. Finding linear structure in large datasets with scalable canonical correlation analysis. Proceedings of the 32th International Conference on Machine Learning (ICML 2015), PLMR. 2015;37:169–78.
16. Kessy A, Lewin A, Strimmer K. Optimal whitening and decorrelation. Am Stat. 2018;72:309–14. https://doi.org/10.1080/00031305.2016.1277159.
17. Zuber V, Strimmer K. High-dimensional regression and variable selection using CAR scores. Stat Appl Genet Molec Biol. 2011;10:34.
18. Adrover JG, Donato SM. A robust predictive approach for canonical correlation analysis. J Multiv Anal. 2015;133:356–76.
19. Martin PGP, Guillou H, Lasserre F, Déjean S, Lan A, Pascussi J-M, Cristobal MS, Legrand P, Besse P, Pineau T. Novel aspects of PPAR$\alpha$-mediated regulation of lipid and xenobiotic metabolism revealed through a multigenomic study. Hepatology. 2007;54:767–77.
20. Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, Xie M, JF McMichael QZ, Wyczalkowski MA, Leiserson MDM, Miller CA, Welch JS, Walter MJ, Wendl MC, Ley TJ, Wilson RK, Raphael BJ, Ding L. Mutational landscape and significance across 12 major cancer types. Nature. 2013;502:333–9.
21. Wan Y-W, Allen GI, Liu Z. TCGA2STAT: simple TCGA data access for integrated statistical analysis in R. Bioinformatics. 2016;32:952–4.
22. Schäfer J, Strimmer K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Stat Appl Genet Molec Biol. 2005;4:32.
23. Bickel PJ, Levina E. Regularized estimation of large covariance matrices. Ann Stat. 2008;36:199–227.
24. Hannart A, Naveau P. Estimating high dimensional covariance matrices: A new look at the Gaussian conjugate framework. J Multiv Anal. 2014;131: 149–62.
25. Touloumis A. Nonparametric Stein-type shrinkage covariance matrix estimators in high-dimensional settings. Comp Stat Data Anal. 2015;83: 251–61.
26. Hastie T, Tibshirani T. Efficient quadratic regularization for expression arrays. Biostatistics. 2004;5:329–40.
27. Zuber V, Duarte Silva AP, Strimmer K. A novel algorithm for simultaneous SNP selection in high-dimensional genome-wide association studies. BMC Bioinformatics. 2012;13:284.