Published in final edited form as:

Biometrika. 2020 September ; 107(3): 609–625. doi:10.1093/biomet/asaa007.

Sparse semiparametric canonical correlation analysis for data of mixed types

GRACE YOON, **RAYMOND J. CARROLL**, **IRINA GAYNANOVA**

Department of Statistics, Texas A&M University, College Station, Texas 77843, U.S.A.

Summary

Canonical correlation analysis investigates linear relationships between two sets of variables, but often works poorly on modern datasets due to high-dimensionality and mixed data types (continuous/binary/zero-inflated). We propose a new approach for sparse canonical correlation analysis of mixed data types that does not require explicit parametric assumptions. Our main contribution is the use of truncated latent Gaussian copula to model the data with excess zeroes, which allows us to derive a rank-based estimator of latent correlation matrix without the estimation of marginal transformation functions. The resulting semiparametric sparse canonical correlation analysis method works well in high-dimensional settings as demonstrated via numerical studies, and application to the analysis of association between gene expression and micro RNA data of breast cancer patients.

Keywords

BIC; Gaussian copula model; Kendall's τ; Latent correlation matrix; Truncated continuous variable; Zero-inflated data

1. Introduction

Canonical correlation analysis investigates linear associations between two sets of variables, and is widely used in various fields including biomedical sciences, imaging and genomics (Hardoon et al., 2004; Chi et al., 2013; Safo et al., 2018). However, sample canonical correlation analysis often performs poorly due to two main challenges: high-dimensionality and non-normality of the data.

In high-dimensional settings, sample canonical correlation analysis is known to overfit the data due to singularity of sample covariance matrices (Hardoon et al., 2004; Guo et al., 2016). Additional regularization is often used to address this challenge. González et al. (2008) focus on ridge regularization of sample covariance matrices to avoid singularity, while more recent methods focus on sparsity regularization of canonical vectors (Parkhomenko et al., 2009; Witten et al., 2009; Chi et al., 2013; Cruz-Cano & Lee, 2014; Wilms & Croux, 2015; Safo et al., 2018). At the same time, with the advancement in technology, it is common to collect data of different types. For example, the Cancer Genome

gyoon@stat.tamu.edu .

Atlas Project contains matched data of mixed types such as gene expression (continuous), mutation (binary) and micro RNA (count) data. While regularized canonical correlation methods work well for Gaussian data, they still are based on sample covariance matrix, and therefore are not appropriate for the analysis in the presence of binary data or data with excess of zero values.

Several approaches have been proposed to address the non-normality of the data. On the one hand, there are completely non-parametric approaches such as kernel canonical correlation analysis (Hardoon et al., 2004). On the other hand, there are parametric approaches building up on probabilistic interpretation of Bach & Jordan (2005). For example, Zoh et al. (2016) develop probabilistic canonical correlation analysis for count data by exploring natural parameter for Poisson distribution. More recently, Agniel & Cai (2017) utilize the normal semi-parametric transformation model for the analysis of mixed types of variables, however the method requires estimation of marginal transformation functions via nonparametric maximum likelihood.

In summary, a significant progress has been made in developing regularized variants of sample canonical correlation analysis that work well in high-dimensional settings. However, these approaches are not suited for mixed data types. At the same time, several methods have been proposed to account for non-normality of the data, however are not designed for high-dimensional settings. More importantly, to our knowledge none of the existing methods explicitly address the case of zero-inflated measurement, which, for example, is common for micro RNA and microbiome abundance data.

To bridge this gap, we propose a semi-parametric approach for sparse canonical correlation analysis, which allows to handle high-dimensional data of mixed types via a common latent Gaussian copula framework. Our work has three main contributions. First, we assume that zeros in the data are observed due to truncation of underlying latent continuous variable, and define corresponding truncated Gaussian copula model. We derive explicit formulas for the bridge functions that connect the Kendall's τ of observed data to the latent correlation matrix for different combinations of data types, and use these formulas to construct a rank-based estimator of the latent correlation matrix for the mixed (continuous/binary/truncated) data. Fan et al. (2016) use bridge function approach in the context of graphical models, however the authors do not consider the truncated variable type. The latter requires derivation of new bridge functions, and those derivations are considerably more involved than corresponding derivations for continuous/binary case. The significant advantage of bridge function technique is that it allows to estimate the latent correlation structure of Gaussian copula without estimating marginal transformation functions, in contrast to Agniel & Cai (2017). Secondly, we use the derived rank-based estimator instead of sample correlation matrix within the sparse canonical correlation analysis framework that is motivated by Chi et al. (2013) and Wilms & Croux (2015). This allows us to take into account the dataset-specific correlation structure in addition to cross-correlation structure. In contrast, Parkhomenko et al. (2009) and Witten et al. (2009) model the variables within each dataset as uncorrelated. We develop an efficient optimization algorithm to solve the corresponding problem. Finally, we propose two types of Bayesian Information Criterion (BIC) for tuning parameter selection, which leads to significant computational saving compared to commonly used cross-validation and permutation techniques (Witten & Tibshirani, 2009). Wilms & Croux (2015) also use BIC in canonical correlation analysis context, however only one criterion is proposed. Two criteria originate from BIC formulation for Gaussian linear models depending on whether the case of known or unknown error variance is considered. We found that both are competitive in our numerical studies, however one criterion works best for variable selection, whereas the other works best for prediction.

2. Background

2·1. Canonical correlation analysis

In this section we review both the classical canonical correlation analysis, and its sparse alternatives. Given two random vectors $\mathbf{X}_1 \in \mathbb{R}^{p_1}$ and $\mathbf{X}_2 \in \mathbb{R}^{p_2}$, let $\Sigma_1 = \mathrm{cov}(\mathbf{X}_1)$, $\Sigma_2 = \mathrm{cov}(\mathbf{X}_2)$ and $\Sigma_{12} = \mathrm{cov}(\mathbf{X}_1, \mathbf{X}_2)$. The population canonical correlation analysis (Hotelling, 1936) seeks linear combinations $w_1^\top \mathbf{X}_1$ and $w_2^\top \mathbf{X}_2$ with maximal correlation:

$$\underset{w_1, w_2}{\text{maximize}} \left\{ w_1^\top \Sigma_{12} w_2 \right\} \text{ subject to } w_1^\top \Sigma_1 w_1 = 1, \;\; w_2^\top \Sigma_2 w_2 = 1 . \tag{1}$$

Problem (1) has a closed form solution via the singular value decomposition of $\Sigma_1^{-1/2}\Sigma_{12}\Sigma_2^{-1/2}$. Given the first pair of singular vectors (u, v), the solutions to (1) can be expressed as $w_1 = \Sigma_1^{-1/2}u$ and $w_2 = \Sigma_2^{-1/2}v$.

The sample canonical correlation analysis replaces Σ_1, Σ_2 and Σ_{12} in (1) by corresponding sample covariance matrices S_1, S_2 and S_{12}. In high-dimensional settings when sample size is small compared to the number of variables, S_1 and S_2 are singular, thus leading to non-uniqueness of solution and poor performance due to overfitting. A common approach to circumvent this challenge is to consider sparse regularization of w_1 and w_2 via the addition of ℓ_1 penalty in the objective function of (1) (Witten et al., 2009; Parkhomenko et al., 2009; Chi et al., 2013; Wilms & Croux, 2015). The sparse canonical correlation analysis is then formulated as

$$\underset{w_1, w_2}{\text{maximize}} \left\{ w_1^\top S_{12} w_2 - \lambda_1 \left\| w_1 \right\|_1 - \lambda_2 \left\| w_2 \right\|_1 \right\} \text{ subject to } w_1^\top S_1 w_1 \le 1, \;\; w_2^\top S_2 w_2 \le 1 . \tag{2}$$

In addition to ℓ_1 penalties, the equality constraints in (1) are replaced with inequality constraints which define convex sets. This generalization is possible since nonzero solutions to (2) satisfy the constraints with equality, see Proposition 1 below.

While problem (2) works well in high-dimensional settings, it still relies on sample covariance matrices, and therefore is not well-suited for skewed or non-continuous data, such as binary or zero-inflated. Further we review the Gaussian copula models that we propose to use to address these challenges.

2·2. Latent Gaussian copula model for mixed data

In this section we review the Gaussian copula model in Liu et al. (2009), and its extension to mixed (continuous/binary) data in Fan et al. (2016).

Definition 1 (Gaussian copula model). *A random vector* $\mathbf{X} = (X_1, \dots, X_p)^\top$ *satisfies Gaussian copula model if there exists a set of monotonically increasing transformations* $f = \left(f_j\right)_{j=1}^{p}$ *satisfying* $f(\mathbf{X}) = (f_1(X_1), \dots, f_p(X_p))^\top \sim \mathrm{N}_p(0,\Sigma)$ *with* $\Sigma_{jj} = 1$. *We denote* $\mathbf{X} \sim \mathrm{NPN}(0,\Sigma, f)$.

Definition 2 (Latent Gaussian copula model for mixed data). *Let* $\mathbf{X}_1 \in \mathbb{R}^{p_1}$ *be continuous and* $\mathbf{X}_2 \in \mathbb{R}^{p_2}$ *be binary random vectors with* $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$. *Then* $\mathbf{X}$ *satisfies the latent Gaussian copula model if there exists a* p_2*-dimensional random vector* $\mathbf{U}_2 = \left(U_{p_1+1}, \dots, U_{p_1+p_2}\right)^\top$ *such that* $\mathbf{U} := (\mathbf{X}_1, \mathbf{U}_2) \sim \mathrm{NPN}(0,\Sigma, f)$ *and* $X_j = I(U_j > C_j)$ *for all* $j = p_1 + 1, \dots, p_1 + p_2$, *where* $I(\cdot)$ *is the indicator function and* $\mathbf{C} = (C_1, \dots, C_p)$ *is a vector of constants. We denote* $\mathbf{X} \sim \mathrm{LNPN}(0, \Sigma, f, \cdot\mathbf{C})$, *where* Σ *is the latent correlation matrix.*

Fan et al. (2016) consider the problem of estimating Σ for the latent Gaussian copula model based on the Kendall's $\boldsymbol{\tau}$. Given the observed data $(X_{j1}, X_{k1}), \dots, (X_{jn}, X_{kn})$ for variables X_j and X_k, Kendall's $\boldsymbol{\tau}$ is defined as

$$\hat{\tau}_{jk} = 2\{n(n-1)\}^{1/2} \sum_{1 \le i < i' \le n} \mathrm{sign}\left(X_{ji} - X_{ji'}\right)\mathrm{sign}\left(X_{ki} - X_{ki'}\right).$$

Since $\hat{\tau}_{jk}$ is invariant under monotone transformation of the data, it is well-suited to capture associations in copula models. Let $\tau_{jk} = \mathbb{E}\left(\hat{\tau}_{jk}\right)$ be the population Kendall's $\boldsymbol{\tau}$. The latent correlation matrix Σ can be connected to the Kendall's $\boldsymbol{\tau}$ via the so-called bridge function F such that $\Sigma_{jk} = F^{-1}(\tau_{jk})$ for all variables j and k. Fan et al. (2016) derive an explicit form of the bridge function for continuous, binary and mixed (continuous/binary) variable pairs, which allows to estimate latent correlation matrix via method of moments. We summarize these results below.

Theorem 1 (Fan et al. (2016)). *Let* $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2) \sim \mathrm{LNPN}(0,\Sigma, f, \mathbf{C})$ *with* p_1*-dimensional continuous* $\mathbf{X}_1$ *and* p_2*-dimensional binary* $\mathbf{X}_2$. *The rank-based estimator of* Σ *is given by the symmetric matrix* $\hat{R}$ *with* $\hat{R}_{jj} = 1$ *and* $\hat{R}_{jk} = \hat{R}_{kj} = F_{jk}^{-1}\left(\hat{\tau}_{jk}\right)$, *where for* $t \in (0, 1)$

$$F_{jk}(t) = \begin{cases} 2\sin^{-1}(t)/\pi & \text{if } 1 \le j < k \le p_1; \\ 2\left\{\Phi_2\left(\Delta_j, \Delta_k, t\right) - \Phi\left(\Delta_j\right)\Phi\left(\Delta_k\right)\right\} & \text{if } p_1 + 1 \le j < k \le p_1 + p_2; \\ 4\Phi_2\left(\Delta_j, 0, t/\sqrt{2}\right) - 2\Phi\left(\Delta_j\right) & \text{if } 1 \le j \le p_1, p_1 + 1 \le k \le p_1 + p_2. \end{cases}$$

Here $\Delta_j = f_j(C_j)$, $\Phi(\cdot)$ *is the cdf of standard normal distribution, and* $\Phi_2(\cdot,\cdot,t)$ *is the cdf of standard bivariate normal distribution with correlation t*

Remark 1. Since $\Delta_j = f_j(C_j)$ is unknown in practice, Fan et al. (2016) propose to use plug-in estimator from the moment equation $\mathbb{E}\left(X_{ij}\right) = 1 - \Phi\left(\Delta_j\right)$ leading to $\hat{\Delta}_j = \Phi^{-1}\left(1 - \bar{X}_j\right)$.

Fan et al. (2016) use these results in the context of Gaussian graphical models, and replace the sample covariance matrix with rank-based estimator $\hat{R}$, which allows to use Gaussian models with skewed con tinuous and binary data. However, Fan et al. (2016) do not consider the case of zero-inflated data, which requires formulation of a new model, and subsequently derivation of new bridge functions.

3. Methodology

3·1. Truncated latent Gaussian copula model

Our goal is to model the zero-inflated data through the latent Gaussian copula models. Two motivating examples are micro RNA and microbiome data, where it is common to encounter large number of zero counts. In both examples it is reasonable to assume that zeros are observed due to truncation of underlying latent continuous variable. More generally, one can think of zeroes as representing the measurement error due to truncation of values below a certain positive threshold. This intuition leads us to consider the following model.

Definition 3 (Truncated latent Gaussian copula model). *A random vector* $\mathbf{X} = (X_1, \ldots, X_d)^\top$ *satisfies truncated Gaussian copula model if there exists a d-dimensional random vector* $\mathbf{U} = (U_1, \ldots, U_d)^\top \sim \mathrm{NPN}(0, \Sigma, f)$ *such that*

$$X_j = I\left(U_j > C_j\right)U_j \quad (j = 1, \ldots, d),$$

where $I(\cdot)$ *is the indicator function and* $\mathbf{C} = (C_1, \ldots, C_d)$ *is a vector of positive constants. We denote* $X \sim \mathrm{TLNPN}(0, \Sigma, f, \mathbf{C})$, *where* Σ *is the latent correlation matrix.*

The methodology in Fan et al. (2016) allows to estimate the latent correlation matrix in the presence of mixed continuous and binary data. Our Definition 3 adds a third type, which we denote as *truncated* for short. To construct a rank-based estimator for Σ as in Theorem 1 in the presence of truncated variables, below we derive an explicit form of the bridge function for all possible combinations of the data types (continuous/binary/truncated). Throughout, we use $\Phi(\cdot)$ for the cdf of standard normal distribution and $\Phi_d(\cdot\cdot\cdot;\Sigma_d)$ for the cdf of standard d-variate normal distribution with correlation matrix Σ_d. All the proofs are deferred to the Appendix A.

Theorem 2. *Let* X_j *be truncated and* X_k *be binary. Then* $\mathbb{E}\left(\hat{\tau}_{jk}\right) = F\left(\Sigma_{jk}; \Delta_j, \Delta_k\right)$, *where*

$$F\left(\Sigma_{jk}; \Delta_j, \Delta_k\right) = 2\left\{1 - \Phi\left(\Delta_j\right)\right\}\Phi\left(\Delta_k\right) - 2\Phi_3\left(-\Delta_j, \Delta_k, 0; \Sigma_{3a}\right) - 2\Phi_3\left(-\Delta_j, \Delta_k, 0; \Sigma_{3b}\right), \Delta_j = f_j\left(C_j\right), \Delta_k = f_k\left(C_k\right)$$

and

$$\Sigma_{3a} = \begin{pmatrix} 1 & -\Sigma_{jk} & 1/\sqrt{2} \\ -\Sigma_{jk} & 1 & -\Sigma_{jk}/\sqrt{2} \\ 1/\sqrt{2} & -\Sigma_{jk}/\sqrt{2} & 1 \end{pmatrix}, \quad \Sigma_{3b} = \begin{pmatrix} 1 & 0 & -1/\sqrt{2} \\ 0 & 1 & -\Sigma_{jk}/\sqrt{2} \\ -1/\sqrt{2} & -\Sigma_{jk}/\sqrt{2} & 1 \end{pmatrix}.$$

Theorem 3. *Let X_j be truncated and X_k be continuous. Then* $\mathbb{E}\left(\hat{\tau}_{jk}\right) = F\left(\Sigma_{jk}; \Delta_j\right)$, *where*

$$F\left(\Sigma_{jk}; \Delta_j\right) = -2\Phi_2\left(-\Delta_j, 0; 1/\sqrt{2}\right) + 4\Phi_3\left(-\Delta_j, 0, 0; \Sigma_3\right), \Delta_j = f_j\left(C_j\right)$$

and

$$\Sigma_3 = \begin{pmatrix} 1 & 1/\sqrt{2} & \Sigma_{jk}/\sqrt{2} \\ 1/\sqrt{2} & 1 & \Sigma_{jk} \\ \Sigma_{jk}/\sqrt{2} & \Sigma_{jk} & 1 \end{pmatrix}.$$

Theorem 4. *Let both X_j and X_k be truncated. Then* $\mathbb{E}\left(\hat{\tau}_{jk}\right) = F\left(\Sigma_{jk}; \Delta_j, \Delta_k\right)$, *where*

$$F\left(\Sigma_{jk}; \Delta_j, \Delta_k\right) = -2\Phi_4\left(-\Delta_j, -\Delta_k, 0, 0; \Sigma_{4a}\right) + 2\Phi_4\left(-\Delta_j, -\Delta_k, 0, 0; \Sigma_{4b}\right), \Delta_j = f_j\left(C_j\right), \Delta_k = f_k\left(C_k\right)$$

and

$$\Sigma_{4a} = \begin{pmatrix} 1 & 0 & 1/\sqrt{2} & -\Sigma_{jk}/\sqrt{2} \\ 0 & 1 & -\Sigma_{jk}/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -\Sigma_{jk}/\sqrt{2} & 1 & -\Sigma_{jk} \\ -\sum_{jk}/\sqrt{2} & 1/\sqrt{2} & -\Sigma_{jk} & 1 \end{pmatrix}$$

and

$$\Sigma_{4b} = \begin{pmatrix} 1 & \Sigma_{jk} & 1/\sqrt{2} & \Sigma_{jk}/\sqrt{2} \\ \Sigma_{jk} & 1 & \Sigma_{jk}/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & \Sigma_{jk}/\sqrt{2} & 1 & \Sigma_{jk} \\ \Sigma_{jk}/\sqrt{2} & 1/\sqrt{2} & \Sigma_{jk} & 1 \end{pmatrix}.$$

We also show that the inverse bridge function exists for all of the cases.

Theorem 5. *For any constants* Δ_j *and* Δ_k, *the bridge functions* $F(\Sigma_{jk})$ *in Theorems* 2–4 *are strictly increasing in* $\Sigma_{jk} \in (-1, 1)$, *and therefore, the inverse function* $F^{-1}(\Sigma_{jk})$ *exists.*

Theorems 2–5 complement the results of Fan et al. (2016) summarized in Theorem 1 by adding three more cases (continuous/truncated, binary/truncated and truncated/truncated), thus allowing to construct rank-based estimator $\hat{R}$ for Σ in the presence of mixed (contunuous/binary/truncated) variables.

Remark 2. Since $\hat{R}$ is not guaranteed to be positive semidefinite, Fan et al. (2016) regularize $\hat{R}$ by projecting it onto the cone of positive semidefinite matrices. We follow this approach using nearPD function in Matrix R package leading to estimator $\hat{R}_p$. Furthermore, we consider

$$\tilde{R} = (1 - \rho)\hat{R}_p + \rho I \tag{3}$$

with a small value of $\rho > 0$, so that $\tilde{R}$ is strictly positive definite. Throughout, we fix $\rho = 0.01$.

Remark 3. As in binary case, $\Delta_j = f_j(C_j)$ is unknown for truncated variables. Similar to Fan et al. (2016), we use a plug-in estimator $\hat{\Delta}_j$ based on the moment equation $\mathbb{E}I\left(X_{ij} > 0\right) = \mathbb{P}\left(X_j > 0\right) = \mathbb{P}\left(f_j\left(U_j\right) > \Delta_j\right) = 1 - \Phi\left(\Delta_j\right)$. Let $n_{\text{nonzero}} = \sum_{i=1}^{n} I\left(X_{ij} > 0\right)$ for $i = 1, \ldots, n$, then we use $\hat{\Delta}_j = \Phi^{-1}\left(1 - n_{\text{nonzero}}/n\right)$.

3·2. Semiparametric sparse canonical correlation analysis

Our proposal is based on formulating sparse canonical correlation analysis using latent correlation matrix from the Gaussian copula model for mixed data. On a population level, let Σ be the latent correlation matrix for $(\mathbf{X}_1, \mathbf{X}_2) \sim \text{LNPN}(0, \Sigma, f, \mathbf{C})$ where each $\mathbf{X}_1$ and $\mathbf{X}_2$ follows one of the three data types: continuous, binary or truncated. In Section 3·1 we derived a rank-based estimator for Σ, which we propose to use within the sparse canonical correlation analysis framework (2).

Given semiparametric estimator $\tilde{R}$ in (3), we propose to find canonical vectors by solving

$$\begin{aligned} &\underset{w_1, w_2}{\text{minimize}}\left\{-w_1^\top \tilde{R}_{12} w_2 + \lambda_1 \left\|w_1\right\|_1 + \lambda_2 \left\|w_2\right\|_1\right\} \quad \text{subject to} \quad w_1^\top \tilde{R}_1 w_1 \le 1, \\ &w_2^\top \tilde{R}_2 w_2 \le 1\,. \end{aligned} \tag{4}$$

While we focus only on the estimation of the first canonical pair, the subsequent canonical pairs can be found sequentially by using the deflation scheme. Let $\tilde{R}_{12}^{(1)} = \tilde{R}_{12}$ and let $\hat{w}_1, \hat{w}_2$ be the $(k - 1)$th estimated canonical pair. To estimate the kth pair for $k > 1$, form

$$\tilde{R}_{12}^{(k)} = \tilde{R}_{12}^{(k-1)} - \left(\hat{w}_1^\top \tilde{R}_{12}^{(k-1)} \hat{w}_2\right) \tilde{R}_1 \hat{w}_1 \hat{w}_2^\top \tilde{R}_2,$$

and solve (4) using $\tilde{R}_{12}^{(k)}$ instead of $\tilde{R}_{12}$.

While problem (4) is not jointly convex in w_1 and w_2, it is biconvex. Therefore, we propose to iteratively optimize over w_1 and w_2. First, consider optimizing over w_1 with w_2 fixed.

Proposition 1. *For a fixed* $w_2 \in \mathbb{R}^{p_2}$*, let*

$$\widehat{w}_1 = \underset{w_1}{\operatorname{argmin}}\left\{-w_1^\top \widetilde{R}_{12} w_2 + \lambda_1 \left\|w_1\right\|_1\right\} \quad \textit{subject to} \quad w_1^\top \widetilde{R}_1 w_1 \le 1\,. \tag{5}$$

This problem is equivalent to finding

$$\widetilde{w}_1 = \underset{w_1}{\operatorname{argmin}}\left\{(1/2) w_1^\top \widetilde{R}_1 w_1 - w_1^\top \widetilde{R}_{12} w_2 + \lambda_1 \left\|w_1\right\|_1\right\}, \tag{6}$$

and then setting $\widehat{w}_1 = 0$ *if* $\widetilde{w}_1 = 0$, *and* $\widehat{w}_1 = \widetilde{w}_1 / \left(\widetilde{w}_1^\top \widetilde{R}_1 \widetilde{w}_1\right)^{1/2}$ if $\widetilde{w}_1 \ne 0$.

Both problems (5) and (6) are convex, but unlike (5), problem (6) is unconstrained. Furthermore, problem (6) is of the same form as the well-studied penalized LASSO problem (Tibshirani, 1996), which can be solved efficiently using for example coordinate-descent algorithm. Hence, the proposed optimization algorithm for (4) can be viewed as a sequence of LASSO problems with rescaling. Given the value of w_2 at iteration t, the updates at iteration $t + 1$ have the form

$$\widetilde{w}_1 = \underset{w_1}{\operatorname{argmin}}\left\{(1/2) w_1^\top \widetilde{R}_1 w_1 - w_1^\top \widetilde{R}_{12} w_2^{(t)} + \lambda_1 \left\|w_1\right\|_1\right\};$$

$$\widehat{w}_1^{(t+1)} = \widetilde{w}_1 / \left(\widetilde{w}_1^\top \widetilde{R}_1 \widetilde{w}_1\right)^{1/2};$$

$$\widetilde{w}_2 = \underset{w_2}{\operatorname{argmin}}\left\{(1/2) w_2^\top \widetilde{R}_2 w_2 - w_2^\top \widetilde{R}_{12}^\top w_1^{(t+1)} + \lambda_2 \left\|w_2\right\|_1\right\};$$

$$\widehat{w}_2^{(t+1)} = \widetilde{w}_2 / \left(\widetilde{w}_2^\top \widetilde{R}_2 \widetilde{w}_2\right)^{1/2}.$$

If a zero solution is obtained at any of the steps, the optimization algorithm stops, and both w_1 and w_2 are returned as zeroes. Otherwise, the algorithm proceeds until convergence, which is guaranteed due to biconvexity of (4) (Gorski et al., 2007).

We further describe coordinate-descent algorithm for (6). Consider the KKT conditions

$$\widetilde{R}_1 w_1 - \widetilde{R}_{12} w_2 + \lambda_1 s_1 = 0,$$

where s_1 is the subgradient of $\|w_1\|_1$. If $\lambda_1 \ge \|\widetilde{R}_{12} w_2\|_\infty$, it follows that $\widetilde{w}_1 = 0$. Otherwise, the ith element of w_1 can be expressed through the other coordinates as

$$w_{1i} = S_{\lambda_1}\left\{\left(\tilde{R}_{12}\right)_i w_2^{(t)} - \left(\tilde{R}_1\right)_{i,\,-i}\left(w_1\right)_{-i}\right\},$$

where $S_\lambda(t) = \mathrm{sign}(t)\,(|t| - \lambda)_+$ is the soft-thresholding operator, $(R_{12})_i$ denotes the ith row of matrix R_{12} and $(R_1)_{i,-i}$ denotes ith row of matrix R_1 without the ith component that is $(R)_{i,-i} = (R_{i1}, \ldots, R_{i,i-1}, R_{i,i+1}, \ldots, R_{ip})$. The coordinate-descent algorithm proceeds by using the above formula to update one coordinate at a time until the convergence to global optimum is achieved. This convergence is guaranteed due to convexity of the objective function and separability of the penalty with respect to coordinates (Tseng, 1988).

Remark 4. Problem (6) allows an alternative interpretation of $\tilde{R}$. Using the definition of $\tilde{R}$ in (3), (6) can be written as

$$\underset{w_1}{\text{minimize}}\left[(1-\rho)(1/2)w_1^\top \hat{R}_1 w_1 - (1-\rho)w_1^\top \hat{R}_{12} w_2^{(t)} + \rho(1/2)w_1^\top w_1 + \lambda_1\left\|w_1\right\|_1\right],$$

which is equivalent to using with $\hat{R}$ elastic net regularization rather than the lasso penalty (Zou & Hastie, 2005).

3·3. Selection of tuning parameters

Cross-validation is a popular approach to select the tuning parameter in LASSO. In our context, however, it amounts to performing a grid search over both λ_1 and λ_2. Moreover, splitting the data as in cross-validation leads to too small number of testing samples fto construct the rank-based estimator of latent correlation matrix. Instead, motivated by Wilms & Croux (2015), we propose to adapt the Bayesian information criterion to the canonical correlation analysis to avoid splitting the data and decrease computational costs.

For Gaussian linear regression model, the Bayesian information criterion (BIC) has the form

$$\text{BIC} = -2\ell + \text{df}\log n,$$

where df indicate the number of parameters in the model, and l is the log-likelihood

$$\ell = \log L = -(n/2)\log\sigma^2 - \sum_{i=1}^{n}\left(y_i - X_i\boldsymbol{\beta}\right)^2/\left(2\sigma^2\right).$$

Two cases can be considered depending on whether the variance σ^2 is known or unknown.

1. If σ^2 is known, and the data are scaled so that $\sigma^2 = 1$, then

$$\text{BIC} = n^{-1}\sum_{i=1}^{n}\left(y_i - X_i\hat{\boldsymbol{\beta}}\right)^2 + \text{df}_{\hat{\boldsymbol{\beta}}}\log n/n.$$

2. If σ^2 is unknown, using $\hat{\sigma}^2_{\text{MLE}} = n^{-1}\sum_{n=1}^{n}\left(y_i - X_i\boldsymbol{\beta}\right)^2$ leads to

$$\text{BIC} = n\log\left\{n^{-1}\sum_{i=1}^{n}\left(y_i - X_i\hat{\boldsymbol{\beta}}\right)^2\right\} + \text{df}_{\hat{\boldsymbol{\beta}}}\log n\,.$$

Wilms & Croux (2015) use criterion 2 for canonical correlation analysis by substituting $\|X_1\hat{w}_1 - X_2w_2\|_2^2/n$ instead of $\sum_{i=1}^{n}\left(y_i - X_i\hat{\boldsymbol{\beta}}\right)^2/n$ for centered X_1 and X_2. Since $\|X_1\hat{w}_1 - X_2w_2\|_2^2/n = w_1^\top S_1w_1 - 2w_1^\top S_{12}w_2 + w_2^\top S_2w_2$, and we use $\tilde{R}$ instead of sample covariance matrix S, we substitute

$$f\left(\hat{w}_1\right) = \hat{w}_1^\top\tilde{R}_1\hat{w}_1 - 2\hat{w}_1^\top\tilde{R}_{12}w_2 + w_2\tilde{R}_2w_2$$

instead of residual sum of squares. Furthermore, motivated by the performance of the adjusted degrees of freedom variance estimator in Reid et al. (2016), we also adjust $f\left(\hat{w}_1\right)$ s for the 2nd criterion leading to

$$\text{BIC1} = f\left(\hat{w}_1\right) + \text{df}_{\hat{w}_1}\log n/n;$$

$$\text{BIC2} = \log\left\{\frac{n}{n - \text{df}_{\hat{w}_1}}f\left(\hat{w}_1\right)\right\} + \text{df}_{\hat{w}_1}\log n/n\,.$$

Here $df_{\hat{w}_1}$ coincide with the size of the support (Tibshirani & Taylor, 2012). BICcriteria for w_2 are defined analogously to w_1.

We use both criteria in evaluating our approach. Given the selected criterion (either BIC1 or BIC2), we apply it sequentially at each step of biconvex optimization algorithm of Section 3·2, and each time select the tuning parameter corresponding to the smallest value of criterion.

4. Simulation Studies

In this section we evaluate the performance of the following methods: (i) Classical canonical correlation analysis based on the sample covariance matrix; (ii) Canonical ridge available in the R package CCA (González et al., 2008); (iii) Sparse canonical correlation analysis of Witten et al. (2009) available in the R package PMA; (iv) Sparse canonical correlation analysis via Kendall's τ proposed in this paper. For our method, we evaluate both types of BICcriteria as described in Section 3·3.

We generate $n = 100$ independent pairs $\left(\mathbf{Z}_1, \mathbf{Z}_2\right) \in \mathbb{R}^{p_1 + p_2}$ following

$$\begin{pmatrix}\mathbf{Z}_1\\ \mathbf{Z}_2\end{pmatrix} \sim \mathrm{N}\left\{\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\Sigma_1 & \rho\Sigma_1 w_1 w_2^\top \Sigma_2\\ \rho\Sigma_2 w_2 w_1^\top \Sigma_1 & \Sigma_2\end{pmatrix}\right\}.$$

We consider two settings for the number of variables: low-dimensional ($p_1 = p_2 = 25$) and high-dimensional ($p_1 = p_2 = 100$). Each canonical vector w_g ($g = 1, 2$) is defined by taking a vector of ones at the coordinates (1, 6, 11, 16, 21) and zeros elsewhere, and normalizing it such that $w_g^\top \Sigma_g w_g = 1$, similar model is used in Chen et al. (2013). The value of canonical correlation is set at $\rho = 0.9$. We use autoregressive structure for $\Sigma_1 = \left\{\gamma^{|j-k|}\right\}_{j,k=1}^{p_1}$ and block-diagonal structures for Σ_2:

$$\Sigma_2 = \begin{pmatrix}\Sigma_\gamma & 0 & \cdots & 0\\ 0 & \Sigma_\gamma & \cdots & 0\\ \vdots & & \ddots & \vdots\\ 0 & \cdots & 0 & \Sigma_\gamma\end{pmatrix},$$

where $\Sigma_\gamma \in R^{d \times d}$ is an equicorrelation matrix with value 1 on the diagonal and γ off the diagonal. We use five blocks of size $d \in \{3, 3, 6, 6, 7\}$ for low-dimensional, and $d \in \{12, 14, 21, 25, 28\}$ for high-dimensional setting. We set $\gamma = 0.7$ for both Σ_1 and Σ_2, similar results are obtained when autoregressive structure is substituted with identity matrix. We further randomly permute the order of variables in each Z_g to remove the covariance-induced ordering.

We consider transformations $\mathbf{U}_g = f_g(\mathbf{Z}_g + c)$ with c being 0 or 1 with equal probability. The choice of c allows to vary the proportion of zero values in truncated and binary variables at 5–80%. We consider three choices for f_g: (copula 0) no transformation, $f_g(z) = z$ for $g = 1, 2$; (copula 1) exponential transformation for $\mathbf{U}_1$, $f_1(z) = \exp(z)$, and no transformation for $\mathbf{U}_2$, $f_2(z) = z$; (copula 2) exponential transformation for $\mathbf{U}_1$, $f_1(z) = \exp(z)$, and cubic transformation for $\mathbf{U}_2$, $f_2(z) = z^3$. Finally, we set $\mathbf{X}_\mathrm{g}$ to be equal to $\mathbf{U}_\mathrm{g}$ for continuous variable type, and dichotomize $\mathbf{U}_g$ at value C to form binary/truncated $\mathbf{X}_g$. We set $C = 0$ for copula 0 and 1, and $C = 1.5$ for copula 2. For each case, we consider three combinations of variable types for $\mathbf{X}_1/\mathbf{X}_2$: truncated/truncated, truncated/continuous and truncated/binary.

To compare the methods performance, we evaluate expected out-of-sample correlation

$$\hat{\rho} = \frac{\hat{w}_1^\top \Sigma_{12} \hat{w}_2}{\left(\hat{w}_1^\top \Sigma_1 \hat{w}_1\right)^{1/2}\left(\hat{w}_2^\top \Sigma_2 \hat{w}_2\right)^{1/2}}, \tag{7}$$

and predictive loss

$$L\left(w_g, \widehat{w}_g\right) = 1 - \frac{|\widehat{w}_g^\top \Sigma_g w_g|}{\left(\widehat{w}_g^\top \Sigma_g \widehat{w}_g\right)^{1/2}} \quad (g = 1, 2); \tag{8}$$

similar loss function is used in Gao et al. (2017). Since $w_g^\top \Sigma_g w_g = 1$, $L\left(w_g, \widehat{w}_g\right) \in [0, 1]$ with $L\left(w_g, \widehat{w}_g\right) = 0$ if $\widehat{w}_g = w_g$. We also evaluate the variable selection performance using the selected model size, true-positive rate and true-negative rate defined as

$$\mathrm{TPR}_g = \frac{\#\left\{j : \widehat{w}_{gj} \neq 0 \text{ and } w_{gj} \neq 0\right\}}{\#\left\{j : w_{gj} \neq 0\right\}}, \quad \mathrm{TNR}_g = \frac{\#\left\{j : \widehat{w}_{gj} = 0 \text{ and } w_{gj} = 0\right\}}{\#\left\{j : w_{gj} = 0\right\}} \quad (g = 1, 2) .$$

The results for truncated/truncated case over 100 replications are presented in Figures 1–3, the results for other cases are qualitatively similar and deferred to Appendix B.

From Figure 1, all methods perform better in absence of data transformation (copula 0) compared to cases where transformation is applied (copula 1 and 2). Similarly, the performance deteriorates with increased dimensions leading to smaller values of $\widehat{\rho}$, larger predictive losses and worse true positive rates. The classical canonical correlation analysis performs especially poor in high-dimensional settings with $\widehat{\rho}$ being very close to 0 and predictive loss being close to 1 for both w_1 and w_2. Canonical ridge works well in copula 0 setting, however its performance is strongly affected in the presence of transformations (copula 1 and 2). Witten's method outperforms canonical ridge in the presence of transformations, however works worse than both variants of our approach. Overall, our method with BIC1 attains the highest values of $\widehat{\rho}$ in low-dimensional settings, whereas BIC2 is the highest in high-dimensional settings. Unlike the classical canonical correlation and canonical ridge, both Witten's and our method perform variable selection. Unexpected to us, the number of selected variables varies significantly across replications for Witten's method (Figure 3), leading to significant variations in true positive and true negative rates. In all cases BIC1 leads to sparsest model and highest true negative rate. On the other hand, since BIC1 sometimes misses true variables, especially in the high-dimensional settings, BIC2 shows more accurate values of $\widehat{\rho}$ and smaller predictive loss (See Figure 1). In summary, BIC1 works better for variable selection, whereas BIC2 works better for prediction.

5. Application To Tcga Data

The Cancer Genome Atlas (TCGA) project collects data from multiple platforms using high-throughput sequencing technologies. We consider gene expression data (p_1 = 891) and micro RNA data (p_2 = 431) for n = 500 matched subjects from TCGA BRCA database. We treat gene expression data as continuous and micro RNA data as truncated continuous. The range of proportions of zero values contained in each variable in micro RNA data is $0 - 49.8\%$. The subjects belong to one of the 5 breast cancer subtypes: Normal, Basal, Her2, LumA and LumB, with 37 subjects having missing subtype information (denoted as NA). The goal of the analysis is to characterize the association between gene expression and micro RNA data, and investigate whether this association is relevant with respect to breast cancer subtypes.

To investigate the performance of our method relative to other approaches, we randomly split the data 100 times. Each time 400 samples are used for training, and the remaining 100 test samples are used to asses the found association via

$$\hat{\rho}_{\text{test}} = \frac{\hat{w}_{1,\,\text{train}}^{\top}\Sigma_{12,\,\text{test}}\hat{w}_{2,\,\text{train}}}{\left(\hat{w}_{1,\,\text{train}}^{\top}\Sigma_{1,\,\text{test}}\hat{w}_{1,\,\text{train}}\right)^{1/2}\left(\hat{w}_{2,\,\text{train}}^{\top}\Sigma_{2,\,\text{test}}\hat{w}_{2,\,\text{train}}\right)^{1/2}}.$$

Here Σ_{test} is evaluated based on the test samples, and is either rank-based estimator $\tilde{R}$ (for our method), or sample covariance matrix (for other methods). We also compare the number of selected genes and micro RNAs, the results are presented in Table 1.

As expected, neither sample canonical correlation analysis nor canonical ridge method perform variable selection. In addition, $\hat{\rho}_{\text{test}}$ is very close to 0 for sample canonical correlation, confirming poor performance of the method. Canonical ridge leads to significantly higher values of $\hat{\rho}_{\text{test}}$ demonstrating the advantage of added regularization, however it still has smaller correlation values compared to other approaches. The method of Witten et al. (2009) leads to higher correlation values compared to both sample canonical correlation analysis and canonical ridge, however it still selects a significant number of variables, with highly varied model sizes across replications. We suspect this is due to the use of permutation-based algorithm for tuning parameter selection, similar behaviour is observed in Section 4. Finally, the values of $\hat{\rho}_{\text{test}}$ are the highest for both variations of our method. At the same time, both variations result in sparsest models with smallest variability in model size across replications. While BIC2 criterion leads to largest out-of-sample correlation value, BIC1 criterion leads to sparsest model. In light of these results and results of Section 4, we conclude that BIC1 works well for variable selection, whereas BIC2 works well for prediction.

We further apply our method with BIC1 criterion using the full set of $n = 500$ samples, leading to the selection of 64 genes and 8 micro RNAs. Figures 4 and 5 show heatmaps of selected variables for each platform, with samples ordered by their respective cancer subtype. The heatmaps show clear separation between Basal and other subtypes, suggesting that found association is relevant to cancer biology.

Some of the selected genes and micro RNAs can be found in recent literature which supports their association with breast cancer. For example, Xiao et al. (2018) identify hsa-miR-452–5p in the analysis of estrogen receptor subtypes of breast cancer, and Manvati et al. (2015) demonstrate negative correlation of hsa-miR-24–2 with both metastasis and increasing nodes in sporadic breast tumours. As for hsa-miR-135b, not only it is reported to be related to breast cancer cell growth (Aakula et al., 2015; Hua et al., 2016), but it is also demonstrated to regulate estrogen receptor α gene ESR1 (Aakula et al., 2015), which coincidentally is among the 64 genes selected by our approach. Some other genes among the selected ones that demonstrate association with breast cancer according to previous research are ERBB4, FOXA1, UGT2B15 and ELF5 (Kim et al., 2016; Hu et al., 2016; Piggin et al., 2016).

6. Discussion

One of the main contributions of this work is the proposed truncated Gaussian copula model for the zero-inflated data, and corresponding development of a rank-based estimator for the latent correlation matrix. While our focus is on canonical correlation analysis, the derived estimator can be used in conjunction with other covariance-based approaches, for example it can be used for constructing graphical models as in Fan et al. (2016) in cases where some or all of the variables have excess of zeroes. Micro RNA data is one example that we have explored in this work, however another prominent example is microbiome abundance data. It would be of interest to further explore the potential of our modeling approach in different application areas.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

Yoon's research was funded by a grant from the National Cancer Institute (T32-CA090301). Carroll's research was supported by a grant from the National Cancer Institute (U01-CA057030). Carroll is also Distinguished Professor, School of Mathematical and Physical Sciences, University of Technology Sydney, Broadway NSW 2007, Australia. Gaynanova's research was supported by NSF grant DMS-1712943.

REFERENCES

Aakula A, Leivonen S-K, Hintsanen P, Aittokallio T, Ceder Y, Børresen-Dale A-L, Perälä M,Östling P & Kallioniemi O (2015). Microrna-135b regulates erα, ar and hif1an and affects breast and prostate cancer cell growth. Molecular oncology 9, 1287–1300. [PubMed: 25907805]

Agniel D & Cai T (2017). Analysis of multiple diverse phenotypes via semiparametric canonical correlation analysis. Biometrics.

Bach FR & Jordan MI (2005). A probabilistic interpretation of canonical correlation analysis.

Chen M, Gao C, Ren Z & Zhou HH (2013). Sparse CCA via Precision Adjusted Iterative Thresholding. arXiv, 1311.6186v1.

Chi EC, Allen GI, Zhou H, Kohannim O, Lange K & Thompson PM (2013). Imaging genetics via sparse canonical correlation analysis. In Biomedical Imaging (ISBI), 2013 IEEE 10th International Symposium on. IEEE.

Cruz-Cano R & Lee M-LT (2014). Fast regularized canonical correlation analysis. Computational Statistics & Data Analysis 70, 88–100.

Fan J, Liu H, Ning Y & Zou H (2016). High dimensional semiparametric latent graphical model for mixed data. Journal of the Royal Statistical Society, Series B.

Gao C, Ma Z & Zhou HH (2017). Sparse CCA: Adaptive estimation and computational barriers. The Annals of Statistics 45, 2074–2101.

González I, Déjean S, Martin PG & Baccini A (2008). CCA: An R package to extend canonical correlation analysis.

Gorski J, Pfeuffer F & Klamroth K (2007). Biconvex sets and optimization with biconvex functions: a survey and extensions. Mathematical Methods of Operations Research 66, 373–407.

Guo Y, Ding X, Liu C & Xue J-H (2016). Sufficient canonical correlation analysis. IEEE Transactions on Image Processing 25, 2610–2619. [PubMed: 27071172]

Hardoon DR, Szedmak S & Shawe-Taylor J (2004). Canonical correlation analysis: An overview with application to learning methods. Neural computation 16, 2639–2664. [PubMed: 15516276]

Hotelling H (1936). Relations between two sets of variates. Biometrika 28, 321–377.

Hu DG, Selth LA, Tarulli GA, Meech R, Wijayakumara D, Chanawong A, Russell R, Caldas C, Robinson JL, Carroll JS et al. (2016). Androgen and estrogen receptors in breast cancer coregulate human udp-glucuronosyltransferases 2b15 and 2b17. Cancer research 76, 5881–5893. [PubMed: 27496708]

Hua K, Jin J, Zhao J, Song J, Song H, Li D, Maskey N, Zhao B, Wu C, Xu H et al. (2016). mir-135b, upregulated in breast cancer, promotes cell growth and disrupts the cell cycle by regulating lats2. International journal of oncology 48, 1997–2006. [PubMed: 26934863]

Kim J-Y, Jung HH, Do I-G, Bae S, Lee SK, Kim SW, Lee JE, Nam SJ, Ahn JS, Park YH et al. (2016). Prognostic value of erbb4 expression in patients with triple negative breast cancer. BMC cancer 16, 138. [PubMed: 26907936]

Liu H, Lafferty J & Wasserman L (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. Journal of Machine Learning Research 10, 2295–2328.

Manvati S, Mangalhara KC, Kalaiarasan P, Srivastava N & Bamezai R (2015). mir-24–2 regulates genes in survival pathway and demonstrates potential in reducing cellular viability in combination with docetaxel. Gene 567, 217–224. [PubMed: 25943634]

Parkhomenko E, Tritchler D & Beyene J (2009). Sparse canonical correlation analysis with application to genomic data integration. Statistical applications in genetics and molecular biology 8, 1–34.

Piggin CL, Roden DL, Gallego-Ortega D, Lee HJ, Oakes SR & Ormandy CJ (2016). ELF5 isoform expression is tissue-specific and significantly altered in cancer. Breast Cancer Research 18, 4. [PubMed: 26738740]

Plackett RL (1954). A Reduction Formula for Normal Multivariate Integrals. Biometrika 41, 351–360.

Reid S, Tibshirani R & Friedman J (2016). A study of error variance estimation in lasso regression. Statistica Sinica, 35–67.

Safo SE, Li S & Long Q (2018). Integrative analysis of transcriptomic and metabolomic data via sparse canonical correlation analysis with incorporation of biological information. Biometrics 74, 300–312. [PubMed: 28482123]

Tibshirani RJ (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Ser. B 58, 267–288.

Tibshirani RJ & Taylor J (2012). Degrees of freedom in lasso problems. Annals of Statistics 40, 1198–1232.

Tseng P (1988). Coordinate Ascent for Maximizing Nondifferentiable Concave Functions. Massachusetts Institute of Technology, Laboratory for Information and Decision Systems.

Wilms I & Croux C (2015). Sparse canonical correlation analysis from a predictive point of view. Biometrical Journal 57, 834–851. [PubMed: 26147637]

Witten DM & Tibshirani RJ (2009). Extensions of sparse canonical correlation analysis with applications to genomic data. Statistical applications in genetics and molecular biology 8, 1–27.

Witten DM & Tibshirani RJ (2011). Penalized classification using Fisher’s linear discriminant. Journal of the Royal Statistical Society, Ser. B 73, 753–772.

Witten DM, Tibshirani RJ & Hastie T (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics 10, 515–534. [PubMed: 19377034]

Xiao B, Zhang W, Chen L, Hang J, Wang L, Zhang R, Liao Y, Chen J, Ma Q, Sun Z et al. (2018). Analysis of the mirna–mrna–lncrna network in human estrogen receptor-positive and estrogen receptor-negative breast cancer based on tcga data. Gene 658, 28–35. [PubMed: 29518546]

Zoh RS, Mallick B, Ivanov I, Baladandayuthapani V, Manyam G, Chapkin RS, Lampe JW & Carroll RJ (2016). PCAN: Probabilistic correlation analysis of two non-normal data sets. Biometrics, n/a–n/a.

Zou H & Hastie T (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67, 301–320.

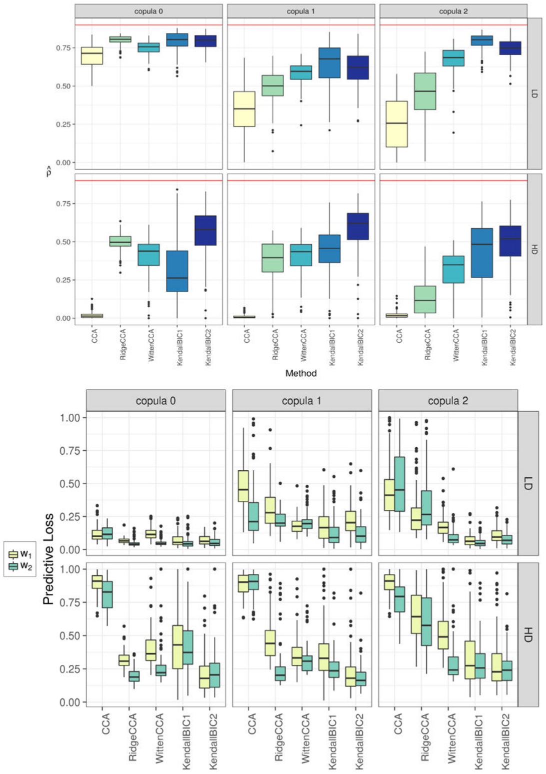

Fig. 1.

Top: The value of $\hat{\rho}$ from (7). The horizontal lines indicate true canonical correlation value ρ = 0.9. Bottom: The value of predictive loss (8). Results over 100 replications. CCA: Sample canonical correlation analysis; RidgeCCA: Canonical ridge of González et al. (2008); WittenCCA: method of Witten et al. (2009); KendallBIC1, Kendall-BIC2: proposed method with tuning parameter selected using either BIC1 or BIC2 criterion; LD: low-dimensional setting ($p_1 = p_2 = 25$); HD: high-dimensional setting ($p_1 = p_2 = 100$).

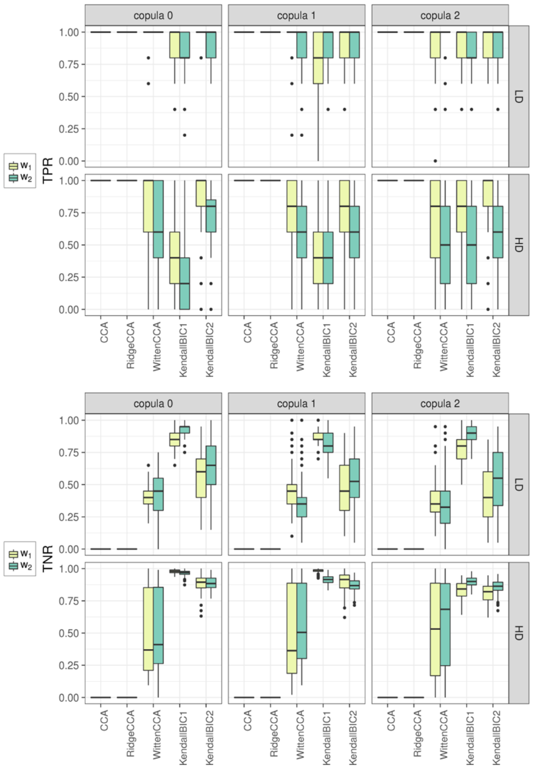

Fig. 2.

Top: True positive rate (TPR); **Bottom**: True negative rate (TNR). Results over 100 replications. CCA: Sample canonical correlation analysis; RidgeCCA: Canonical Ridge of González et al. (2008); WittenCCA: method of Witten et al. (2009); KendallBIC1, KendallBIC2: proposed method with tuning parameter selected using either BIC1 or BIC2 criterion; LD: low-dimensional setting ($p_1 = p_2 = 25$); HD: high-dimensional setting ($p_1 = p_2 = 100$).

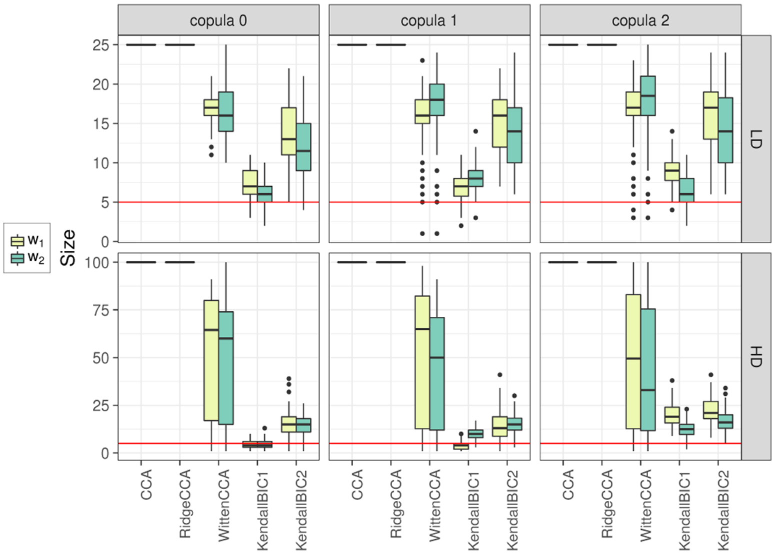

Fig. 3.

Selected model size over 100 replications. The horizontal lines indicate true model size 5. CCA: Sample canonical correlation analysis; RidgeCCA: Canonical Ridge of González et al. (2008); WittenCCA: method of Witten et al. (2009); KendallBIC1, KendallBIC2: proposed method with tuning parameter selected using either BIC1 or BIC2 criterion; LD: low-dimensional setting ($p_1 = p_2 = 25$); HD: high-dimensional setting ($p_1 = p_2 = 100$).

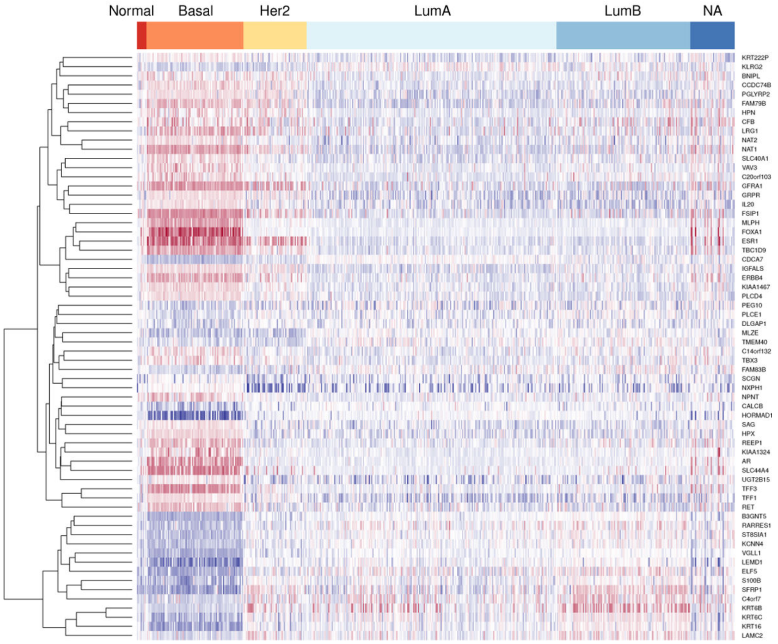

Fig. 4.

A heatmap of 64 genes selected by the proposed approach when using BIC1 criterion. Dissimilarity measure is set as $1 - \tau^2$ with τ being the Kendall's τ, and the Ward linkage is used.

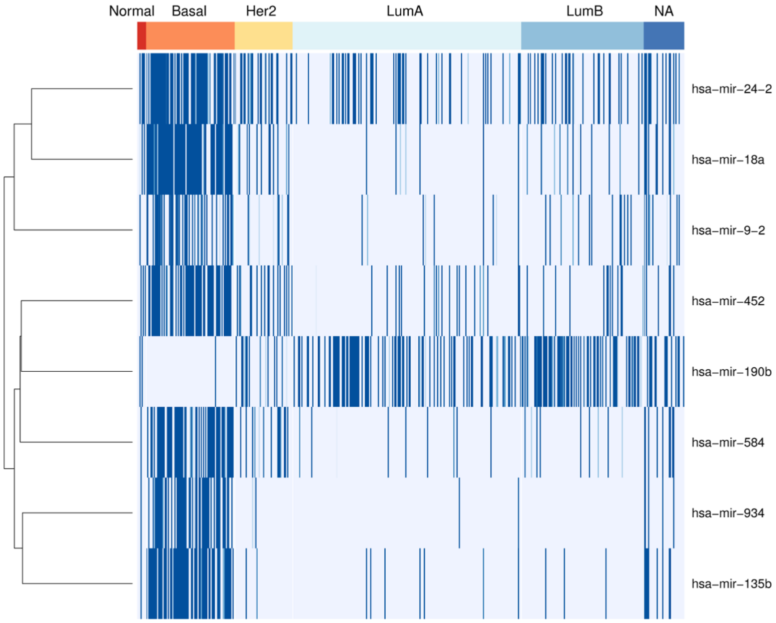

Fig. 5.

A heatmap of 8 micro RNAs selected by the proposed approach when using BIC1 criterion. Dissimilarity measure is set as $1 - \tau^2$ with τ being the Kendall's τ, and the Ward linkage is used. Colors are assigned based on variable-specific quantiles.

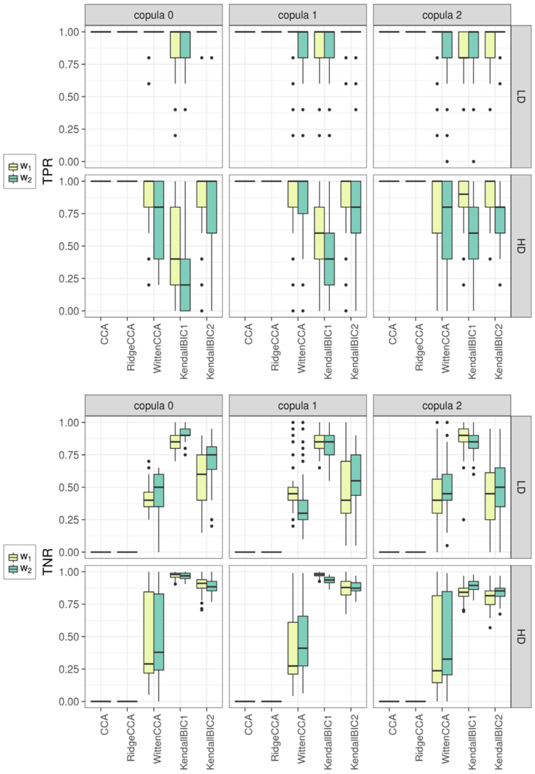

Fig. 6.

Top: The value of $\hat{\rho}$ from (7). The horizontal lines indicate true canonical correlation value ρ = 0.9. **Bottom**: The value of predictive loss (8). Results over 100 replications. CCA: Sample canonical correlation analysis; RidgeCCA: Canonical ridge of González et al. (2008); WittenCCA: method of Witten et al. (2009); KendallBIC1, Kendall-BIC2: proposed method with tuning parameter selected using either BIC1 or BIC2 criterion; LD: low-dimensional setting ($p_1 = p_2 = 25$); HD: high-dimensional setting ($p_1 = p_2 = 100$).

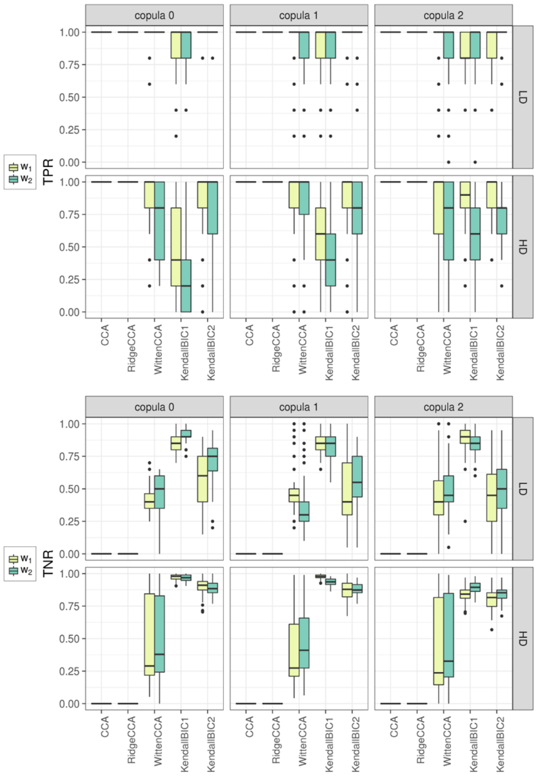

Fig. 7.

Top: True positive rate (TPR); **Bottom**: True negative rate (TNR). Results over 100 replications. CCA: Sample canonical correlation analysis; RidgeCCA: Canonical Ridge of González et al. (2008); WittenCCA: method of Witten et al. (2009); KendallBIC1, KendallBIC2: proposed method with tuning parameter selected using either BIC1 or BIC2 criterion; LD: low-dimensional setting ($p_1 = p_2 = 25$); HD: high-dimensional setting ($p_1 = p_2 = 100$).

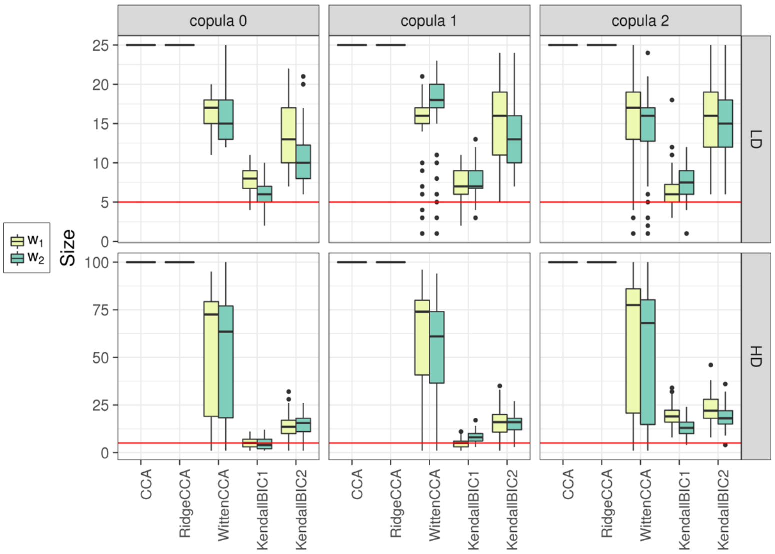

Fig. 8.

Selected model size over 100 replications. The horizontal lines indicate true model size 5. CCA: Sample canonical correlation analysis; RidgeCCA: Canonical Ridge of González et al. (2008); WittenCCA: method of Witten et al. (2009); KendallBIC1, KendallBIC2: proposed method with tuning parameter selected using either BIC1 or BIC2 criterion; LD: low-dimensional setting ($p_1 = p_2 = 25$); HD: high-dimensional setting ($p_1 = p_2 = 100$).

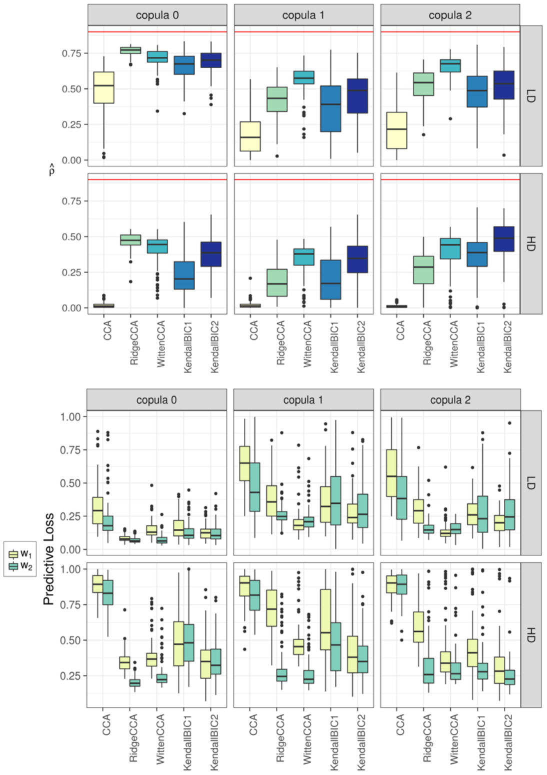

Fig. 9.

Top: The value of $\hat{\rho}$ from (7). The horizontal lines indicate true canonical correlation value ρ = 0.9. **Bottom**: The value of predictive loss (8). Results over 100 replications. CCA: Sample canonical correlation analysis; RidgeCCA: Canonical ridge of González et al. (2008); WittenCCA: method of Witten et al. (2009); KendallBIC1, Kendall-BIC2: proposed method with tuning parameter selected using either BIC1 or BIC2 criterion; LD: low-dimensional setting ($p_1 = p_2 = 25$); HD: high-dimensional setting ($p_1 = p_2 = 100$).

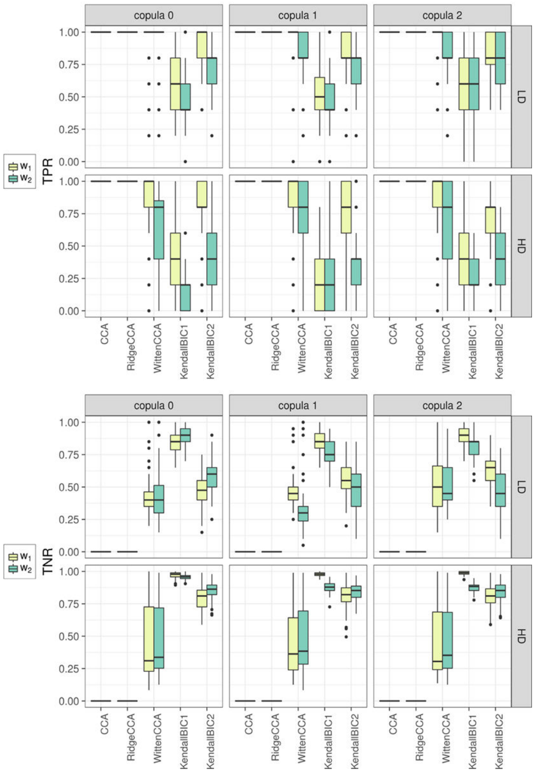

Fig. 10.

Top: True positive rate (TPR); **Bottom**: True negative rate (TNR). Results over 100 replications. CCA: Sample canonical correlation analysis; RidgeCCA: Canonical Ridge of González et al. (2008); WittenCCA: method of Witten et al. (2009); KendallBIC1, Kendall-BIC2: proposed method with tuning parameter selected using either BIC1 or BIC2 criterion; LD: low-dimensional setting ($p_1 = p_2 = 25$); HD: high-dimensional setting ($p_1 = p_2 = 100$).

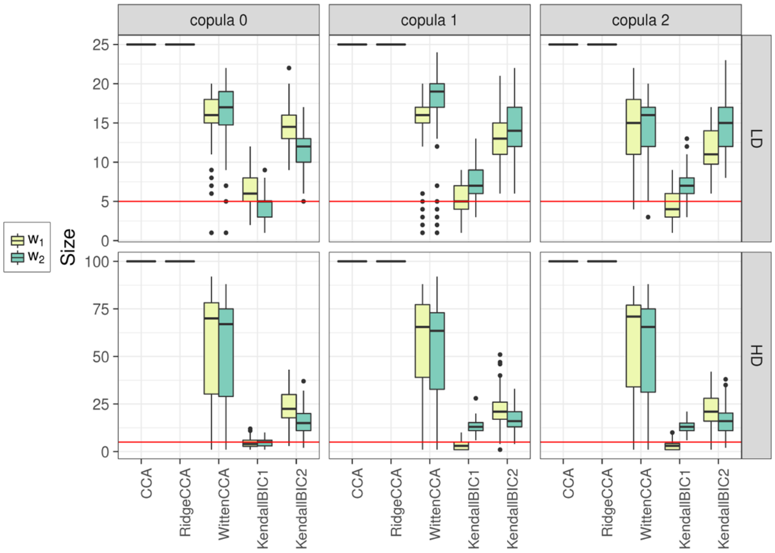

Fig. 11.

Selected model size over 100 replications. The horizontal lines indicate true model size 5. CCA: Sample canonical correlation analysis; RidgeCCA: Canonical Ridge of González et al. (2008); WittenCCA: method of Witten et al. (2009); KendallBIC1, KendallBIC2: proposed method with tuning parameter selected using either BIC1 or BIC2 criterion; LD: low-dimensional setting ($p_1 = p_2 = 25$); HD: high-dimensional setting ($p_1 = p_2 = 100$).

Table 1.

Mean support sizes and values of $\widehat{\rho}_{\text{test}}$' over 100 random splits of breast cancer data, standard deviation is given in parentheses

Method	**Selected Genes**	**Selected micro RNAs**	$\widehat{\rho}_{\text{test}}$
CCA	891	431	0·0219
	(0·00)	(0·00)	(0·111)
RidgeCCA	891	431	0·704
	(0·00)	(0·00)	(0·129)
WittenCCA	368·91	179·86	0·787
	(195·38)	(100·95)	(0·0448)
KendallBICl	83·73	6·11	0·888
	(23·43)	(1·95)	(0·0438)
KendallBIC2	106·03	105·90	0·926
	(10·86)	(10·20)	(0·231)

CCA: Sample canonical correlation analysis; RidgeCCA: Canonical Ridge of González et al. (2008); WittenCCA: method of Witten et al. (2009); KendallBIC1, KendallBIC2: proposed method with either BIC1 or BIC2 criterion.