# Minimax quasi-Bayesian estimation in sparse canonical correlation analysis via a Rayleigh quotient function

Qiuyun Zhu and Yves Atchadé*
Department of Mathematics and Statistics, Boston University

**Abstract**

Canonical correlation analysis (CCA) is a popular statistical technique for exploring relationships between datasets. In recent years, the estimation of sparse canonical vectors has emerged as an important but challenging variant of the CCA problem, with widespread applications. Unfortunately, existing rate-optimal estimators for sparse canonical vectors have high computational cost. We propose a quasi-Bayesian estimation procedure that not only achieves the minimax estimation rate, but also is easy to compute by Markov Chain Monte Carlo (MCMC). The method builds on ([34]) and uses a re-scaled Rayleigh quotient function as the quasi-log-likelihood. However, unlike ([34]), we adopt a Bayesian framework that combines this quasi-log-likelihood with a spike-and-slab prior to regularize the inference and promote sparsity. We investigate the empirical behavior of the proposed method on both continuous and truncated data, and we demonstrate that it outperforms several state-of-the-art methods. As an application, we use the proposed methodology to maximally correlate clinical variables and proteomic data for better understanding the Covid-19 disease.

*Keywords:* Sparse CCA, Minimax estimation, quasi-Bayesian inference, Markov chain Monte Carlo, simulated annealing, simulated tempering, Covid-19

*The authors gratefully acknowledge *NSF grant DMS 2015485.* The authors are grateful to Roger Zoh for very helpful discussions.

# 1 Introduction

Canonical correlation analysis (CCA) is a statistical technique –dating back at least to [16] – that is used to maximally correlate multiple datasets for joint analysis. The technique has become a fundamental tool in biomedical research where technological advances have made it possible to observe fundamental biological phenomena from multiple viewpoints — the so-called multi-omic datasets ([38, 24, 27]). Over the past two decades, limited sample size and growing dimensionality in these datasets, and the search for meaningful biological interpretations, have led to the development of sparse CCA ([37, 38, 26, 36, 15]), where a sparsity assumption is imposed on the canonical vectors.

Statistically optimal estimation of sparse CCA has been recently considered in the literature. ([11]) derived the minimax rate of estimation of sparse CCA, and proposed a two-stage estimation procedure that achieves the rate. ([34]) uses a generalized Rayleigh quotient approach to propose a two-stage estimator that also achieves the minimax rate. These two rate-optimal estimation procedures share the same limitation, that is, high computational cost. Specifically, in both approaches, each iteration of the first-stage optimization problem has a computational cost of $O(p^3)$, where $p$ is the joint number of variables in the datasets. Furthermore, the two-stage nature of these estimators can also be a problem in practice, since it can be hard to set the required stopping criterion of the first-stage solver that guarantees a good behavior of the final estimator.

We address these issues by proposing a conceptually simple, yet rate-optimal quasi-Bayesian estimator for sparse CCA. More specifically, building on ([34]), we propose a quasi-Bayesian approach that employs a re-scaled version of the Rayleigh quotient function as the quasi-log-likelihood together with a spike and slab prior to obtain a quasi-posterior distribution. The method is agnostic to the covariance matrix estimators used in con-

structing the Rayleigh quotient function. For example, we observe in our experiments that both the sample covariance matrix estimator and the Kendall's-tau-based covariance matrix estimator ([39]) can be used to construct the Rayleigh quotient function, and these matrices are allowed to be singular. Although we do not pursue this here, one can straightforwardly extend our method to solve other generalized eigenvalue problems in the same spirit as ([34]). In fact, at a high level, our method can be viewed as an improved version of simulated annealing ([18, 3]) for minimizing the Rayleigh quotient under a sparsity constraint. As such, it can be easily extended to tackle other similarly challenging non-convex statistical optimization problems with sparsity constraints.

We analyze the proposed estimator and derive its convergence rate (see Theorem 2). In the particular case where sample covariance matrices are used to estimate the Rayleigh quotient, we show that the estimator achieves the minimax rate for sparse CCA estimation, under some modest sample size conditions.

We propose a Markov Chain Monte Carlo algorithm based on simulated tempering to sample from the quasi-posterior distribution, and compute the estimator. At stationarity, the proposed algorithm has a per-iteration cost of $O(\bar{s}^2 p)$, where $\bar{s}$ is the underlying sparsity level of the posterior distribution. In all our numerical experiments, we have observed that $\bar{s}$ is of the same order as $s_\star$, namely the true sparsity level of the principal canonical vectors, leading to a very small percentage of false-positives. Furthermore, we show empirically that for sufficiently large sample size, the mixing time of the algorithm scales linearly in $p$. As a result, our estimator has a much lower computational cost than the Rifle estimator in ([34]). We also compare our method with the popular mixedCCA estimator in ([39]). The results show that although our method is computationally slower than mixedCCA, it produces statistically better estimates. We note that the estimation rate of mixedCCA is

currently unknown.

The paper is organized as follows. In Section 2 we introduce our estimation procedure and derive its convergence rate. In Section 3 we detail a simulated tempering algorithm to sample from the resulting quasi-posterior distribution. In Section 4, we study the behavior of the proposed method on both continuous and truncated data, and compare it with other methods. In Section 5, we apply the method to a case study, where one aims to correlate clinical and proteomic data from Covid-19 patients, for a better understanding of the disease. Our analysis identifies that Alpha-1-acid glycoprotein 1 (AGP 1) plays an important role in the progression of Covid-19 into a severe illness.

A Python code is available from `https://github.com/rachelwho/Sparse-CCA`.

# 2 Quasi-Bayesian sparse CCA using a Rayleigh quotient function

Let $(X, Y) \in \mathbb{R}^{p_x} \times \mathbb{R}^{p_y}$ be a pair of high-dimensional zero-mean random vectors with joint distribution $f$ and covariance matrices $\Sigma_x \stackrel{\text{def}}{=} \mathbb{E}(XX^{\mathrm{T}})$, $\Sigma_y \stackrel{\text{def}}{=} \mathbb{E}(YY^{\mathrm{T}})$ and $\Sigma_{xy} \stackrel{\text{def}}{=} \mathbb{E}(XY^{\mathrm{T}})$. Let $(v_{x\star}, v_{y\star}) \in \mathbb{R}^{p_x} \times \mathbb{R}^{p_y}$ be a pair of principal canonical vectors of $f$, that is, a vector pair that solves the following optimization problem:

$$\max_{v_x \in \mathbb{R}^{p_x},\, v_y \in \mathbb{R}^{p_y}} \; v_x^T \Sigma_{xy} v_y \quad \text{s.t.} \quad v_x^{\mathrm{T}} \Sigma_x v_x = v_y^{\mathrm{T}} \Sigma_y v_y = 1. \tag{1}$$

Since we are only interested in the directions of $v_{x\star}^{\mathrm{T}}$ and $v_{y\star}^{\mathrm{T}}$, we set $\theta_\star \stackrel{\text{def}}{=} \frac{(v_{x\star}^{\mathrm{T}}, v_{y\star}^{\mathrm{T}})^{\mathrm{T}}}{\|(v_{x\star}^{\mathrm{T}}, v_{y\star}^{\mathrm{T}})^{\mathrm{T}}\|_2}$ (so that $\|\theta_\star\|_2 = 1$) to be our main parameter of interest. The parameter $\theta_\star$ is identifiable only up to a change of sign, and hence, we shall focus on the estimation of the related projector

$\theta_\star\theta_\star^{\mathrm{T}}$. Let us define $p \stackrel{\mathrm{def}}{=} p_x + p_y$, and the matrices

$$A \stackrel{\mathrm{def}}{=} \begin{bmatrix} 0 & \Sigma_{xy} \\ \Sigma_{xy}^{\mathrm{T}} & 0 \end{bmatrix}, \quad B \stackrel{\mathrm{def}}{=} \begin{bmatrix} \Sigma_x & 0 \\ 0 & \Sigma_y \end{bmatrix} \quad \text{and} \quad \Sigma \stackrel{\mathrm{def}}{=} A + B = \begin{bmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{xy}^{\mathrm{T}} & \Sigma_y \end{bmatrix}. \tag{2}$$

Using simple arguments, we notice that the problem in (1) is equivalent to the following generalized eigenvalue problem (GEP):

$$\max_{\theta=(v_x^{\mathrm{T}},v_y^{\mathrm{T}})^{\mathrm{T}}\in\mathbb{R}^p} \theta^{\mathrm{T}} A\theta \quad \text{s.t.} \quad \theta^{\mathrm{T}} B\theta = 2. \tag{3}$$

Clearly, finding a solution of (3) is equivalent to finding a solution of

$$\max_{\theta=(v_x^{\mathrm{T}},v_y^{\mathrm{T}})^{\mathrm{T}}\in\mathbb{R}^p} \mathsf{R}(\theta) \stackrel{\mathrm{def}}{=} \frac{\theta^{\mathrm{T}} A\theta}{\theta^{\mathrm{T}} B\theta}, \tag{4}$$

where we convene that $0/0 = 0$. The objective function $\mathsf{R}(\cdot)$ in (4) is known as the *(generalized) Rayleigh quotient* of $A$ and $B$. The reformulation in (4) suggests a way to estimate the sparse canonical vectors by directly targeting the Rayleigh quotient, and this idea was first proposed in ([34]). Note that solving (4) requires specifying matrices $A$ and $B$, which are typically unknown in practice. Instead, given $n$ i.i.d. samples $\mathbf{Z} \stackrel{\mathrm{def}}{=} \{(X_i, Y_i)\}_{i=1}^n$ from $f$, one first constructs estimators of $\Sigma_x$, $\Sigma_y$ and $\Sigma_{xy}$, denoted by $\hat{\Sigma}_x$, $\hat{\Sigma}_y$ and $\hat{\Sigma}_{xy}$, respectively, and then construct estimators of $A$ and $B$ (denoted by $\hat{A}$ and $\hat{B}$, respectively) from $\hat{\Sigma}_x$, $\hat{\Sigma}_y$, and $\hat{\Sigma}_{xy}$ in the same way as in (2). In Section 4, we will provide some examples of constructing $\hat{\Sigma}_x$, $\hat{\Sigma}_y$, and $\hat{\Sigma}_{xy}$. Based on $\hat{A}$ and $\hat{B}$, one then solves (4) with the Rayleigh quotient $\mathsf{R}(\cdot)$ replaced by its sample version $\mathsf{R}_n(\cdot; \mathbf{Z})$, which is defined as

$$\mathsf{R}_n(\theta; \mathbf{Z}) \stackrel{\mathrm{def}}{=} \frac{\theta^{\mathrm{T}} \hat{A}\theta}{\theta^{\mathrm{T}} \hat{B}\theta}, \quad \forall\, \theta \in \mathbb{R}^p.$$

To guarantee that the Rayleigh quotient $\mathsf{R}_n(\cdot; \mathbf{Z})$ is well-defined, we maintain the following assumption throughout this work.

**H 1.** *For all* $\theta \in \mathbb{R}^p$, $|\theta^{\mathrm{T}} \hat{A} \theta| \leq \theta^{\mathrm{T}} \hat{B} \theta$.

**Remark 1.** H1 implies that $\theta^{\mathrm{T}} \hat{A} \theta = 0$ whenever $\theta^{\mathrm{T}} \hat{B} \theta = 0$, in which case we have $\mathsf{R}_n(\theta; \mathbf{Z}) = 0/0 = 0$. We note that H1 naturally holds when $\hat{\Sigma}_x$, $\hat{\Sigma}_y$, and $\hat{\Sigma}_{xy}$ are sample covariance matrices. Indeed, if $\hat{\Sigma}_x = n^{-1} \sum_{i=1}^n X_i X_i^{\mathrm{T}}$, $\hat{\Sigma}_y = n^{-1} \sum_{i=1}^n Y_i Y_i^{\mathrm{T}}$, and $\hat{\Sigma}_{xy} = n^{-1} \sum_{i=1}^n X_i Y_i^{\mathrm{T}}$, then for $\theta = (v_x^{\mathrm{T}}, v_y^{\mathrm{T}})^{\mathrm{T}}$, and by Cauchy-Schwarz's inequality

$$
\begin{aligned}
|\theta^{\mathrm{T}} \hat{A} \theta| = \left| \tfrac{2}{n} \textstyle\sum_{i=1}^n v_x^{\mathrm{T}} X_i Y_i^{\mathrm{T}} v_y \right| &\leq \tfrac{2}{n} \left\{ \textstyle\sum_{i=1}^n \langle v_x, X_i \rangle^2 \right\}^{1/2} \left\{ \textstyle\sum_{i=1}^n \langle v_y, Y_i \rangle^2 \right\}^{1/2} \\
&= 2 \sqrt{v_x^{\mathrm{T}} \hat{\Sigma}_x v_x} \sqrt{v_y^{\mathrm{T}} \hat{\Sigma}_y v_y} \leq v_x^{\mathrm{T}} \hat{\Sigma}_x v_x + v_y^{\mathrm{T}} \hat{\Sigma}_y v_y = \theta^{\mathrm{T}} \hat{B} \theta.
\end{aligned}
$$

It is worth mentioning that in high-dimensional regimes where $p > n$, the constructed estimators $\hat{\Sigma}_x$ and $\hat{\Sigma}_y$ (e.g., sample covariance matrices) are usually singular, thereby making a direct maximization of $\mathsf{R}_n$ challenging. Similarly, other classical CCA algorithms based on eigen-decomposition of $\hat{B}^{-1} \hat{A}$, or the singular value decomposition of $\hat{\Sigma}_x^{-1/2} \hat{\Sigma}_{x,y} \hat{\Sigma}_y^{-1/2}$ (see e.g., ([21, 1])) are also difficult to use under these regimes. Furthermore, these classical methods do not yield sparse estimates of the canonical correlation vectors.

([34]) addressed these issues by maximizing $\mathsf{R}_n(\cdot; \mathbf{Z})$ under a sparsity constraint. The authors show that this maximization problem can be solved provided that a good initial value that is sufficiently close to global maxima is provided. However, finding such a good initial value is very costly. Furthermore, the Rayleigh quotient typically admits several local maxima (as well as local minima and saddle points) that correspond to other canonical vectors, making direct maximization of $\mathsf{R}_n$ very challenging.

## 2.1 A Quasi-Bayesian approach

We propose a quasi-Bayesian framework that turns maximizing the Rayleigh quotient into a Bayesian procedure. More precisely, we propose using

$$\theta \mapsto \sigma_n \mathsf{R}_n(\theta; \mathbf{Z}) \tag{5}$$

as the quasi-log-likelihood, where $\sigma_n > 0$ is a scaling parameter. We combine this quasi-log-likelihood with a spike-and-slab prior distribution, which is a common choice for Bayesian sparse modeling ([13]). Specifically, given a variable selection parameter $\delta \in \Delta \stackrel{\text{def}}{=} \{0,1\}^p$, we let the conditional distribution of $\theta$ given $\delta$ be

$$\pi(\theta|\delta) = \prod_{j=1}^{p} \pi(\theta_j|\delta), \quad \text{where} \quad \theta_j|\delta \;=\; \theta_j|\delta_j \sim \begin{cases} \mathbf{N}(0, \rho_1^{-1}), & \text{if } \; \delta_j = 1 \\ \mathbf{N}(0, \rho_0^{-1}), & \text{if } \; \delta_j = 0 \end{cases}, \tag{6}$$

where $\rho_0 > \rho_1 > 0$ are precision parameters. Given some parameter $\mathsf{u} > 1$ and integer $s \geq 1$, the prior distribution of $\delta$ is taken as the independent product of Bernoulli distribution $\mathsf{Ber}(1/(1+p^{\mathsf{u}}))$ conditioned to stay in the set $\Delta_s \stackrel{\text{def}}{=} \{\delta \in \Delta : \; \|\delta\|_0 \leq s\}$. More specifically,

$$\pi(\delta) \propto \mathbf{1}_{\Delta_s}(\delta) \prod_{j=1}^{p} \left(\frac{1}{1+p^{\mathsf{u}}}\right)^{\delta_j} \left(\frac{p^{\mathsf{u}}}{1+p^{\mathsf{u}}}\right)^{1-\delta_j} \quad \propto \; \mathbf{1}_{\Delta_s}(\delta) \left(\frac{1}{p^{\mathsf{u}}}\right)^{\|\delta\|_0}, \qquad \forall\, \delta \in \Delta. \tag{7}$$

If we combine the spike-and-slab prior with the quasi-log-likelihood in (5), we then obtain the quasi-posterior distribution

$$\Pi(\delta, \mathrm{d}\theta|\mathbf{Z}) \propto \mathbf{1}_{\Delta_s}(\delta) \left(\frac{1}{p^{\mathsf{u}}}\sqrt{\frac{\rho_1}{\rho_0}}\right)^{\|\delta\|_0} \exp\left(-\frac{\rho_1}{2}\|\theta_\delta\|_2^2 - \frac{\rho_0}{2}\|\theta - \theta_\delta\|_2^2 + \sigma_n \mathsf{R}_n(\theta_\delta; \mathbf{Z})\right) \mathrm{d}\theta, \tag{8}$$

where $\theta_\delta$ is the component-wise product of $\theta$ and $\delta$, $\|\cdot\|_2$ is the Euclidean norm. Note that in this posterior distribution, the parameter $\theta$ is typically dense. However, since $\delta$ is sparse,

so is $\theta_\delta$. We note that the Rayleigh quotient $\mathsf{R}_n$ can take value $+\infty$ when its numerator is non-zero while its denominator is zero. If this happens over a set with non-zero Lebesgue measure, then (8) is not well-defined. H1 rules out these cases.

The spike-and-slab prior shown in (6) and (7) is fairly standard, and goes back at least to ([13]). However the way it is combined with the pseudo-likelihood to yield (8) is non-standard, and follows from ([2]). The key feature of this approach is that the parameter $\theta$ enters the quasi-likelihood only through its sparsified form $\theta_\delta$ (see (8)). This decouples the active components (namely those corresponding to $\delta_j = 1$) and the non-active components (namely those corresponding to $\delta_j = 0$), and is particularly attractive from the computational standpoint. The approach should be viewed as an approximation of the point-mass spike-and-slab prior ([23]), using the pseudo-prior device in ([5]). We refer the reader to ([2]) for more details. However, we point out that the posterior contraction theory developed in ([2]) cannot be applied to our setting.

### 2.1.1 Hyper-parameter tuning

The posterior distribution $\Pi$ is very robust to the choice of $\rho_1$ and $\mathsf{u}$, and we recommend choosing $\rho_1 \approx 1$ and $\mathsf{u} \in (1, 2]$ for best performance. The parameter $\rho_0$ has no effect on the statistical recovery of the selected components of $\theta$, but can adversely impact the MCMC mixing if its value is too large. We suggest setting $\rho_0 \sim n$, in order to match the posterior variance of the selected components that are actually zero (false-positives), and the posterior variance of the true-negatives.

The sparsity level $s$ is an upper-bound on the true sparsity of the signal, which is typically unknown. We observe that if $\delta_1, \ldots, \delta_p \overset{i.i.d.}{\sim} \mathsf{Ber}(1/(1+p^{\mathsf{u}}))$, then by Chernoff's inequality (see e.g., [35, Theorem 2.3.1]), for any $s_0 \geq \exp(1)$, we have $\mathbb{P}(\|\delta\|_0 > s_0) \leq$

$(1/p)^{(\mathsf{u}-1)s_0}$ . This suggests simply choosing $s = p$ in (6), and the resulting prior distribution would still automatically concentrate on sets $\Delta_{s_0}$, for $s_0$ small. We made this choice in all our numerical implementations. We found that the resulting posterior distribution is always automatically sparse, and learns the true sparsity of the signal. However, for the theoretical analysis of the method we will assume that a sparsity level $s$ is given such that $n \geq c_0 s \log(p)$, for some absolute constant $c_0$. We discuss the choice of $\sigma_n$ below after the statement of Theorem 2.

## 2.2 Connection with simulated annealing

Our methodology can be viewed as a principled version of simulated annealing algorithm ([18, 3]) for computing the Rifle estimator of ([34]). Given $s \geq 1$, let $\Theta_s \stackrel{\text{def}}{=} \{\theta \in \mathbb{R}^p : \|\theta\|_0 \leq s\}$. Let $\sigma_t > 0$ be given such that $\lim_{t\to\infty} \sigma_t = +\infty$, and define

$$\Pi_t(\mathrm{d}\theta) \propto e^{\sigma_t \mathsf{R}_n(\theta;\mathbf{Z})} \mathbf{1}_{\Theta_s}(\theta)\mathrm{d}\theta, \tag{9}$$

where $\mathrm{d}\theta$ denotes the extension of the Lebesgue measure to the set $\Theta_s$. The maximization problem tackled by the authors of Rifle in ([34]) is $\max_{\theta\in\Theta_s} \mathsf{R}_n(\theta;\mathbf{Z})$. A simulated annealing solution to this problem consists in simulating a non-homogeneous Markov chain with sequence of transition kernels $\{\mathcal{M}_k,\ k \geq 1\}$, such that $\mathcal{M}_k$ has invariant distribution $\Pi_{t_k}$. As $\sigma_{t_k} \to \infty$, the distribution $\Pi_{t_k}$ puts most of its probability mass around the global modes of $\mathsf{R}_n$, and the resulting Markov chain behaves similarly (under appropriate conditions). There are however several limitations to simulated annealing in this particular setting. First, the set $\Theta_s$ is a union of a large number of subsets with varying dimensions. Therefore, sampling from $\Pi_{t_k}$ (that is, designing a good Markov kernel $\mathcal{M}_k$ with invariant distribution $\Pi_{t_k}$) is actually non-trivial. Second, the convergence of simulated annealing is

known to be highly dependent on the choice of the sequence $\{\sigma_{t_k},\ k \geq 1\}$. Our approach circumvents the first issue by working with a relaxation of $\Theta_s$, using the spike-and-slab prior. We circumvent the second issue by connecting the annealing schedule to the sample size $n$ ($\sigma_{t_k} = \sigma_n$, see details below), in such a way that the fluctuations in the resulting distribution $\Pi_{t_k}$ matches the statistical uncertainty of the underlying CCA problem.

## 2.3 Rate of convergence

Although the Rayleigh quotient $\mathsf{R}_n(\cdot; \mathbf{Z})$ may possess multiple local modes, we show in this section that most of the probability mass of the quasi-posterior distribution $\Pi(\cdot|\mathbf{Z})$ are located around $\{\pm\theta_\star\}$. For $M, N \in \mathbb{R}^{q\times q}$, we define

$$\langle M, N\rangle_{\mathsf{F}} \stackrel{\text{def}}{=} \mathsf{Tr}(M^{\mathrm{T}}N), \quad \|M\|_{\mathsf{F}} \stackrel{\text{def}}{=} \sqrt{\langle M, M\rangle_{\mathsf{F}}}, \quad \text{and} \quad \|M\|_{\mathsf{op}} \stackrel{\text{def}}{=} \sup_{u\in\mathbb{R}^q:\ \|u\|_2=1} \|Mu\|_2.$$

For $J \subseteq [1:q] \stackrel{\text{def}}{=} \{1, \ldots, q\}$, let $M_{J,J}$ denote the submatrix $(M_{ij})_{i,j\in J}$. Given $k \geq 1$, we let

$$\lambda_{\mathsf{min}}(M, k) \stackrel{\text{def}}{=} \min_{u\in\mathbb{R}^q:\ \|u\|_2=1, \|u\|_0\leq k} u^{\mathrm{T}}Mu, \quad \text{and} \quad \lambda_{\mathsf{max}}(M, k) \stackrel{\text{def}}{=} \max_{u\in\mathbb{R}^q:\ \|u\|_2=1, \|u\|_0\leq k} u^{\mathrm{T}}Mu.$$

Given an integer $\alpha \geq 1$, we set

$$\lambda^{(\alpha)}_{\mathsf{max}}(M, s) \stackrel{\text{def}}{=} \max_{J\subseteq[1:q]:\ \|J\|_0=s} \max_{\substack{A\in\mathbb{R}^{s\times s}:\ \|A\|_{\mathsf{F}}=1 \\ \mathsf{Rank}(A)\leq\alpha}} |\langle M_{J,J}, A\rangle| .$$

We first make the following basic assumption without which the sparse CCA problem would not be well defined.

**H 2.** *The joint density $f$ possesses positive definite covariance matrices $\Sigma_x$, $\Sigma_y$, and $\Sigma$, and a principal canonical vector pair $\theta_\star = (v_{x\star}^{\mathrm{T}}, v_{y\star}^{\mathrm{T}})^{\mathrm{T}}$, ($\|\theta_\star\|_2 = 1$) with density level*[1]

[1]Throughout this work, the density level of a vector refers to the proportion of its non-zero elements.

$s_\star \stackrel{\text{def}}{=} \|\theta_\star\|_0$. *Furthermore, the difference between the largest and second largest eigenvalue of* $S \stackrel{\text{def}}{=} B^{-1/2}\Sigma B^{-1/2}$ *(denoted by* $\mathsf{gap}$*), is positive.*

Our main assumption on the data generation process is the following.

**H3.** *The dataset* $\mathbf{Z} \stackrel{\text{def}}{=} \{(X_i, Y_i)\}_{i=1}^n$ *and the integer* $s \geq s_\star$ *are such that the estimators* $\hat{\Sigma}_x$, $\hat{\Sigma}_y$, $\hat{\Sigma}$ *satisfy the following.*

1. *For some absolute constants* $0 < \underline{\kappa} \leq \bar{\kappa}$,

$$\min\left(\lambda_{\mathsf{min}}(\hat{\Sigma}_x, s+s_\star),\ \lambda_{\mathsf{min}}(\hat{\Sigma}_y, s+s_\star),\ \lambda_{\mathsf{min}}(\hat{\Sigma}, s+s_\star)\right) \geq \underline{\kappa},$$

$$\max\left(\lambda_{\mathsf{max}}(\hat{\Sigma}_x, s+s_\star),\ \lambda_{\mathsf{max}}(\hat{\Sigma}_y, s+s_\star), \lambda_{\mathsf{max}}(\hat{\Sigma}, s+s_\star)\right) \leq \bar{\kappa}.$$

2. *For some constant* $\mathsf{r}_1$ *(depending possibly on* $n, p$*),*

$$\max\left(\lambda^{(2)}_{\mathsf{max}}(\hat{\Sigma}_x - \Sigma_x, s+s_\star),\ \lambda^{(2)}_{\mathsf{max}}(\hat{\Sigma}_y - \Sigma_y, s+s_\star),\ \lambda^{(2)}_{\mathsf{max}}(\hat{\Sigma} - \Sigma, s+s_\star)\right) \leq \mathsf{r}_1.$$

**Theorem 2.** *Assume H1-H3, and suppose that* $p \geq \max(c_0, s_\star \exp(1))$*, for some absolute constant* $c_0$*. Choose* $\sigma_n$ *such that* $1 \leq \sigma_n \leq p$*, and* $\mathsf{u} > 1$ *such that* $p^{\mathsf{u}-1} > 2$*. Set*

$$\epsilon \stackrel{\text{def}}{=} \frac{\mathsf{r}_1}{\mathsf{gap}}. \tag{10}$$

*There exists some absolute constant* $C_0$ *that depends only on* $\underline{\kappa}$ *and* $\bar{\kappa}$*, such that the following holds. For all* $M > C_0$ *such that*

$$\frac{M^2}{8\mathsf{gap}}\left(\frac{\underline{\kappa}}{\bar{\kappa}}\right)^2 \sigma_n \mathsf{r}_1^2 \geq s_\star(\mathsf{u}+1)\log(p), \tag{11}$$

*we have*

$$\Pi\left((\delta,\theta):\ \left\|\frac{\theta_\delta\theta_\delta^{\mathrm{T}}}{\|\theta_\delta\|_2^2} - \theta_\star\theta_\star^{\mathrm{T}}\right\|_{\mathsf{F}} > M\epsilon|\mathbf{Z}\right)\ \leq 2e^{-\frac{M^2}{8\mathsf{gap}}\left(\frac{\underline{\kappa}}{\bar{\kappa}}\right)^2\sigma_n\mathsf{r}_1^2} \leq \frac{2}{p^{s_\star(\mathsf{u}+1)}}.$$

*Proof.* See Section S-1.1 in supplementary material. □

The main conclusion of the theorem is that the posterior $\Pi(\cdot|\mathbf{Z})$ contracts around $\theta_\star\theta_\star^{\mathrm{T}}$ at the rate at least $M\epsilon$. Furthermore, setting

$$\widehat{\mathcal{P}} \stackrel{\text{def}}{=} \int_{\Delta_s\times\mathbb{R}^p} \frac{\theta_\delta\theta_\delta^{\mathrm{T}}}{\|\theta_\delta\|_2^2}\Pi(\mathrm{d}\delta,\mathrm{d}\theta|\mathbf{Z}),$$

the result implies that $\widehat{\mathcal{P}}$ (as a frequentist estimator) estimates $\theta_\star\theta_\star^{\mathrm{T}}$ at the rate $M\epsilon$. Indeed, we have

$$\left\|\widehat{\mathcal{P}} - \theta_\star\theta_\star^{\mathrm{T}}\right\|_{\mathsf{F}} \leq \int_{\Delta_s\times\mathbb{R}^p} \left\|\frac{\theta_\delta\theta_\delta^{\mathrm{T}}}{\|\theta_\delta\|_2^2} - \theta_\star\theta_\star^{\mathrm{T}}\right\|_{\mathsf{F}} \Pi\left(\mathrm{d}\delta,\mathrm{d}\theta|\mathbf{Z}\right) \leq M\epsilon + 4e^{-\frac{M^2}{8\mathsf{gap}}\left(\frac{\underline{\kappa}}{\bar{\kappa}}\right)^2 n\mathsf{r}_1^2}. \tag{12}$$

We note that Theorem 2 applies to any given dataset $\mathbf{Z}$ and estimators $\hat{\Sigma}_x$, $\hat{\Sigma}_y$ that satisfy H3, regardless of how they are formed. In the particular case where $\hat{\Sigma}_x, \hat{\Sigma}_y$ and $\hat{\Sigma}$ are covariances matrices, we show in Proposition 3 below that if $f$ is a sub-Gaussian distribution, then H3 holds with high probability. Furthermore $\mathsf{r}_1 = C_0\sqrt{(s+s_\star)\log(p)/n}$. In that case the condition in (11) becomes

$$\frac{M^2}{8\mathsf{gap}}\left(\frac{\underline{\kappa}}{\bar{\kappa}}\right)^2 C_0^2\left(\frac{\sigma_n}{n}\right)(s+s_\star)\log(p) \geq s_\star(\mathsf{u}+1)\log(p),$$

which is easily satisfied when the scaling parameter $\sigma_n$ satisfies $n = O(\sigma_n)$, as $n\to\infty$. In this case the convergence rate of $\widehat{\mathcal{P}}$ towards $\theta_\star\theta_\star^{\mathrm{T}}$ is

$$\epsilon = \frac{1}{\mathsf{gap}}\sqrt{\frac{(s+s_\star)\log(p)}{n}}, \tag{13}$$

which achieves the minimax rate of the CCA problem, as derived in ([11]), by taking $s$ as some constant multiple of $s_\star$. Further increasing $\sigma_n$ has no impact on this rate, but of course, makes $\Pi$ more concentrated around the modes of $\mathsf{R}_n(\cdot;\mathbf{Z})$, thereby making the MCMC computation more challenging. This suggests that the choice $\sigma_n \propto n$ is the right scaling.

**Remark 2.1.** *The discussion so far has focused on estimating the projector $\theta_\star\theta_\star^{\mathrm{T}}$. If the vector $\theta_\star$ itself is needed, we are able to construct an estimator of $\theta_\star$ from the projector estimator $\widehat{\mathcal{P}}$. Specifically, let $v_1(\widehat{\mathcal{P}})$ denote the leading eigenvector of $\widehat{\mathcal{P}}$, then from the Davis-Kahan theorem (see e.g., [35, Theorem 4.5.5]), we have*

$$\min\Big(\|v_1(\widehat{\mathcal{P}})-\theta_\star\|_2,\ \|v_1(\widehat{\mathcal{P}})+\theta_\star\|_2\Big)\le 2^{3/2}\|\widehat{\mathcal{P}}-\theta_\star\theta_\star^{\mathrm{T}}\|\le 2^{3/2}\|\widehat{\mathcal{P}}-\theta_\star\theta_\star^{\mathrm{T}}\|_{\mathsf{F}}, \tag{14}$$

*and $\|\widehat{\mathcal{P}}-\theta_\star\theta_\star^{\mathrm{T}}\|_{\mathsf{F}}$ can be bounded as in* (12).

### 2.3.1 On Assumption H3

It is well-known that Assumption H3-(1) holds true in the particular case of covariance matrices of sub-Gaussian random vectors, provided that the sample size satisfies $n\ge c_0(s+s_\star)\log(p)$, for some absolute constant $c_0$. See for instance [28] Theorem 1, or [12] Lemma 6.5 for the Gaussian case, and [31] Theorem 3.2 for more general sub-Gaussian distributions. Under roughly the same sample size conditions, H3-(2) is also known to hold as we show below.

**Proposition 3.** *Suppose that $\mathbf{Z}\overset{\text{def}}{=}\{(X_i,Y_i)\}_{i=1}^n$ are i.i.d. random vectors from a mean-zero sub-Gaussian distribution $f$, with sub-Gaussian norm $K\overset{\text{def}}{=}\sup\{\|\langle Z,u\rangle\|_{\psi_2},\ u\in\mathbb{R}^p,\ \|u\|_2=1\}$, where $\|\cdot\|_{\psi_2}$ refers to the sub-Gaussian norm of a random variable. Let $\hat\Sigma_x=n^{-1}\sum_{i=1}^n X_iX_i^{\mathrm{T}}$, $\hat\Sigma_y=n^{-1}\sum_{i=1}^n Y_iY_i^{\mathrm{T}}$, and $\hat\Sigma=n^{-1}\sum_{i=1}^n Z_iZ_i^{\mathrm{T}}$. There exist absolute constants $c_0,C>1$, such that for all $1\le s\le p$, and all $n\ge 4c_0s\log(p)$,*

$$\max\Big(\lambda_{\mathit{max}}^{(\alpha)}(\hat\Sigma_x-\Sigma_x,s),\ \lambda_{\mathit{max}}^{(\alpha)}(\hat\Sigma_y-\Sigma_y,s),\ \lambda_{\mathit{max}}^{(\alpha)}(\hat\Sigma-\Sigma,s)\Big)\le CK^2\lambda_{\mathit{max}}(\Sigma,s)\sqrt{\frac{c_0\alpha s\log(p)}{n}},$$

*with probability $1-2p^{-(c_0-1)s}$.*

*Proof.* See Section S-2 in supplementary material. □

### 2.3.2 Bayesian inference

We have developed a method that employs a quasi-posterior distribution to produce a frequentist estimator. The idea of using a Bayesian framework to produce frequentist estimators is of course well-established in statistical decision theory ([29]). The extension to quasi-likelihood functions is also not new ([22, 6, 7]). An important statistical question here is whether one can use the full quasi-posterior distribution $\Pi(\cdot|\mathbf{Z})$ to carry inference on $\theta_\star\theta_\star^T$, for instance through credible sets. The difficulty is the lack of calibration of the quasi-likelihood function (we could easily replace $\sigma_n$ by $2\sigma_n$ as a scaling factor in the Rayleigh quotient). To address this issue some authors have developed post-processing methods to match samples from the quasi-posterior distribution to the corresponding frequentist central limit theorem distribution ([4, 33]). However these methods rely crucially on the Bernstein-von Mises theorem and the central limit theorem that are only well-understood in fixed-dimensional settings. Extending these ideas to the (high/growing)-dimensional setting remains largely open. We leave this question as a possible future research. Currently we do not advocate the use of our quasi-posterior distribution for Bayesian inference on $\theta_\star$.

# 3 Computation using Markov Chain Monte Carlo

As shown in Section 2.3, by re-scaling (annealing) the Rayleigh quotient function, we have created a posterior distribution $\Pi(\cdot|\mathbf{Z})$ that puts most of its probability mass around its global mode (located near $\{\pm\theta_\star\}$). However, the annealing also significant decreases the accessibility of the global mode starting from other parts of the space. To effectively deal with this configuration, we propose a Markov Chain Monte Carlo sampling strategy based on simulated tempering ([14, 20]). Given $K$ temperatures $1 = t_1 < t_2 < \ldots < t_K$, and

$K$ positive weights $c_1, \ldots, c_K$, we introduce an extended distribution on $\mathsf{X} \stackrel{\text{def}}{=} \Delta \times \mathbb{R}^p \times \{1, \ldots, K\}$, which is

$$\bar{\Pi}(\delta, \mathrm{d}\theta, k|\mathbf{Z}) \propto \frac{1}{c_k} \exp\left(\frac{\mathsf{a}}{t_k}\|\delta\|_0 - \frac{\rho_1}{2t_k}\|\theta_\delta\|_2^2 - \frac{\rho_0}{2t_k}\|\theta - \theta_\delta\|_2^2 + \frac{\sigma_n}{t_k}\mathsf{R}_n(\theta_\delta; \mathbf{Z})\right) \mathrm{d}\theta. \tag{15}$$

We recover the distribution (8) as the conditional distribution of $(\delta, \theta)$ given $k = 1$ in (15). To sample from (15), we use a simulated tempering Metropolis-Hastings-within-Gibbs strategy that is described in Algorithm 1 in the supplementary material S-3. The algorithm is very fast and scales well with the dimension $p$, and iteration $k$ of the algorithm has computational cost $O(p\|\delta^{(k)}\|_0^2)$.

Algorithm 1 generates a Markov chain $\{X^{(t)},\ t \geq 0\}$, where $X^{(t)} = (\delta^{(t)}, \theta^{(t)}, k^{(t)}) \in \mathsf{X}$ that is phi-irreducible aperiodic with invariant distribution given by (15). The pairs $(\delta^{(t)}, \theta^{(t)})$ at times $t$ where $\{k^{(t)} = 1\}$ then give the desired approximate samples from $\Pi(\cdot|\mathbf{Z})$. For the investigation of mixing of the algorithm, please see the supplementary material.

# 4 Numerical studies

We perform a simulation study that compares our approach to the frequentist methods Rifle in [34] and mixedCCA in [39]. We investigate the behavior of these methods in two settings: (i) continuous datasets, where we use sample covariance matrix estimator and (ii) mixed datasets, where we use Kendall's-tau-based estimator as proposed in [39]. The Python codes for our method is available from `https://github.com/rachelwho/Sparse-CCA`.

## 4.1 Simulated data generation

We simulate the datasets using the following model from [34]. Specifically, we let $p_x = p_y = p/2$, and consider two $(p/2)$-dimensional random vectors $X$ and $Y$ with joint distribution $(X, Y) \sim \mathbf{N}(0, \Sigma)$. Here we let

$$\Sigma = \begin{pmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{xy}^{\mathrm{T}} & \Sigma_y \end{pmatrix} \quad \text{and} \quad \Sigma_{xy} = \frac{\lambda_1 \Sigma_x v_{x\star} v_{y\star}^{\mathrm{T}} \Sigma_y}{\sqrt{v_{x\star}^{\mathrm{T}} \Sigma_x v_{x\star}} \sqrt{v_{y\star}^{\mathrm{T}} \Sigma_y v_{y\star}}},$$

where $0 < \lambda_1 < 1$ is the largest generalized eigenvalue, and $v_{x\star}$ and $v_{y\star}$ are the principal canonical vectors. The structures of $\Sigma_x$ and $\Sigma_y$ vary across different experimental settings, and will be described in the subsequent sections. Clearly, $(v_{x\star}, v_{y\star})$ is the maximizer of the Rayleigh quotient in (4), and $\lambda_1$ is the maximum value. Then we generate $n$ samples $\mathbf{Z} = \{(X^{(i)}, Y^{(i)})\}_{i=1}^n$ from $\mathbf{N}(0, \Sigma)$.

## 4.2 Comparison with other methods

We compare our method to two other methods, namely Rifle [34] and mixedCCA [39]. We investigate the behaviors of these methods in two settings. In the first setting, we use continuous datasets and compare our method with both Rifle and mixedCCA. In the second setting, we use mixed datasets and compare our method with mixedCCA (since Rifle is only designed for the continuous datasets).

### 4.2.1 Description of Rifle and mixedCCA

Before presenting our experimental results, let us briefly describe the other two methods, namely Rifle and mixedCCA. Rifle is a two-stage algorithm, where in the first stage, it (approximately) solves a convex relaxation of the problem in (1) to produce an initial

estimate of the singular vectors $(v_x^{\mathrm{T}}, v_y^{\mathrm{T}})^{\mathrm{T}}$, which are then refined in the second stage using gradient ascent on the Rayleigh quotient $\mathsf{R}_n(\cdot; \mathbf{Z})$, with a truncation step such that only the $m$ entries with the largest absolute values are kept (and the remaining entries are set to zero). Here $m$ is a user-specified parameter that indicates the desired sparsity level of the estimated principle canonical vectors $(v_x, v_y)$ – similar to $s$ above. Note that since the first stage involves solving a matrix optimization problem, its computational time is typically much higher than that of the second stage. As a different approach, mixedCCA proposes a novel and robust estimator $\hat{\Sigma}$ for the covariance matrix $\Sigma$, namely the Kendall's-tau-based estimator, and estimates the canonical vectors $(v_x, v_y)$ by solving the following convex problem:

$$\max_{v_x, v_y} v_x^{\mathrm{T}} \hat{\Sigma}_{xy} v_y - \lambda_1 \|v_x\|_1 - \lambda_2 \|v_y\|_1, \quad \text{s.t. } v_x^{\mathrm{T}} \hat{\Sigma}_x v_x \leq 1, \;\; v_y^{\mathrm{T}} \hat{\Sigma}_y v_y \leq 1, \tag{16}$$

where $\lambda_1$ and $\lambda_2$ are positive regularization parameters that need to be selected. Problem (16) is then solved via a sequence of LASSO problems.

### 4.2.2 Comparison with continuous datasets

We randomly generated 100 continuous datasets using the model in Section 4.1, with the covariance matrices $\Sigma_x$ and $\Sigma_y$ constructed in a similar way to [39]. Specifically, we set the sample size $n = 200$ and the dimension $p = 500$, and let $\Sigma_x$ and $\Sigma_y$ have the same structure, namely a block-diagonal matrix with five blocks of dimensions $\{d_1, ..., d_5\}$, respectively, and the $(j, j')$-th element of each block takes value $0.7^{|j-j'|}$. We set $\{d_1, ..., d_5\} = \{25, 50, 83, 50, 42\}$ for $\Sigma_x$ and $\{d_1, ..., d_5\} = \{83, 50, 62, 31, 24\}$ for $\Sigma_y$. In addition, we let $\lambda_1 = 0.8, (v_{x\star})_j = (v_{y\star})_j = 1/\sqrt{3}$ for $j \in \{1, 6, 11\}$, and $(v_{x\star})_j = (v_{y\star})_j = 0$ otherwise. Therefore, the true density level is $s_\star = 6$. In constructing the Rayleigh quotient $\mathsf{R}_n(\cdot; \mathbf{Z})$, we used the sample covariance matrices as estimators of $\Sigma_x$, $\Sigma_y$ and $\Sigma_{xy}$.

In Algorithm 2, we let the set of temperatures be $\{1, 1/0.9, 1/0.8, 1/0.7, 1/0.6\}$, and only recorded the iterations corresponding to temperature 1. For comparison, we used the implementation of Rifle in the R package Rifle, and set the parameter $m = 2s_\star = 12$. As pointed out in [34], the first stage is computationally expensive to run. In addition, we empirically found that when the sample size $n$ is not sufficiently large, either the estimated $v_x$ or $v_y$ from the first stage of Rifle is often zero vector, which caused us serious problems in running the second stage. Because of these issues, we evaluated separately the two stages of Rifle, which we call Rifle1 and Rifle2, respectively. We ran Rifle1 with default parameters, and ran Rifle2 starting from a solution generated by perturbing the ground-truth $(v_{x\star}^{\mathrm{T}}, v_{y\star}^{\mathrm{T}})^{\mathrm{T}}$, where the perturbation was drawn from a centered Gaussian with standard deviation 0.2. We used the implementation of mixedcca in the R package mixedCCA, where $\lambda_1$ and $\lambda_2$ were selected using two different criteria, namely BIC1 and BIC2. For this reason, we shall call the resulting algorithms mixedCCA-BIC1 and mixedCCA-BIC2, respectively. All the other parameters in Rifle and mixedCCA were set to the default values in the R packages. Both our algorithm and mixedCCA used the starting point found in the R package of mixedCCA. The output of each algorithm was normalized to have unit Euclidean norms.

**Comparison of running times.** We first compare the computational efficiency of different algorithms. Since these algorithms converge to possibly different estimators, we first ran each algorithm for a maximum iteration of 2000 to obtain the "limit point" of the sequence generated by each algorithm, denoted by $(\hat{v}_x^{\mathrm{T}}, \hat{v}_y^{\mathrm{T}})^{\mathrm{T}}$. Then, we terminated each algorithm if it either reached 1000 iterations or the estimate $(v_x^{\mathrm{T}}, v_y^{\mathrm{T}})^{\mathrm{T}}$ satisfies $\max\{|\mathsf{error}(v_x) - \mathsf{error}(\hat{v}_x)|, |\mathsf{error}(v_y) - \mathsf{error}(\hat{v}_y)|\} \leq 1 \times 10^{-4}$. As mentioned above, we treated the two stages of Rifle separately. We estimated the computation time for

**Table 1:** The computation times of all algorithms averaged across 100 continuous datasets.

| Algorithm | Simulated tempering | MixedCCA-BIC1 | MixedCCA-BIC2 | Rifle1 | Rifle2 |
|---|---|---|---|---|---|
| Running times (s) | 12 | 9.6 | 10.9 | 276 | 2.2 |

Rifle1 using the default stopping criterion as in [34], and estimated the running time of Rifle2 (starting from the perturbed ground-truth) using the termination criterion described above.

We repeatedly ran these algorithms on 100 different simulated datasets, and show the averaged estimated computation times of the algorithms in Table 1. The results confirm the high computational cost of Rifle. The results also show that our proposed estimator remains computationally competitive compared to mixedCCA, even though it is based on MCMC.

**Comparison of statistical efficiency.** We measure the quality of the estimated principle canonical vectors $v_x$ and $v_y$ using three metrics. The first one is the squared-$l_2$ errors of $v_x$ and $v_y$ to the ground-truth $v_{x\star}$ and $v_{y\star}$, respectively. Specifically, we have

$$\mathsf{error}(v_x) \stackrel{\text{def}}{=} \min\left(\|v_x - v_{x\star}\|_2^2,\ \|v_x + v_{x\star}\|_2^2\right), \tag{17}$$

and $\mathsf{error}(v_y)$ is defined similarly. The other two metrics are true-positive rate (TPR) and true-negative rate (TNR), which measure the quality of variable selection by the estimated $v_x$ and $v_y$. For $v_x$, its TPR and TNR are defined as

$$\mathsf{TPR}(v_x) \stackrel{\text{def}}{=} \frac{|\{j : (v_x)_j \neq 0, (v_{x\star})_j \neq 0\}|}{|\{j : (v_{x\star})_j \neq 0\}|} \ \text{ and } \ \mathsf{TNR}(v_x) \stackrel{\text{def}}{=} \frac{|\{j : (v_x)_j = 0, (v_{x\star})_j = 0\}|}{|\{j : (v_{x\star})_j = 0\}|}, \tag{18}$$

respectively, and for $v_y$, its TPR and TNR are defined similarly.

To estimate these metrics we run all the algorithms for 1000 iterations, well beyond their convergence times. For each algorithm, we plot the quality of the estimated $v_x$ and $v_y$ (measured by error, TPR and TNR) averaged across 100 datasets, and the results are shown in Figure 1. Note that for Rifle, we only plot its second stage, which has a better starting point (namely, the randomly perturbed ground-truth) as compared to the other two algorithms.

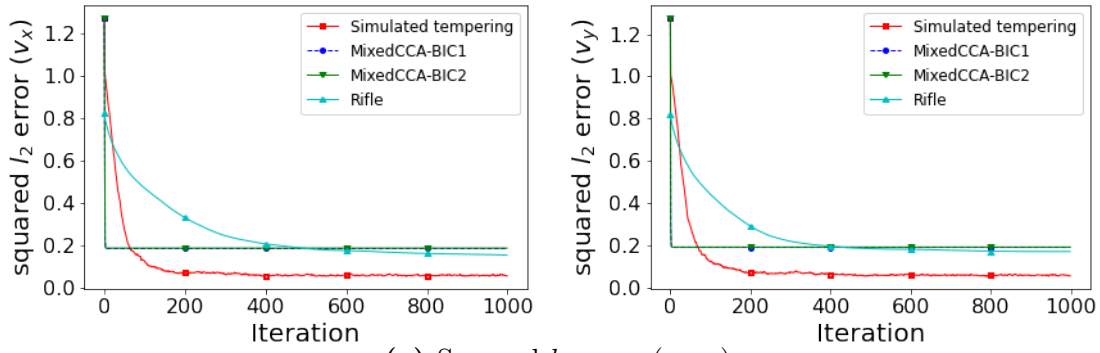
From both Figure 1 and Table 1, we see that our algorithm not only outperforms Rifle in terms of the quality of estimated $v_x$ and $v_y$ (across all the three metrics), but also enjoys much shorter running time. Compared with MixedCCA, although our algorithm has slightly longer computational time, the quality of estimated $v_x$ and $v_y$ from our algorithm is better, and the advantage is especially significant in terms of error and TPR.

### 4.2.3 Comparison in a mixed data setting

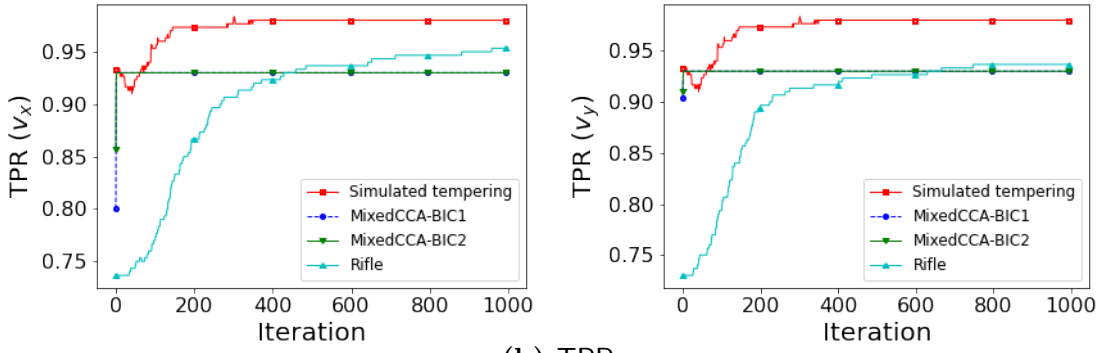
In many applications, particularly bio-medical ones, researchers often face the challenge that one of the variables $X$ or $Y$ is not observed directly, but only through its truncated or quantized version. Specifically, we consider the truncated latent Gaussian copula model of ([39]), which extends both the Gaussian copula model ([19]) and the latent Gaussian copula model ([9]).

**Definition 4** (Gaussian copula model)**.** A random vector $Z = (Z_1, \ldots, Z_p)^{\mathrm{T}}$ is a realization of the Gaussian copula model, if there exists a transformation $h : \mathbb{R}^p \to \mathbb{R}^p$ such that $h(Z) = (h_1(Z_1), \ldots, h_p(Z_p))^{\mathrm{T}} \sim \mathbf{N}(0, \Sigma)$ and for each $j = 1, \ldots, p$, transformation $h_j : \mathbb{R} \to \mathbb{R}$ is monotonically increasing. We write this as $Z \sim \mathbf{NPN}(0, \Sigma, h)$.

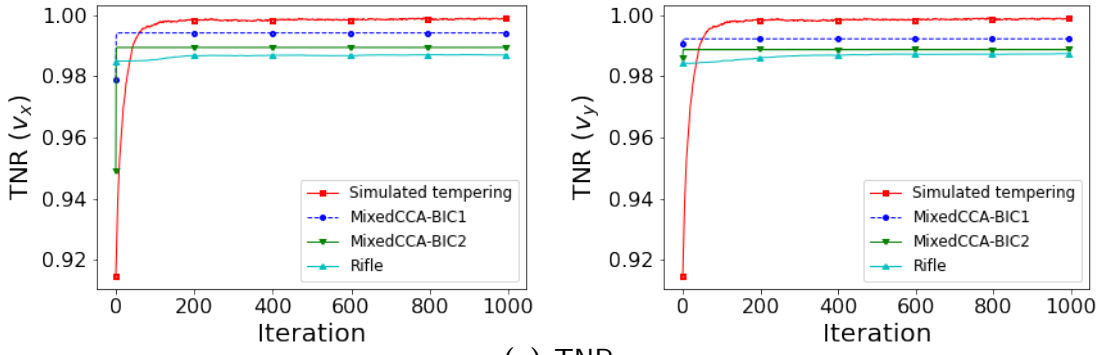
**Definition 5** (Truncated Gaussian copula model)**.** The random vector $(X^{\mathrm{T}}, Y^{\mathrm{T}})^{\mathrm{T}}$, where $X \in \mathbb{R}^{p_x}$ and $Y \in \mathbb{R}^{p_y}$, is a realization of a latent Gaussian copula model with truncation if




**(a)** Squared-$l_2$ error (error)




**(b)** TPR




**(c)** TNR

**Figure 1:** Comparison of the quality of estimated $v_x$ and $v_y$ by all the algorithms in terms of (a) squared-$l_2$ error (error), (b) TPR and (c) TNR. The results are averaged over 100 continuous datasets. To better compare TPR and TNR, we show the results starting from the first iteration, since the initial points are usually not sparse. The performances of mixedCCA-BIC1 and mixedCCA-BIC2 are indistinguishable on plots (a) and (b).

there exists a random vector $U \in \mathbb{R}^{p_y}$ such that $(X, U) \sim \mathbf{NPN}(0, \Sigma, h)$ and $Y_j = I(U_j > C_j)(U_j - C_j) + C_j$ for all $j = 1, \ldots, p_y$, where $C = (C_1, \ldots, C_{p_y})$ is a truncation parameter. We write $(X, Y) \sim \mathbf{TNPN}(0, \Sigma, h, C)$.

Taking $h$ as the identity map, suppose that we are interested in the sparse CCA of $(X, U) \sim \mathbf{N}(0, \Sigma)$, but we observe only independent copies of $(X, Y)$, where $Y_j = I(U_j > C_j)(U_j - C_j) + C_j$, for truncation levels $C = (C_1, \ldots, C_{p_y})$. Clearly, the classical Pearson sample covariance estimator cannot be used to estimate $\Sigma$. Nevertheless, building on ([9]), ([39]) showed that consistent estimators for $\Sigma_x$, $\Sigma_y$ and $\Sigma_{xy}$ can be constructed from independent replications of $(X, Y)$ using a Kendall's-tau covariance. Based on those estimates one can readily apply our Rayleigh quotient approach to obtain the sparse canonical correlation vectors of $\Sigma$. We compare our estimator with MixedCCA. In this mixed data setting, and unlike the continuous data setting, we found out that the two methods have comparable performances, with a slight advantage to our method in terms of statistical recovery, and a slight advantage to MixedCCA in terms of computational speed. We illustrate this below in a low sample size regime.

We randomly generated 100 mixed datasets in a similar way as in Section 4.2.2, except with an additional truncation step on the random vector $Y$. Specifically, we set the sample size $n = 180$ and the dimension $p = 200$, and let $\Sigma_x$ and $\Sigma_y$ each have five diagonal blocks of dimensions $\{d_1, ..., d_5\}$, respectively, and the $(j, j')$-th element of each block takes value $0.7^{|j-j'|}$. We set $\{d_1, ..., d_5\} = \{10, 20, 33, 20, 17\}$ for $\Sigma_x$ and $\{d_1, ..., d_5\} = \{33, 20, 25, 12, 10\}$ for $\Sigma_y$. In addition, we let $v_{x\star}$ and $v_{y\star}$ have the same structures as in Section 4.2.2 (so that the true density level $s_\star = 6$), and set $\lambda_1 = 0.8$. Let $\mathsf{truc}(\cdot; C)$ be the (elementwise) truncation operator at level $C$, such that given any vector $y$, $\mathsf{truc}(y; C)_j = y_j$ if $y_j > C$ and $\mathsf{truc}(y; C)_j = C$ otherwise. (In particular, we can recover the continuous data setting for

$C$ negatively large.) For each dataset, we generated $n$ samples from $(X, \mathsf{truc}(U; C))$, where $(X, U) \sim \mathbf{N}(0, \Sigma)$.

We ran Algorithm 2 with the set of temperatures $\{1, 1/0.9, 1/0.8, 1/0.7, 1/0.6\}$, that we compare with both mixedCCA-BIC1 and mixedCCA-BIC2 in terms of the running time and the statistical performances, as measured in Section 4.2.2. To evaluate the convergence times, we first run both algorithms for $N = 10,000$ iterations to obtain their respective"limit points".

The statistical performances of these algorithms (as measured by error, TPR and TNR) over the 100 mixed datasets generated as above are shown in Figure 2, and Figure 3 and Table 2. Because TPR and TNR are discrete values, we show the results of TPR and TNR in terms of mean and standard deviation. The computation times are recorded in Table 3. Due to the low sample size, both methods are prone to producing poor estimates that we consider as outliers. The boxplots in Figure 2 and Figure 3 report the distributions of $\mathsf{error}(v_x)$ and $\mathsf{error}(v_y)$ respectively, with and without these outliers (by removing the points outside of the whiskers of the boxplots).

In the low-truncation regime ($C = -2$) we recover the same conclusion as in the continuous data setting that our method outperforms mixedCCA. In the high-truncation setting ($C = 0$), our method still slightly outperforms mixedCCA, particularly in the recovery of $v_y$. The performance in terms of TPR and TNR are mostly similar, but again with a slight advantage to our method. However, here the computational time of our estimator is noticeably higher than mixedCCA as shown in Table 3.

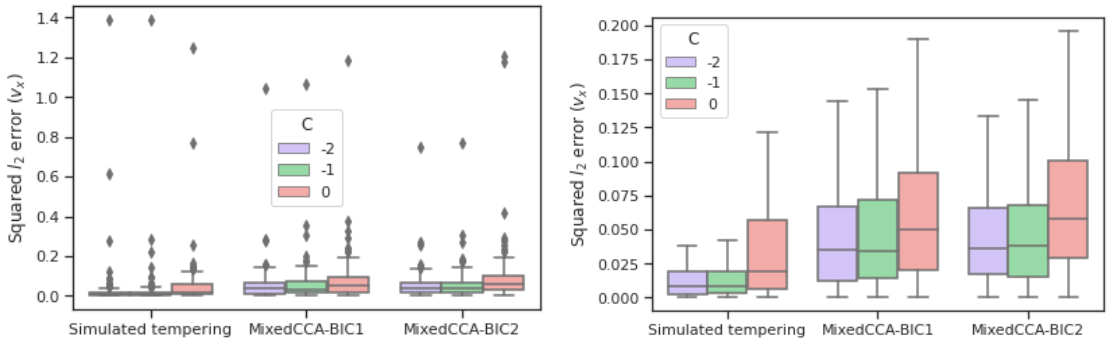



**Figure 2:** Squared-$l_2$ error (error) of estimated $v_x$ by all the algorithms for different truncation levels $C$, with outliers (Left) and without outliers (Right)

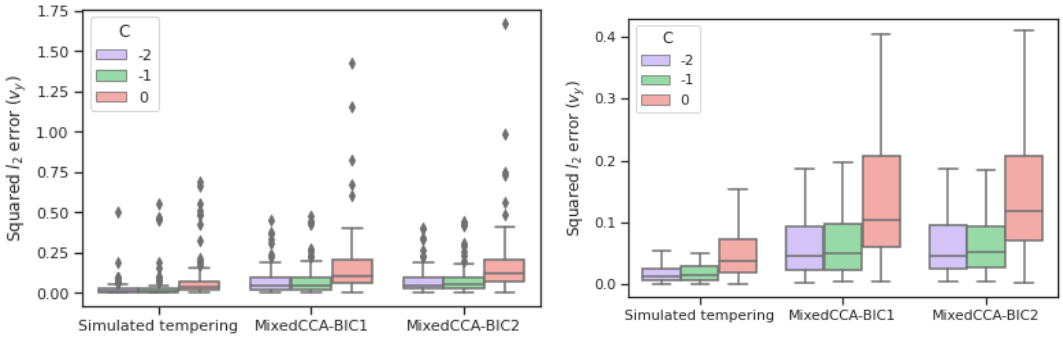



**Figure 3:** Squared-$l_2$ error (error) of estimated $v_y$ by all the algorithms for different truncation levels $C$, with outliers (Left) and without outliers (Right)

| TPR | $v_x$ | | | $v_y$ | | |
|---|---|---|---|---|---|---|
| $C$ (Truncation level) | -2 | -1 | 0 | -2 | -1 | 0 |
| Simulated tempering | 0.99 (0.07) | 0.99 (0.07) | **0.99** (0.07) | 1.00 (0.03) | 0.99 (0.07) | 0.98 (0.09) |
| MixedCCA-BIC1 | 1.00 (0.03) | 0.99 (0.07) | **0.99** (0.07) | 1.00 (0.03) | 0.99 (0.07) | **0.99** (0.07) |
| MixedCCA-BIC2 | **1.00** (0.00) | **1.00** (0.00) | **0.99** (0.07) | **1.00** (0.00) | **1.00** (0.00) | **0.99** (0.07) |

**(a)** TPR

| TNR | $v_x$ | | | $v_y$ | | |
|---|---|---|---|---|---|---|
| $C$ (Truncation level) | -2 | -1 | 0 | -2 | -1 | 0 |
| Simulated tempering | **1.00** (0.01) | **1.00** (0.01) | **0.99** (0.01) | **1.00** (0.00) | **1.00** (0.00) | **0.99** (0.01) |
| MixedCCA-BIC1 | 0.99 (0.01) | 0.99 (0.01) | 0.98 (0.01) | 0.98 (0.01) | 0.98 (0.01) | 0.98 (0.01) |
| MixedCCA-BIC2 | 0.98 (0.02) | 0.97 (0.02) | 0.96 (0.05) | 0.97 (0.02) | 0.97 (0.02) | 0.95 (0.08) |

**(b)** TNR

**Table 2:** Mean (and standard deviation) of TPR and TNR of our method and mixedCCA for different values of truncation level $C$.

# 5 Principal canonical correlation of clinical and proteomic data in Covid-19 patients

Covid-19 is an infectious disease that is rapidly sweeping through the world. The disease is caused by a severe acute respiratory syndrome coronavirus (SARS-CoV-2). There is currently an intense global effort to better understand the virus and find cures and vaccines. We use our methodology to re-analysis a data set produced by [8] that aims to identify biomarkers for early detection of severely ill Covid-19 patients[2]. To that end, the

[2]For reasons that are still poorly understood, about 80% of patients infected by SARS-CoV-2 experience mild to no symptoms, whereas in about 20% of the cases, patients become severely ill.

| Method | Computation Time (s) | | |
|---|---|---|---|
| $C$ (Truncation level) | -2 | -1 | 0 |
| Simulated tempering | 6.04 | 7.79 | 16.58 |
| MixedCCA-BIC1 | 1.54 | 0.81 | 2.28 |
| MixedCCA-BIC2 | 2.15 | 4.8 | 6.19 |

**Table 3:** The computation times of all algorithms averaged across 100 continuous datasets for different values of truncation level $C$.

study enrolled 86 patients (some non-Covid-19 patients, and among the Covid-19 patients, some that developed mild symptoms, and some that became severely ill). The exact protocol for recruiting these patients is unclear. For each patient they measured three (3) physical characteristics (sex, age, and body mass index), twelve (12) clinical variables as routinely measured from blood samples (white blood cells count, lymphocytes count, C-reactive protein, etc...). Furthermore, the serum of each patient is analyzed by liquid mass spectrometry-based proteomics to quantify their proteome and metabolome. In [8], the data is used to build a statistical model to predict whether or not a Covid-19 patient will progress to a severe state of illness. The dataset of [8] is freely available from the journal website.

We use canonical correlation analysis to re-analyze the data. A common working assumption is that SARS-CoV-2 induces patterns of molecular changes that can be detected in the sera of patients. Canonical correlation analysis may help identify these patterns. To do this we focus on the proteomic data, and we estimate the principal sparse canonical correlation between the physical and clinical variables on one hand and the proteomic variables on the other. See for instance [30] for a similar analysis on tuberculosis and malaria.

We pre-process the data by removing all the proteins for which 50% or more values are missing, leading to a total of $p_y = 513$ proteins, and $p_x = 16$ clinical and physical variables. The sample size $n = 86$. Liquid mass spectrometry-based proteomics typically produces a large quantity of missing values ([17, 25]). We make the assumption here that the missing values are driven mainly by detection limit truncation ([17]). We apply both our algorithm and mixedCCA to this problem, with the same parameter setting as in the simulation test on the mixed datasets (cf. Section 4.2.3). We run our algorithm for $N = 10,000$ iterations. Since we do not know the true canonical pair, we will focus on the estimated canonical correlation to measure the performance of two algorithms. In terms of the estimated canonical correlation, both our algorithm and mixedCCA takes less than 1 second to converge.

Our estimate of the principal canonical vectors of first dataset ($v_{x\star}$) has only one selected component (corresponding to C-reactive protein – CRP) with estimated inclusion probability of $\Pi(\delta_j = 1|\mathbf{Z}) = 0.99$. All other physical and clinical variables have inclusion probabilities smaller than 0.1. We found also that the principal canonical vectors of the proteomic data is also driven by a single protein (P02763, also known as Alpha-1-acid glycoprotein 1 or AGP 1), with estimated inclusion probability of $\Pi(\delta_j = 1|\mathbf{Z}) = 0.89$. All other proteins have inclusion probability smaller than 0.1. Fig. 4 shows the traceplot of the estimated canonical correlation $\hat{\rho}$ between the two data set, as well as the boxplot and autocorrelation function of the MCMC output (after burning in 3/4 of iterations) of the coefficients of CRP and AGP 1 in the quasi-posterior distribution. The fast decay of the autocorrelation functions show a good mixing of the MCMC sampler.

MixedCCA also selects CRP for the clinical dataset and AGP 1 for the proteomic dataset, but both BIC1 and BIC2 criterion select many other variables. mixedCCA-BIC1 also selects

glucose for clinical dataset and 3 other variables for the proteomic dataset, with estimated canonical correlation 0.90. mixedCCA-BIC1 selects 8 other variables for clinical dataset and 3 additional variables for the proteomic dataset with estimated canonical correlation 0.93. Although the estimated canonical correlation of mixedCCA is larger than the estimated canonical correlation (0.80) in our algorithm, the highly sparse nature of the estimated canonical vectors estimated from our method is striking.

Several studies have observed the predictive power of C-reative protein (CRP) in the progression of Covid-19 into a severe illness (see for instance [32] for a meta-analysis). This suggests that the correlation detected in our analysis between the two datasets is indeed driven by the progression of Covid-19 into a severe illness. Therefore, our analysis suggests that protein AGP 1 may also be playing an important role in the progression of Covid-19 into a severe illness. In Fig. 5, we present the boxplot of CRP and AGP 1 by group of patients. We can see that severe covid patients will have higher value of CRP and AGP 1, compared to non-covid and non-severe patients. We learn from Uniprot[3], that AGP 1 functions as transport protein in the blood stream, and appears to function in modulating the activity of the immune system during the acute-phase reaction. Furthermore, AGP 1 appears on the list of differentially expressed proteins in the sera of severely ill Covid-19 patients designed by [8], and also appeared in the literature as playing a role in the immune system's response to malaria ([10]).

# 6 Conclusion

In this work, we have developed a minimax optimal estimation procedure for sparse canonical correlation analysis using a quasi-Bayesian framework. Our method can be further

[3]https://www.uniprot.org

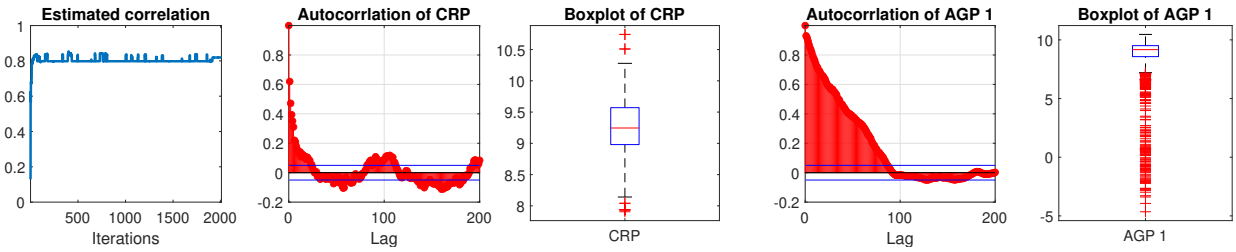


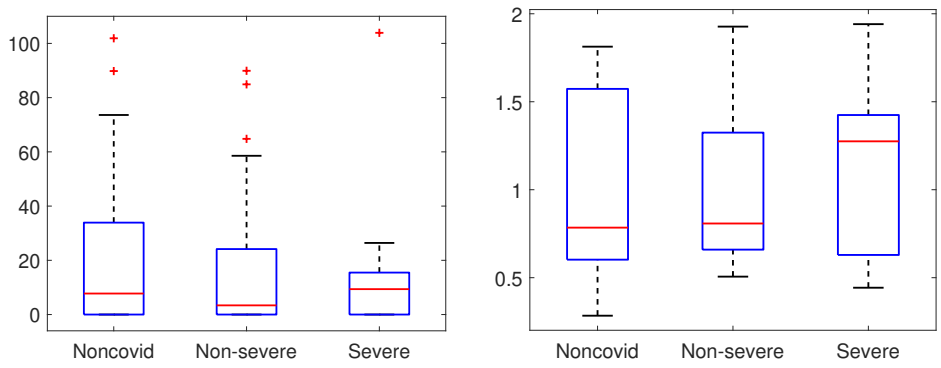

**Figure 4:** From left to right: The first plot is the trace plot of estimated canonical correlation; The second and third plot is the autocorrelation and boxplot of the coefficient of CRP from MCMC output; The fourth and fifth plot is the autocorrelation and boxplot of the coefficient of AGP 1 from MCMC output.

extended to capture more than one canonical vector, either by deflation, or by reformulating the problem as a higher dimensional canonical correlation analysis estimation problem as in [34]. Furthermore, one can straightforwardly extend our method to solve other generalized eigenvalue problems that arise in other statistical problems, as for instance in Fisher discriminant analysis. At a higher level, the method developed in this work can be viewed as a more statistical implementation of simulated annealing for optimization under sparsity constraints. As such, it can be applied more widely to solve non-convex optimization problems with sparsity constraints.

## SUPPLEMENTARY MATERIAL

**Proofs and technical details:** It contains the proofs of Theorem 2 and Proposition 3, as well as the description and investigation of the MCMC algorithms. (.pdf file)

**Conflict of Interest.** The authors report there are no competing interests to declare.




**(a)** CRP **(b)** AGP

**Figure 5:** Boxplots of (a) CRP and (b) AGP 1 by group of patients.

# References

[1] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *Proc. ICML*, pages 1247–1255, 2013.

[2] Yves Atchadé and Anwesha Bhattacharyya. An approach to large-scale quasi-bayesian inference with spike-and-slab priors, 2019.

[3] Dimitris Bertsimas and John Tsitsiklis. Simulated Annealing. *Stat. Sci.*, 8(1):10 – 15, 1993.

[4] P. G. Bissiri, C. C. Holmes, and S. G. Walker. A general framework for updating belief distributions. *J. R. Stat. Soc. Ser. B*, 78(5):1103–1130, 2016.

[5] B. P. Carlin and S. Chib. Bayesian model choice via markov chain monte carlo methods. *J. Roy. Stat. Soc. B*, 57(3):473–484, 1995.

[6] Olivier Catoni. *Statistical learning theory and stochastic optimization*, volume 1851 of

*Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2004. Lecture notes from the 31st Summer School on Probability Theory held in Saint-Flour, July 8–25, 2001.

[7] Victor Chernozhukov and Han Hong. An MCMC approach to classical estimation. *J. Econometrics*, 115(2):293–346, 2003.

[8] Bo Shen et al. Proteomic and metabolomic characterization of covid-19 patient sera. *Cell*, 182(1):59 – 72.e15, 2020.

[9] Jianqing Fan, Han Liu, Yang Ning, and Hui Zou. High dimensional semiparametric latent graphical model for mixed data. *J. R. Stat. Soc. Ser. B*, 79(2):405–421, 2017.

[10] M J Friedman. Control of malaria virulence by alpha 1-acid glycoprotein (orosomucoid), an acute-phase (inflammatory) reactant. *Proceedings of the National Academy of Sciences*, 80(17):5421–5424, 1983.

[11] Chao Gao, Zongming Ma, Zhao Ren, and Harrison H. Zhou. Minimax estimation in sparse canonical correlation analysis. *Ann. Statist.*, 43(5):2168–2197, 10 2015.

[12] Chao Gao, Zongming Ma, and Harrison H. Zhou. Sparse cca: Adaptive estimation and computational barriers. *Ann. Statist.*, 45(5):2074–2101, 10 2017.

[13] Edward I. George and Robert E. McCulloch. Approaches for bayesian variable selection. *Statistica Sinica*, 7(2):339–373, 1997.

[14] Charles J. Geyer and Elizabeth A. Thompson. Annealing markov chain monte carlo with applications to ancestral inference. *J. Amer. Stat. Assoc.*, 90(431):909–920, 1995.

[15] David Hardoon and John Shawe-Taylor. Sparse canonical correlation analysis. *Machine Learning*, 83:331–353, 06 2011.

[16] H Hotelling. Relations between two sets of variates. *Biometrika*, 1936.

[17] Yuliya V. Karpievitch, Ashoka D. Polpitiya, Gordon A. Anderson, Richard D. Smith, and Alan R. Dabney. Liquid chromatography mass spectrometry-based proteomics: Biological and technological aspects. *Ann. Appl. Stat.*, 4(4):1797–1823, 12 2010.

[18] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by Simulated Annealing. *Sci.*, 220(4598):671–680, May 1983.

[19] Han Liu, John Lafferty, and Larry Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, 10, 04 2009.

[20] Jun S Liu. *Monte Carlo strategies in scientific computing*. Springer Science & Business Media, 2008.

[21] Kantilal Varichand Mardia, John T. Kent, and John M. Bibby. *Multivariate analysis*. Probability and mathematical statistics. Acad. Press, London [u.a.], 1979.

[22] David A. McAllester. Some pac-bayesian theorems. *Machine Learning*, 37(3):355–363, 1999.

[23] T. J. Mitchell and J. J. Beauchamp. Bayesian variable selection in linear regression. *J. Amer. Stat. Assoc.*, 83(404):1023–1032, 1988.

[24] Qianxing Mo, Ronglai Shen, Cui Guo, Marina Vannucci, Keith S Chan, and Susan G Hilsenbeck. A fully bayesian latent variable model for integrative clustering analysis of multi-type omics data. *Biostatistics*, 19(1):71–86, 2017.

[25] Jonathon J. O'Brien, Harsha P. Gunawardena, Joao A. Paulo, Xian Chen, Joseph G. Ibrahim, Steven P. Gygi, and Bahjat F. Qaqish. The effects of nonignorable missing data on label-free mass spectrometry proteomics experiments. *Ann. Appl. Stat.*, 12(4):2075–2095, 12 2018.

[26] Elena Parkhomenko, David Tritchler, and Joseph Beyene. Sparse canonical correlation analysis with application to genomic data integration. *Statistical applications in genetics and molecular biology*, 8:Article 1, 2009.

[27] Nimrod Rappoport and Ron Shamir. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic acids research*, 46(20):10546–10562, 2018.

[28] Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated gaussian designs. *J. Mach. Learn. Res.*, 11:2241–2259, 2010.

[29] Christian P. Robert. *The Bayesian choice*. Springer-Verlag, New York, second edition, 2001.

[30] Juho Rousu, Daniel D. Agranoff, Olugbemiro Sodeinde, John Shawe-Taylor, and Delmiro Fernandez-Reyes. Biomarker discovery by sparse canonical correlation analysis of complex clinical phenotypes of tuberculosis and malaria. *PLOS Computational Biology*, 9(4):1–10, 04 2013.

[31] Mark Rudelson and Shuheng Zhou. Reconstruction from anisotropic random measurements. *IEEE Trans. Inf. Theor.*, 59(6):3434–3447, June 2013.

[32] Bikash R. Sahu, Raj Kishor Kampa, Archana Padhi, and Aditya K. Panda. C-reactive protein: A promising biomarker for poor prognosis in covid-19 infection. *Clinica Chimica Acta*, 509:91 – 94, 2020.

[33] Benjamin A. Shaby. The open-faced sandwich adjustment for mcmc using estimating functions. *Journal of Computational and Graphical Statistics*, 23(3):853–876, 10 2014.

[34] Kean Ming Tan, Zhaoran Wang, Han Liu, and Tong Zhang. Sparse generalized eigenvalue problem: Optimal statistical rates via truncated rayleigh flow. *J. R. Stat. Soc. Ser. B*, 80(5):1057–1086, 2018.

[35] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

[36] Sandra Waaijenborg and Aeilko Zwinderman. Sparse canonical correlation analysis for identifying, connecting and completing gene-expression networks. *BMC bioinformatics*, 10:315, 09 2009.

[37] Ami Wiesel, Mark Kliger, and Alfred Hero. A greedy approach to sparse canonical correlation analysis. 02 2008.

[38] Daniela M Witten and Robert J Tibshirani. Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical applications in genetics and molecular biology*, 8(1):1–27, 2009.

[39] Grace Yoon, Raymond J. Carroll, and Irina Gaynanova. Sparse semiparametric canonical correlation analysis for data of mixed types. *arXiv: Methodology*, 2018.