# A Mathematical Programming Approach to Sparse Canonical Correlation Analysis

Lavinia Amorosi[a,*], Tullia Padellini[b]

[a]*Department of Statistical Sciences, Sapienza University of Rome, Italy*
[b]*School of Public Health, Imperial College London*

**Abstract**

Recent developments in the interplay between Operational Research and Statistics allowed to exploit advances in Mixed Integer Optimization (MIO) solvers to improve the quality of statistical analysis. In this work we tackle Canonical Correlation Analysis (CCA), a dimensionality reduction method that summarises multiple data sources jointly, retaining their dependency structure. We propose a new technique to encode Sparsity in CCA by means of a new mathematical programming formulation which allows to obtain an exact solution using readily available solvers (like Gurobi). We show a preliminary investigation of the performance of our proposal through a simulation study, which highlights the potential of our approach.

*Keywords:* Canonical Correlation Analysis, Mixed Integer Optimization, Sparsity

## 1. Introduction

Despite the great deal of overlap between Operational Research (OR) algorithms and Data Science (DS) methods, the interaction between the two fields is still under-exploited. In recent years there have been more efforts on both sides to borrow from each others literature, bridging the gap between the two disciplines.

*Corresponding author
*Email addresses:* lavinia.amorosi@uniroma1.it (Lavinia Amorosi), t.padellini@imperial.ac.uk (Tullia Padellini)

From the OR side, Machine Learning (ML) techniques have been used to improve solver performances. Among the contributions in this direction we recall as examples [1], which recasts the solver's algorithm selection for Mixed-Integer Quadratic Programming (MIQP) as a classification procedure; [2], which adopts supervised learning to predict if a Mixed-Integer Programming (MIP) instance will be solved within a given time limit; [3], which provides a survey on ML approaches to variable and node selection in Branch-and-Bound algorithms.

On the other side, the one this work falls in, OR methods have been instrumental in both developing and improving fitting procedures of statistical models. Among the most established strategies to improve statistical fitting procedure by means of advances in the OR literature, we recall [4] where a mixed integer quadratic optimization (MIQO) approach has been adopted for designing high quality linear regression models balancing many competing desired properties; [5] which presents a novel Mixed-Integer Linear Programming (MILP) formulation of the optimal classification tree problem or [6] that proposes a Branch-and-Bound algorithm, based on principles of Mixed-Integer Optimization (MIO), to solve the Sparse Principal Component Analysis (SPCA) to optimality, just to mention a few. Indeed, improvements in hardware and in optimization methods in the last 30 years produced impressive advances in MIO solver performances. Consequently, several statistical problems that have a natural MIO formulation considered intractable in the past, now can be solved to optimality.

In this work, inspired by [4], we adopt a mathematical programming approach to recast sparsity in Canonical Correlation Analysis (CCA), a dimensionality reduction method that allows to summarise multiple sources of data simultaneously while retaining their dependency structure. We propose a new fitting procedure which, exploiting advances of MIO, provides an exact solution for Sparse CCA, a problem which, to the best of our knowledge is only approximated in the DS literature. More in details, we implement our proposal in Gurobi 9.0, whose latest update includes a bilinear solver for problems with quadratic constraints. This new solver has the noticeable advantage of being able to find a global optimum, which guarantees the exactness of the solution

to our problem. The remainder of this paper is organized as follows. Section 2 reviews the OR literature on sparsity for different statistical methods and the general MIO framework in which the sparsity can be embedded. Section 3 first recalls the CCA technique and the meaning of sparsity in this context and then presents the proposed MIQO formulation. Section 4 reports the results obtained testing the model on simulated data sets.

## 2. State of the Art

Nowadays large quantities of data, related to every field, like in marketing, social networks, telecommunications, medicine and health care and others, are available. However, in order to extract useful knowledge from them, it is necessary to summarise such huge amount of information and discard all non-informative elements. To this end the concept of sparsity is very important in statistical models. Indeed, sparsity is the property which guarantees that only a reduced number of parameters (or predictors) play an important role in the statistical model under consideration, thus greatly improving its interpretability. It is therefore no coincidence that in the research area that studies the use of advances in operational research techniques to improve the performance of statistical models, sparsity was one of the first properties to be analyzed under a "modern optimization lens".

In this section we focus on the articles related to sparsity embedded in optimization approaches proposed to improve the quality of several statistical models and on the main contributions of this work. After the seminal work [4], by Bertsimas and King, on MIQO formulations of linear regression, in [6] a MIO approach to sparse principal component analysis has been presented. The authors proposed a Branch-and-Bound algorithm able to prove optimality or to find high quality solutions in seconds, depending on the sample size, explaining a higher portion of variance with respect to other existing methods in literature. In [7] the sparse regression problem has been reformulated as a pure integer convex optimization problem. A cutting plane algorithm to solve it on instances with

number of samples and regressors in the 100000s is presented. Moreover, the authors observed that the sparse regression problem has the property that, as the sample size ($n$) increases, the problem becomes easier in the resolution perfectly recovering the support of the true signal, faster than LASSO, whereas for small $n$ values, their approach takes a large amount of time to solve the problem. Similarly, in [8] the sparse principal component analysis model has also been reformulated as a convex mixed-integer semidefinite optimization problem. The authors designed a cutting-plane method which solves to optimality instances with 10 selected covariates from 300 variables and provides solutions with small gap for larger scale instances. Moreover, they proposed two convex relaxations and randomized rounding schemes to obtain feasible solutions of high quality (very close to the optimal one) within minutes for a number of variables p = 100s or hours for p = 1000s. In [9] the authors formulated the cardinality constrained maximum likelihood problem for the sparse inverse covariance matrix as a MIO model and proposed a combination of outer-approximation algorithm and first-order methods to solve it. Their approach provides near optimal solutions fast, and a guarantee on the solutions suboptimality if the method is terminated early. It delivers near optimal solutions in a matter of seconds, and provably optimal solutions in a matter of minutes for p in the 100s and k in the 10s. The algorithm also provides high-quality solutions to problems in the 1000s, but a certificate of optimality is more computationally expensive for those sizes. In [10] Blanquero et al. studied sparse optimal randomized classification trees adopting a new optimization approach based on the one presented in [11] with oblique cuts. This is a continuous optimization model which is able to find a trade-off between accuracy and global sparsity. Indeed, the problem of building optimal decision trees is NP-complete and for this reason in literature greedy procedures have been proposed in which at each branch node of the tree, some purity criterion is (locally) optimized like CARTs. However, these approaches provide good local sparsity making classic decision trees locally easy to interpret but this is not true at a global level. Mathematical optimization is again the key to build optimal, deterministic and randomized classification

trees and address issues like global sparsity as shown in [10]. Another recent example to be mentioned among the works in literature focusing on sparsity in statistical models dealt with an optimization approach is [12]. It focuses on the sparse polynomial regression problem, also named sparse hierarchical regression problem, which consists in determining a polynomial of degree r that depends on at most k inputs counting at most l monomial terms minimizing the sum of squares of its prediction errors. For this problem the authors presented a two-step approach: first a fast input ranking heuristic discards the irrelevant inputs, then a cutting plane method solves the integer nonlinear optimization model used to formulate the remaining reduced sparse hierarchical regression problem. The experimental results show that the proposed method is able to deal with problems of practical size (n=10000 and p=1000), for which its computational complexity is on par with LASSO, outperforming heuristic methods in both finding all relevant non-linearities as well as rejecting obfuscating ones. In this paper, we study another statistical model, Canonical Correlation Analysis, for which we embed sparsity by adopting a mathematical programming approach. To the best of our knowledge this is still an uncovered topic by the existing literature related to OR advances application for improving fitting and quality of statistical methods. The main contributions of this work are a new MIO formulation for Sparse Canonical Correlation Analysis and experimental results based on simulated dataset showing that the proposed approach is able to discern between informative and non informative variables.

### *2.1. Mixed Integer Optimization and Sparsity*

In this section we give a glimpse of mixed integer optimization (MIO) and of its use in modelling sparsity.

The general formulation of a MIO problem is as follows:

$$\min x^T Q x + q^T x \tag{1}$$

s.t.

$$Ax = b \tag{2}$$

$$l \leq x \leq u \tag{3}$$

$$x^T Q_t x + q_t^T x \leq b_t \quad t = 1, .., p \tag{4}$$

$$x_i \in \{0, 1\} \text{ or } \in Z \quad \forall i \in J \tag{5}$$

$$x_i \in R_+ \quad \forall i \notin J \tag{6}$$

where $q \in R^m$, $b \in R^k$, $Q \in R^{mxm}$ (positive semidefinite), $Q_t \in R^{mxm}$, $q_t \in R^m$ and $b_t \in R$.

MIO models with a quadratic objective but without quadratic constraints are called Mixed Integer Quadratic Optimization (MIQO) problems. MIO models with quadratic constraints are called Mixed Integer Quadratically Constrained Optimization (MIQCO) problems. Models without any quadratic features and linear objective and constraints are referred to as Mixed Integer Linear Optimization (MILO) problems. This framework is flexible enough to include a vast class of statistical problems, typically characterized by different matrices $Q$, as well as different penalisations, voted to enforce some desirable statistical properties. We focus in particular on the case of sparsity, that is when we require some of the variables to be set to 0, following the approach defined by Bertsimas. According to [13], the notion of sparsity can in fact be formalized by considering the variable $x = (y, z)$ by composed of two subsets of equal dimension. The first one $y = (y_1, \ldots, y_p) \in R^p$ represents the statistical parameters of interest, while the second one $z = (z_1, \ldots, z_p) \in \{0, 1\}^p$ can be taken as auxiliary variables, which allow to identify which elements of the first subset are different than 0. Under these assumptions, the problem (1)-(6) can be rewritten to enforce statistical sparsity as follows:

$$\min y^T Q y + q^T y \tag{7}$$

s.t.

$$Iz = k \tag{8}$$

$$Ay = b \tag{9}$$

$$-Mz < y < Mz \tag{10}$$

$$x^T Q_t x + q_t^T x \leq b_t \quad t = 1, .., p \tag{11}$$

The presence of sparsity is guaranteed by constraints (10), which imply some of the elements of $y$ are automatically set to zero, while the amount of sparsity is regulated by constraint (8).

In the past decades, because of the difficulty in scaling solution algorithms for MIO problems, the research to deal with statistical sparsity focused on methods which solve convex approximation of the original problem. We now briefly recall the most famous example of this approach, the LASSO [14], in which a penalty based on the $L_1$ norm is used to regularize the standard least square fit of a regression problem.

Given a sample of $n$ observations, the standard linear regression model can be written as

$$y = X\beta + \varepsilon$$

where $y = (y_1, \ldots, y_n)$ is the response (or *dependent*) variable, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ is the stochastic error term and $X$ is $(n \times p)$ the matrix containing the values of $p$ independent variables (also known as *covariates*, or *predictors*) observed on the sample and $\beta = (\beta_1, \ldots, \beta_p)$ their corresponding coefficients. The standard approach to fit this model is to minimize the error term, typically l2.

$$\min_{\beta} \quad (y - X\beta)^T (y - X\beta) \tag{12}$$

Sparsity is enforced in the coefficients $\beta$ by adding the following constraint

$$\sum_{i=1}^{p} I\{\beta_i \neq 0\} \leq k \tag{13}$$

The rationale for inducing sparsity in this kind of model is, in fact, reducing the number of predictors by assessing which independent variables are relevant for recovering the dependent one. Each element $\beta_i$ of the coefficient $\beta$ represents the impact of the $i-th$ covariate on the response $y$. If $\beta_i$ is 0, the corresponding covariate does not play any role in explaining the dependent variable $y$. Solving

12-13 was not considered to be approachable directly, as it is a NP hard problem. The basic idea of the LASSO thus consisted in approximating it by

$$\min_{\beta}(y - X\beta)^T(y - X\beta) \tag{14}$$

s.t.

$$\sum_{i=1}^{p} |\beta_i| \leq k \tag{15}$$

While the LASSO rightfully enjoys popularity in the statistical community due to its good predictive properties, and depending on the signal-to-noise ratio in the data, it may even outperform an exact solution of 12-13 [15], it still provides only an approximated solution to the original problem of sparsity. However, [13] showed that it is possible to obtain an exact solution to the problem 12-13 by adding a binary variable $z$ and reformulating it as a MI(Q)O problem:

$$\min_{\beta,z}(y - X\beta)^T(y - X\beta) \tag{16}$$

s.t.

$$-Cz_i \leq \beta_i \leq Cz_i \qquad i = 1, \ldots, p \tag{17}$$

$$z_i \in \{0, 1\} \qquad i = 1, \ldots, p \tag{18}$$

$$\sum_{i=1}^{p} z_i \leq k \tag{19}$$

where the parameters $C$ and $k$ represent respectively the maximum value coefficients $\beta_i$ can take and the maximum number of non zero $\beta$ coefficients. Variable $z_i$ is equal to 0 when the corresponding coefficient $\beta_i$ is 0 and 1 otherwise. Constraint 17 insures that if $z_i$ is 0, $\beta_i$ is 0 as well. It is easy to see that (16)-(19) is a special case of the framework of (7)-(11). In the last 30 years the computational power of MIO solvers increased together with the speed improvements in hardware (supercomputers). Consequently, the use of MIO approaches for statistical problems is no longer practically irrelevant. Moreover, MIO solvers provide lower bound improvements during the resolution process. This means that, even if the MIO solver is stopped before reaching the optimal solution,

it provides a certificate of suboptimality of the current feasible solution. Thus, in Section 3.1 we present the bilinear quadratically constrained mixed integer optimization model which incorporates the sparsity property for CCA and, in Section 4, the experimental results obtained solving it on simulated data through Gurobi 9.0.

## 3. A MIQO formulation for Canonical Correlation Analysis

Canonical Correlation Analysis (CCA) is a multivariate statistical technique designed to investigate the relationship between two sets of variables. The basic framework of CCA is that where we have two datasets $X_1 \in R^{n \times p_1}$ and $X_2 \in R^{n \times p_2}$, taken to be i.i.d. realization of two random vectors $x_1 \in R^{p_1}, x_2 \in R^{p_2}$. Each component of these vectors, and therefore each column of the datasets, represents a different source of information or attribute. Example of interest in CCA are the cases of genomic data [16], microbiome data [17] and EEG [18]. In its original formulation, given two scaled datasets $X_1$ and $X_2$, CCA's goal is to find a lower dimensional linear representation $X_1 w_1 = Y_1$ and $X_2 w_2 = Y_2$ so that the resulting transformed variables $Y_1 \in R^n$ and $Y_2 \in R^n$ are maximally correlated. After imposing that $X_1$ and $X_2$ are centered, i.e. they have columnwise 0 mean, the empirical correlation between the two new variables $Y_1$ and $Y_2$ can be expressed as

$$\rho_{1,2} = \frac{w_1^t X_1^t X_2 w_2}{\sqrt{w_1^t X_1^t X_1 w_1}\sqrt{w_2^t X_2^t X_2 w_2}}$$

The coefficients $w_1 \in R^{p_1}$ and $w_2 \in R^{p_2}$ that guarantee that the correlation between $Y_1$ and $Y_2$ is maximized, can thus be found by solving the following optimization problem

$$\max_{w_1, w_2} \frac{w_1^t X_1^t X_2 w_2}{\sqrt{w_1^t X_1^t X_1 w_1}\sqrt{w_2^t X_2^t X_2 w_2}} \tag{20}$$

s.t.

$$w_1^t X_1^t X_1 w_1 = w_2^t X_2^t X_2 w_2 = 1 \tag{21}$$

where the constraints 21 are imposed to ensure identifiability, since solutions to problem 20 are invariant with respect to orthogonal rotation. Notice that, as a consequence of these constraints, formulation 20-21 can be simplified as:

$$\max_{w_1,w_2} w_1^t X_1^t X_2 w_2 \tag{22}$$

s.t.

$$w_1^t X_1^t X_1 w_1 = w_2^t X_2^t X_2 w_2 = 1 \tag{23}$$

The new variables $Y_1$ and $Y_2$, usually called *Canonical Variables*, provide a vector representation of the original matrices $X_1$ and $X_2$, and can be used as lower dimensional proxies. The *Canonical Vectors* (or *Weights*) $w_1$ and $w_2$ assess the importance of each column component of respectively $X_1$ and $X_2$ in explaining the association with the other dataset, and are thus instrumental for the interpretation of the canonical variables and for untangling the relationship between $X_1$ and $X_2$. Elements of $w_1$ and $w_2$ that are large (in absolute values) indicate which columns, i.e. attributes, are highly relevant in explaining the linear association between $X_1$ and $X_2$. On the other hand, values of $w_1$ and $w_2$ that are close to 0, identify attributes that can be neglected for explaining the dependence between the two sets.

In the context of CCA, inducing sparsity consists of forcing some elements of the coefficients $w_1$ and $w_2$ to be 0. Considering only a subset of the original attributes in our analysis, thus constraining the number of non-zero coefficients to be smaller than a maximum acceptable value, can lead to more informative Canonical Weights and Variables. This especially true when $p_1$ and $p_2$ become large, and the interpretability of the weights, as well as the Canonical Variables, becomes harder.

As in the case of the LASSO for regression problems, sparsity has most often be induced by adding a regularization term based on the $L_1$ norm in CCA as well. The most common formulation for Sparse CCA in the statistical and machine learning literature is thus the following:

$$\max_{w_1,w_2} w_1^t X_1^t X_2 w_2 \tag{24}$$

s.t.

$$w_1^t X_1^t X_1 w_1 = w_2^t X_2^t X_2 w_2 = 1 \tag{25}$$

$$\sum_{i=1}^{p_j} |w_{j,i}| \leq k_j \tag{26}$$

Several algorithms have been introduced to efficiently solve this convex optimization problem. For example [19] propose an approach based on a penalized matrix decomposition and the resulting regularized version of the singular value decomposition. In [20], the solution to the problem 24-26 is obtained by means of a linearized Bregman iterative method. Similar problems to 24-26 is also considered in [21] and more recently in [22] and in [23]. In a Bayesian fashion, probabilistic approaches to sparse CCA have also been considered [24]. Methods that tackle sparsity by means of $L_0$ loss, and are thus more similar in spirit to our work, are far less common in the literature. Among these, [25] introduces a provable algorithm for estimating canonical vector with an $L_0$ penalization, but it is limited by the assumption that each of the two data objects $X_1$ and $X_2$, although dependent from the other object, consists of independent variables. [26] introduces a two stage procedure to solve sparse CCA, and provides an algorithm to identify active entries of the $L_0$-penalized canonical vectors, i.e. non-zero elements of $w_1$ and $w_2$. Values of these active entries must be then computed through a separate procedure. Finally, [27] suggests an alternating iterative algorithm to solve the $L_0$ sparse CCA formulation by using a sparse projection strategy.

### *3.1. The MIQO formulation*

In order to obtain an exact solution to Sparse Canonical Correlation, we recast problem 24-26 as a MIQO, adding auxiliary binary variables in the fashion of [13]. The resulting mathematical programming formulation is as follows:

$$\max_{w_1, w_2, z_1, z_2} w_1^t X_1^t X_2 w_2 \tag{27}$$

s.t.

$$w_1^t X_1^t X_1 w_1 = w_2^t X_2^t X_2 w_2 = 1 \tag{28}$$

$$-C_j z_{j,i} \leq w_{j,i} \leq C_j z_{j,i} \qquad i = 1, \ldots, p_j \;\; j = 1, 2 \tag{29}$$

$$z_{j,i} \in \{0, 1\} \qquad i = 1, \ldots, p_j \;\; j = 1, 2 \tag{30}$$

$$\sum_{i=1}^{p_j} z_{j,i} \leq k_j \qquad j = 1, 2 \tag{31}$$

where $w_1 \in R^{p_1}$ and $w_2 \in R^{p_2}$ are sets of continuous variables bounded by $[-C_1, C_1]$ and $[-C_2, C_2]$ respectively and $z_1 \in \{0, 1\}^{p_1}$ and $z_2 \in \{0, 1\}^{p_2}$ are sets of binary variables representing which elements of $w_1$ and $w_2$ are non-zero. Parameters $C_1$ and $C_2$ are positive constants defining the range of variation of $w_1$ and $w_2$, while $k_1$ and $k_2$ represent the maximum number of non-zero components of $w_1$ and $w_2$. Since the objective function (27) is bilinear in $w_1$ and $w_2$ and because of constraints 28, the formulation 27-31 is a bilinear quadratically constrained model and can be solved through the adoption of the most recent Gurobi module for this class of problems.

## 4. Experimental Results & Discussion

We evaluate the performance of our proposal on simulated data. Following [19] we generate data according to the model

$$X_i = u w_i^t + \varepsilon_i \qquad i = 1, 2$$

where each component of the error matrix $\varepsilon_i$ follows a Normal distribution with mean 0 and standard deviation 0.1. We consider 2 different scenarios w.r.t the number of variables $p_1$ and $p_2$ and test them at different levels of sparsity: a *standard* one and *extreme* one. Details are given in Table 1.

We generate the $u$ vector from a $n$ variate normal distribution, with 0 mean vector and identity and variance-covariance matrix. We also consider 3 different values for the sample size $n$: $50, 150, 500$.

Even though still very limited, results shown in Figures 1-4 highlight the potential of our proposal. The sparsity pattern appears to be captured well in almost

| | Scenario 1 | Scenario 2 |
|---|---|---|
| | $p_1 = 5\ p_2 = 3$ | $p_1 = 20\ p_2 = 9$ |
| Sparse | $k_1 = 3\ k_2 = 2$ | $k_1 = 4\ k_2 = 3$ |
| Very Sparse | $k_1 = 1\ k_2 = 1$ | $k_1 = 1\ k_2 = 1$ |

Table 1: Dimension and sparsity of the Canonical Vectors to be recovered in the simulation study.

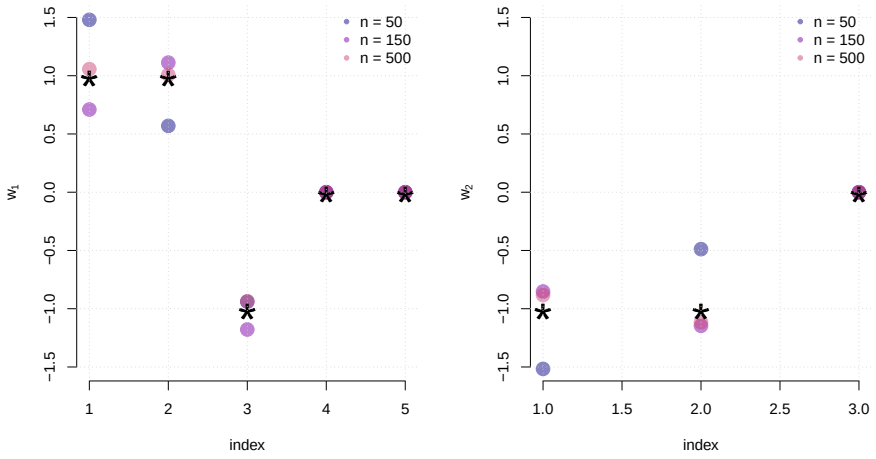


Figure 1: Estimated canonical weights for Scenario 1 - Sparse setting. Black stars represent the true value of the coefficients, while colored dots represent estimates

all settings. Moreover the direction of the association, represented by the sign of the Canonical Vectors, is always correctly identified.

Interestingly, the model performances seem to be affected by the sample size more than the sparsity level, and they are reassuring even in presence of extreme sparsity. Figure 4 shows in fact how the model seems to be able to deal with a sparsity level of more than 90%. As we could expect, the estimated values, denoted by the coloured circles, become closer to the real values, indicated by the black star, as the sample size $n$ increases. This is because when the sample size is small, the empirical covariance $X_1^t X_2$ in the objective function (27) is

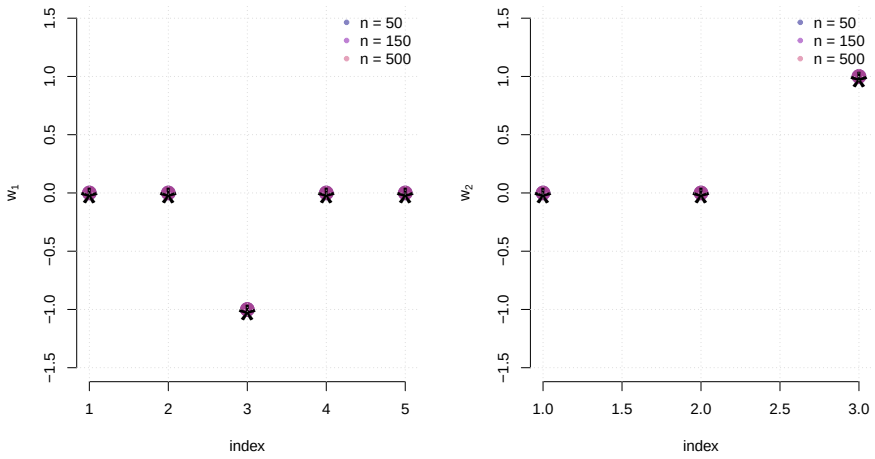


Figure 2: Estimated canonical weights for Scenario 1 - Very Sparse setting. Black stars represent the true value of the coefficients, while colored dots represent estimates

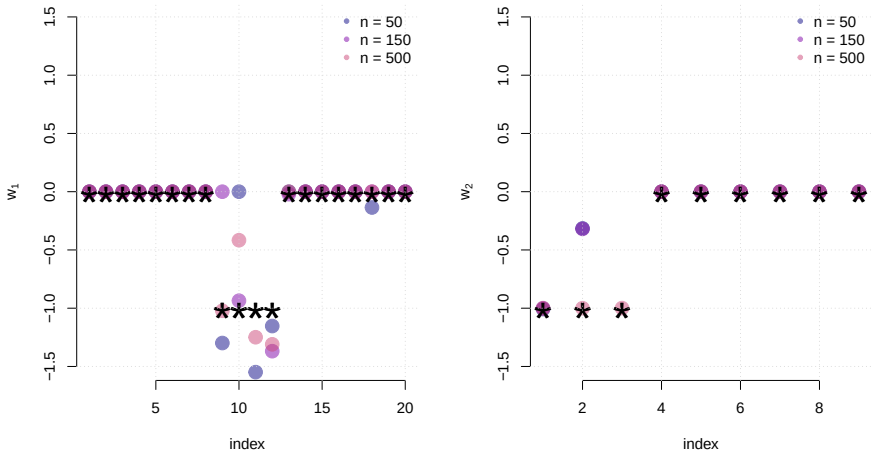


Figure 3: Estimated canonical weights for Scenario 2 - Sparse setting. Black stars represent the true value of the coefficients, while colored dots represent estimates

a less reliable measure of dependence between the random variables we are trying to summarize. As the sample size increases, $X_1^t X_2$ better captures the

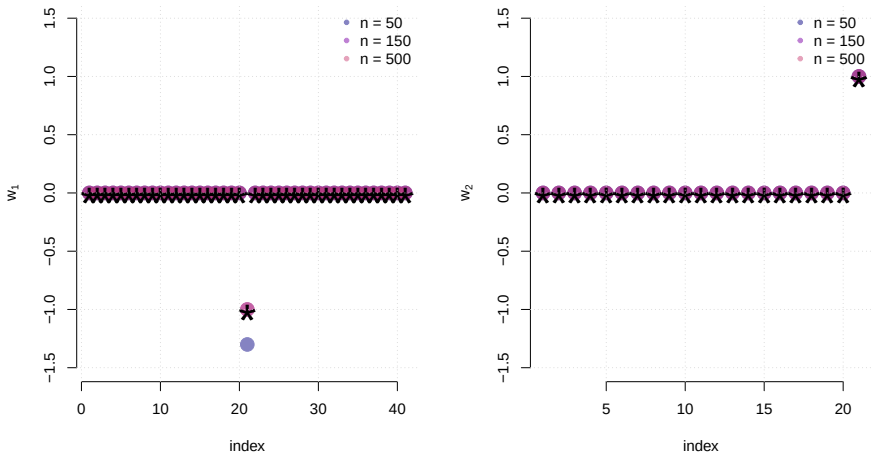


Figure 4: Estimated canonical weights for Scenario 2 - Very Sparse setting. Black stars represent the true value of the coefficients, while colored dots represent estimates

association between the random mechanisms generating the different datasets, and the Canonical Vectors obtained by maximising (27) are less affected by spurious association created by the Gaussian noise.

In conclusion, additional investigation is needed, but these preliminary results suggest that our approach is promising and encourage us to further explore it in more challenging settings as well as real data examples.

## References

[1] P. Bonami, A. Lodi, G. Zarpellon, Learning a classification of mixed-integer quadratic programming problems, Integration of Constraint Programming, Artificial Intelligence, and Operations Research. CPAIOR 2018. Lecture Notes in Computer Science 10848.

[2] M. Fischetti, A. Lodi, G. Zarpellon, Learning milp resolution outcomes before reaching time-limit, International Conference on Integration of Con-

straint Programming, Artificial Intelligence, and Operations Research. CPAIOR 2019. Lecture Notes in Computer Science 11494.

[3] A. Lodi, G. Zarpellon, On learning and branching: a survey, TOP 25 (2017) 207–236.

[4] D. Bertsimas, A. King, Or forum—an algorithmic approach to linear regression, Operations Research 64 (1) (2016) 2–16.

[5] D. Bertsimas, J. Dunn, Optimal classification trees, Machine Learning 106 (7) (2017) 1039–1082.

[6] L. Berk, D. Bertsimas, Certifiably optimal sparse principal component analysis, Mathematical Programming Computation 11 (3) (2019) 381–420.

[7] D. Bertsimas, B. Van Parys, Sparse high-dimensional regression: Exact scalable algorithms and phase transitions, Annals of Statistics 48 (1) (2020) 300–323.

[8] D. Bertsimas, R. Cory-Wright, J. Pauphilet, Solving large-scale sparse pca to certifiable (near) optimality, arXiv:2005.05195.

[9] D. Bertsimas, J. Lamperski, J. Pauphilet, Certifiably optimal sparse inverse covariance estimation, Mathematical Programming 184.

[10] R. Blanquero, E. Carrizosa, C. Molero-Río, D. Romero Morales, Sparsity in optimal randomized classification trees, European Journal of Operations Research 284.

[11] R. Blanquero, E. Carrizosa, C. Molero-Rıo, D. Romero Morales, Optimal randomized classification trees, in: Technical Report, 2018.

[12] D. Bertsimas, B. Van Parys, Sparse hierarchical regression with polynomials, Machine Learning 109.

[13] D. Bertsimas, A. King, R. Mazumder, Best subset selection via a modern optimization lens, The annals of statistics (2016) 813–852.

[14] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society: Series B (Methodological) 58 (1) (1996) 267–288.

[15] T. Hastie, R. Tibshirani, R. Tibshirani, et al., Best subset, forward stepwise or lasso? analysis and recommendations based on extensive comparisons, Statistical Science 35 (4) (2020) 579–592.

[16] D. M. Witten, R. J. Tibshirani, Extensions of sparse canonical correlation analysis with applications to genomic data, Statistical applications in genetics and molecular biology 8 (1).

[17] J. Chen, F. D. Bushman, J. D. Lewis, G. D. Wu, H. Li, Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis, Biostatistics 14 (2) (2013) 244–258.

[18] M. Nakanishi, Y. Wang, Y.-T. Wang, T.-P. Jung, A comparison study of canonical correlation analysis based methods for detecting steady-state visual evoked potentials, PloS one 10 (10) (2015) e0140703.

[19] D. M. Witten, R. Tibshirani, T. Hastie, A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis, Biostatistics 10 (3) (2009) 515–534.

[20] D. Chu, L.-Z. Liao, M. K. Ng, X. Zhang, Sparse canonical correlation analysis: New formulation and algorithm, IEEE transactions on pattern analysis and machine intelligence 35 (12) (2013) 3050–3065.

[21] D. R. Hardoon, J. Shawe-Taylor, Sparse canonical correlation analysis, Machine Learning 83 (3) (2011) 331–353.

[22] X. Chen, L. Han, J. Carbonell, Structured sparse canonical correlation analysis, in: Artificial intelligence and statistics, PMLR, 2012, pp. 199–207.

[23] W. Wang, Y.-H. Zhou, Imbalanced sparse canonical correlation analysis (2020). arXiv:2004.10231.

[24] Q. Zhu, Y. Atchade, On bayesian sparse canonical correlation analysis via rayleigh quotient framework (2020). arXiv:2010.08627.

[25] M. Asteris, A. Kyrillidis, O. Koyejo, R. Poldrack, A simple and provable algorithm for sparse diagonal cca, in: International Conference on Machine Learning, PMLR, 2016, pp. 1148–1157.

[26] O. S. Solari, J. B. Brown, P. J. Bickel, Sparse canonical correlation analysis via concave minimization, arXiv preprint arXiv:1909.07947.

[27] M. Wenwen, L. Juan, S. Zhang, et al., Sparse weighted canonical correlation analysis, Chinese Journal of Electronics 27 (3) (2018) 459–466.