Variable selection for generalized canonical correlation analysis

ARTHUR TENENHAUS*
SUPELEC, Plateau de molon, 3 rue Joliot-Curie, 91192 Gif-sur-Yvette Cedex, France
arthur.tenenhaus@supelec.fr

CATHY PHILIPPE
CNRS-IGR-Paris XI university, UMR8203, 94805 Villejuif cedex, France

VINCENT GUILLEMOT
NEUROSPIN, I2BM, CEA saclay, 91191 Gif-sur-Yvette cedex, France

KIM-ANH LE CAO
Queensland Facility for Advanced Bioinformatics, University of Queensland, 306 Carmody Road, St Lucia, QLD 4072, Australia

JACQUES GRILL
CNRS-IGR-Paris XI university, UMR8203, 94805 Villejuif cedex, France

VINCENT FROUIN
NEUROSPIN, I2BM, CEA saclay, 91191 Gif-sur-Yvette cedex, France

SUMMARY
Regularized generalized canonical correlation analysis (RGCCA) is a generalization of regularized canonical correlation analysis to 3 or more sets of variables. RGCCA is a component-based approach which aims to study the relationships between several sets of variables. The quality and interpretability of the RGCCA components are likely to be affected by the usefulness and relevance of the variables in each block. Therefore, it is an important issue to identify within each block which subsets of significant variables are active in the relationships between blocks. In this paper, RGCCA is extended to address the issue of variable selection. Specifically, sparse generalized canonical correlation analysis (SGCCA) is proposed to combine RGCCA with an $\ell_1$-penalty in a unified framework. Within this framework, blocks are not necessarily fully connected, which makes SGCCA a flexible method for analyzing a wide variety of practical problems. Finally, the versatility and usefulness of SGCCA are illustrated on a simulated dataset and on a 3-block dataset which combine gene expression, comparative genomic hybridization, and a qualitative phenotype measured on a set of 53 children with glioma. SGCCA is available on CRAN as part of the RGCCA package.

Keywords: Generalized canonical correlation analysis; Multiblock data analysis; Variable selection.

*To whom correspondence should be addressed.

© The Author 2014. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.
Many biological and medical studies now involve multiple genomic assays along with rich phenotypes from patient files or imaging. For instance, in the oncology study (Curtis and others, 2012), gene expression, comparative genomic hybridization, and survival times are measured on the same patients and considered jointly in an integrated analysis. In the neuroscience studies from Le Floch and others (2012), some endophenotypes extracted from neuroimaging are considered in addition to genetic measurements and behavioral phenotypes. Such structured datasets may be considered as endophenotypes extracted from neuroimaging are considered in addition to genetic measurements and behavioral phenotypes. Section 4 discusses some well-known methods that are just special cases of SGCCA.

1. Introduction

Many biological and medical studies now involve multiple genomic assays along with rich phenotypes from patient files or imaging. For instance, in the oncology study (Curtis and others, 2012), gene expression, comparative genomic hybridization, and survival times are measured on the same patients and considered jointly in an integrated analysis. In the neuroscience studies from Le Floch and others (2012), some endophenotypes extracted from neuroimaging are considered in addition to genetic measurements and behavioral phenotypes. Such structured datasets may be considered as endophenotypes extracted from neuroimaging are considered in addition to genetic measurements and behavioral phenotypes. Section 4 discusses some well-known methods that are just special cases of SGCCA.

Canonical correlation analysis (CCA) is a standard approach to studying the relationships between two blocks of variables only, and numerous \( \ell_1 \) and/or \( \ell_2 \) regularized extensions of the CCA have been proposed when the number of variables \( p_j \) exceeds the number of observations \( n \) for any \( j \)th block (e.g. Vinod, 1976; Waaijenborg and others, 2008; Parkhomenko and others, 2009; Le Cao and others, 2009; Witten and others, 2009; Lykou and Whittaker, 2010; Hardoon and Shawe-Taylor, 2011).

Regularized generalized canonical correlation analysis (RGCCA) as proposed in Tenenhaus and Tenenhaus (2011) is a framework for studying associations between more than 2 blocks. The aim of RGCCA is to extract the information which is shared by the \( J \) blocks of variables taking into account an a priori graph of connections between blocks. In this framework, a design matrix \( C = \{c_{jk}\} \) defines this graph of connections: \( c_{jk} \) is equal to 1 if blocks \( X_j \) and \( X_k \) are connected or is equal to 0 otherwise. Considering a set of shrinkage constants \( \tau_1, \ldots, \tau_J \) taking value in \([0, 1]\), RGCCA is defined as the following optimization problem:

\[
\begin{align*}
\arg\max_{a_1, a_2, \ldots, a_J} & \quad \sum_{j,k=1; j \neq k}^J c_{jk} g(\text{cov}(X_j a_j, X_k a_k)) \\
\text{subject to} & \quad (1 - \tau_j) \text{var}(X_j a_j) + \tau_j \|a_j\|_2^2 = 1, \quad j = 1, \ldots, J.
\end{align*}
\]

In this optimization problem, \( g \) may be defined as \( g(x) = x \) (Horst scheme proposed in Kramer, 2007), \( g(x) = |x| \) (centroid scheme proposed in Wold, 1985) or \( g(x) = x^2 \) (factorial scheme proposed in Lohmöller, 1989). The vector \( a_j \) (respectively \( y_j = X_j a_j \)) is referred to as an outer weight vector (respectively outer component). The Horst scheme penalizes structural negative correlation between components while both centroid and factorial schemes can be viewed as attractive alternatives that enable two components to be negatively correlated. The optimization problem (1.1) is limited to these 3 schemes since they are the most used in the multiblock and partial least squares (PLS) literature.

Biomedical data are known to be measurements of intrinsically parsimonious processes. In order to account for this parsimony and to improve the interpretability of the resulting RGCCA model, an important issue is to identify subsets of variables from each block which are active in the relation between connected blocks. This variable selection step can be achieved by adding, within the RGCCA optimization problem (1.1), a penalty promoting sparsity. For that purpose, an \( \ell_1 \) penalization on the outer weight vectors \( a_1, \ldots, a_J \) is applied which induces a sparse RGCCA model that gives rise to sparse generalized canonical correlation analysis (SGCCA).

This paper is organized as follows: The general optimization problem for SGCCA is presented in Section 2 and the associated algorithm is derived. The monotone convergence of the algorithm is proven (see supplementary material available at Biostatistics online). In Section 3, the versatility and usefulness of SGCCA are demonstrated on a simulated dataset and on a 3-block dataset which combine gene expression, comparative genomic hybridization and a qualitative phenotype. Section 4 discusses some well-known methods that are just special cases of SGCCA.
2. Sparse Generalized Canonical Correlation Analysis

The SGCCA optimization problem and the associated algorithm are presented in this section.

2.1 Optimization problem for Sparse GCCA

From now on, all \( \tau_j \) equal 1, which means that the constraints are applied on the length of the \( a_j \). Adding an \( \ell_1 \) penalty on \( a_1, \ldots, a_J \), SGCCA is defined as the following optimization problem:

\[
\begin{align*}
\argmax_{a_1, a_2, \ldots, a_J} & \quad \sum_{j,k=1,j\neq k}^{J} c_{jk} g(\text{cov}(X_j a_j, X_k a_k)) \\
\text{subject to} & \quad \|a_j\|_2 = 1 \quad \text{and} \quad \|a_j\|_1 \leq s_j, \quad j = 1, \ldots, J,
\end{align*}
\]

(2.1)

where \( s_j \) is a positive constant that determines the amount of sparsity for \( a_j, j = 1, \ldots, J \). The smaller \( s_j \), the larger the degree of sparsity for \( a_j \).

**Proposition 1** The solution of optimization problem (2.1) satisfies

\[
a_j = \frac{S((1/n)X_j'z_j, \lambda_{1j})}{\|S((1/n)X_j'z_j, \lambda_{1j})\|_2}, \quad j = 1, \ldots, J,
\]

(2.2)

where

\[
z_j = \sum_{k=1,k\neq j}^{J} c_{jk} w\left(\frac{1}{n}a_j'X_j'X_k a_k\right)X_k a_k.
\]

(2.3)

with \( w(x) = (1/\varphi)g'(x) \) with \( \varphi = 1 \) for both the horst and centroid scheme, and \( \varphi = 2 \) for the Factorial scheme; \( S \) is the soft-thresholding operator defined by \( S(a, \lambda) = \text{sign}(a) \max(0, |a| - \lambda) \) and \( \lambda_{1j} \) chosen such that \( \|a_j\|_1 \leq s_j \).

**Proof of Proposition 1.** First, we re-write the criterion (2.1) using Lagrange multipliers

\[
\mathcal{L} = \sum_{j,k=1,k\neq j}^{J} c_{jk} g\left(\frac{1}{n}a_j'X_j'X_k a_k\right) - \varphi \left[ \sum_{j=1}^{J} \frac{\lambda_{2j}}{2} \|a_j\|_2^2 - 1 \right] - \sum_{j=1}^{J} \lambda_{1j}(\|a_j\|_1 - s_j)
\]

(2.4)

where \( \lambda_{11}, \ldots, \lambda_{1J}, \lambda_{21}, \ldots, \lambda_{2J} \) are the Lagrange multipliers. We may suppose that \( (1/n)a_j'X_j'X_k a_k \) is different from zero, since if it were not the case, we would just set the design coefficient \( c_{jk} \) to zero. Therefore, we may consider the derivative \( g' \) of \( g \) when \( g \) is the absolute value. Considering the partial subgradients of the Lagrangian function with respect to \( a_j \) yields the following equation for SGCCA:

\[
\partial_{a_j} \mathcal{L} = \sum_{k=1,k\neq j}^{J} c_{jk} g'\left(\frac{1}{n}a_j'X_j'X_k a_k\right) \frac{1}{n}X_j'X_k a_k - \varphi \left[ \lambda_{2j}a_j + \lambda_{1j}y_j \right], \quad j = 1, \ldots, J,
\]

(2.5)

where \( y_{jk} \), the kth element of \( y_j \), is the subgradient of \( \sum_{k=1}^{\rho_j} |a_{jk}| \) with respect to \( a_{jk} \), and is defined by:

\( y_{jk} = \text{sign}(a_{jk}) \) if \( a_{jk} \neq 0 \) and \( y_{jk} \in [-1, +1] \) if \( a_{jk} = 0 \).

Let us introduce the inner components \( z_j \) defined as:

\[
z_j = \sum_{k=1,k\neq j}^{J} c_{jk} w\left((1/n)a_j'X_j'X_k a_k\right)X_k a_k
\]

where \( w(x) = (1/\varphi)g'(x) \) with \( \varphi = 1 \) for both the Horst and centroid scheme, and \( \varphi = 2 \) for the factorial scheme.
These inner components play a central role in the SGCCA algorithm to be described and enable us to simplify (2.5) as follows:

\[
\partial_a \mathcal{L} = \frac{1}{n} X_j^t z_j - \lambda_{2j} a_j - \lambda_{1j} y_j, \quad j = 1, \ldots, J. \tag{2.6}
\]

From the definition of the subgradient and from (2.6),

\[
\partial_a \mathcal{L} = \begin{cases}
\frac{1}{n} X_j^t z_j - \lambda_{2j} a_j - \lambda_{1j} \text{sign}(a_{jk}) & \text{if } a_{jk} \neq 0, \\
\frac{1}{n} X_j^t z_j - \frac{1}{n} X_j^t z_j + \lambda_{1j} & \text{if } a_{jk} = 0,
\end{cases} \tag{2.7}
\]

where \( x_{jk} \) represents the \( k \)-th column of \( X_j \). At the optimum, we must have \( 0 \in \partial_a \mathcal{L} \) and we get:

\[
\begin{align*}
\frac{1}{n} x_{jk}^t z_j - \lambda_{2j} a_j - \lambda_{1j} \text{sign}(a_{jk}) &= 0 & \text{if } a_{jk} \neq 0, \\
\left| \frac{1}{n} x_{jk}^t z_j \right| &< \lambda_{1j} & \text{if } a_{jk} = 0.
\end{align*} \tag{2.8}
\]

From (2.8), the following equality holds: \( \text{sign}((1/n)x_{jk}^t z_j) = \text{sign}(a_{jk}) \), which yields:

\[
a_{jk} = \begin{cases}
\frac{1}{\lambda_{2j}} \left( \frac{1}{n} x_{jk}^t z_j - \lambda_{1j} \text{sign} \left( \frac{1}{n} x_{jk}^t z_j \right) \right) & \text{if } \left| \frac{1}{n} x_{jk}^t z_j \right| < \lambda_{1j} \\
0 & \text{if } \left| \frac{1}{n} x_{jk}^t z_j \right| \geq \lambda_{1j}
\end{cases} = \begin{cases}
\frac{1}{\lambda_{2j}} \text{sign} \left( \frac{1}{n} x_{jk}^t z_j \right) \left( \frac{1}{n} x_{jk}^t z_j - \lambda_{1j} \right) & \text{if } \left| \frac{1}{n} x_{jk}^t z_j \right| < \lambda_{1j} \\
0 & \text{if } \left| \frac{1}{n} x_{jk}^t z_j \right| \geq \lambda_{1j}
\end{cases} \tag{2.9}
\]

or in compact form as follows:

\[
a_{jk} = \frac{1}{\lambda_{2j}} \text{sign} \left( \frac{1}{n} x_{jk}^t z_j \right) \max \left( 0, \left| \frac{1}{n} x_{jk}^t z_j \right| - \lambda_{1j} \right). \tag{2.10}
\]

\( a_j \) can be written in matrix notation: \( a_j = (1/\lambda_{2j}) S((1/n)X_j^t z_j, \lambda_{1j}) \) where \( \lambda_{2j} \) is chosen such that \( \|a_j\|_2 = 1 \) and \( \lambda_{1j} \) chosen such that \( \|a_j\|_1 \leq s_j \). This yields the following final expression for \( a_j \):

\[
a_j = \frac{S((1/n)X_j^t z_j, \lambda_{1j})}{\|S((1/n)X_j^t z_j, \lambda_{1j})\|_2}, \quad j = 1, \ldots, J, \tag{2.11}
\]

which concludes the proof of Proposition 1.

From (2.10), we conclude that each \( a_{jk} \) is canceled out and does not contribute to the construction of the block component \( y_j \), if the covariance between \( x_{jk} \) and \( z_j \) is below a given threshold \( \lambda_{1j} \) (where \( \lambda_{1j} \) is defined such that \( \|a_j\|_1 \leq s_j \)).

Note that (2.11) is also solution of the following convex problem:

\[
\max_{a_j} \text{cov}(X_j a_j, z_j) \quad \text{s.t.} \quad \|a_j\|_2 \leq 1 \quad \text{and} \quad \|a_j\|_1 \leq s_j. \tag{2.12}
\]

This result is central to the proof of Proposition 2 given further.
In addition, the following equality holds (see Tenenhaus and Tenenhaus, 2011 for the proof):

\[
\sum_{j,k; j \neq k} c_{jk} g(\text{cov}(X_j a_j, X_k a_k)) = \sum_{j=1}^{J} \text{cov}(X_j a_j, z_j).
\] (2.13)

This equality shows that maximizing the SGCCA objective function boils down to maximizing the right-hand term of (2.13). This alternative expression suggests a multiconvex formulation for the optimization (2.1), leading us to propose an iterative procedure, which will be presented in the next section. Indeed, for a fixed vector \( z_j \), the solution is obtained by maximizing with respect to \( a_j \) (through the convex optimization (2.12)) sequentially each element of the sum using (2.11).

### 2.2 Algorithm for SGCCA

To obtain a monotonically convergent algorithm for optimization problem (2.1) (i.e. the bounded objective function \( \sum_{j,k=1; j \neq k} c_{jk} g(\text{cov}(X_j a_j, X_k a_k)) \) to be maximized increases at each step of the iterative procedure), a sequence of operations similar to the ones used by Wold (1985), Tenenhaus and Tenenhaus (2011) is described in Algorithm 1. The procedure begins with an arbitrary choice of initial \( \ell_2 \)-normalized \( a_1^0, \ldots, a_J^0 \). Steps A and B are then iterated until convergence of the bounded criterion that is guaranteed by inequality (2.15) given in Proposition 2 and the corresponding proof that can be found in supplementary material available at Biostatistics online.

**Proposition 2** Let \((a_1^s, \ldots, a_J^s), s = 0, 1, 2, \ldots\) be a sequence of outer weight vectors generated by the Algorithm 1 for optimization problem (2.1). Let \( h \) be the function defined by:

\[
h(a_1, \ldots, a_J) = \sum_{j,k=1; j \neq k} c_{jk} g(\text{cov}(X_j a_j, X_k a_k)).
\] (2.14)

The following inequalities hold:

\[
\forall s \ h(a_1^s, \ldots, a_J^s) \leq h(a_1^{s+1}, \ldots, a_J^{s+1}).
\] (2.15)

**Proof.** The reader is referred to supplementary material available at Biostatistics online for the proof. □

The essential feature of this algorithm is that each replacement is optimal (each update results from the convex optimization problem (2.12) and thus necessarily increases the value of the objective function) and sequential (that is to say that \( a_j^s \) must be replaced by \( a_j^{s+1} \) before replacing \( a_j^{s+1} \)). Algorithm 1 is found to be very stable and usually reaches a convergence tolerance within a few iterations (see Section 3.1). Moreover, it is worth mentioning that Algorithm 1 can easily handle missing data simply by skipping the missing elements in the computation of inner products.

#### 2.2.1 Multiple components for SGCCA

Algorithm 1 yields one component per block. Next SGCCA dimensions can be obtained by deflation on the previous components. Let us detail the procedure for the second dimension. Let \( y_{1j} \) be the first outer component associated with \( X_j \). The second dimension is obtained by replacing in optimization problem (2.1) each matrix \( X_j \) by the residual matrices \( X_{1j} \) of the regression between the original blocks \( X_j \) on \( y_{1j} \). The resulting outer components are, by construction, orthogonal within each block. The algorithm for computing more than 1 component per block is described
Algorithm 1 Sparse generalized canonical correlation analysis

Require: \( J \) blocks \( X_1, \ldots, X_J \), \( J \) \( \ell_1 \) constraints \( s_1, \ldots, s_J \), a design matrix \( C \) and the scheme
Choose \( J \) arbitrary normalized vectors \( a_0^j, \ldots, a_J^j \)
repeat (\( s = 0, 1, 2, \ldots \))
  for \( j = 1, 2, \ldots, J \) do
  \( \triangleright \) Step A. Compute the inner components
  \[ z_j^s \leftarrow \sum_{k=1}^{j-1} c_{jk} w[\text{cov}(X_j a_j^s, X_k a_k^{s+1})]X_k a_k^{s+1} + \sum_{k=j+1}^J c_{jk} w[\text{cov}(X_j a_j^s, X_k a_k^s)]X_k a_k^s \]
  \( \triangleright \) Step B. Compute the outer weight
  \[ a_{s+1}^j \leftarrow S\left(\frac{1}{n} X_j' z_j^s, \lambda_{1j}\right) \]
  Where \( S \) denotes the soft-thresholding operator defined above and \( \lambda_{1j} = 0 \) if \( \|a_j^{s+1}\|_1 \leq s_j \) or \( \lambda_{1j} \) chosen such that \( \|a_j^s\|_1 = s_j \) (binary search) with \( 0 \leq s_j \leq \sqrt{p_j} \)
end for
until \( h(a_{s+1}^1, \ldots, a_{s+1}^J) - h(a_1^s, \ldots, a_J^s) \leq \varepsilon \)
return \( a_{s+1}^1, \ldots, a_{s+1}^J \)

Algorithm 2 Computing the (h+1)th block components for SGCCA

Require: \( X_{1h} = X_1, \ldots, X_{Jh} = X_J \) (convention: \( X_{10} = X_1, \ldots, X_{J0} = X_J \)), \( y_{1h}, \ldots, y_{Jh} \)
\( \triangleright \) Compute \( J \) res. matrices
\[ X_{1,h+1} \leftarrow \left( I_{p_1} - \frac{y_{1h} y_{1h}'}{y_{1h}' y_{1h}} \right) X_{1h}, \ldots, X_{J,h+1} \leftarrow \left( I_{p_J} - \frac{y_{Jh} y_{Jh}'}{y_{Jh}' y_{Jh}} \right) X_{Jh} \]
return \( X_{1,h+1}, \ldots, X_{J,h+1} \)

in Algorithm 2. It can be noted that in the multiblock framework there are many different ways of deflation (Vivien and Sabatier, 2003). In this paper, we only consider the so-called symmetric way which consists in deflating each block with its own component. Moreover, we stress that the number of components per block could differ from one block to another: this is made possible by choosing not to deflate all the blocks in Algorithm 2. We mention that Algorithm 1 and Algorithm 2 are available on CRAN as part of the RGCCA package (Tenenhaus and Guillemot, 2013).

3. Results

3.1 Simulation study

SGCCA is evaluated on a simple simulated example. The objective of this simulation is 2-fold: (i) to show the reliability of the SGCCA algorithm at distinguishing between the elements of the outer weight vectors that are truly nonzero and those that are not and (ii) to study the convergence properties of the SGCCA algorithm.

3.1.1 Sensitivity/specificity of SGCCA. In this simulation, 3 blocks are considered. Each \( n \times p_j \) block \( X_j, j = 1, \ldots, 3 \) is generated according to the following model:

\[
X_j = u_j w_j' + E_j, \quad j = 1, \ldots, 3, \tag{3.1}
\]
Variable selection for generalized canonical correlation analysis

(a) (b) (c) Fig. 1. Sensitivity (Se) and specificity (Sp) of the variable selection and impact of the mis-specification of the design matrix. C1 (respectively C2) corresponds to the correctly specified (respectively misspecified) design matrix. (a) X1, (b) X2, and (c) X3.

where the outer components u1, u2, and u3 are 50-vectors drawn from a multivariate normal distribution with 0 mean and predetermined correlation structure. In this experiment, cor(u1, u3) = cor(u2, u3) = 0.7 and cor(u1, u2) = 0. In addition, w1 is 200-vector, w2 is 500-vector, and w3 is 700-vector. Only the first 75 elements of w1, w2, and w3 were nonzero and are drawn from a uniform distribution in the interval of [−0.3, −0.2] ∪ [0.2, 0.3]. Ej is a 50 × pj residual matrix where each element is drawn from a normal distribution with mean equal to 0.2. Since each block Xj is strongly unidimensional with associated component uj, only the first dimension of each block is studied.

A first experiment consists in running SGCCA on X1, X2, and X3 associated with the design matrix C1 defined by: c13 = c23 = 1 and c12 = 0 (which encodes the correlation structure between components) and the centroid scheme. The regularization parameters are chosen so that the number of nonzero elements per outer weight vector is approximately equal to 75. We may note that for a fixed sj, the number of nonzero elements in wj appears to be not very sensitive to modifications of sk, k ≠ j. The resulting sparse estimates are then compared with the true w1, w2, and w3 in terms of sensitivity (no. of nonzero elements in aj which are nonzero in wj / no. of nonzero in wj) and specificity (no. of zero elements in aj which are zero in wj / no. of zero in wj). This experiment was repeated independently 10 000 times and the resulting boxplots of sensitivity and specificity associated with C1 are reported in Figure 1. The sparse estimates of w1, w2, and w3 are fairly accurate at selecting active variables.

In a second experiment, the impact of the mis-specification of the design matrix is evaluated. SGCCA was run on the same 10 000 simulations described above but the correlation structure between components was incorrectly encoded in the design matrix C2 as follows: c13 = c23 = c12 = 1. The resulting boxplots of sensitivity/specificity are reported in Figure 1. A small degradation of the sensitivity and specificity can be observed from Figure 1. However, SGCCA with misspecified design remains fairly comparable with the results obtained with the correctly specified design matrix. The rationale behind these results is that the covariance-based model tends to find components with large variance that explain well their own block (first priority), while taking into account the correlations with neighboring components (second priority). Indeed, it has been shown that the variance terms of the covariance criteria dominates over the correlation term (Tenenhaus and Tenenhaus, 2011). To some extent, this unbalanced compromise between variance and correlation may justify the stability of the variable selection procedure for SGCCA even with misspecified design matrix.
3.1.2 Convergence property of the SGCCA algorithm. To check whether the global optimum solution is reached or not, a 3-block dataset is simulated according to (3.1). Data simulation were repeated independently 100 times. For each of these 100 3-block simulated dataset, SGCCA was run 100 times with different random initial weights. Over the 10 000 runs, SGCCA led to the global optimum solution in 99% of cases, in 7.76 iterations (stopping criteria: $10^{-16}$) at around 0.28 s on midrange laptop computer. Initialization by the first right-singular vectors of $X_j$ can be used instead of random initial weights and in that case, SGCCA led to the global solution each time in 6.21 iterations (stopping criteria: $10^{-16}$) at around 0.27 s on midrange laptop computer.

3.2 Application to pediatric glioma data

The versatility of SGCCA is illustrated for the analysis of genomic experiments on pediatric high-grade gliomas (pHGG). An interesting aspect lies in the possibility of taking into account some hypotheses on connections between blocks. The design matrix $C$ enables the specification of this structural connection between blocks. The different designs make it possible to analyze the 3-block glioma dataset from different viewpoints as discussed in Section 3.2.3.

3.2.1 Biological problem. Brain tumors are the most common solid tumors in children and have the highest mortality rate of all pediatric cancers. Despite advances in multimodality therapy, children with pHGG invariably have an overall survival of around 20% at 5 years. Depending on their location (e.g. brainstem, central nuclei, or supratentorial), pHGG present different characteristics in terms of radiological appearance, histology, and prognosis. Our hypothesis is that pHGG have different genetic origins and oncogenic pathways depending on their location. Thus, the biological processes involved in the development of the tumor may be different from one location to another, as it has been frequently suggested. The possibility of assigning different gene expression signatures for each location has been shown in Puget and others (2012). In this paper, integrated analysis of gene expression and chromosomal imbalances is proposed to identify more precisely the biological processes and genes that differentiate these pHGG for the different locations. This may ultimately justify specific treatment for each location of the tumor.

3.2.2 Description of the data. Pretreatment frozen tumor samples were obtained from 53 children with newly diagnosed pHGG from Necker Enfants Malades (Paris, France) (Puget and others, 2012). The 53 tumors are divided into 3 locations: supratentorial (HEMI), central nuclei (MIDL), and brain stem (DIPG). The final dataset is organized in 3 blocks of variables defined for the 53 tumors: the first block $X_1$ provides the expression of 15 702 genes (GE). The second block $X_2$ contains the imbalances of 1229 segments (CGH) of chromosomes. $X_3$ is a block of dummy variables describing the categorical variable location denoted ($loc$) hereafter. The dummy variable HEMI has been left out because of redundancy with MIDL and DIPG. All the variables have been standardized (zero mean and unit variance).

3.2.3 Path diagram. Three designs have been considered for the glioma application and are reported in Figure 2.

Design 1 assumes all blocks to be connected to each other. From a biological viewpoint, SGCCA combined with this design could be relevant for identifying GE and CGH variables corresponding to the loss of one allele of a gene and overexpression of the remaining allele. Design 2 is commonly used in many applications and is oriented toward the prediction of the location. $X_1$ (GE) and $X_2$ (CGH) are called first-order blocks and are often considered as predictor blocks. $X_3$ is called second-order block and is often considered
Variable selection for generalized canonical correlation analysis 577

Fig. 2. Path diagram. (a) Design 1: All blocks are connected. (b) Design 2: \(X_1\) (GE) is connected to \(X_3\) (location) \((c_{13} = 1)\), \(X_2\) (CGH) is connected to \(X_3\) \((c_{23} = 1)\), and \(X_1\) is not connected to \(X_2\) \((c_{12} = 0)\). This design is oriented toward the prediction of the location: (c) Design 3: \(X_1\) (GE) is connected to \(X_2\) (CGH) \((c_{12} = 1)\), \(X_1\) (CGH) is connected to \(X_3\) \((c_{13} = 1)\), and \(X_2\) is not connected to \(X_3\) \((c_{23} = 0)\).

as a response block. This design does not impose any correlation between GE and CGH blocks. From a biological viewpoint, SGCCA with Design 2 tends to highlight situations where genes associated with location can be identified even if there is no relation between CGH and GE. This could correspond to a gene of a region gained in CGH but methylated and therefore completely downregulated in GE. Design 3 mimics the “central dogma” of molecular biology which asserts that “DNA makes RNA makes protein”. SGCCA with Design 3 tends to highlight events starting with the chromosomal imbalances and their consequences on gene expression. These 3 designs are compared in the next sections.

3.2.4 Predictive performance. The efficiency of SGCCA is demonstrated by the prediction of the location of the tumor in the brain (3 modalities) from GE \((X_1)\) and CGH \((X_2)\) data. To evaluate the predictive performance of SGCCA for each design/scheme configuration (9 scenarios), we performed a stratified 10-fold Monte-Carlo cross-validation (MCCV). For each iteration \(b\), \(b = 1, \ldots, 10\), we randomly selected 9/10 of the data to obtain the training set \(X_{b} = [X_{1b}, X_{2b}, X_{3b}]\) and the test set \(X_{-b}\). We determined the outer weight vectors \(a_{b} = (a_{b1}, a_{b2}, a_{b3})\) from \(X_{b}\) using Algorithm 1 and the resulting outer components \(y_{b1} = X_{1b}a_{b1}\), \(y_{b2} = X_{2b}a_{b2}\), and \(y_{b3} = X_{3b}a_{b3}\) are constructed. The test error rate is then estimated using the following 3-step protocol: (i) build a discriminant model \(L\) (using a Bayesian discriminant analysis) of the qualitative variable \(loc_{b}\) based on the components \(y_{b1}\) and \(y_{b2}\), (ii) build the components for the test set \(X_{-b}\) as \(y_{1-b} = X_{1-b}a_{1b}\) and \(y_{2-b} = X_{2-b}a_{2b}\), and (iii) predict the labels with \(L\) based on \((y_{1-b}, y_{2-b})\) and compare with the expected labels. The sparsity parameters (one for each block) are chosen for each of the 10 MCCV iterations using an internal 5-fold CV loop: the parameters that minimize the prediction error across the 5-fold are retained.

The test error rates across the 10 MCCV folds for SGCCA (Designs 1–3, 3 schemes), \(\ell_1\)-linear discriminant analysis (\(\ell_1\)-LDA) (Witten and Tibshirani, 2011), supervised CCA (Witten and Tibshirani, 2009), and RGCCA (Tenenhaus and Tenenhaus, 2011) are compared and reported in Figure 3. The test error rates when considering one component (respectively 2 components) per block is presented in Figure 3(a) (respectively Figure 3(b)).
Fig. 3. Boxplot of the test error rate across the 10-fold for $\ell_1$-linear discriminant analysis (Witten and Tibshirani, 2011), supervised CCA (Witten and Tibshirani, 2009), SGCCA (Designs 2 and 3 schemes), SGCCA (Designs 1 and 3 schemes), RGCCA ($\tau_1 = \tau_2 = 1$ and $\tau_3 = 0$) (with Design 2). (a) Boxplot of the test error rate across the 10-fold when considering one component. (b) Boxplot of the test error rate across the 10-fold when considering two components.

The small differences between factorial and centroid schemes confirm that block components are not very sensitive to the choice of the factorial or centroid scheme (Noonan and Wold, 1982). However, more striking differences between the 3 designs are observed, underlining the importance of guiding the choice of the design by *a priori* on the connections between blocks. SGCCA with Design 2 yields the best predictive performances for whichever scheme, which confirms that this structural connection between blocks is oriented toward the prediction of the location. It is worth mentioning that the multi-CCA approach proposed in Witten and Tibshirani (2009) is recovered by considering SGCCA combined with the Horst scheme and Design 1. Boxplots of the test error rate across the 10-fold (1 and 2 components) for multi-CCA are exactly the same as those obtained with Horst-SGCCA with Design 1 (results not shown). The predictive performance of SGCCA is very close to the one obtained with $\ell_1$-LDA (Witten and Tibshirani, 2011) and supervised CCA (Witten and Tibshirani, 2009). Nevertheless, we emphasize that (i) $\ell_1$-LDA does not take into account the block structure and (ii) even if supervised CCA was proposed in Witten and Tibshirani (2009) to deal with the “first–second” order blocks structure, the second order block must be restricted to only one variable. Moreover, this second order block is not fully considered in the modelization of the relationships between blocks but rather used to preprocess the first order blocks. Finally, the predictive performances of SGCCA are also comparable with those obtained with RGCCA but benefit from the $\ell_1$-constraint to identify the small subset of significant variables that are active in the relationships between connected blocks.

3.2.5 Stability of the variable selection procedure. At each iteration of the MCCV algorithm, different sets of variables are selected. The predictive performance described in Figure 3 can be associated with: (i) an indicator of stability and (ii) the number of variables which contribute to the construction of the components (length of the signature). To evaluate the stability of the signatures, we propose using the
Table 1. Stability and mean of the length of the signatures for SGCCA, ℓ₁-LDA, and supervised CCA across the 10-fold

<table>
<thead>
<tr>
<th>Method</th>
<th>Fleiss’ κ (GE)</th>
<th>Length of the GE signature</th>
<th>Fleiss’ κ (CGH)</th>
<th>Length of the CGH signature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised CCA</td>
<td>0.130</td>
<td>455.3</td>
<td>0.116</td>
<td>36.5</td>
</tr>
<tr>
<td>ℓ₁-LDA</td>
<td><strong>0.476</strong></td>
<td><strong>9790.0</strong></td>
<td>0.322</td>
<td>480.9</td>
</tr>
<tr>
<td>Horst SGCCA (Design 1)</td>
<td>0.103</td>
<td>132.2</td>
<td>0.071</td>
<td>35.9</td>
</tr>
<tr>
<td>Factorial SGCCA (Design 1)</td>
<td>0.071</td>
<td>79.0</td>
<td>0.014</td>
<td>59.6</td>
</tr>
<tr>
<td>Centroid SGCCA (Design 1)</td>
<td>0.137</td>
<td>73.6</td>
<td>0.105</td>
<td>22.6</td>
</tr>
<tr>
<td>Horst SGCCA (Design 2)</td>
<td><strong>0.468</strong></td>
<td><strong>61.1</strong></td>
<td><strong>0.296</strong></td>
<td><strong>33.6</strong></td>
</tr>
<tr>
<td>Factorial SGCCA (Design 2)</td>
<td>0.439</td>
<td>42.0</td>
<td>0.343</td>
<td>37.6</td>
</tr>
<tr>
<td>Centroid SGCCA (Design 2)</td>
<td><strong>0.478</strong></td>
<td><strong>40.6</strong></td>
<td><strong>0.317</strong></td>
<td><strong>34.8</strong></td>
</tr>
<tr>
<td>Horst SGCCA (Design 3)</td>
<td>0.071</td>
<td>83.6</td>
<td>0.074</td>
<td>40.7</td>
</tr>
<tr>
<td>Factorial SGCCA (Design 3)</td>
<td>0.061</td>
<td>118.3</td>
<td>0.026</td>
<td>49.2</td>
</tr>
<tr>
<td>Centroid SGCCA (Design 3)</td>
<td>0.040</td>
<td>75.5</td>
<td>0.035</td>
<td>40.2</td>
</tr>
</tbody>
</table>

For each method, Fleiss’ κ was used to evaluate the stability of the 10 signatures obtained across the 10-fold. Some remarkable values are in bold.

Fleiss’ κ indicator defined in Fleiss (1971). For each variable, one can count how many times the variable is selected/not selected across the 10 iterations. These frequencies are summarized by the Fleiss’ κ score that measures the agreement among the 10 iterations. The Fleiss’ κ score is always ≤ 1, and the higher the value of κ is, the more stable are the methods with respect to sampling. The length of the signature and the κ-score of the various options of SGCCA, ℓ₁-LDA, and supervised CCA are reported in Table 1.

Whichever design and scheme, SGCCA selects fewer variables than ℓ₁-LDA and supervised CCA. The stability of the signatures for SGCCA with Design 2 and ℓ₁-LDA is equivalent but the lengths of the signatures are drastically shorter for SGCCA with Design 2, even if the prediction performance is kept at a comparable level. We stress that the length of the signatures for ℓ₁-LDA yields artificially high values for their corresponding κ-scores.

3.2.6 Visualization. Two components are built per block (GE1 and GE2 for block $X_1$) and (CGH1 and CGH2 for block $X_2$), and the graphical displays of tumors labeled according to their location are presented in Figure 4.

We observe that the GE block contains much more discriminative information than the CGH block. From Figures 4(b)–(d), the second components seem to bring additional discriminative information but Figure 3 reveals that this additional information does not improve the predictive performance.

To conclude this section, the reader is referred to supplementary material available at Biostatistics online for the biological interpretations of the signatures unveiled by SGCCA.

4. Discussion and conclusion

SGCCA for covariance-based model under sparsity. In multiblock data analysis, all blocks are assumed to be connected and many criteria were proposed with the objective of finding block components that satisfy some kind of covariance optimality. The sum of the covariance (SUMCOV) was proposed in Van de Geer (1984), the sum of squares of the covariance (SSQCOV) was proposed in Hanafi and Kiers (2006), and the sum of the absolute value of the covariance (SABSCOV) by Kramer (2007). Note that when $J = 2$, SUMCOV, SSQCOV, and SABSCOV yield the same solution and depending on the strategy of deflation...
leads to Tucker’s inter-battery factor analysis (Tucker, 1958) (symmetric deflation) and PLS regression (Wold and others, 1983) (non-symmetric deflation). Note that other covariance-based methods combined with hierarchical structure are discussed in Tenenhaus and Tenenhaus (2011). It is quite remarkable that SGCCA offers a sparse counterpart for all the covariance-based methods cited above and brings (i) a unified statistical framework for sparse multiblock data analysis and (ii) a unified implementation strategy. We note that a sparse version of SUMCOV was already proposed in Witten and Tibshirani (2009). In that paper, connections between sparse SUMCOV and sparse methods for principal component analysis (PCA)
Variable selection for generalized canonical correlation analysis

(including Jolliffe and others, 2003 and Zou and others, 2006) were discussed. Again, SGCCA provides an efficient implementation for these sparse PCA approaches.

Choice for $\tau_j = 1$. From the viewpoint of the optimization problem (1.1), the regularization parameters $\tau_j \in [0, 1], j = 1, \ldots, J$ enable a smooth interpolation between the maximization of the covariance (all $\tau_j s' = 1$) and the maximization of the correlation (all $\tau_j s' = 0$). The covariance-based model ($\tau_j = 1$) that underlies SGCCA first tends to find components with large variance that explain well their own block (PCA criteria), while taking into account the correlations with neighboring components. This unbalanced compromise between variance and correlation is mandatory for stable variable selection. Moreover, in case of multi-collinearity within blocks or when the number of observations is smaller than the number of variables, the sample covariance matrix $S_{jj} = (1/n)X_j'X_j$ is a poor estimation of the true covariance matrix. The usual strategy for finding a better estimation is to consider the class of linear combinations $\{\hat{\Sigma}_{jj} = \tau_j I + (1 - \tau_j)S_{jj}\}$ of the identity matrix $I$ and the sample covariance matrix $S_{jj}$ (e.g. Schäfer and Strimmer, 2005). In this study, by setting all $\tau_j$ equal to 1, we implicitly assume that for each block, the true covariance matrix is estimated by the identity. To some extent, this also justifies the stability of the variable selection in SGCCA. This also highlights the fact that when $\tau_j = 1$, neither matrix inversion nor diagonalization is required.

Structured variable selection. As a conclusion, work in progress extends SGCCA so that it could exploit a pre-given structure between variables within blocks via structured and sparsity-inducing penalties (Lofstedt and others, 2014). Such structured penalties lead to new challenges on optimization techniques that were previously addressed in the two-block setting in Chen and others (2012).

SUPPLEMENTARY MATERIAL


ACKNOWLEDGMENTS

We thank the 2 anonymous referees and the associate editor for their comments which have greatly improved our manuscript. Conflict of Interest: None declared.

FUNDING

This work was supported by grants from the French National Research Agency (ANR GENIM; grant ANR-10-BLAN-0128) and (ANR Investissement d Avenir BRAINOMICS; grant ANR-10-BINF-04).

REFERENCES


A. TENENHAUS AND OTHERS


[Received November 29, 2012; revised January 8, 2014; accepted for publication January 9, 2014]