Statistical Applications in Genetics and Molecular Biology

Volume 8, *Issue* 1 2009 *Article* 28

Extensions of Sparse Canonical Correlation Analysis with Applications to Genomic Data

Daniela M. Witten[*] Robert J. Tibshirani[†]

[*] Stanford University, dwitten@stanford.edu

[†] Stanford University, tibs@stat.stanford.edu

Extensions of Sparse Canonical Correlation Analysis with Applications to Genomic Data*

Daniela M. Witten and Robert J. Tibshirani

Abstract

In recent work, several authors have introduced methods for sparse canonical correlation analysis (sparse CCA). Suppose that two sets of measurements are available on the same set of observations. Sparse CCA is a method for identifying sparse linear combinations of the two sets of variables that are highly correlated with each other. It has been shown to be useful in the analysis of high-dimensional genomic data, when two sets of assays are available on the same set of samples. In this paper, we propose two extensions to the sparse CCA methodology. (1) Sparse CCA is an unsupervised method; that is, it does not make use of outcome measurements that may be available for each observation (e.g., survival time or cancer subtype). We propose an extension to sparse CCA, which we call sparse supervised CCA, which results in the identification of linear combinations of the two sets of variables that are correlated with each other and associated with the outcome. (2) It is becoming increasingly common for researchers to collect data on more than two assays on the same set of samples; for instance, SNP, gene expression, and DNA copy number measurements may all be available. We develop sparse multiple CCA in order to extend the sparse CCA methodology to the case of more than two data sets. We demonstrate these new methods on simulated data and on a recently published and publicly available diffuse large B-cell lymphoma data set.

KEYWORDS: sparse canonical correlation analysis, gene expression, microarray, DNA copy number, CGH, SNP, lasso, fused lasso

*We thank Andrew Beck, Patrick Brown, Jonathan Pollack, Robert West, and anonymous reviewers for helpful comments. Daniela Witten was supported by a National Defense Science and Engineering Graduate Fellowship. Robert Tibshirani was partially supported by National Science Foundation Grant DMS-9971405 and National Institutes of Health Contract N01-HV-28183.

1 Introduction

Canonical correlation analysis (CCA), due to Hotelling (1936), is a classical method for determining the relationship between two sets of variables. Given two data sets $\mathbf{X}_1$ and $\mathbf{X}_2$ of dimensions $n \times p_1$ and $n \times p_2$ on the same set of n observations, CCA seeks linear combinations of the variables in $\mathbf{X}_1$ and the variables in $\mathbf{X}_2$ that are maximally correlated with each other. That is, $\mathbf{w}_1 \in \mathbb{R}^{p_1}$ and $\mathbf{w}_2 \in \mathbb{R}^{p_2}$ maximize the *CCA criterion*, given by

$$\text{maximize}_{\mathbf{w}_1,\mathbf{w}_2} \mathbf{w}_1^T \mathbf{X}_1^T \mathbf{X}_2 \mathbf{w}_2 \text{ subject to } \mathbf{w}_1^T \mathbf{X}_1^T \mathbf{X}_1 \mathbf{w}_1 = \mathbf{w}_2^T \mathbf{X}_2^T \mathbf{X}_2 \mathbf{w}_2 = 1, \quad (1)$$

where we assume that the columns of $\mathbf{X}_1$ and $\mathbf{X}_2$ have been standardized to have mean zero and standard deviation one. In this paper, we will refer to $\mathbf{w}_1$ and $\mathbf{w}_2$ as the canonical vectors (or weights), and we will refer to $\mathbf{X}_1\mathbf{w}_1$ and $\mathbf{X}_2\mathbf{w}_2$ as the canonical variables.

In recent years, CCA has gained popularity as a method for the analysis of genomic data. It has become common for researchers to perform multiple assays on the same set of patient samples; for instance, DNA copy number (or comparative genomic hybridization, CGH), gene expression, and single nucleotide polymorphism (SNP) data might all be available. Examples of studies involving two or more genomic assays on the same set of samples include Hyman et al. (2002), Pollack et al. (2002), Morley et al. (2004), Stranger et al. (2005), and Stranger et al. (2007). In the case of, say, DNA copy number and gene expression measurements on a single set of patient samples, one might wish to perform CCA in order to identify genes whose expression is correlated with regions of genomic gain or loss. However, genomic data is characterized by the fact that the number of features generally greatly exceeds the number of observations; for this reason, CCA cannot be applied directly.

To circumvent this problem, Parkhomenko et al. (2007), Waaijenborg et al. (2008), Parkhomenko et al. (2009), Le Cao et al. (2009), and Witten et al. (2009) have proposed methods for *penalized CCA*. In this paper, we will restrict ourselves to the criterion proposed in Witten et al. (2009), which takes the form

$$\begin{aligned}&\text{maximize}_{\mathbf{w}_1,\mathbf{w}_2} \mathbf{w}_1^T \mathbf{X}_1^T \mathbf{X}_2 \mathbf{w}_2 \\ &\text{subject to } ||\mathbf{w}_1||^2 \le 1, ||\mathbf{w}_2||^2 \le 1, P_1(\mathbf{w}_1) \le c_1, P_2(\mathbf{w}_2) \le c_2\end{aligned} \quad (2)$$

where P_1 and P_2 are convex penalty functions. Since P_1 and P_2 are generally chosen to yield $\mathbf{w}_1$ and $\mathbf{w}_2$ sparse, we call this the *sparse CCA criterion*. This criterion follows from applying penalties to $\mathbf{w}_1$ and $\mathbf{w}_2$ and also from assuming that the covariance matrix of the features is diagonal; that is, we replace $\mathbf{w}_1^T \mathbf{X}_1^T \mathbf{X}_1 \mathbf{w}_1$ and $\mathbf{w}_2^T \mathbf{X}_2^T \mathbf{X}_2 \mathbf{w}_2$ in the CCA criterion with $\mathbf{w}_1^T \mathbf{w}_1$ and

$\mathbf{w}_2^T\mathbf{w}_2$. The sparse CCA criterion results in $\mathbf{w}_1$ and $\mathbf{w}_2$ unique, even when $p_1, p_2 \gg n$, for appropriate choices of P_1 and P_2.

It has been shown that sparse CCA can be used to identify genes that have expression that is correlated with regions of DNA copy number change (Waaijenborg et al. 2008, Witten et al. 2009), to identify genes that have expression that is correlated with SNPs (Parkhomenko et al. 2009), and to identify sets of genes on two different microarray platforms that have correlated expression (Le Cao et al. 2009). However, some questions remain:

1. Sometimes, in addition to data matrices $\mathbf{X}_1 \in \mathbb{R}^{n\times p_1}$ and $\mathbf{X}_2 \in \mathbb{R}^{n\times p_2}$, a vector of outcome measurements in $\mathbb{R}^n$ is also available. For instance, a survival time might be known for each patient. CCA and sparse CCA are *unsupervised* methods; that is, they do not make use of an outcome. However, if outcome measurements are available, then one might seek sets of variables in the two data sets that are correlated with each other and associated with the outcome.

2. More than two sets of variables on the same set of observations might be available. For instance, it is becoming increasingly common for researchers to collect gene expression, SNP, and DNA copy number measurements on the same set of patient samples. In this case, an extension of sparse CCA to the case of more than two data sets is required.

In this paper, we develop extensions to sparse CCA that address these situations and others.

The rest of this paper is organized as follows. Section 2 contains methods for sparse CCA when the data consist of matrices $\mathbf{X}_1$ and $\mathbf{X}_2$. In Section 2.1, we present details of the sparse CCA method from Witten et al. (2009), and in Section 2.2, we explain the connections between that method and those of Waaijenborg et al. (2008), Le Cao et al. (2009), and Parkhomenko et al. (2009). The remainder of Section 2 contains some extensions of sparse CCA for two sets of features on a single set of observations. Section 3 contains an explanation of *sparse multiple CCA*, an extension of sparse CCA to the case of K data sets $\mathbf{X}_1, ..., \mathbf{X}_K$ with features on a single set of samples. In Section 4, we present *sparse supervised CCA*, a method for performing sparse CCA when the data consist of matrices $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{y}$, a vector containing an outcome measurement for each sample. Section 5 contains the discussion. Throughout the paper, methods are applied to the diffuse large B-cell lymphoma (DLBCL) data set of Lenz et al. (2008), which consists of 17350 gene expression measurements and 386165 DNA copy number measurements for 203 patients. For each patient, two clinical outcomes are available: a possibly censored survival time, as well as the subtype of DLBCL to which that patient's disease belongs.

2 Sparse CCA

2.1 The sparse CCA method

The sparse CCA criterion was given in Equation (2) for general penalty functions P_1 and P_2. We will be interested in two specific forms of these penalty functions:

- P_1 is an L_1 (or *lasso*) penalty; that is, $P_1(\mathbf{w}_1) = ||\mathbf{w}_1||_1$. This penalty will result in $\mathbf{w}_1$ sparse for c_1 chosen appropriately. We assume that $1 \leq c_1 \leq \sqrt{p_1}$.

- P_1 is a *fused lasso* penalty (see e.g. Tibshirani et al. 2005), of the form $P_1(\mathbf{w}_1) = \sum_j |w_{1j}| + \sum_j |w_{1j} - w_{1(j-1)}|$. This penalty will result in $\mathbf{w}_1$ sparse and smooth, and is intended for cases in which the features in $\mathbf{X}_1$ have a natural ordering along which smoothness is expected.

In order to indicate the form of the penalties P_1 and P_2 in use, we will refer to the method as sparse CCA(P_1, P_2). That is, if both penalties are L_1, then we will call this sparse CCA(L_1, L_1), and if P_1 is an L_1 penalty and P_2 a fused lasso penalty, then we will call it sparse CCA(L_1, FL) (where "FL" indicates fused lasso). Note that when P_1 and P_2 are L_1 or fused lasso penalties, the resulting canonical vectors are unique, even when $p_1, p_2 \gg n$. Witten et al. (2009) propose the use of sparse CCA(L_1, FL) in the case where $\mathbf{X}_1$ corresponds to gene expression measurements and $\mathbf{X}_2$ corresponds to copy number measurements (ordered by position along the chromosomes); this is related to the proposal of Tibshirani & Wang (2008) for estimating copy number for a single CGH sample.

Now, consider the criterion (2) with P_1 and P_2 convex penalty functions. With $\mathbf{w}_1$ fixed, the criterion is convex in $\mathbf{w}_2$, and with $\mathbf{w}_2$ fixed, it is convex in $\mathbf{w}_1$. The objective function of this *biconvex* criterion increases in each step of a simple iterative algorithm.

Algorithm for sparse CCA:

1. Initialize $\mathbf{w}_2$ to have L_2 norm 1.

2. Iterate the following two steps until convergence:

(a) $\mathbf{w}_1 \leftarrow \arg\max_{\mathbf{w}_1} \mathbf{w}_1^T \mathbf{X}_1^T \mathbf{X}_2 \mathbf{w}_2$ subject to $||\mathbf{w}_1||^2 \leq 1, P_1(\mathbf{w}_1) \leq c_1$.

(b) $\mathbf{w}_2 \leftarrow \arg\max_{\mathbf{w}_2} \mathbf{w}_1^T \mathbf{X}_1^T \mathbf{X}_2 \mathbf{w}_2$ subject to $||\mathbf{w}_2||^2 \leq 1, P_2(\mathbf{w}_2) \leq c_2$.

If P_1 is an L_1 penalty, then the update has the form

$$\mathbf{w}_1 \leftarrow \frac{S(\mathbf{X}_1^T \mathbf{X}_2 \mathbf{w}_2, \Delta_1)}{||S(\mathbf{X}_1^T \mathbf{X}_2 \mathbf{w}_2, \Delta_1)||_2}, \tag{3}$$

where $\Delta_1 = 0$ if this results in $||\mathbf{w}_1||_1 \leq c_1$; otherwise, $\Delta_1 > 0$ is chosen so that $||\mathbf{w}_1||_1 = c_1$. Here, $S(\cdot)$ denotes the soft-thresholding operator; that is, $S(a, c) = \text{sgn}(a)(|a| - c)_+$. Soft-thresholding arises in the update due to the L_1 penalty and the assumption that the covariance matrices are independent. Δ_1 can be chosen by a binary search. If P_1 is instead a fused lasso penalty, then a slightly modified version of the sparse CCA criterion yields the update step

$$\mathbf{w}_1 \leftarrow \text{argmin}_{\mathbf{w}_1}\{\frac{1}{2}||\mathbf{X}_1^T \mathbf{X}_2 \mathbf{w}_2 - \mathbf{w}_1||^2 + \lambda_1 \sum_j |w_{1j}| + \lambda_2 \sum_j |w_{1j} - w_{1(j-1)}|\}, \tag{4}$$

which can be computed using software implementing fused lasso regression. $\mathbf{w}_2$ can be updated analogously.

Methods for selecting tuning parameter values and assessing significance of the resulting canonical vectors are presented in Appendix A. The above algorithm is easily extended to obtain multiple canonical vectors, as described in Witten et al. (2009) and summarized in Appendix B. However, to simplify interpretation of the examples presented in this paper, we will only consider the first canonical vectors $\mathbf{w}_1$ and $\mathbf{w}_2$, as given in the criterion (2).

2.2 Connections with other sparse CCA proposals

This paper extends the sparse CCA proposal of Witten et al. (2009). As mentioned earlier, the Witten et al. (2009) method is closely related to a number of other methods for sparse CCA. We briefly review those methods here.

Waaijenborg et al. (2008) first recast classical CCA as an iterative regression procedure; then an elastic net penalty is applied in order to obtain penalized canonical vectors. An approximation of the iterative elastic net procedure results in an algorithm that is similar to that of Witten et al. (2009) in the case of L_1 penalties on $\mathbf{w}_1$ and $\mathbf{w}_2$. However, Waaijenborg et al. (2008) do not appear to be exactly optimizing a criterion.

Parkhomenko et al. (2009) develop an iterative algorithm for estimating the singular vectors of $\mathbf{X}_1^T\mathbf{X}_2$. At each step, they regularize the estimates of the singular vectors by soft-thresholding. Though they do not explicity state a criterion, it appears that they are approximately optimizing a criterion that is related to (2) with L_1 penalties. However, they use the Lagrange form, rather than the bound form, of the constraints on $\mathbf{w}_1$ and $\mathbf{w}_2$. Their algorithm is closely related to that of Witten et al. (2009), though extra normalization steps are required due to computational problems with the Lagrange form of the constraints. The algorithm of Le Cao et al. (2009) is also closely related to those of Parkhomenko et al. (2009) and Witten et al. (2009), though again Le Cao et al. (2009) use the Lagrange form, rather than the bound form, of the penalties.

Hence, the Waaijenborg et al. (2008), Parkhomenko et al. (2009), Le Cao et al. (2009) and Witten et al. (2009) methods are all closely related; we pursue the criterion (2) in this paper.

2.3 Sparse CCA with nonnegative weights

The sparse CCA method will result in canonical vectors $\mathbf{w}_1$ and $\mathbf{w}_2$ that are sparse, if the penalties P_1 and P_2 are chosen appropriately. However, the nonzero elements of $\mathbf{w}_1$ and $\mathbf{w}_2$ may be of different signs. In some cases, one might seek a sparse weighted average of the features in $\mathbf{X}_1$ that is correlated with a sparse weighted average of the features in $\mathbf{X}_2$. Then one will want to additionally restrict the elements of $\mathbf{w}_1$ and $\mathbf{w}_2$ to be nonnegative (or nonpositive). If we require the elements of $\mathbf{w}_1$ and $\mathbf{w}_2$ to be nonnegative, the sparse CCA criterion becomes

$$\begin{aligned} \text{maximize}_{\mathbf{w}_1,\mathbf{w}_2} \quad & \mathbf{w}_1^T\mathbf{X}_1^T\mathbf{X}_2\mathbf{w}_2 \text{ subject to } ||\mathbf{w}_1||^2 \leq 1, ||\mathbf{w}_2||^2 \leq 1, \\ & w_{1j} \geq 0, w_{2j} \geq 0, P_1(\mathbf{w}_1) \leq c_1, P_2(\mathbf{w}_2) \leq c_2, \end{aligned} \tag{5}$$

and the resulting algorithm is as follows:

Algorithm for sparse CCA with nonnegative weights:

1. Initialize $\mathbf{w}_2$ to have L_2 norm 1.
2. Iterate the following two steps until convergence:
   (a) $\mathbf{w}_1 \leftarrow \arg\max_{\mathbf{w}_1} \mathbf{w}_1^T\mathbf{X}_1^T\mathbf{X}_2\mathbf{w}_2$ subject to $||\mathbf{w}_1||^2 \leq 1, w_{1j} \geq 0, P_1(\mathbf{w}_1) \leq c_1$.

(b) $\mathbf{w}_2 \leftarrow \arg\max_{\mathbf{w}_2} \mathbf{w}_1^T \mathbf{X}_1^T \mathbf{X}_2 \mathbf{w}_2$ subject to $||\mathbf{w}_2||^2 \leq 1, w_{2j} \geq 0, P_2(\mathbf{w}_2) \leq c_2$.

Consider the criterion (5) with $\mathbf{w}_1$ fixed; we can write the optimization problem for $\mathbf{w}_2$ with $\mathbf{X}_2^T \mathbf{X}_1 \mathbf{w}_1 = \mathbf{a}$ as

$$\text{minimize}_{\mathbf{w}_2} - \mathbf{a}^T \mathbf{w}_2 \text{ subject to } ||\mathbf{w}_2||^2 \leq 1, w_{2j} \geq 0, P_2(\mathbf{w}_2) \leq c_2. \tag{6}$$

Assume that P_2 is an L_1 penalty. It is obvious that if $a_j \leq 0$, then $w_{2j} = 0$. For j such that $a_j > 0$, w_{2j} can be found by solving the optimization problem

$$\text{minimize}_{w_{2j}:a_j>0} - \sum_{j:a_j>0} a_j w_{2j} \text{ subject to } \sum_{j:a_j>0} w_{2j}^2 \leq 1, \sum_{j:a_j>0} |w_{2j}| \leq c_2. \tag{7}$$

This can be solved using the following update for $\mathbf{w}_2$:

$$\mathbf{w}_2 \leftarrow \frac{S((\mathbf{X}_2^T \mathbf{X}_1 \mathbf{w}_1)_+, \Delta_2)}{||S((\mathbf{X}_2^T \mathbf{X}_1 \mathbf{w}_1)_+, \Delta_2)||_2}, \tag{8}$$

where $\Delta_2 = 0$ if this results in $||\mathbf{w}_2||_1 \leq c_2$; otherwise, $\Delta_2 > 0$ is chosen so that $||\mathbf{w}_2||_1 = c_2$. An analogous update step can be derived for $\mathbf{w}_1$ if P_1 is an L_1 penalty.

2.4 Application of sparse CCA to the DLBCL data

We demonstrate the sparse CCA method on the lymphoma data set of Lenz et al. (2008), which consists of gene expression and array CGH measurements on 203 patients with DLBCL. The data set is publicly available at `http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE11318`. There are 17350 gene expression measurements and 386165 copy number measurements. (In the raw data set, more gene expression measurements are available. However, we limited the analysis to genes for which we knew the chromosomal location, and we averaged expression measurements for genes for which multiple measurements were available.) For computational reasons, sets of adjacent CGH spots on each chromosome were averaged before all analyses were performed. In previous research, gene expression profiling has been used to define three subtypes of DLBCL, called germinal center B-cell-like (GCB), activated B-cell-like (ABC), and primary mediastinal B-cell lymphoma (PMBL) (Alizadeh et al. 2000, Rosenwald et al. 2002). For each of the 203 observations, survival time and DLBCL subtype are known.

For chromosome i, we performed sparse CCA(L_1, FL) using $\mathbf{X}_1$ equal to expression data of genes on all chromosomes and $\mathbf{X}_2$ equal to DNA copy number data on chromosome i. Tuning parameter values were chosen by permutations; details are given in Appendix A. P-values obtained using the method in Appendix A, as well as the chromosomes on which the genes corresponding to nonzero $\mathbf{w}_1$ weights are located, can be found in Table 1. Canonical vectors found on almost all chromosomes were significant, and for the most part, *cis* interactions were found. Cis interactions are those for which the regions of DNA copy number change and the sets of genes with correlated expression are located on the same chromosome. The presence of cis interactions is not surprising because copy number gain on a given chromosome could naturally result in increased expression of the genes that were gained.

We used the CGH and expression canonical variables as features in a multivariate Cox proportional hazards model to predict survival. Note that $\mathbf{X}_1\mathbf{w}_1$ and $\mathbf{X}_2\mathbf{w}_2$ are vectors in $\mathbb{R}^n$. We also used the canonical variables as features in a multinomial logistic regression to predict cancer subtype. The resulting p-values are shown in Table 1. The Cox proportional hazards models predicting survival from the canonical variables were not significant on most chromosomes. However, on many chromosomes, the canonical variables were highly predictive of DLBCL subtype. Boxplots showing the canonical variables as a function of DLBCL subtype are displayed in Figure 1 for chromosomes 6 and 9. For chromosome 9, Figure 2 shows $\mathbf{w}_2$, the canonical vector corresponding to copy number, as well as the raw copy number for the samples with largest and smallest (absolute) value in the canonical variable for the CGH data. It is not surprising that there are many significant p-values for the prediction of cancer subtype in Table 1, since the subtypes are defined using gene expression, and it was found in Lenz et al. (2008) that the subtypes are characterized by regions of copy number change.

We can also compare the sparse CCA canonical variables obtained on the DLBCL data to the first principal components that arise if principal components analysis (PCA) is performed separately on the expression data and on the copy number data. PCA and sparse CCA were performed using all of the gene expression data, and the CGH data on chromosome 3; Figure 3 shows the resulting canonical variables and principal components. Sparse CCA results in CGH and expression canonical variables that are highly correlated with each other, due to the form of the sparse CCA criterion. PCA results in principal components that are far less correlated with each other, and, as a result, may yield better separation between the three subtypes. But PCA does not allow for an integrated interpretation of the expression and CGH data together.

In this section, we assessed the association between the canonical variables

Chr.	P-Value	Chr. of Genes w/Nonzero Weights	P-Value w/Surv.	P-Value w/Subtype
1	0	1 1 1	0.346978	0.020115
2	0	2 2	0.139221	0.000336
3	0	3 3 3	0.000395	0
4	0	4 4	0.473788	0.127702
5	0	5 5 5 5 5	0.607717	0.015423
6	0	6 6 6	0.421753	6e-05
7	0	7 7	0.530759	0
8	0	8 8	0.698352	0.000375
9	0	9 9 9 9 9 9 9 9	0.435872	0
10	0	10 10 10 10 10	0.123747	7e-06
11	0	11 11 11	0.344889	0.000265
12	0	12 12	0.436956	0.000557
14	0.02	1 14	0.006953	0
15	0	15 15 15 15 15 15 15 15	0.000552	3.5e-05
16	0	16 16	0.298069	0.029809
17	0	17 17 17 17 17 17 17	0.047078	0.775681
18	0	18 18	0.00619	5e-06
19	0	19 19 10	0.046235	0
20	0	20 20 20 2 3 20 20 20	0.922793	0.004198
21	0	21 21	0.309212	0.012914
22	0.06	22 1 1	0.441799	0.000272

Table 1: **Column 1**: *The number indicates the chromosome to which the CGH data corresponds. Sparse CCA was performed using all gene expression measurements, and CGH data from chromosome* i *only.* **Column 2**: *In almost every case, the canonical vectors found were highly significant. P-values were obtained using the permutation approach given in Appendix A.* **Column 3**: *For the most part, CGH measurements on chromosome* i *were found to be correlated with the expression of sets of genes on chromosome* i*. That is, the canonical vectors found for the expression data had non-zero elements only for genes on chromosome* i*.* **Columns 4 and 5**: *P-values are reported for the Cox proportional hazards and multinomial logistic regression models that use the canonical variables to predict survival and cancer subtype.*

found using sparse CCA and the clinical outcomes in order to determine if the results of sparse CCA have biological significance. However, in general, if a clinical outcome of interest is available, then the sparse sCCA approach of Section 4 may be appropriate.

2.5 Connection with nearest shrunken centroids

Consider now a new setting in which we have n observations on p features, and each observation belongs to one of two classes. Let $\mathbf{X}_1$ denote the $n \times p$ matrix

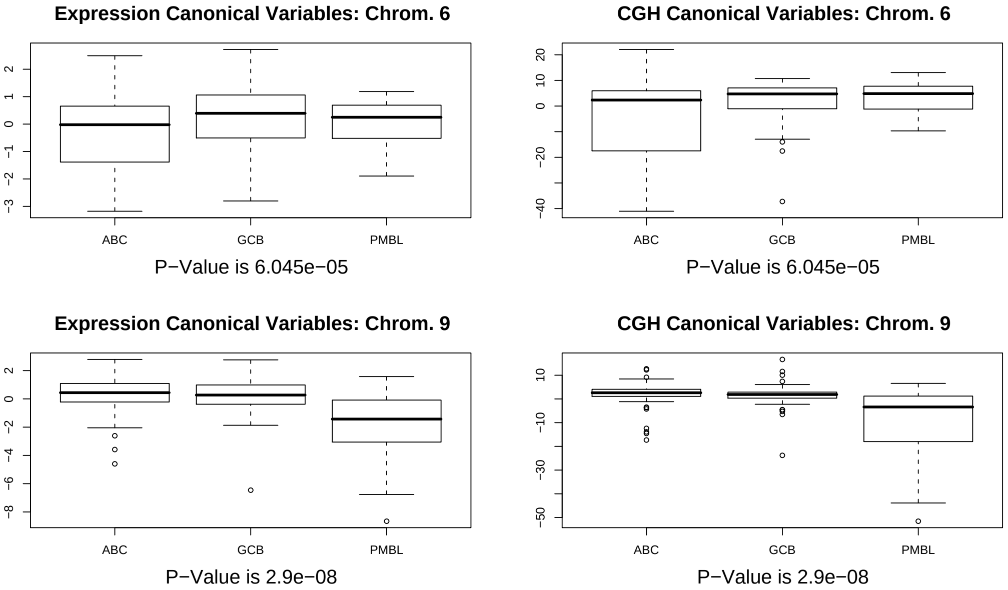

Figure 1: *Sparse CCA was performed using CGH data on a single chromosome and all gene expression measurements. For chromosomes 6 and 9, the gene expression and CGH canonical variables, stratified by cancer subtype, are shown. It is clear that the values of the canonical variables differ by subtype. P-values reported are replicated from Table 1; they reflect the extent to which the canonical variables predict cancer subtype in a multinomial logistic regression model.*

of observations by features, and let $\mathbf{X}_2$ be a binary $n \times 1$ matrix indicating class membership of each observation of $\mathbf{X}_1$. In this section, we will show that sparse CCA applied to $\mathbf{X}_1$ and $\mathbf{X}_2$ yields a canonical vector $\mathbf{w}_1$ that is closely related to the nearest shrunken centroids solution (NSC, Tibshirani et al. 2002, Tibshirani et al. 2003).

Assume that each column of $\mathbf{X}_1$ has been standardized to have mean zero and pooled within-class standard deviation equal to one. NSC is a high-dimensional classification method that involves defining "shrunken" class centroids based on only a subset of the features; each test set observation is then classified to the nearest shrunken centroid. We first explain the NSC method, applied to data $\mathbf{X}_1$. For $1 \leq k \leq 2$, we define vectors $\mathbf{d}_k, \mathbf{d}'_k, \overline{\mathbf{X}}'_{1k} \in \mathbb{R}^p$ as follows:

$$\mathbf{d}_k = \frac{\overline{\mathbf{X}}_{1k}}{m_k}, \quad \mathbf{d}'_k = S(\mathbf{d}_k, \delta), \quad \overline{\mathbf{X}}'_{1k} = m_k \mathbf{d}'_k. \tag{9}$$

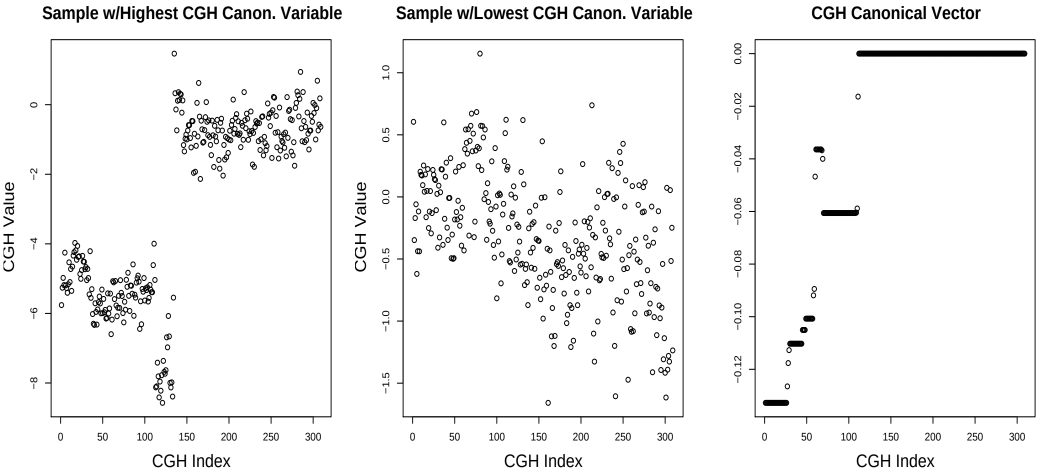

Figure 2: *Sparse CCA was performed using CGH data on chromosome 9, and all gene expression measurements. The samples with the highest and lowest absolute values in the CGH canonical variable are shown, along with the canonical vector corresponding to the CGH data. As expected, the sample with the highest CGH canonical variable is highly correlated with the CGH canonical vector, and the sample with the lowest CGH canonical variable shows little correlation. The sample with highest CGH canonical variable is of subtype PMBL, and the sample with lowest canonical variable is of subtype ABC. The CGH data on chromosome 9 consists of 309 features, of which 111 have non-zero weights in the right-hand panel.*

Here, $\overline{\mathbf{X}}_{1k}$ is the mean vector of the features in $\mathbf{X}_1$ over the observations in class k, and $m_k = \sqrt{\frac{1}{n_k} - \frac{1}{n}}$ where n_k is the number of observations in class k. $\overline{\mathbf{X}}'_{1k}$ is the shrunken centroid for class k obtained using tuning parameter $\delta \geq 0$. As in Section 2.1, S is the soft-thresholding operator.

Now, consider the effect of applying sparse CCA with L_1 penalties to data $\mathbf{X}_1$ and $\mathbf{X}_2$. Rescale $\mathbf{X}_2$ so that the class 1 values are $\frac{1}{n_1}$ and the class 2 values are $-\frac{1}{n_2}$. The sparse CCA criterion is

$$\begin{aligned}&\text{maximize}_{\mathbf{w}_1,\mathbf{w}_2} \mathbf{w}_1^T \mathbf{X}_1^T \mathbf{X}_2 \mathbf{w}_2 \\ &\text{subject to } ||\mathbf{w}_1||^2 \leq 1, ||\mathbf{w}_2||^2 \leq 1, ||\mathbf{w}_1||_1 \leq c_1, ||\mathbf{w}_2||_1 \leq c_2.\end{aligned} \tag{10}$$

Since $\mathbf{w}_2 \in \mathbb{R}^1$, the constraints on its value result in $\mathbf{w}_2 = 1$. The criterion reduces to

$$\text{maximize}_{\mathbf{w}_1} \mathbf{w}_1^T \mathbf{X}_1^T \mathbf{X}_2 \text{ subject to } ||\mathbf{w}_1||^2 \leq 1, ||\mathbf{w}_1||_1 \leq c_1, \tag{11}$$

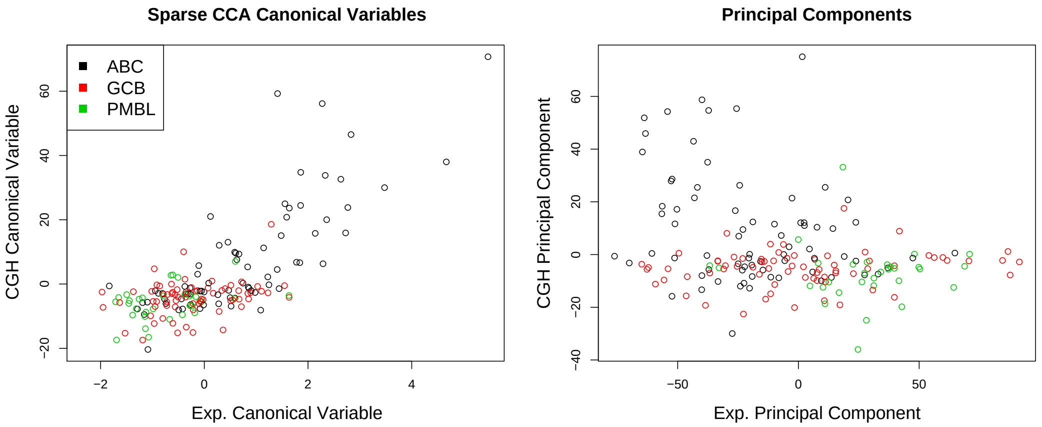

Figure 3: *Sparse CCA and PCA were performed using CGH data on chromosome 3, and all gene expression measurements. The resulting canonical variables and principal components are shown. The CGH and expression canonical variables are highly correlated with each other. Both sparse CCA and PCA result in some separation between the three DLBCL subtypes, although PCA results in better separation because the first principal components of the CGH and expression data are less correlated with each other.*

which can be rewritten as

$$\text{maximize}_{\mathbf{w}_1} (\overline{\mathbf{X}}_{11} - \overline{\mathbf{X}}_{12})^T \mathbf{w}_1 \text{ subject to } ||\mathbf{w}_1||^2 \leq 1, ||\mathbf{w}_1||_1 \leq c_1. \tag{12}$$

The solution to (12) is

$$\mathbf{w}_1 = \frac{S(\overline{\mathbf{X}}_{11} - \overline{\mathbf{X}}_{12}, \Delta)}{||S(\overline{\mathbf{X}}_{11} - \overline{\mathbf{X}}_{12}, \Delta)||_2} = \frac{S((1 + \frac{n_1}{n_2})\overline{\mathbf{X}}_{11}, \Delta)}{||S((1 + \frac{n_1}{n_2})\overline{\mathbf{X}}_{11}, \Delta)||_2} \tag{13}$$

where $\Delta = 0$ if that results in $||\mathbf{w}_1||_1 \leq c_1$; otherwise, $\Delta > 0$ is chosen so that $||\mathbf{w}_1||_1 = c_1$. So sparse CCA yields a canonical vector that is proportional to the shrunken centroid $\overline{\mathbf{X}}'_{11}$ when the tuning parameters for NSC and sparse CCA are chosen appropriately.

3 Sparse multiple CCA

3.1 The sparse multiple CCA method

CCA and sparse CCA can be used to perform an integrative analysis of two data sets with features on a single set of samples. But what if more than two such data sets are available? A number of approaches for generalizing CCA to more than two data sets have been proposed in the literature, and some of these extensions are summarized in Gifi (1990). We will focus on one of these proposals for multiple-set CCA.

Let the K data sets be denoted $\mathbf{X}_1, ..., \mathbf{X}_K$; data set k contains p_k variables, and each variable has mean zero and standard deviation one as in previous sections. Then, the single-factor multiple-set CCA criterion involves finding $\mathbf{w}_1, ..., \mathbf{w}_K$ that maximize

$$\sum_{i<j} \mathbf{w}_i^T \mathbf{X}_i^T \mathbf{X}_j \mathbf{w}_j \text{ subject to } \mathbf{w}_k^T \mathbf{X}_k^T \mathbf{X}_k \mathbf{w}_k = 1 \; \forall k, \tag{14}$$

where $\mathbf{w}_k \in \mathbb{R}^{p_k}$. It is easy to see that when $K = 2$, then multiple-set CCA simplifies to ordinary CCA. We can develop a method for sparse multiple CCA by imposing sparsity constraints on this natural formulation for multiple-set CCA. In the spirit of our criterion for sparse CCA with two sets of variables (2), we assume that the features are independent within each data set: that is, $\mathbf{X}_k^T \mathbf{X}_k = \mathbf{I}$ for each k. Then, our criterion for *sparse multiple CCA* (sparse mCCA) is as follows:

$$\text{maximize}_{\mathbf{w}_1,...,\mathbf{w}_K} \sum_{i<j} \mathbf{w}_i^T \mathbf{X}_i^T \mathbf{X}_j \mathbf{w}_j \text{ subject to } ||\mathbf{w}_i||^2 \leq 1, P_i(\mathbf{w}_i) \leq c_i \; \forall i, \tag{15}$$

where P_i are convex penalty functions. Then, $\mathbf{w}_i$ is the canonical vector associated with $\mathbf{X}_i$. If P_i is an L_1 or fused lasso penalty and c_i is chosen appropriately, then $\mathbf{w}_i$ will be sparse.

It is not hard to see that just as (2) is *biconvex* in $\mathbf{w}_1$ and $\mathbf{w}_2$, (15) is *multiconvex* in $\mathbf{w}_1, ..., \mathbf{w}_K$. That is, with $\mathbf{w}_j$ held fixed for all $j \neq i$, (15) is convex in $\mathbf{w}_i$. This suggests an iterative algorithm that increases the objective function of (15) at each iteration.

Algorithm for sparse mCCA:

1. For each i, fix an initial value of $\mathbf{w}_i \in \mathbb{R}^{p_k}$.

2. Repeat until convergence: For each i, let

$$\mathbf{w}_i \leftarrow \text{argmax}_{\mathbf{w}_i} \mathbf{w}_i^T \mathbf{X}_i^T (\sum_{j \neq i} \mathbf{X}_j \mathbf{w}_j) \text{ subject to } ||\mathbf{w}_i||^2 \leq 1, P_i(\mathbf{w}_i) \leq c_i. \tag{16}$$

For instance, if P_i is an L_1 penalty, then the update takes the form

$$\mathbf{w}_i \leftarrow \frac{S(\mathbf{X}_i^T(\sum_{j\neq i}\mathbf{X}_j\mathbf{w}_j), \Delta_i)}{||S(\mathbf{X}_i^T(\sum_{j\neq i}\mathbf{X}_j\mathbf{w}_j), \Delta_i)||_2}, \tag{17}$$

where $\Delta_i = 0$ if this results in $||\mathbf{w}_i||_1 \leq c_i$; otherwise, $\Delta_i > 0$ is chosen such that $||\mathbf{w}_i||_1 = c_i$.

We demonstrate the performance of sparse mCCA on a simple simulated example. Data were generated according to the model

$$\mathbf{X}_i = \mathbf{u}\mathbf{w}_i^T + \epsilon_i, 1 \leq i \leq 3 \tag{18}$$

where $\mathbf{u} \in \mathbb{R}^{50}$, $\mathbf{w}_1 \in \mathbb{R}^{100}$, $\mathbf{w}_2 \in \mathbb{R}^{200}$, $\mathbf{w}_3 \in \mathbb{R}^{300}$. Only the first 20, 40, and 60 elements of $\mathbf{w}_1$, $\mathbf{w}_2$, and $\mathbf{w}_3$ were nonzero, respectively. Sparse mCCA was run on this data, and the resulting estimates of $\mathbf{w}_1$, $\mathbf{w}_2$, and $\mathbf{w}_3$ are shown in Figure 4.

A permutation algorithm for selecting tuning parameter values and assessing significance of sparse mCCA can be found in Appendix A. In addition, an algorithm for obtaining multiple sparse mCCA factors is given in Appendix B.

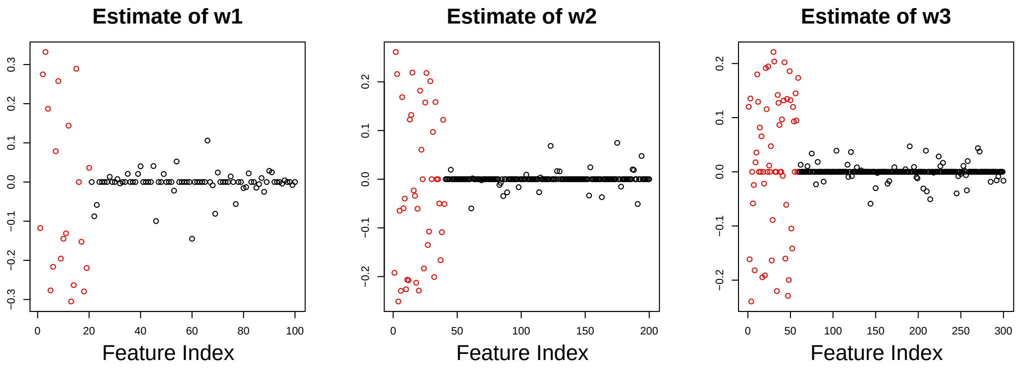

Figure 4: *Three data sets* $\mathbf{X}_1$, $\mathbf{X}_2$, *and* $\mathbf{X}_3$ *are generated under a simple model, and sparse mCCA is performed. The resulting estimates of* $\mathbf{w}_1$, $\mathbf{w}_2$, *and* $\mathbf{w}_3$ *are fairly accurate at distinguishing between the elements of* $\mathbf{w}_i$ *that are truly nonzero (red) and those that are not (black). From left to right, the three canonical vectors shown have 57, 67, and 92 nonzero elements.*

3.2 Application of sparse mCCA to DLBCL copy number data

If CGH measurements are available on a set of patient samples, then one may wonder whether copy number changes in genomic regions on separate chromosomes are correlated. For instance, certain genomic regions may tend to be coamplified or codeleted.

To answer this question for a single pair of chromosomes, we can perform sparse CCA(FL, FL) with two data sets, $\mathbf{X}_1$ and $\mathbf{X}_2$, where $\mathbf{X}_1$ contains the CGH measurements on the first chromosome of interest and $\mathbf{X}_2$ contains the CGH measurements on the second chromosome of interest. If copy number change on the first chromosome is independent of copy number change on the second chromosome, then we expect the corresponding p-value obtained using the method of Appendix A not to be small. A small p-value indicates that copy number changes on the two chromosomes are more correlated with each other than one would expect due to chance. However, in general, there are $\binom{24}{2}$ pairs of chromosomes that can be tested for correlated patterns of amplification and deletion; this leads to a multiple testing problem and excessive computation. Instead, we take a different approach, using sparse mCCA. We apply sparse mCCA to data sets $\mathbf{X}_1, ..., \mathbf{X}_{24}$, where $\mathbf{X}_i$ contains the CGH measurements on chromosome i. A fused lasso penalty is used on each data set. The goal is to identify correlated regions of gain and loss across the entire genome.

This method is applied to the DLBCL data set mentioned previously. We first denoise the samples by applying the fused lasso to each sample individually, as in Tibshirani & Wang (2008). We then perform sparse mCCA on the resulting smoothed CGH data. The canonical vectors that result are shown in Figure 5. From the figure, one can conclude that complex patterns of gain and loss tend to co-occur. It is unlikely that a single sample would display the entire pattern found; however, samples with large values in the canonical variables most likely contain some of the patterns shown in the figure.

4 Sparse supervised CCA

In Section 2.4, we determined that on the lymphoma data, many of the canonical variables obtained using sparse CCA are highly associated with tumor subtype (and for some chromosomes, the canonical variables are also associated with survival time). Though an outcome was available, we took an unsupervised approach in performing sparse CCA. In this section, we will develop a framework to directly make use of an outcome in sparse CCA. Our method for *sparse supervised CCA* (sparse sCCA) is closely related to the *supervised principal components analysis* (supervised PCA) method of Bair & Tibshirani (2004) and Bair et al. (2006), and so we begin with an overview of that method.

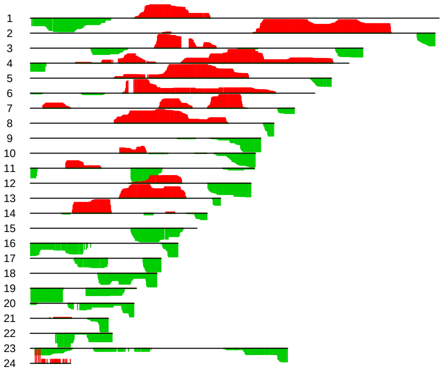

Figure 5: *Sparse mCCA was performed on the DLBCL copy number data, treating each chromosome as a separate "data set", in order to identify genomic regions that are coamplified and/or codeleted. The canonical vectors* $\mathbf{w}_1, ..., \mathbf{w}_{24}$ *are shown. Positive values of the canonical vectors are shown in red, and negative values are in green.*

4.1 Supervised PCA

Principal components regression (PCR; see e.g. Massy 1965) is a method for predicting an outcome $\mathbf{y} \in \mathbb{R}^n$ from a data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$. Assume that the columns of $\mathbf{X}$ have been standardized. Then, PCR involves regressing $\mathbf{y}$ onto the first few columns of $\mathbf{XV}$, where $\mathbf{X} = \mathbf{UDV}^T$ is the singular value decomposition of $\mathbf{X}$. Since $\mathbf{V}$ is estimated in an unsupervised manner, it is not guaranteed that the first few columns of $\mathbf{XV}$ will predict $\mathbf{y}$ well, even if some of the features in $\mathbf{X}$ are correlated with $\mathbf{y}$.

Bair & Tibshirani (2004) and Bair et al. (2006) propose the use of supervised PCA, which is a semisupervised approach. Their method can be described simply:

1. On training data, the features that are most associated with the outcome $\mathbf{y}$ are identified.

2. PCR is performed using only the features identified in the previous step.

Theoretical results regarding the performance of this method under a latent variable model are presented in Bair et al. (2006).

4.2 The sparse supervised CCA method

Suppose that a quantitative outcome is available; that is, we have $\mathbf{y} \in \mathbb{R}^n$ in addition to $\mathbf{X}_1$ and $\mathbf{X}_2$. Then we might seek linear combinations of the variables in $\mathbf{X}_1$ and $\mathbf{X}_2$ that are highly correlated with each other and associated with the outcome.

We define the criterion for *supervised CCA* as follows:

$$\begin{aligned} \text{maximize}_{\mathbf{w}_1,\mathbf{w}_2} \quad & \mathbf{w}_1^T\mathbf{X}_1^T\mathbf{X}_2\mathbf{w}_2 \text{ subject to } ||\mathbf{w}_1||^2 \leq 1, ||\mathbf{w}_2||^2 \leq 1, \\ & w_{1j} = 0 \; \forall j \notin Q_1, w_{2j} = 0 \; \forall j \notin Q_2 \end{aligned} \tag{19}$$

where Q_1 is the set of features in $\mathbf{X}_1$ that are most correlated with $\mathbf{y}$, and Q_2 is the set of features in $\mathbf{X}_2$ that are most correlated with $\mathbf{y}$. The number of features in Q_1 and Q_2, or alternatively the correlation threshold for features to enter Q_1 and Q_2, can be treated as a tuning parameter or can simply be fixed. If $\mathbf{X}_1 = \mathbf{X}_2$, then the criterion (19) simplifies to supervised PCA; that is, $\mathbf{w}_1$ and $\mathbf{w}_2$ are equal to each other and to the first principal component of the subset of the data containing only the features that are most associated with the outcome.

sCCA can be easily extended to give sparse sCCA,

$$\begin{aligned} \text{maximize}_{\mathbf{w}_1,\mathbf{w}_2} \quad & \mathbf{w}_1^T\mathbf{X}_1^T\mathbf{X}_2\mathbf{w}_2 \text{ subject to } ||\mathbf{w}_1||^2 \leq 1, ||\mathbf{w}_2||^2 \leq 1, P_1(\mathbf{w}_1) \leq c_1, \\ & P_2(\mathbf{w}_2) \leq c_2, w_{1j} = 0 \; \forall j \notin Q_1, w_{2j} = 0 \; \forall j \notin Q_2, \end{aligned} \tag{20}$$

where as usual, P_1 and P_2 are convex penalty functions.

We have discussed the possibility of $\mathbf{y}$ being a quantitative outcome (e.g. tumor diameter), but other options exist as well. For instance, $\mathbf{y}$ could be a time to event (e.g. a possibly censored survival time) or a class label (for instance, DLBCL subtype). Our definition of sparse sCCA must be generalized in order to accommodate other outcome types. If $\mathbf{y}$ is a survival time, then for each feature, we can compute the score statistic (or Cox score) for the univariate Cox proportional hazards model that uses that feature to predict the outcome. Only features with sufficiently high (absolute) Cox scores will be in the sets Q_1 and Q_2. In the case of a multiple class outcome, only features with a sufficiently high F-statistic for a one-way ANOVA will be in Q_1 and Q_2. Other outcome types could be incorporated in an analogous way. The algorithm for sparse sCCA can be written as follows:

Algorithm for sparse sCCA:

1. Let $\tilde{\mathbf{X}}_1$ and $\tilde{\mathbf{X}}_2$ denote the submatrices of $\mathbf{X}_1$ and $\mathbf{X}_2$ consisting of the features in Q_1 and Q_2. Q_1 and Q_2 are calculated as follows:

(a) In the case of an L_1 penalty on $\mathbf{w}_i$, Q_i is the set of indices of the features in $\mathbf{X}_i$ that have highest univariate association with the outcome.

(b) In the case of a fused lasso penalty on $\mathbf{w}_i$, the vector of univariate associations between the features in $\mathbf{X}_i$ and the outcome is smoothed using the fused lasso. The resulting smoothed vector is thresholded to obtain the desired number of nonzero cofficients. Q_i contains the indices of the coefficients that are nonzero after thresholding.

2. Perform sparse CCA using data $\tilde{\mathbf{X}}_1$ and $\tilde{\mathbf{X}}_2$.

Note that the fused lasso case is treated specially because one wishes for the features included in $\tilde{\mathbf{X}}_i$ to be contiguous, so that smoothness in the resulting $\mathbf{w}_i$ weights will translate to smoothness in the weights of the original variable set. Algorithms for tuning parameter selection and assessment of significance, as well as a method for obtaining multiple canonical vectors, are given in the Appendix.

We explore the performance of sparse sCCA with a quantitative outcome on a toy example. Data are generated according to the model

$$\mathbf{X}_1 = \mathbf{u}\mathbf{w}_1^T + \epsilon_1, \quad \mathbf{X}_2 = \mathbf{u}\mathbf{w}_2^T + \epsilon_2, \quad \mathbf{y} = \mathbf{u}, \tag{21}$$

with $\mathbf{u} \in \mathbb{R}^{50}$, $\mathbf{w}_1 \in \mathbb{R}^{500}$, $\mathbf{w}_2 \in \mathbb{R}^{1000}$, $\epsilon_1 \in \mathbb{R}^{50\times500}$, $\epsilon_2 \in \mathbb{R}^{50\times1000}$. 50 elements of $\mathbf{w}_1$ and 100 elements of $\mathbf{w}_2$ are non-zero. The first canonical vectors of sparse CCA and sparse sCCA (using L_1 penalties) were computed for a range of values of c_1 and c_2. In Figure 6, the resulting number of true positives (features that are nonzero in $\mathbf{w}_1$ and $\mathbf{w}_2$ and also in the estimated canonical vectors) are shown on the y-axis, as a function of the number of nonzero elements of the canonical vectors. It is clear that greater numbers of true positives are obtained when the outcome is used. In Figure 7, the canonical variables obtained using sparse CCA and sparse sCCA are plotted against the outcome. The canonical variables obtained using sparse sCCA are correlated with the outcome, and those obtained using sparse CCA are not. Note that under the model (21), in the absence of noise, the canonical variables are proportional to $\mathbf{u}$; therefore, sparse sCCA more accurately uncovers the true canonical variables.

In theory, one could choose Q_1 and Q_2 in Step 1 of the sparse sCCA algorithm to contain fewer than n features; then, ordinary CCA could be performed instead of sparse CCA in Step 2. However, we recommend using a less stringent cutoff for Q_1 and Q_2 in Step 1, and instead performing further feature selection in Step 2 via sparse CCA.

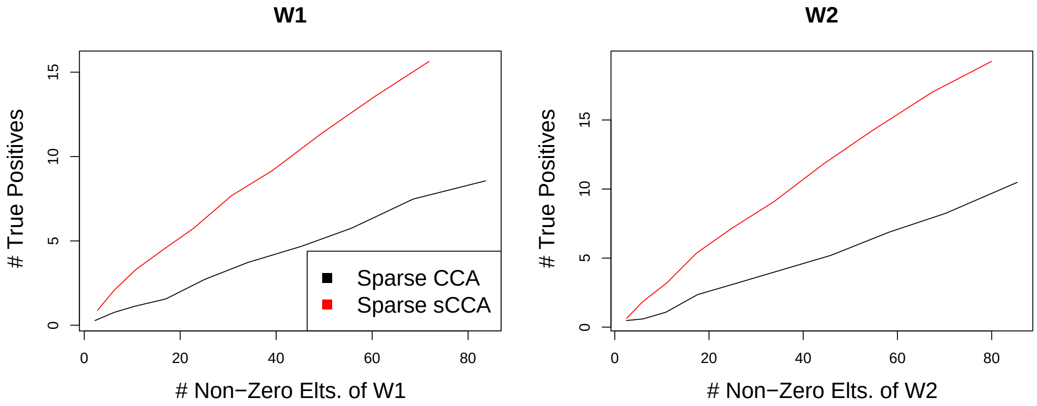

Figure 6: *Sparse CCA and sparse sCCA were performed on a toy example, for a range of values of the tuning parameters in the sparse CCA criterion. The number of true positives in* $\mathbf{w}_1$ *and* $\mathbf{w}_2$ *is shown as a function of the number of nonzero elements in the estimates of the canonical vectors.*

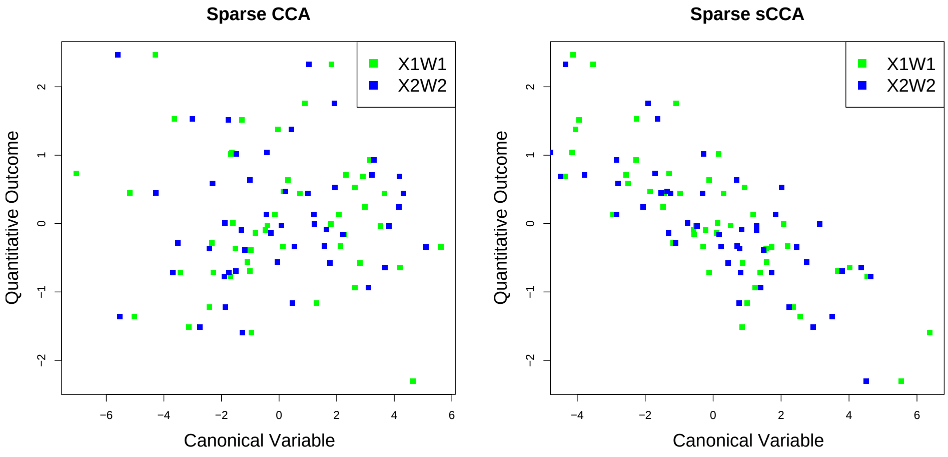

Figure 7: *Sparse CCA and sparse sCCA were performed on a toy example. The canonical variables obtained using sparse sCCA are highly correlated with the outcome; those obtained using sparse CCA are not.*

4.3 Connection with sparse mCCA

Given $\mathbf{X}_1$, $\mathbf{X}_2$, and a two-class outcome $\mathbf{y}$, one could perform sparse mCCA by treating $\mathbf{y}$ as a third data set. This would yield a different but related method for performing sparse sCCA in the case of a two-class outcome.

Note that the outcome $\mathbf{y}$ is a matrix in $\mathbb{R}^{n\times 1}$. We code the two classes (of n_1 and n_2 observations, respectively) as $\frac{\lambda}{n_1}$ and $-\frac{\lambda}{n_2}$. Assume that the columns of $\mathbf{X}_1$ and $\mathbf{X}_2$ have mean zero and pooled within-class standard deviation equal to one. Consider the sparse mCCA criterion with L_1 penalties, applied to data sets $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{y}$:

$$\begin{aligned}\text{maximize}_{\mathbf{w}_1,\mathbf{w}_2,\mathbf{w}_3} \quad & \mathbf{w}_1^T\mathbf{X}_1^T\mathbf{X}_2\mathbf{w}_2 + \mathbf{w}_1^T\mathbf{X}_1^T\mathbf{y}\mathbf{w}_3 + \mathbf{w}_2^T\mathbf{X}_2^T\mathbf{y}\mathbf{w}_3 \\ & \text{subject to } ||\mathbf{w}_i||^2 \le 1, ||\mathbf{w}_i||_1 \le c_i \ \forall i. \end{aligned} \tag{22}$$

Note that since $\mathbf{w}_3 \in \mathbb{R}^1$, it follows that $\mathbf{w}_3 = 1$. So we can re-write the criterion (22) as

$$\begin{aligned}&\text{maximize}_{\mathbf{w}_1,\mathbf{w}_2}\mathbf{w}_1^T\mathbf{X}_1^T\mathbf{X}_2\mathbf{w}_2 + \mathbf{w}_1^T\mathbf{X}_1^T\mathbf{y} + \mathbf{w}_2^T\mathbf{X}_2^T\mathbf{y} \\ &\text{subject to } ||\mathbf{w}_1||^2 \le 1, ||\mathbf{w}_2||^2 \le 1, ||\mathbf{w}_1||_1 \le c_1, ||\mathbf{w}_2||_1 \le c_2. \end{aligned} \tag{23}$$

Now, this criterion is biconvex and leads naturally to an iterative algorithm. However, this is not the approach that we take with our sparse sCCA method. Instead, notice that

$$\mathbf{w}_1^T\mathbf{X}_1^T\mathbf{y} = \lambda(\overline{\mathbf{X}}_{11} - \overline{\mathbf{X}}_{12})^T\mathbf{w}_1 = \lambda\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\mathbf{t}_1^T\mathbf{w}_1, \tag{24}$$

where $\overline{\mathbf{X}}_{1k} \in \mathbb{R}^p$ is the mean vector of the observations in $\mathbf{X}_1$ that belong to class k, and where $\mathbf{t}_1 \in \mathbb{R}^p$ is the vector of two-sample t-statistics testing whether the classes defined by $\mathbf{y}$ have equal means within each feature of $\mathbf{X}_1$. Similarly, $\mathbf{w}_2^T\mathbf{X}_2^T\mathbf{y} = \lambda\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\mathbf{t}_2^T\mathbf{w}_2$ for $\mathbf{t}_2$ defined analogously. So we can rewrite (23) as

$$\begin{aligned}&\text{maximize}_{\mathbf{w}_1,\mathbf{w}_2}\mathbf{w}_1^T\mathbf{X}_1^T\mathbf{X}_2\mathbf{w}_2 + \lambda\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}(\mathbf{t}_1^T\mathbf{w}_1 + \mathbf{t}_2^T\mathbf{w}_2) \\ &\text{subject to } ||\mathbf{w}_1||^2 \le 1, ||\mathbf{w}_2||^2 \le 1, ||\mathbf{w}_1||_1 \le c_1, ||\mathbf{w}_2||_1 \le c_2. \end{aligned} \tag{25}$$

As λ increases, the elements of $\mathbf{w}_1$ and $\mathbf{w}_2$ that correspond to large $|\mathbf{t}_1|$ and $|\mathbf{t}_2|$ values increase in absolute value relative to those that correspond to smaller $|\mathbf{t}_1|$ and $|\mathbf{t}_2|$ values.

Rather than adopting the criterion (25) for sparse sCCA, our sparse sCCA criterion results from assigning nonzero weights only to the elements of $\mathbf{w}_1$ and

$\mathbf{w}_2$ corresponding to large $|\mathbf{t}_1|$ and $|\mathbf{t}_2|$. We prefer our proposed sparse sCCA algorithm because it is simple, generalizes to the supervised PCA method when $\mathbf{X}_1 = \mathbf{X}_2$, and extends easily to non-binary outcomes.

4.4 Application of sparse sCCA to DLBCL data

We evaluate the performance of sparse sCCA on the DLBCL data set, in terms of the association of the resulting canonical variables with the survival and subtype outcomes. We repeatedly split the observations into training and test sets (75% / 25%). Let $(\mathbf{X}_1^{train}, \mathbf{X}_2^{train}, \mathbf{y}^{train})$ denote the training data, and let $(\mathbf{X}_1^{test}, \mathbf{X}_2^{test}, \mathbf{y}^{test})$ denote the test data. ($\mathbf{y}$ can denote either the survival time or the cancer subtype.) We perform sparse sCCA analysis on the training data. As in Section 2.4, for each chromosome, sparse sCCA is run using CGH measurements on that chromosome, and all available gene expression measurements. An L_1 penalty is applied to the expression data, and a fused lasso penalty is applied to the CGH data. Let $\mathbf{w}_1^{train}, \mathbf{w}_2^{train}$ denote the canonical vectors obtained. We then use $\mathbf{X}_1^{test}\mathbf{w}_1^{train}$ and $\mathbf{X}_2^{test}\mathbf{w}_2^{train}$ as features in a Cox proportional hazards model or a multinomial logistic regression model to predict $\mathbf{y}^{test}$. The resulting p-values are shown in Figure 8 for both the survival and subtype outcomes; these are compared to the results obtained if the analysis is repeated using unsupervised sparse CCA on the training data. On the whole, for the subtype outcome, the p-values obtained using sparse sCCA are much smaller than those obtained using sparse CCA. The canonical variables obtained using sparse CCA and sparse sCCA with the survival outcome are not significantly associated with survival. In this example, sparse CCA was performed so that 20% of the features in $\mathbf{X}_1$ and $\mathbf{X}_2$ were contained in Q_1 and Q_2 in the sparse sCCA algorithm.

5 Discussion

As it becomes more commonplace for biomedical researchers to perform multiple assays on the same set of patient samples, methods for the integrative analysis of two or more high-dimensional data sets will become increasingly important. The sparse CCA methods previously proposed in the literature (Parkhomenko et al. 2007, Waaijenborg et al. 2008, Parkhomenko et al. 2009, Le Cao et al. 2009, Witten et al. 2009) provide an attractive framework for performing an integrative analysis of two data sets. In this paper, we have developed extensions to sparse CCA that can be used to apply the method to the case of more than two data sets, and to incorporate an

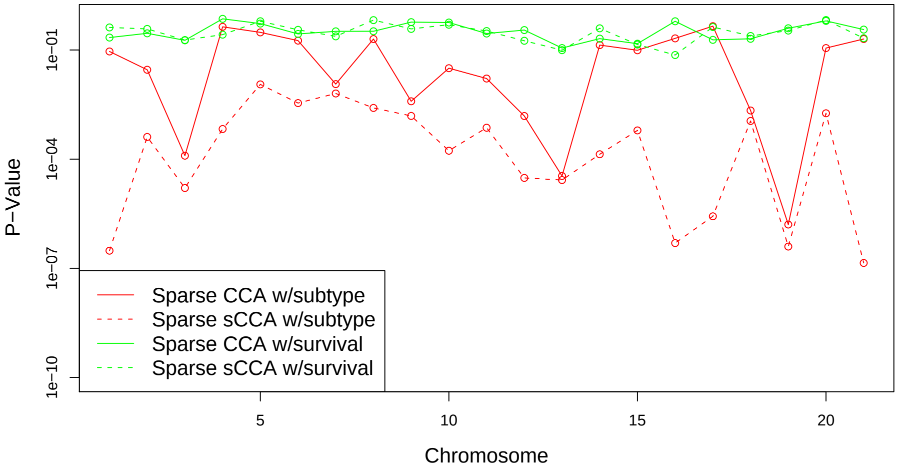

Figure 8: *On a training set, sparse CCA and sparse sCCA were performed using CGH measurements on a single chromosome, and all available gene expression measurements. The resulting test set canonical variables were used to predict survival time and DLBCL subtype. Median p-values (over training set / test set splits) are shown.*

outcome into the analysis.

The methods proposed in this paper will be available on CRAN as part of the `PMA` (Penalized Multivariate Analysis) package.

APPENDIX

A Tuning parameter selection and calculation of p-values

We first present a permutation-based algorithm for selection of tuning parameters and calculation of p-values for sparse CCA. Note that a number of methods have been proposed in the literature for selecting tuning parameters for sparse CCA (see e.g. Waaijenborg et al. 2008, Parkhomenko et al. 2009, Witten et al. 2009). The method proposed here has the advantage over the proposals of Waaijenborg et al. (2008) and Parkhomenko et al. (2009) that it does not require splitting a possibly small set of samples into a training set and a test set. Witten et al. (2009) present a method for tuning parameter selection for their penalized matrix decomposition; however, it does not extend in a straightforward way to the sparse CCA case.

Algorithm to select tuning parameters and determine significance for sparse CCA:

1. For each tuning parameter value (generally this will be a two-dimensional vector) T_j being considered:
   (a) Compute $\mathbf{w}_1$ and $\mathbf{w}_2$, the canonical vectors using data $\mathbf{X}_1$ and $\mathbf{X}_2$ and tuning parameter T_j. Compute $d_j = \text{Cor}(\mathbf{X}_1\mathbf{w}_1, \mathbf{X}_2\mathbf{w}_2)$.
   (b) For $i \in 1, ..., N$, where N is some large number of permutations:
      i. Permute the rows of $\mathbf{X}_1$ to obtain the matrix $\mathbf{X}_1^i$, and compute canonical vectors $\mathbf{w}_1^i$ and $\mathbf{w}_2^i$ using data $\mathbf{X}_1^i$ and $\mathbf{X}_2$ and tuning parameter T_j.
      ii. Compute $d_j^i = \text{Cor}(\mathbf{X}_1^i\mathbf{w}_1^i, \mathbf{X}_2\mathbf{w}_2^i)$.
   (c) Calculate the p-value $p_j = \frac{1}{N}\sum_{i=1}^{N} 1_{d_j^i \geq d_j}$.
2. Choose the tuning parameter T_j corresponding to the smallest p_j. Alternatively, one can choose the tuning parameter T_j for which $(d_j - \frac{1}{N}\sum_i d_j^i)/\text{sd}(d_j^i)$ is largest, where $\text{sd}(d_j^i)$ indicates the standard deviation of $d_j^1, ..., d_j^N$. The resulting p-value is p_j.

Since multiple tuning parameters T_j are considered in the above algorithm, a strict cut-off for the p-value p_j should be used in order to avoid problems associated with multiple testing.

Given the above algorithm, the analogous method for selecting tuning parameters and determining significance for sparse sCCA is straightforward. For simplicity, we assume that the threshold for features to enter Q_1 and Q_2 in the sparse sCCA algorithm is fixed (not a tuning parameter).

Algorithm to select tuning parameters and determine significance for sparse sCCA:

1. For each tuning parameter (generally this will be a two-dimensional vector) T_j being considered:
   (a) Compute $\mathbf{w}_1$ and $\mathbf{w}_2$, the supervised canonical vectors using data $\mathbf{X}_1, \mathbf{X}_2$, and $\mathbf{y}$ and tuning parameter T_j. Compute $d_j = \text{Cor}(\mathbf{X}_1\mathbf{w}_1, \mathbf{X}_2\mathbf{w}_2)$.

   (b) For $i \in 1, ..., N$, where N is some large number of permutations:
      i. Permute the rows of $\mathbf{X}_1$ and $\mathbf{X}_2$ separately to obtain the matrices $\mathbf{X}_1^i$ and $\mathbf{X}_2^i$, and compute supervised canonical vectors $\mathbf{w}_1^i$ and $\mathbf{w}_2^i$ using data $\mathbf{X}_1^i$, $\mathbf{X}_2^i$, $\mathbf{y}$, and tuning parameter T_j.
      ii. Compute $d_j^i = \text{Cor}(\mathbf{X}_1^i \mathbf{w}_1^i, \mathbf{X}_2^i \mathbf{w}_2^i)$.
   (c) Calculate the p-value $p_j = \frac{1}{N} \sum_{i=1}^N 1_{d_j^i \geq d_j}$.

2. Choose the tuning parameter T_j corresponding to the smallest p_j. Alternatively, one can choose the tuning parameter T_j for which $(d_j - \frac{1}{N} \sum_i d_j^i)/\text{sd}(d_j^i)$ is largest, where $\text{sd}(d_j^i)$ indicates the standard deviation of $d_j^1, ..., d_j^N$. The resulting p-value is p_j.

Note that in the permutation step, we permute the rows of $\mathbf{X}_1$ and $\mathbf{X}_2$ without permuting $\mathbf{y}$; this means that under the permutation null distribution, $\mathbf{y}$ is not correlated with the columns of $\mathbf{X}_1$ and $\mathbf{X}_2$.

We can similarly use the following permutation-based algorithm to assess the significance of the canonical vectors obtained using sparse mCCA:

Algorithm to select tuning parameters and determine significance for sparse mCCA:

1. For each tuning parameter (generally this will be a K-dimensional vector) T_j being considered:
   (a) Compute $\mathbf{w}_1, ..., \mathbf{w}_K$, the canonical vectors using data $\mathbf{X}_1, ..., \mathbf{X}_K$ and tuning parameter T_j. Compute $d_j = \sum_{s<t} \text{Cor}(\mathbf{X}_s \mathbf{w}_s, \mathbf{X}_t \mathbf{w}_t)$.
   (b) For $i \in 1, ..., N$, where N is some large number of permutations:
      i. Permute the rows of $\mathbf{X}_1, ..., \mathbf{X}_K$ separately to obtain the matrices $\mathbf{X}_1^i, ..., \mathbf{X}_K^i$, and compute canonical vectors $\mathbf{w}_1^i, ..., \mathbf{w}_K^i$ using data $\mathbf{X}_1^i, ..., \mathbf{X}_K^i$ and tuning parameter T_j.
      ii. Compute $d_j^i = \sum_{s<t} \text{Cor}(\mathbf{X}_s^i \mathbf{w}_s^i, \mathbf{X}_t^i \mathbf{w}_t^i)$.
   (c) Calculate the p-value $p_j = \frac{1}{N} \sum_{i=1}^N 1_{d_j^i \geq d_j}$.

2. Choose the tuning parameter T_j corresponding to the smallest p_j. Alternatively, one can choose the tuning parameter T_j for which $(d_j - \frac{1}{N} \sum_i d_j^i)/\text{sd}(d_j^i)$ is largest, where $\text{sd}(d_j^i)$ indicates the standard deviation of $d_j^1, ..., d_j^N$. The resulting p-value is p_j.

B Extension of methods to obtain multiple canonical vectors

We first review the method of Witten et al. (2009) for obtaining multiple sparse CCA canonical vectors. Note that the sparse CCA algorithm uses the cross-product matrix $\mathbf{Y} = \mathbf{X}_1^T\mathbf{X}_2$ and does not require knowledge of $\mathbf{X}_1$ and $\mathbf{X}_2$ individually.

Algorithm for obtaining J sparse CCA factors:

```
1. Let $\mathbf{Y}^1 \leftarrow \mathbf{X}_1^T\mathbf{X}_2$.
2. For $j \in 1, ..., J$:
    (a) Compute $\mathbf{w}_1^j$ and $\mathbf{w}_2^j$ by applying the single-factor sparse CCA algorithm to data $\mathbf{Y}^j$.
    (b) $\mathbf{Y}^{j+1} \leftarrow \mathbf{Y}^j - (\mathbf{w}_1^{j^T}\mathbf{Y}^j\mathbf{w}_2{}^j)\mathbf{w}_1^j\mathbf{w}_2^{j^T}$.
3. $\mathbf{w}_1^j$ and $\mathbf{w}_2^j$ are the $j^{th}$ canonical vectors.
```

To obtain J sparse sCCA factors, submatrices $\tilde{\mathbf{X}}_1$ and $\tilde{\mathbf{X}}_2$ are formed from the features most associated with the outcome; the algorithm for obtaining J sparse CCA factors is then applied to this new data.

To obtain J sparse mCCA factors, note that the sparse mCCA algorithm requires knowledge only of the $\binom{K}{2}$ cross-product matrices of the form $\mathbf{X}_s^T\mathbf{X}_t$ with $s < t$, rather than the raw data $\mathbf{X}_s$ and $\mathbf{X}_t$.

Algorithm for obtaining J sparse mCCA factors:

```
1. For each $1 \leq s < t \leq K$, let $\mathbf{Y}_{st}^1 \leftarrow \mathbf{X}_s^T\mathbf{X}_t$.
2. For $j \in 1, ..., J$:
    (a) Compute $\mathbf{w}_1^j, ..., \mathbf{w}_K^j$ by applying the single-factor sparse mCCA algorithm to data $\mathbf{Y}_{st}^j$ for $1 \leq s < t \leq K$.
    (b) $\mathbf{Y}_{st}^{j+1} \leftarrow \mathbf{Y}_{st}^j - (\mathbf{w}_s^{j^T}\mathbf{Y}_{st}^j\mathbf{w}_t^j)\mathbf{w}_s^j\mathbf{w}_t^{j^T}$.
3. $\mathbf{w}_1^j, ..., \mathbf{w}_K^j$ are the $j^{th}$ canonical vectors.
```

References

Alizadeh, A., Eisen, M., Davis, R. E., Ma, C., Lossos, I., Rosenwald, A., Boldrick, J., Sabet, H., Tran, T., Yu, X., Powell, J., Marti, G., Moore, T., Hudson, J., Lu, L., Lewis, D., Tibshirani, R., Sherlock, G., Chan, W., Greiner, T., Weisenburger, D., Armitage, K., Warnke, R., Levy, R., Wilson, W., Grever, M., Byrd, J., Botstein, D., Brown, P. & Staudt, L. (2000), 'Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling', *Nature* **403**, 503–511.

Bair, E., Hastie, T., Paul, D. & Tibshirani, R. (2006), 'Prediction by supervised principal components', *J. Amer. Statist. Assoc.* **101**, 119–137.

Bair, E. & Tibshirani, R. (2004), 'Semi-supervised methods to predict patient survival from gene expression data', *PLOS Biology* **2**, 511–522.

Gifi, A. (1990), *Nonlinear multivariate analysis*, Wiley, Chichester, England.

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

Hyman, E., Kauraniemi, P., Hautaniemi, S., Wolf, M., Mousses, S., Rozenblum, E., Ringner, M., Sauter, G., Monni, O., Elkahloun, A., Kallioniemi, O.-P. & Kallioniemi, A. (2002), 'Impact of DNA amplification on gene expression patterns in breast cancer', *Cancer Research* **62**, 6240–6245.

Le Cao, K., Pascal, M., Robert-Granie, C. & Philippe, B. (2009), 'Sparse canonical methods for biological data integration: application to a cross-platform study', *BMC Bioinformatics* **10**.

Lenz, G., Wright, G., Emre, N., Kohlhammer, H., Dave, S., Davis, R., Carty, S., Lam, L., Shaffer, A., Xiao, W., Powell, J., Rosenwald, A., Ott, G., Muller-Hermelink, H., Gascoyne, R., Connors, J., Campo, E., Jaffe, E., Delabie, J., Smeland, E., Rimsza, L., Fisher, R., Weisenburger, D., Chano, W. & Staudt, L. (2008), 'Molecular subtypes of diffuse large B-cell lymphoma arise by distinct genetic pathways', *Proc. Natl. Acad. Sci.* **105**, 13520–13525.

Massy, W. (1965), 'Principal components regression in exploratory statistical research', *Journal of the American Statistical Association* **60**, 234–236.

Morley, M., Molony, C., Weber, T., Devlin, J., Ewens, K., Spielman, R. & Cheung, V. (2004), 'Genetic analysis of genome-wide variation in human gene expression', *Nature* **430**, 743–747.

Parkhomenko, E., Tritchler, D. & Beyene, J. (2007), 'Genome-wide sparse canonical correlation of gene expression with genotypes', *BMC Proceedings* **1**, S119.

Parkhomenko, E., Tritchler, D. & Beyene, J. (2009), 'Sparse canonical correlation analysis with application to genomic data integration', *Statistical Applications in Genetics and Molecular Biology* **8**, 1–34.

Pollack, J., Sorlie, T., Perou, C., Rees, C., Jeffrey, S., Lonning, P., Tibshirani, R., Botstein, D., Borresen-Dale, A. & Brown, P. (2002), 'Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors', *Proceedings of the National Academy of Sciences* **99**, 12963–12968.

Rosenwald, A., Wright, G., Chan, W. C., Connors, J. M., Campo, E., Fisher, R. I., Gascoyne, R. D., Muller-Hermelink, H. K., Smeland, E. B. & Staudt, L. M. (2002), 'The use of molecular profiling to predict survival after chemotherapy for diffuse large B-cell lymphoma', *The New England Journal of Medicine* **346**, 1937–1947.

Stranger, B., Forrest, M., Clark, A., Minichiello, M., Deutsch, S., Lyle, R., Hunt, S., Kahl, B., Antonarakis, S., Tavare, S., Deloukas, P. & Dermitzakis, E. (2005), 'Genome-wide associations of gene expression variation in humans', *PLOS Genetics* **1(6)**, e78.

Stranger, B., Forrest, M., Dunning, M., Ingle, C., Beazley, C., Thorne, N., Redon, R., Bird, C., de Grassi, A., Lee, C., Tyler-Smith, C., Carter, N., Scherer, S., Tavare, S., Deloukas, P., Hurles, M. & Dermitzakis, E. (2007), 'Relative impact of nucleotide and copy number variation on gene expression phenotypes', *Science* **315**, 848–853.

Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. (2002), 'Diagnosis of multiple cancer types by shrunken centroids of gene expression', *Proc. Natl. Acad. Sci.* **99**, 6567–6572.

Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. (2003), 'Class prediction by nearest shrunken centroids, with applications to DNA microarrays', *Statistical Science* pp. 104–117.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. & Knight, K. (2005), 'Sparsity and smoothness via the fused lasso', *J. Royal. Statist. Soc. B.* **67**, 91–108.

Tibshirani, R. & Wang, P. (2008), ‘Spatial smoothing and hotspot detection for CGH data using the fused lasso’, *Biostatistics* **9**, 18–29.

Waaijenborg, S., Verselewel de Witt Hamer, P. & Zwinderman, A. (2008), ‘Quantifying the association between gene expressions and DNA-markers by penalized canonical correlation analysis’, *Statistical Applications in Genetics and Molecular Biology* **7**.

Witten, D., Tibshirani, R. & Hastie, T. (2009), ‘A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis’, *Biostatistics* .