Sparse canonical correlation analysis, with applications to genomic data

Daniela M. Witten and Robert Tibshirani

June 15, 2009
Suppose that we have $n$ observations on $p_1 + p_2$ variables, and the variables are naturally partitioned into two groups of $p_1$ and $p_2$ variables, respectively.
The framework

- Suppose that we have $n$ observations on $p_1 + p_2$ variables, and the variables are naturally partitioned into two groups of $p_1$ and $p_2$ variables, respectively.
- Let $X_1 \in \mathbb{R}^{n \times p_1}$ correspond to the first set of variables, and let $X_2 \in \mathbb{R}^{n \times p_2}$ correspond to the second set of variables.
The framework

- Suppose that we have $n$ observations on $p_1 + p_2$ variables, and the variables are naturally partitioned into two groups of $p_1$ and $p_2$ variables, respectively.

- Let $X_1 \in \mathbb{R}^{n \times p_1}$ correspond to the first set of variables, and let $X_2 \in \mathbb{R}^{n \times p_2}$ correspond to the second set of variables.

- Assume that the columns of $X_1$ and $X_2$ have been standardized to have mean zero and standard deviation one.
The data

\[
\begin{bmatrix}
1 & \cdots & p_1 \\
1 & \vdots & \vdots \\
n & \vdots & \vdots \\
\end{bmatrix} \quad \begin{bmatrix}
1 & \cdots & p_2 \\
1 & \vdots & \vdots \\
n & \vdots & \vdots \\
\end{bmatrix}
\]
Canonical correlation analysis

Canonical correlation analysis (CCA) is a classical method in statistics.
Canonical correlation analysis

Canonical correlation analysis (CCA) is a classical method in statistics.

We can seek $w_1 \in \mathbb{R}^{p_1}$ and $w_2 \in \mathbb{R}^{p_2}$ that maximize correlation between $X_1w_1$ and $X_2w_2$; that is,

$$\text{maximize}_{w_1,w_2} w_1^T X_1^T X_2 w_2 \text{ subject to } w_1^T X_1^T X_1 w_1 = w_2^T X_2^T X_2 w_2 = 1.$$
Canonical correlation analysis (CCA) is a classical method in statistics.

We can seek $\mathbf{w}_1 \in \mathbb{R}^{p_1}$ and $\mathbf{w}_2 \in \mathbb{R}^{p_2}$ that maximize correlation between $\mathbf{X}_1 \mathbf{w}_1$ and $\mathbf{X}_2 \mathbf{w}_2$; that is,

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

CCA is not appropriate when $p_1, p_2 \approx n$ or $p_1, p_2 \gg n$. 
The data when $p_1, p_2 \gg n$
Sparse Canonical Correlation Analysis

\begin{align*}
\text{maximize}_{w_1, w_2} & \quad w_1^T X_1^T X_2 w_2 \\
\text{subject to} & \quad w_1^T X_1^T X_1 w_1 = w_2^T X_2^T X_2 w_2 = 1.
\end{align*}

- We impose $L_1$ constraints onto $w_1$ and $w_2$: that is, $\|w_1\|_1 \leq c_1$ and $\|w_2\|_1 \leq c_2$.
- We assume that $X_1^T X_1 = I$ and $X_2^T X_2 = I$. 
Sparse Canonical Correlation Analysis

The *sparse CCA criterion* is

\[
\text{maximize}_{w_1, w_2} w_1^T X_1^T X_2 w_2 \\
\text{subject to } \|w_1\|^2 \leq 1, \|w_2\|^2 \leq 1, \|w_1\|_1 \leq c_1, \|w_2\|_1 \leq c_2.
\]
The sparse CCA criterion is

\[
\text{maximize}_{\mathbf{w}_1, \mathbf{w}_2} \mathbf{w}_1^T \mathbf{X}_1^T \mathbf{X}_2 \mathbf{w}_2 \\
\text{subject to } \|\mathbf{w}_1\|^2 \leq 1, \|\mathbf{w}_2\|^2 \leq 1, \|\mathbf{w}_1\|_1 \leq c_1, \|\mathbf{w}_2\|_1 \leq c_2.
\]

For \(c_1\) and \(c_2\) small, this results in \(\mathbf{w}_1\) and \(\mathbf{w}_2\) sparse: many of the elements of \(\mathbf{w}_1\) and \(\mathbf{w}_2\) will exactly equal zero.
Sparse CCA Algorithm

1. Begin with an initial value for \( \mathbf{w}_2 \in \mathbb{R}^{p_2} \).

2. Iterate until convergence:

\[
\mathbf{w}_1 \leftarrow \arg\max_{\mathbf{w}_1} \mathbf{w}_1^T \mathbf{X}_1^T \mathbf{X}_2 \mathbf{w}_2 \text{ subject to } ||\mathbf{w}_1||^2 \leq 1, ||\mathbf{w}_1||_1 \leq c_1.
\]

\[
\mathbf{w}_2 \leftarrow \arg\max_{\mathbf{w}_2} \mathbf{w}_1^T \mathbf{X}_1^T \mathbf{X}_2 \mathbf{w}_2 \text{ subject to } ||\mathbf{w}_2||^2 \leq 1, ||\mathbf{w}_2||_1 \leq c_2.
\]
Sparse CCA Algorithm

1. Begin with an initial value for $w_2 \in \mathbb{R}^{p_2}$.
2. Iterate until convergence:

   $w_1 \leftarrow \arg \max_{w_1} w_1^T X_1^T X_2 w_2$ subject to $\|w_1\|^2 \leq 1, \|w_1\|_1 \leq c_1$.

   $w_2 \leftarrow \arg \max_{w_2} w_1^T X_1^T X_2 w_2$ subject to $\|w_2\|^2 \leq 1, \|w_2\|_1 \leq c_2$.

Each update takes the form $w_1 \leftarrow \frac{S(X_1^T X_2 w_2, \Delta_1)}{\|S(X_1^T X_2 w_2, \Delta_1)\|_2}$ where $\Delta_1 \geq 0$ is chosen so that $\|w_1\|_1 = c_1$. Here, $S$ is the soft-thresholding operator: $S(x, a) = \text{sgn}(x)(|x| - a)_+$. 
Recently, it has become increasingly common for researchers to use multiple assays to obtain measurements on a single set of patient samples.

For instance, a data set might consist of gene expression and DNA copy number measurements on the same set of samples.
Genomic Data

Recently, it has become increasingly common for researchers to use multiple assays to obtain measurements on a single set of patient samples.

For instance, a data set might consist of gene expression and DNA copy number measurements on the same set of samples.

The Question: Can we identify sets of genes whose expression is correlated with regions of copy number change?
DNA copy number change
DNA copy number change
Gene expression

- DNA
  - Transcription
  - mRNA
  - Translation
  - Protein
Gene expression + copy number data

\[ X_1 = \text{Gene Expr.} \]

\[ X_2 = \text{Copy Number} \]
Gene expression + copy number data

We want to find regions of DNA copy number change that are highly correlated with the expression of a set of genes.
We want to find regions of DNA copy number change that are highly correlated with the expression of a set of genes.

That is, we want $\mathbf{w}_1 \in \mathbb{R}^{p_1}$, $\mathbf{w}_2 \in \mathbb{R}^{p_2}$ sparse such that $\text{Cor}(\mathbf{X}_1 \mathbf{w}_1, \mathbf{X}_2 \mathbf{w}_2)$ is high.
Gene expression + copy number data

We want to find regions of DNA copy number change that are highly correlated with the expression of a set of genes.

That is, we want \( w_1 \in \mathbb{R}^{p_1}, w_2 \in \mathbb{R}^{p_2} \) sparse such that \( \text{Cor}(X_1 w_1, X_2 w_2) \) is high.

We’ll make a statement like \( 0.3 \times \text{(gene 1 expression)} + 0.2 \times \text{(gene 2 expression)} - 4 \times \text{(gene 3 expression)} \) is highly correlated with genomic loss on part of chromosome 3.
Idea

To find regions of copy number change on chromosome 1 that are correlated with gene expression sets anywhere on the genome, run sparse CCA using copy number data on chromosome 1 and expression measurements for all genes.
We used a publicly-available breast cancer data set of Chin et al. (2006).

- 89 samples with breast cancer.
- 19672 gene expression measurements.
- 2149 DNA copy number measurements.
Copy number change on chromosome 1
Genes correlated w/copy # change on chrom. 1

<table>
<thead>
<tr>
<th>$i$</th>
<th>Gene</th>
<th>Chromosome</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>jumping translocation breakpoint</td>
<td>1</td>
<td>0.039</td>
</tr>
<tr>
<td>2</td>
<td>translocated promoter region (to activated MET oncogene)</td>
<td>1</td>
<td>0.153</td>
</tr>
<tr>
<td>3</td>
<td>glyceronephosphate O-acyltransferase</td>
<td>1</td>
<td>0.255</td>
</tr>
<tr>
<td>4</td>
<td>NADH dehydrogenase (ubiquinone) Fe-S protein 2</td>
<td>1</td>
<td>0.265</td>
</tr>
<tr>
<td>5</td>
<td>nucleoporin 133kD</td>
<td>1</td>
<td>0.007</td>
</tr>
<tr>
<td>6</td>
<td>geranylgeranyl diphosphate synthase 1</td>
<td>1</td>
<td>0.131</td>
</tr>
<tr>
<td>7</td>
<td>rab3 GTPase-activating protein, non-catalytic subunit (150kD)</td>
<td>1</td>
<td>0.283</td>
</tr>
<tr>
<td>8</td>
<td>peroxisomal biogenesis factor 11B</td>
<td>1</td>
<td>0.154</td>
</tr>
<tr>
<td>9</td>
<td>phosphatidylinositol glycan, class C</td>
<td>1</td>
<td>0.124</td>
</tr>
<tr>
<td>10</td>
<td>tubulin-specific chaperone e</td>
<td>1</td>
<td>0.069</td>
</tr>
<tr>
<td>11</td>
<td>protoporphyrinogen oxidase</td>
<td>1</td>
<td>0.052</td>
</tr>
<tr>
<td>12</td>
<td>tuftelin 1</td>
<td>1</td>
<td>0.037</td>
</tr>
<tr>
<td>13</td>
<td>papillary renal cell carcinoma (translocation-associated)</td>
<td>1</td>
<td>0.055</td>
</tr>
<tr>
<td>14</td>
<td>splicing factor 3b, subunit 4, 49kD</td>
<td>1</td>
<td>0.469</td>
</tr>
<tr>
<td>15</td>
<td>UDP-Gal:betaGlcNAc beta 1,4- galactosyltransferase</td>
<td>1</td>
<td>0.27</td>
</tr>
<tr>
<td>16</td>
<td>hypothetical protein FLJ12671</td>
<td>1</td>
<td>0.229</td>
</tr>
<tr>
<td>17</td>
<td>hypothetical protein HSPC155</td>
<td>1</td>
<td>0.168</td>
</tr>
<tr>
<td>18</td>
<td>mitochondrial ribosomal protein L24</td>
<td>1</td>
<td>0.195</td>
</tr>
<tr>
<td>19</td>
<td>HSPC003 protein</td>
<td>1</td>
<td>0.391</td>
</tr>
<tr>
<td>20</td>
<td>hypothetical protein FLJ10876</td>
<td>1</td>
<td>0.091</td>
</tr>
<tr>
<td>21</td>
<td>CGI-78 protein</td>
<td>1</td>
<td>0.154</td>
</tr>
<tr>
<td>22</td>
<td>chromosome 1 open reading frame 27</td>
<td>1</td>
<td>0.133</td>
</tr>
<tr>
<td>23</td>
<td>hypothetical protein My014</td>
<td>1</td>
<td>0.278</td>
</tr>
</tbody>
</table>
Results, continued

1. Similar results for other chromosomes.
Results, continued

1. Similar results for other chromosomes.
2. We can use a permutation approach to estimate a p-value for the \( w_1 \) and \( w_2 \) obtained.
1. Similar results for other chromosomes.
2. We can use a permutation approach to estimate a p-value for the $w_1$ and $w_2$ obtained.
3. We described the identification of cis effects. To identify trans effects, run sparse CCA using copy number data on chromosome 1 and expression measurements for all genes NOT on chromosome 1.
Conclusions

- A method for the integrative analysis of genomic data sets.
- An extension of sparse CCA leads to sparse multiple CCA, for the case of $K > 2$ data sets on a single set of observations.
- If an outcome measurement (e.g. survival time, cancer subtype) is available for each observation, than sparse supervised CCA can be used.
- Software available in R PMA package and Excel AddIn ”Correlate” (implemented by Sam Gross)
References


