Sparse Additive Functional and kernel CCA

Sivaraman Balakrishnan*
Kriti Puniyani*
John Lafferty
*Carnegie Mellon University  University of Chicago

Presented by Miao Liu
5/3/2013
Canonical correlation analysis (CCA, Hotelling 1936)

• Given \((X_1, Y_1), \ldots, (X_n, Y_n)\) with \(X \in \mathbb{R}^{p_1}\), \(Y \in \mathbb{R}^{p_2}\)

• Form the design matrices \(X \in \mathbb{R}^{n \times p_1}\), \(Y \in \mathbb{R}^{n \times p_2}\)

• Find \(u \in \mathbb{R}^{p_1}\) and \(v \in \mathbb{R}^{p_2}\) that are the solution to the sample version of the problem of maximizing the correlation between \(u^T X\) and \(v^T Y\), i.e.

\[
\arg \max_{u,v} \frac{1}{n} u^T X^T Y v
\]

\[
\text{s.t.} \quad \frac{1}{n} u^T X^T X u \leq 1, \quad \frac{1}{n} v^T Y^T Y v \leq 1,
\]
Motivation

• CCA is a valuable dimension reduction tool
• Increasing attractive in genomic data analysis [Witten et al. 2009]
  – DNA copy number (or comparative genomic hybridization, CGH)
  – Gene expression
  – single nucleotide polymorphism (SNP) information
• Classical formulation of CCA is not meaningful for genomic data
  – The sample covariance matrices $X^T X$ and $Y^T Y$ are singular
• Sparse CCA
  – Lead to more interpretable models, reduced computational cost
• Existing methods are restricted in finding linear combination of the variables
The general nonparametric analogue of (1) is

\[
\arg\max_{f,g} \frac{1}{n} \sum_{i=1}^{n} f(X_i)g(Y_i)
\]

subject to

\[
\frac{1}{n} \sum_{i=1}^{n} f^2(X_i) \leq 1 \quad \frac{1}{n} \sum_{i=1}^{n} g^2(Y_i) \leq 1
\]

-- \(f\) and \(g\) are restricted to belong to an approximate class of smooth functions.
• Kernel CCA [Bach & Jordan 2003]
  – Using “kernel trick” to allow nonparametric modeling of correlations
  – Regularization for smoothness of $f$ and $g$ in RKHS
  – Suffers from the curse of dimensionality
• Additive models [Hastie & Tibshirani 1986]

\[ f(x_1, x_2, \ldots, x_{p_1}) = \sum_{j=1}^{p_1} f_j(x_j) \]  \hspace{1cm} (3)

\[ g(y_1, y_2, \ldots, y_{p_2}) = \sum_{k=1}^{p_2} g_k(y_k) \]  \hspace{1cm} (4)

  – In regression, no longer require the sample size to be exponential in the dimension
  – Only have strong statistical properties in low dimensions
  – Recent works shows sparse additive models for regression can be efficiently estimated even when $p>n$ [Ravikumar et al., 2009, Koltchinskii&Yuan, 2010; Meier et al., 2009; Raskutti et al., 2010]
Contributions

• Two additive nonparametric formulation of CCA
  – Over a family of RKHS
  – Over Sobolev spaces without a reproducing kernel
    • In the low-dimensional setting without sparsity, this formulation is closely related to the Alternating Conditional Expectations (ACE) formulation of nonparametric regression [Breiman & Friedman 1985]

• Risk consistency guarantees for the global risk minimizer in the high-dimensional regime where \( \min(p_1, p_2) > n \)

• Propose Marginal thresholding to reduce the dimensionality
Considerations

• An important consideration: the sparse nonparametric CCA is biconvex, but not jointly convex in $f$ and $g$
  – In the absence of the sparsity constraints the linear problem (1) reduces to a generalized eigenvalue problem
  – Over RKHS, the problem without sparsity is a generalized eigenvalue problem
  – Over Sobolev space, the problem reduces to an eigenvalue problem w.r.t conditional expectation operators

• Solving the nonconvex sparse CCA problem
  – Use the solution to nonsparse version of the problem to initialize sparse CCA [Witten et al. 2009, Parkhomenko et al. 2007]
  – Random initializations [Chen & Liu 2012]
Notations

- \( \mathcal{F}_j \subset L_2(\mu(x_j)) \) a Reproducing kernel Hilbert space of univariate functions on the domain \( X_j \) for each \( j = 1, \ldots, p_1 \), similarly \( \mathcal{G}_k \subset L_2(\mu(y_k)) \)

- Assume \( \mathbb{E}[f_j(X_j)] = 0 \) and \( \mathbb{E}[g_k(Y_k)] = 0 \) for all \( f_j \in \mathcal{F}_j, g_k \in \mathcal{G}_k \) for each \( j \) and \( k \).

- The sets of additive functions of \( x \) and \( y \)

\[
\mathcal{F} = \{ f = \sum_{j=1}^{p_1} f_j(x_j) | f_j \in \mathcal{F}_j \}
\]
\[
\mathcal{G} = \{ g = \sum_{k=1}^{p_2} g_k(y_k) | g_k \in \mathcal{G}_k \}
\]

- Given \( n \) independent tuples \((X_i, Y_i)_{i=1}^n\) where \( X_i = \{X_{i1}, \ldots, X_{ip_1}\} \), \( Y_i = \{Y_{i1}, \ldots, Y_{ip_2}\} \)

- Two norms

\[
\|f_j\|_{\mathcal{F}_j} = \sqrt{\langle f_j, f_j \rangle_{\mathcal{F}_j}} \quad \|f_j\|_2 = \sqrt{\frac{1}{n} \sum_{i=1}^{n} f_j^2(X_{ij})}
\]
Sparse additive kernel CCA

- **Infinite dimensional optimization**

\[
\max_{f \in \mathcal{F}, g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} f(X_i)g(Y_i) \quad \text{subject to}
\]

\[
\frac{1}{n} \sum_{i=1}^{n} f^2(X_i) + \gamma_f \sum_{j=1}^{p_1} \|f_j\|_{\mathcal{F}_j}^2 \leq 1
\]

\[
\frac{1}{n} \sum_{i=1}^{n} g^2(Y_i) + \gamma_g \sum_{k=1}^{p_2} \|g_k\|_{\mathcal{G}_k}^2 \leq 1
\]

\[
\sum_{j=1}^{p_1} \|f_j\|_2 \leq C_f
\]

\[
\sum_{k=1}^{p_2} \|g_k\|_2 \leq C_g.
\]

- **Finite dimensional optimization**

\[
\max_{\alpha, \beta} \frac{1}{n} \left( \sum_{j=1}^{p_1} K_{xj} \alpha_j \right) \left( \sum_{k=1}^{p_2} K_{yk} \beta_k \right) \quad \text{subject to}
\]

\[
\frac{1}{n} \left( \sum_{j=1}^{p_1} K_{xj} \alpha_j \right)^T \left( \sum_{j=1}^{p_1} K_{xj} \alpha_j \right) + \gamma_f \sum_{j=1}^{p_1} \alpha_j^T K_{xj} \alpha_j \leq 1
\]

\[
\frac{1}{n} \left( \sum_{k=1}^{p_2} K_{yk} \beta_k \right)^T \left( \sum_{k=1}^{p_2} K_{yk} \beta_k \right) + \gamma_g \sum_{k=1}^{p_2} \beta_k^T K_{yk} \beta_k \leq 1
\]

\[
\sum_{j=1}^{p_1} \sqrt{\frac{1}{n} \alpha_j^T K_{xj}^T K_{xj} \alpha_j} \leq C_f, \quad \sum_{k=1}^{p_2} \sqrt{\frac{1}{n} \beta_k^T K_{yk}^T K_{yk} \beta_k} \leq C_g.
\]

Here \( \alpha \) is an \((n \times p_1)\) matrix, \( \alpha_j \) is its \( j \)-th column, \( \beta \) is an \((n \times p_2)\) matrix and \( \beta_k \) is its \( k \)-th column.

- The problem (6) is not convex, but biconvex
- An biconvex optimization procedure will be derived, which only guarantees local optimal
- Without the sparsity penalty, the problem reduces to an additive form of kernel CCA [Bach & Jordan, 2003]
- Experiment shows this method works for \( p_1, p_2 < n \)
Notations

- The set of smooth functions

\[ \mathcal{F}_j = \{ f_j \in \mathcal{S}_j : f_j = \sum_{k=0}^{\infty} \beta_{jk} \psi_{jk}, \sum_{k=0}^{\infty} \beta_{jk}^2 k^4 \leq C^2 \} \]

- Denote \( \mathcal{S}_j \) the subspace of \( \mu(x_j) \) measurable functions with mean 0, with the usual inner product \( \langle f_j, f'_j \rangle = \mathbb{E} \left( f_j(X_j)f'_j(X_j) \right) \)
Sparse additive functional CCA (SA-FCCA)

\[
\max_{f \in \mathcal{F}, g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} f(X_i)g(Y_i) 
\]

\[
\text{s.t.} \quad \frac{1}{n} \sum_{j=1}^{p_1} \sum_{i=1}^{n} f_j^2(X_{ij}) \leq 1, \quad \sum_{j=1}^{p_1} \|f_j\|_2 \leq C_f
\]

\[
\frac{1}{n} \sum_{k=1}^{p_2} \sum_{i=1}^{n} g_k^2(Y_{ik}) \leq 1, \quad \sum_{k=1}^{p_2} \|g_k\|_2 \leq C_g
\]

• Three differences to (2)
  – No regularization for smoothness, but instead work directly over a Sobolev space of smooth functions
  – Constrain the sum of the variance of the individual \(f_j\)s.
  – No longer appeal to the representer theorem
• biconvex backfitting procedure for SA-FCCA
  – A simple example, \( g \) is fixed and both constraints on \( f \) are tight

\[
\max_f \min_{\lambda \geq 0, \gamma \geq 0} \mathbb{E}[f(X)g(Y)] - \lambda(\|f\|_2^2 - 1) - \gamma(\|f\|_1 - C_f).
\]

\[
\|f\|_1 = \sum_{j=1}^{p_1} \sqrt{\mathbb{E}(f_j^2(x_j))}, \quad \|f\|_2^2 = \sum_{j=1}^{p_1} \mathbb{E}(f_j^2(x_j))
\]

For simplicity, consider the case when \( \lambda, \gamma > 0 \), and denote \( a \equiv g(Y) \).

– A coordinate ascent style procedure where we optimize over \( f_j \) holding the other functions fixed can be derived
Algorithm 1 Biconvex backfitting for SA-FCCA

input \{ (X_i, Y_i) \}, parameters C_f, C_g, initial g(Y_i)

1. Compute smoothing matrices S_j and T_k.
2. Fix g. For each j, set \( f_j \leftarrow \frac{S_j g}{\lambda} \) where \( \lambda = \sqrt{\sum_{j=1}^{p_1} (g^T S_j^T S_j g)} \)
3. if \( \sum_{j=1}^{p_1} ||f_j||_2 \leq C_f \) break
   else let \( F_m \) denote the functions with maximum \( ||.||_2 \) norm. Set all other functions to 0.
   For each \( f \in F_m \), set \( f \leftarrow \frac{C_f f}{||F_m||f||_2} \).
   If \( \sum_{j=1}^{p_1} ||f_j||_2^2 \leq 1 \), break
   else set \( f_j \leftarrow \left( 1 - \frac{\gamma}{\sqrt{||S_j g||_2}} \right) \frac{S_j g}{\lambda} \) where \( \lambda = \sqrt{\sum_{j=1}^{p_1} \left( 1 - \frac{\gamma}{\sqrt{||S_j g||_2}} \right) S_j g ||^2_2} \) and \( \gamma \) is chosen so that \( \sum_{j=1}^{p_1} \sqrt{g^T S_j^T S_j g} = C_f \).
4. Center by setting each \( f_j \leftarrow f_j - \text{mean}(f_j) \).
5. Fix f and repeat above to update g. Iterate both updates till convergence.

output Final functions f, g
Marginal Thresholding

• The biconvex optimization may be trapped at local optimal, to mitigate this issue
  – First run the algorithms without any sparsity constraint
  – Use the nonsparse solution as initialization and incorporate the sparsity penalties.
• Such initialization work well for low dimensional problems.
• To extend the algorithm to the high dimensional scenario, the marginal thresholding is proposed.
• For each pair of variables \((X_i, Y_j)\) (one \(X\) and one \(Y\) covariate)
  – Fit marginal functions to \((X_i, Y_j)\) by optimizing the criteria in (6) or (7) without the sparsity constraints
  – Compute the correlation on held out data, with a matrix \(M\) of size \(p_1 \times p_2\), the \((i, j)\) entry representing the correlation between \(f_i(X_i)\) and \(g_j(Y_j)\)
  – Threshold the entries of \(M\) to obtain a subset of variables on which to run SA-FCCA and SA-KCCA
Main theoretical results

• The quantity of interest

\[ \Theta_n = \sup_{f_j, g_k} \left| \frac{1}{n} \sum_{i=1}^{n} f_j(X_{ij})g_k(Y_{ik}) - \mathbb{E}(f_j(X_j)g_k(Y_k)) \right| \]

where \( f_j, g_k \in \mathcal{C}, j \in \{1, \ldots, p_1\}, k \in \{1, \ldots, p_2\} \).

• To study marginal thresholding, for each covariate \( X_j \), denote

\[ \alpha_j = \sup_{f_j, g_k \in \mathcal{C}, k \in \{1, \ldots, p_2\}} \mathbb{E}(f_j(X_j)g_k(Y_k)) \]

with \( \mathbb{E}(f_j^2) \leq 1, \mathbb{E}(g_k^2) \leq 1 \).

A covariate \( X_j \) is considered irrelevant if \( \alpha_j = 0 \).

• Similarly, each \( Y_k \) is associated with \( \beta_k \) defined analogously.
Lemma 5.1 (Uniform bound over RKHS)

Assume \( \sup_x |K(x, x)| \leq M < \infty \), for functions \( f_j(x) = \sum_{i=1}^n \alpha_{ij} K(x, X_{ij}), g_k(y) = \sum_{i=1}^n \beta_{ik} K_y(y, Y_{ik}) \)

\[
\mathbb{P} \left( \Theta_n \geq \zeta + C \sqrt{\frac{\log \left( \frac{p_1 p_2}{\delta} \right)}{n}} \right) \leq \delta
\]

where \( C \) is a constant depending only on \( M \), and \( \zeta = \max_{j,k} \frac{2}{n} \mathbb{E}_{X \sim x_j, Y \sim y_k} \sqrt{\sum_{i=1}^n K(X_{ij}, X_{ij}) K(Y_{ik}, Y_{ik})} \)

- \( \zeta \) is independent of \( p_1 \) and \( p_2 \)
- Under the assumption \( K \) is bounded \( \zeta = O(1/\sqrt{n}) \)
- Depends only logarithmically on \( p_1 \) and \( p_2 \) (weak dependence), which is the main reason for guaranteeing consistency even when \( p_1, p_2 > n \)
Lemma 5.2 (Uniform bound for Sobolev spaces)
Assume \( \|f\|_\infty \leq M \leq \infty \), then

\[
\mathbb{P}
\left( \Theta_n \geq \frac{C_1}{\sqrt{n}} + C_2 \sqrt{\frac{\log ((p_1p_2)/\delta)}{n}} \right) \leq \delta
\]

where \( C_1 \) and \( C_2 \) depend only on \( M \).

- Lemma 5.1 is proved via a Rademacher summarization argument of [Bartlett&Mendelson 2002]
- Lemma 5.2 is based on a bounded on the bracketing integral of the Sobolev space [Ravikumar et al. 2009]
- Lemma 5.1 and 5.2 provide values at which to threshold the marginal covariates
Theorem 5.3  Given $\mathbb{P}(\Theta_n \geq \epsilon) \leq \delta$.

1. With probability at least $1 - \delta$, marginal thresholding at $\epsilon$ has no false inclusions.
2. Further, if we have that $\alpha_j$ or $\beta_k \geq 2\epsilon$ then under the same $1 - \delta$ probability event marginal thresholding at $\epsilon$ correctly includes the relevant covariate $X_j$ or $Y_k$. 
Theorem 5.4 (Persistence) If $p_1 p_2 \leq e^{n^\xi}$ for some $\xi < 1$ and $C_f C_g = o(n^{(1-\xi)/2})$, then SA-FCCA and SA-KCCA are persistent over their respective function classes.

- The optimization procedure is persistent if

\[ \text{cov}(f^*, g^*) - \text{cov}(\hat{f}, \hat{g}) \to 0 \text{ even if } p_1, p_2 > n. \]
Experiments

• Non-linear correlations
  – Run on data with n=150 samples in p1=15, p2=15 dimensions, where only one relevant variable is present in X and Y (the other dimensions are Gaussian noise)

• Marginal thresholding
  – n=150, p1=150, p2=150
  – Data generation  
    \[ f_i(X_i) = \cos\left(\frac{\pi}{2}X_i\right), \quad i \in \{1, 3\}, \quad f_i(X_i) = X_i^2, \quad i \in \{2, 4\} \]
    \[ Y_j = \sum_{i=1; i \neq j}^{4} f_i(X_i) + \mathcal{N}(0, 0.1^2) \quad j \in \{1, 2, 3, 4\} \]

• Application to DLBCL data
  – 1500 CGH measurements from chromosome 1 of the data
  – 500 gene expression measurements from genes on chromosome 1 and 2 of the data
Non-linear correlations

<table>
<thead>
<tr>
<th>Model</th>
<th>Test correlation</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Precision/Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SA-FCCA</td>
<td>SA-KCCA</td>
<td>SCCA</td>
<td>KCCA</td>
<td></td>
<td>SA-FCCA</td>
</tr>
<tr>
<td>$Y = X^2$</td>
<td>0.96</td>
<td><strong>0.99</strong></td>
<td>0.05</td>
<td>0.44</td>
<td></td>
<td>1/1</td>
</tr>
<tr>
<td>$Y = \text{abs}(X)$</td>
<td><strong>0.98</strong></td>
<td><strong>0.99</strong></td>
<td>0.06</td>
<td>0.35</td>
<td></td>
<td>1/1</td>
</tr>
<tr>
<td>$Y = \cos(X)$</td>
<td>0.94</td>
<td><strong>0.99</strong></td>
<td>0.071</td>
<td>0.04</td>
<td></td>
<td>1/1</td>
</tr>
<tr>
<td>$\log(Y) = \sin(X)$</td>
<td>0.91</td>
<td><strong>0.93</strong></td>
<td>0.22</td>
<td>0.09</td>
<td></td>
<td>1/1</td>
</tr>
<tr>
<td>$Y = X$</td>
<td><strong>0.99</strong></td>
<td><strong>0.99</strong></td>
<td><strong>0.99</strong></td>
<td><strong>0.98</strong></td>
<td></td>
<td>1/1</td>
</tr>
</tbody>
</table>

Figure 1. Test correlations, and precision and recall for identifying relevant variables for the four different methods. SA-FCCA and SA-KCCA find strong correlations in the data, in both linear and non-linear settings. In all five data sets, SA-FCCA and SA-KCCA are always able to find the relevant variables.
# Marginal thresholding

<table>
<thead>
<tr>
<th>Method</th>
<th>Test correlation</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>SA-FCCA</td>
<td>0.94</td>
<td>1</td>
<td>0.785</td>
</tr>
<tr>
<td>SA-KCCA</td>
<td>0.98</td>
<td>0.95</td>
<td>0.8</td>
</tr>
<tr>
<td>SCCA</td>
<td>0.02</td>
<td>0.02</td>
<td>0.36</td>
</tr>
<tr>
<td>KCCA</td>
<td>0.07</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

*Table 2.* Test correlations, precision and recall for identifying the correct relevant variables for the four different methods ($n = 150$, $p_1 = 150$, $p_2 = 150$). Marginal thresholding was used for selecting relevant variables before running SA-FCCA and SA-KCCA.
Marginal thresholding

Figure 3. Regularization paths for non-linear correlations in the data, for SA-FCCA, SA-KCCA and SCCA resp. The paths for the relevant variables (in $X$ and $Y$) are shown in red, the irrelevant variables are shown in blue.
Application to DLBCL data

Figure 2. DLBCL data: The top row shows two of the functions $f_i(X_i)$ with non-zero norms for $X$ in red, and the bottom row shows two functions $g_j(Y_j)$ with non-zero norms for $Y$ in blue.