Sparse Additive Functional and Kernel CCA

Sivaraman Balakrishnan SBALAKRI@CS.CMU.EDU
School of Computer Science, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213 USA

Kriti Puniyani KPUNIYAN@CS.CMU.EDU
School of Computer Science, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213 USA

John Lafferty LAFFERTY@GALTON.UCHICAGO.EDU
Department of Statistics and Department of Computer Science, University of Chicago, Chicago, IL 60637 USA

Abstract

Canonical Correlation Analysis (CCA) is a classical tool for finding correlations among the components of two random vectors. In recent years, CCA has been widely applied to the analysis of genomic data, where it is common for researchers to perform multiple assays on a single set of patient samples. Recent work has proposed sparse variants of CCA to address the high dimensionality of such data. However, classical and sparse CCA are based on linear models, and are thus limited in their ability to find general correlations. In this paper, we present two approaches to high-dimensional nonparametric CCA, building on recent developments in high-dimensional nonparametric regression. We present estimation procedures for both approaches, and analyze their theoretical properties in the high-dimensional setting. We demonstrate the effectiveness of these procedures in discovering nonlinear correlations via extensive simulations, as well as through experiments with genomic data.

1. Introduction

Canonical correlation analysis (Hotelling, 1936), is a classical method for finding correlations between the components of two random vectors $X \in \mathbb{R}^{p_1}$ and $Y \in \mathbb{R}^{p_2}$. Given a set of n paired observations $(X_1, Y_1), \ldots, (X_n, Y_n)$, we form the design matrices $\mathbb{X} \in \mathbb{R}^{n \times p_1}$ and $\mathbb{Y} \in \mathbb{R}^{n \times p_2}$ and find vectors $u \in \mathbb{R}^{p_1}$ and $v \in \mathbb{R}^{p_2}$ that are solutions to the optimization

$$\begin{aligned} &\arg\max_{u,v} \quad \frac{1}{n} u^T \mathbb{X}^T \mathbb{Y} v \qquad (1) \\ &\text{s.t.} \quad \frac{1}{n} u^T \mathbb{X}^T \mathbb{X} u \leq 1 \qquad \frac{1}{n} v^T \mathbb{Y}^T \mathbb{Y} v \leq 1, \end{aligned}$$

where the columns of $\mathbb{X}$ and $\mathbb{Y}$ have been standardized to have mean zero and standard deviation one. This is the sample version of the problem of maximizing the correlation between the linear combinations $u^T X$ and $v^T Y$, assuming the random variables have mean zero.

CCA can serve as a valuable dimension reduction tool, allowing one to quickly zoom in on interesting phenomena shared by multiple data sets. This tool is increasingly attractive in genomic data analysis, where researchers perform multiple assays per item. For instance, data including DNA copy number (or comparative genomic hybridization, CGH), gene expression, and single nucleotide polymorphism (SNP) information can be collected on a common set of patients. Witten et al. (2009) present examples of recent studies involving such data.

When the data are high dimensional, as is often the case for genomic data, the classical formulation of CCA is not meaningful, since the sample covariance matrices $\mathbb{X}^T\mathbb{X}$ and $\mathbb{Y}^T\mathbb{Y}$ are singular. This has motivated different approaches to *sparse* CCA, which regularizes (1) by suitable sparsity-inducing ℓ_1 penalties (Witten et al., 2009; Witten & Tibshirani, 2009; Parkhomenko et al., 2007; Chen & Liu, 2012). Sparsity can lead to more interpretable models, reduced computational cost, and favorable statistical properties for high dimensional data. Existing methods for CCA are, however, restricted in that they attempt to find linear combinations of the variables—interesting correlations need not be linear. The need for this flexibility motivates the nonparametric approaches we consider in this paper.

Appearing in *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland, UK, 2012.

The general nonparametric analogue of (1) is

$$\begin{aligned} &\arg\max_{f,g} \quad \frac{1}{n}\sum_{i=1}^{n} f(X_i)g(Y_i) \qquad (2)\\ &\text{s.t.} \quad \frac{1}{n}\sum_{i=1}^{n} f^2(X_i) \le 1 \qquad \frac{1}{n}\sum_{i=1}^{n} g^2(Y_i) \le 1 \end{aligned}$$

where f and g are restricted to belong to an appropriate class of smooth functions. Bach & Jordan (2003) introduce a version of this called kernel CCA by applying the "kernel trick" to the CCA problem. Kernel CCA allows flexible nonparametric modeling of correlations, solving (2) with additional regularization to enforce smoothness of the functions f and g in appropriate reproducing kernel Hilbert spaces. However, this general nonparametric model suffers from the curse of dimensionality, as the number of samples required for consistency grows exponentially with the dimension. It is thus necessary to further restrict the complexity of possible functions. We consider the class of additive models which can be written as

$$f(x_1, x_2, \ldots, x_{p_1}) = \sum_{j=1}^{p_1} f_j(x_j) \qquad (3)$$

$$g(y_1, y_2, \ldots, y_{p_2}) = \sum_{k=1}^{p_2} g_k(y_k) \qquad (4)$$

in terms of univariate component functions (Hastie & Tibshirani, 1986). In the regression setting, such models no longer require the sample size to be exponential in the dimension; however, they only have strong statistical properties in low dimensions. Recently, several authors have shown how 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).

In this paper we propose two additive nonparametric formulations of CCA, one over a family of RKHSs and another over Sobolev spaces without a reproducing kernel. In the low-dimensional setting where we do not enforce sparsity, the formulation over Sobolev spaces is closely related to the Alternating Conditional Expectations (ACE) formulation of nonparametric regression due to Breiman & Friedman (1985). In addition to formulating algorithms for the optimizations, we provide risk consistency guarantees for the global risk minimizer in the high dimensional regime where $\min(p_1, p_2) > n$.

An important consideration is that sparse nonparametric CCA is biconvex, but not jointly convex in f and g. This is true even for the linear CCA model, which is a special case of the model we propose. In the absence of the sparsity constraints the linear problem reduces to a generalized eigenvalue problem which can be efficiently solved. This remains true in the nonparametric case as well. Over an RKHS, the problem without sparsity is a generalized eigenvalue problem where Gram matrices replace the data covariance matrices. In the population setting over the Sobolev spaces we consider, Breiman & Friedman (1985) show that the problem reduces to an eigenvalue problem with respect to conditional expectation operators.

Returning to the nonconvex sparse CCA problem, Witten et al. (2009) and Parkhomenko et al. (2007) suggest using the solution to the nonsparse version of the problem to initialize sparse CCA; Chen & Liu (2012) use several random initializations. As we show in simulations, both approaches can lead to poor results, even in the linear case. To address this issue, we propose and study a simple marginal thresholding step to reduce the dimensionality, in the spirit of the diagonal thresholding of Johnstone & Lu (2009) and the SURE screening of Fan & Song (2010). This results in a three step procedure where after preprocessing we use the nonsparse version of our problem to determine a good initialization for the sparse formulation.

In Sections 2 and 3 we briefly describe the additive Sobolev and RKHS function spaces over which we work, introduce our two nonparametric CCA formulations, and discuss their optimization. In Section 4 we address the non-convexity of the formulations and initialization strategies. In Section 5 we summarize the theoretical guarantees of these procedures when $p_1, p_2 > n$ and in Section 6 we describe some simulations and real data experiments.

2. Sparse additive kernel CCA

Recall the linear CCA problem (1). We will now derive its additive generalization over RKHSs. Let $\mathcal{F}_j \subset L_2(\mu(x_j))$ be a reproducing kernel Hilbert space of univariate functions on the domain of X_j, and let $\mathcal{G}_k \subset L_2(\mu(y_k))$ be a reproducing kernel Hilbert space of univariate functions on the domain Y_k, for each $j = 1, \ldots, p_1$ and $k = 1, \ldots, p_2$. We assume that $\mathbb{E}[f_j(X_j)] = 0$ and $\mathbb{E}[g_k(Y_k)] = 0$ for all $f_j \in \mathcal{F}_j$, and $g_k \in \mathcal{G}_k$ for each j and k. This is necessary to enforce model identifiability. In practice, we will always work with centered Gram matrices to enforce this (see Bach & Jordan (2003)).

Denote by $\mathcal{F} = \{f = \sum_{j=1}^{p_1} f_j(x_j) | f_j \in \mathcal{F}_j\}$ and $\mathcal{G} = \{g = \sum_{k=1}^{p_2} g_k(y_k) | g_k \in \mathcal{G}_k\}$ the sets of additive functions of x and y, respectively.

We are given n independent tuples of the form $(X_i, Y_i)_{i=1}^n$ where $X_i = \{X_{i1}, \ldots, X_{ip_1}\}$ and $Y_i = \{Y_{i1}, \ldots, Y_{ip_2}\}$, and positive definite kernel functions on each covariate of X and Y. We denote the Gram matrix for the j^{th} X covariate by K_{xj} and for the k^{th} Y covariate by K_{yk}.

We will need to regularize the CCA problem to enforce smoothness and sparsity of the functions. The two norms

$$\|f_j\|_{\mathcal{F}_j} = \sqrt{\langle f_j, f_j \rangle_{\mathcal{F}_j}} \qquad \|f_j\|_2 = \sqrt{\tfrac{1}{n}\textstyle\sum_{i=1}^n f_j^2(X_{ij})}$$

play an important role in our approach. We can now formulate the *sparse additive kernel CCA* (SA-KCCA) problem as

$$\max_{f \in \mathcal{F}, g \in \mathcal{G}} \ \frac{1}{n}\sum_{i=1}^n f(X_i) g(Y_i) \quad \text{subject to} \tag{5}$$

$$\frac{1}{n}\sum_{i=1}^n f^2(X_i) + \gamma_f \sum_{j=1}^{p_1} \|f_j\|_{\mathcal{F}_j}^2 \le 1 \qquad \sum_{j=1}^{p_1} \|f_j\|_2 \le C_f$$

$$\frac{1}{n}\sum_{i=1}^n g^2(Y_i) + \gamma_g \sum_{k=1}^{p_2} \|g_k\|_{\mathcal{G}_k}^2 \le 1 \qquad \sum_{k=1}^{p_2} \|g_k\|_2 \le C_g.$$

for given regularization parameters γ_f, γ_g, C_f and C_g. As with the group LASSO, constraining $\sum_j \|f_j\|_2$ encourages sparsity amongst the functions f_j Ravikumar et al. (2009). As stated, this is an infinite dimensional optimization problem over Hilbert spaces. However, a straightforward application of the representer theorem shows that it is equivalent to the following finite dimensional optimization problem:

$$\max_{\alpha, \beta} \ \frac{1}{n}\left(\sum_{j=1}^{p_1} K_{xj}\alpha_j\right)\left(\sum_{k=1}^{p_2} K_{yk}\beta_k\right) \quad \text{subject to} \tag{6}$$

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

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

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

Here α is an $(n \times p_1)$ matrix, α_j is its j^{th} column, β is an $(n \times p_2)$ matrix and β_k is its k^{th} column.

The problem (6) is not convex. However, if we fix the function g (or equivalently the coefficients β) the problem is convex in f (equivalently α), and vice-versa. This *biconvexity* leads to a natural optimization strategy for (6) which we describe below. However, this procedure only guarantees convergence to a local optimum and in practice we still need to be able to find a good initialization.

In the absence of the sparsity penalty the problem becomes an additive form of kernel CCA (Bach & Jordan, 2003). One could also consider alternative formulations that, for instance, separate the smoothness and variance constraints. One attractive feature of our formulation is that without the sparsity constraint the problem can be reduced to a generalized eigenvalue computation which can be solved optimally. This leads us to a strategy of biconvex optimization that mirrors the linear algorithm of Witten et al. (2009); specifically, initialize by solving the problem without the sparsity constraints, fix α and optimize for β and vice-versa until convergence. As our experiments will show this is indeed a good strategy when $p_1, p_2 < n$. However, new ideas, to be described in Section 4, are necessary to scale this to the high dimensional setting where $p_1, p_2 > n$.

3. Sparse additive functional CCA

We now formulate an optimization problem for sparse additive functional CCA (SA-FCCA), and derive a scalable backfitting procedure for this problem. Here we work directly over the Hilbert spaces $L_2(\mu(x))$ and $L_2(\mu(y))$. We will denote by $\mathcal{S}_j$ the subspace of $\mu(x_j)$ measurable functions with mean 0, with the usual inner product $\langle f_j, f_j' \rangle = \mathbb{E}\left(f_j(X_j) f_j'(X_j)\right)$, and similarly $\mathcal{T}_k$ for the functions of y.

To enforce smoothness we consider functions lying in a ball in a second order Sobolev space. We further assume the functions are uniformly bounded, and the measures μ are supported on a compact subset of a Euclidean space with Lebesgue measure λ. For a fixed uniformly bounded, orthonormal basis ψ_{jk} with respect to λ we have

$$\mathcal{F}_j = \left\{ f_j \in \mathcal{S}_j : f_j = \sum_{k=0}^{\infty} \beta_{jk}\psi_{jk}, \sum_{k=0}^{\infty} \beta_{jk}^2 k^4 \le C^2 \right\}$$

and similarly for $\mathcal{G}_k$. We will call these the *smooth* functions, and denote by $\mathcal{F}$ and $\mathcal{G}$ the set of smooth additive functions over the respective Hilbert spaces.

Our formulation of *sparse additive functional CCA* is the optimization

$$\max_{f \in \mathcal{F},\ g \in \mathcal{G}} \frac{1}{n}\sum_{i=1}^n f(X_i) g(Y_i) \tag{7}$$

$$\text{s.t.} \quad \frac{1}{n}\sum_{j=1}^{p_1}\sum_{i=1}^{n} f_j^2(X_{ij}) \le 1, \qquad \sum_{j=1}^{p_1}\|f_j\|_2 \le C_f$$
$$\frac{1}{n}\sum_{k=1}^{p_2}\sum_{i=1}^{n} g_k^2(Y_{ik}) \le 1, \qquad \sum_{k=1}^{p_2}\|g_k\|_2 \le C_g$$

where the $\|.\|_2$ norm is defined as in additive kernel CCA. This problem is superficially similar to (2); however, there are three important differences. First, we don't regularize for smoothness but instead work directly over a Sobolev space of smooth functions. Secondly, we do not constrain the variance of the function f. Instead, in the spirit of "diagonal penalized CCA" of Witten et al. (2009) we constrain the sum of the variances of the individual f_js. This choice is made primarily because it leads to backfitting updates that have a particularly simple and intuitive form. Perhaps most importantly, we can no longer appeal to the representer theorem since we are not working over RKHSs.

We study the population version of this problem to derive a biconvex backfitting procedure to directly optimize this criterion. The sample version of the algorithm is described in Algorithm 1, and a complete derivation is part of the supplementary material. To gain some intuition for this procedure we describe one special case of the population algorithm, where g is fixed and both constraints on f are tight. Consider the Lagrangian problem

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

The norms are defined as $\|f\|_1 = \sum_{j=1}^{p_1}\sqrt{\mathbb{E}(f_j^2(x_j))}$ and $\|f\|_2^2 = \sum_{j=1}^{p_1}\mathbb{E}(f_j^2(x_j))$. For simplicity, consider the case when $\lambda, \gamma > 0$, and denote $a \equiv g(Y)$.

We now can derive a coordinate ascent style procedure where we optimize over f_j holding the other functions fixed. The Fréchet derivative w.r.t. f_j in the direction η gives one of the KKT conditions $\mathbb{E}[(a - 2\lambda f_j - \gamma\nu_j)\eta] = 0$ for all η in the Hilbert space $\mathcal{H}_j$, where the subdifferential is $\nu_j = \frac{f_j}{\sqrt{\mathbb{E}(f_j^2)}}$ if $\sqrt{\mathbb{E}(f_j^2)}$ is not 0, and is the set $\{u_j \in \mathcal{H}_j \,|\, \mathbb{E}(u_j^2) \le 1\}$ if $\sqrt{\mathbb{E}(f_j^2)} = 0$.

Using iterated expectations the KKT condition can be written as $\mathbb{E}[(\mathbb{E}(a \,|\, X_j) - 2\lambda f_j - \gamma\nu_j)\eta] = 0$. Denote $E(a \,|\, X_j) \equiv P_j$. In particular, if we consider $\eta = \mathbb{E}[(\mathbb{E}(a \,|\, X_j) - 2\lambda f_j - \gamma\nu_j]$, we can see that $\mathbb{E}[(\mathbb{E}(a \,|\, X_j) - 2\lambda f_j - \gamma\nu_j)] = 0$, i.e., $\mathbb{E}(a \,|\, X_j) - 2\lambda f_j - \gamma\nu_j = 0$ almost everywhere.

Then if $\sqrt{\mathbb{E}(P_j^2)} \le \gamma$, we have $f_j = 0$, and we arrive at the following soft thresholding update:

$$f_j = \frac{1}{2\lambda}\left[1 - \frac{\gamma}{\sqrt{\mathbb{E}(P_j^2)}}\right]_+ P_j.$$

Now, going back to the constrained version, we need to select γ and λ so that the two constraints are tight. To get the sample version of this update we replace the conditional expectation P_j by an estimate $S_j a$, where S_j is a locally linear smoother.

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** $\sum_{j=1}^{p_1}\|f_j\|_2 \le C_f$, break
   else let $\mathcal{F}_m$ denote the functions with maximum $\|.\|_2$ norm. Set all other functions to 0. For each $f \in \mathcal{F}_m$, set $f \leftarrow \frac{C_f f}{|\mathcal{F}_m|\|f\|_2}$. If $\sum_{j=1}^{p_1}\|f_j\|_2^2 \le 1$, break
   else set $f_j \leftarrow \left(1 - \frac{\gamma}{\sqrt{\|S_j g\|_2}}\right)_+ \frac{S_j g}{\lambda}$ where $\lambda = \sqrt{\sum_{j=1}^{p_1}\left\|\left(1 - \frac{\gamma}{\sqrt{\|S_j g\|_2}}\right)_+ S_j g\right\|_2^2}$ and γ 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

4. Marginal Thresholding

The formulations of SA-KCCA and SA-FCCA above are not jointly convex, but are biconvex. Hence, iterative optimization algorithms may not be guaranteed to reach the globally optimal solution. To address this issue, we first run the algorithms without any sparsity constraint. The resulting nonsparse collections of functions are then used as initializations for the algorithm that incorporates the sparsity penalties. While such initialization works well for low dimensional problems, as p increases, the performance of the estimator goes down (Figure 1). To extend the algorithms to the high dimensional scenario, we propose marginal thresholding as a screening method to reject irrelevant variables and run the SA-FCCA and SA-KCCA models on the reduced dimensionality problem. For each pair

Init	p=10	p=25	p=50
Random	0.05	0.009	-0.02
Non-sparse	**0.97**	0.62	0.26

Table 1. Test correlation from functions estimated by SA-FCCA for $n = 75$ samples, where $Y_1 = X_1^2$, all other dimensions are Gaussian noise. Random initializations don't work well for all data sizes. Initializing with the non-sparse formulation works well when $n > p$, but fails as $p \geq n$.

of variables X_i and Y_j, we fit marginal functions to that pair by optimizing the criteria in either Equation (6) or Equation (7) *without* the sparsity constraints since we only consider one X and one Y covariate at a time. We then compute the correlation on held out data. This constructs a matrix M of size $p_1 \times p_2$ with (i, j) entry of the matrix representing an estimate of the marginal correlation between $f_i(X_i)$ and $g_j(Y_j)$. We then threshold the entries of M to obtain a subset of variables on which to run SA-FCCA and SA-KCCA. Theorem 5.3 discusses the theoretical properties of marginal thresholding as a screening procedure, and Section 6.2 presents results on marginal thresholding for high dimensional problems.

5. Main theoretical results

In this section we will characterize both the functional and kernel marginal thresholding procedures and study the theoretical properties of the estimators (6) and (7). We will state the main theorems and defer all proofs to the supplementary material.

The theoretical characterization of these procedures relies on *uniform* large deviation inequalities for the covariance between functions. For simplicity in this section we will assume all the univariate spaces are identical. In the RKHS case we restrict our attention to functions in a ball of a *constant* radius in the Hilbert space associated with a reproducing kernel K. In the functional case the univariate space is a second order Sobolev space where the integral of the square of the second derivative is bounded by a *constant*. With some abuse of notation we will denote these spaces $\mathcal{C}$. We are interested in controlling the quantity

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

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

All results extend to the case when each covariate is endowed with a possibly distinct function space.

Lemma 5.1 (Uniform bound over RKHS) *Assume* $\sup_x |K(x,x)| \leq M < \infty$, *for functions* $f_j(x) = \sum_{i=1}^{n} \alpha_{ij} K_x(x, X_{ij})$, $g_k(y) = \sum_{i=1}^{n} \beta_{ik} K_y(y, Y_{ik})$

$$\mathbb{P}\left(\Theta_n \geq \underbrace{\zeta + C\sqrt{\frac{\log\left((p_1 p_2)/\delta\right)}{n}}}_{\epsilon} \right) \leq \delta$$

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

Note that ζ is independent of the dimensions p_1 and p_2 and that under the assumption that K is bounded, $\zeta = O(1/\sqrt{n})$. In some cases however this term can be much smaller. The second term depends only logarithmically on p_1 and p_2 and this *weak* dependence is the main reason our proposed procedures are consistent even when $p_1, p_2 > n$.

Lemma 5.2 (Uniform bound for Sobolev spaces) *Assume* $\|f\|_\infty \leq M \leq \infty$, *then*

$$\mathbb{P}\left(\Theta_n \geq \underbrace{\frac{C_1}{\sqrt{n}} + C_2\sqrt{\frac{\log\left((p_1 p_2)/\delta\right)}{n}}}_{\epsilon} \right) \leq \delta$$

where C_1 *and* C_2 *depend only on* M.

Lemma 5.1 is proved via a Rademacher symmetrization argument of Bartlett & Mendelson (2002) (see also Gretton et al. (2004)) while Lemma 5.2 is based on a bound on the bracketing integral of the Sobolev space (see Ravikumar et al. (2009)). The Rademacher bound gives a distribution dependent bound which can in some cases lead to faster rates.

We are now ready to characterize the marginal thresholding procedure described in Section 4. To study marginal thresholding we need to define *relevant* and *irrelevant* covariates. 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 $\alpha_j = 0$ and relevant if $\alpha_j > 0$. Similarly, for each Y_k we associate β_k defined analogously.

Now, assume that for every pair of covariates, we find the maximizer of the SA-FCCA or SA-KCCA objective over the given sample, over the appropriate class $\mathcal{C}$ and with $\mathbb{E}(f_j^2) \leq 1, \mathbb{E}(g_k^2) \leq 1$. Recall that for marginal thresholding we do not enforce sparsity. The global maximization of the SA-KCCA objective can be efficiently carried out since it is equivalent to a generalized eigenvalue problem. For SA-FCCA however, the backfitting procedure is only guaranteed to find the global maximizer in the population setting.

Theorem 5.3 *Given* $\mathbb{P}(\Theta_n \geq \epsilon) \leq \delta$.

1. *With probability at least* $1-\delta$, *marginal thresholding at* ϵ *has no false inclusions.*
2. *Further, if we have that* α_j *or* $\beta_k \geq 2\epsilon$ *then under the same* $1-\delta$ *probability event marginal thresholding at* ϵ *correctly includes the relevant covariate* X_j *or* Y_k.

The importance of Lemmas 5.1 and 5.2 is that they provide values at which to threshold the marginal covariances. In particular, notice that the minimum sample covariance that can be reliably detected, with no false inclusions, falls rapidly with n and approaches zero even when $p_1, p_2 > n$.

In the spirit of early results on the LASSO of Juditsky & Nemirovski (2000); Greenshtein & Ritov (2004) we will establish the risk consistency or *persistence* of the empirical maximizers of the two objectives. Although we cannot guarantee that we find these empirical maximizers due to the non-convexity this result shows that with good initialization the formulations (6) and (7) can lead to solutions which have good statistical properties in high dimensions.

For SA-KCCA we will assume that our algorithm maximizes

$$\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=1}^{p_1}\mu_j f_j(X_{ij})\right]\left[\sum_{k=1}^{p_2}\gamma_k g_k(Y_{ik})\right]$$

over the classes

$$\mathcal{F} = \left\{ f : f(x) = \sum_{j=1}^{p_1}\mu_j f_j(x_j), \mathbb{E}f_j = 0, \mathbb{E}f_j^2 = 1, \right. \\ \left. \|\mu\|_1 \leq C_f, \|\mu\|_2^2 + \gamma_f \sum_{j=1}^{p_1}\|f_j\|_{\mathcal{H}}^2 \leq 1 \right\}$$

$$\mathcal{G} = \left\{ g : g(x) = \sum_{k=1}^{p_2}\gamma_k g_k(y_k), \mathbb{E}g_k = 0, \mathbb{E}g_k^2 = 1, \right. \\ \left. \|\gamma\|_1 \leq C_g, \|\gamma\|_2^2 + \gamma_g \sum_{k=1}^{p_2}\|f_k\|_{\mathcal{H}}^2 \leq 1 \right\}$$

and for SA-FCCA we will assume that our algorithm maximizes the same objective over the same class without the RKHS constraint but which are instead in a Sobolev ball of constant radius. Denote these solutions $(\hat{f}, \hat{g})$.

We will compare to an *oracle* which maximizes the population covariance

$$\mathrm{cov}(f,g) \equiv \mathbb{E}\left[\sum_{j=1}^{p_1}\mu_j f_j(x_j)\right]\left[\sum_{k=1}^{p_2}\gamma_k g_k(y_k)\right]$$

Denote this maximizer by (f^*, g^*). Our main result will show that these procedures are *persistent*, i.e., $\mathrm{cov}(f^*, g^*) - \mathrm{cov}(\hat{f}, \hat{g}) \to 0$ even if $p_1, p_2 > n$.

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.*

6. Experiments

6.1. Non-linear correlations

We compare SA-FCCA and SA-KCCA with two models, sparse additive linear CCA (SCCA) (Witten et al., 2009) and kernel CCA (KCCA) (Bach & Jordan, 2003). Figure 1 shows the performance of each model, when run on data with $n = 150$ samples in $p_1 = 15$, $p_2 = 15$ dimensions, where only one relevant variable is present in X and Y (the remaining dimensions are Gaussian random noise). We report two metrics to measure whether the correct correlations are being captured by the different methods - (a) test correlation on 200 samples, using the estimated functions, and (b) precision and recall in identifying the correct variables involved in the correlation estimation. Each result is averaged over 10 repeats of the experiment. Since KCCA uses all data dimensions in finding correlations, its precision and recall are not reported.

When the relationship between the relevant variables is linear, all methods identify the correct variables and have high test correlation. While KCCA should be able to identify non-linear correlations, since it is strongly affected by the curse of dimensionality, it has poor test correlation even in $p = 15$ dimensions.

Both SA-FCCA and SA-KCCA correctly identify the relevant variables in all cases, and have high test correlation.

6.2. Marginal thresholding

We now test the efficiency of marginal thresholding by running an experiment for $n = 150$, $p_1 = 150$, $p_2 = 150$. We generate multiple relevant variables as:

$$f_i(X_i) = \cos\left(\frac{\pi}{2}X_i\right),\ i \in \{1,3\},\ f_i(X_i) = X_i^2,\ i \in \{2,4\}$$

$$Y_j = \sum_{i=1; i\neq j}^{4} f_i(X_i) + \mathcal{N}(0, 0.1^2) \quad j \in \{1,2,3,4\}$$

Thus, there are four relevant variables in each data set. X and Y are sampled from a uniform distribution, and standardized before computing $f_i(X_i)$. Each $f_i(X_i)$ is

Model	Test correlation SA-FCCA	SA-KCCA	SCCA	KCCA	Precision/Recall SA-FCCA	SA-KCCA	SCCA
$Y = X^2$	0.96	**0.99**	0.05	0.44	1/1	1/1	0.28/0.14
$Y = \text{abs}(X)$	**0.98**	**0.99**	0.06	0.35	1/1	1/1	0/0
$Y = \cos(X)$	0.94	**0.99**	0.071	0.04	1/1	1/1	0.1/0.1
$\log(Y) = \sin(X)$	0.91	**0.93**	0.22	0.09	1/1	1/1	0.71/0.66
$Y = X$	**0.99**	**0.99**	**0.99**	**0.98**	1/1	1/1	1/1

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.

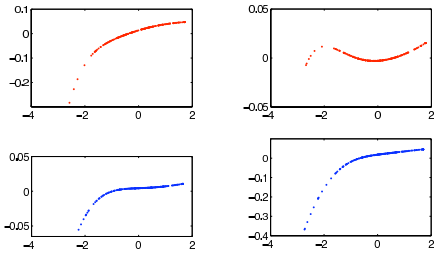

Figure 2. DLBCL data : The top row shows two of the functions $f_i(X_i)$ with non-zero norms for X in red, and the bottom row shows two functions $g_j(Y_j)$ with non-zero norms for Y in blue.

also standardized before computing Y_j. We repeat the experiment by generating data 10 times, and report results in Table 2. Bandwidth in the different methods was selected using a plug-in estimator of the median distance between points in a single dimension. The sparsity and smoothness parameters for all methods were tuned using permutation tests, as described in Witten et al. (2009), assuming that $C_f = C_g = C$, and $\gamma_f = \gamma_g = \gamma$.

We ran marginal thresholding by splitting the data into equal sized train and held out data, fitting marginal functions on the train data, computing functional correlation on the held out data, and picking a threshold so that $n/5$ elements of the thresholded correlation matrix are non-zero. We found that in all experiments, marginal thresholding always selected the relevant variables for the subsampled data. Table 2 shows the precision, recall and test correlations for the different methods. As can be expected, SA-FCCA and SA-KCCA are able to correctly identify the relevant variables, and the estimated functions have high correlation on test data.

We visualize the effect of the parameter tuning by plotting regularization paths, as the sparsity parameter is varied (n=100, p_1=p_2=12). For SA-FCCA and SA-KCCA, the norm of each function is plotted, and for sparse linear CCA, the absolute values of the entries of u and v are shown. Figure 3 shows how, unlike SCCA, SA-FCCA and SA-KCCA are able to separate the relevant and non-relevant variables over the entire range of the sparsity parameter.

6.3. Application to DLBCL data

We apply our non-linear CCA models to a data set of comparative genomic hybridization (CGH) and gene expression measurements from 203 diffuse large B-cell lymphoma (DLBCL) biopsy samples (Lenz, 2008). We obtained 1500 CGH measurements from chromosome

Method	Test correlation	Precision	Recall
SA-FCCA	**0.94**	1	0.785
SA-KCCA	**0.98**	0.95	0.8
SCCA	0.02	0.02	0.36
KCCA	0.07	N/A	N/A

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

1 of the data, and 1500 gene expression measurements from genes on chromosome 1 and 2 of the data. The data was standardized,and Winsorized so that the data lies within two times the mean absolute deviation.

We used marginal thresholding to reduce the dimensionality of the problem, and then ran SA-FCCA. Permutation tests were used to pick an appropriate bandwidth and sparsity parameter, as described in Witten et al. (2009). We found that the model picked interesting non-linear relationships between CGH and gene expression data. Figure 2 shows the functions extracted by the SA-FCCA model from this data. Even though this data has been previously analyzed using linear models, we do not necessarily expect gene expression measurements from Affymetrix chips to be linearly correlated with array CGH measurements, even if the specific CGH mutation is truly affecting the gene expression. Further, the extracted functions in Figure 2 suggest that the changes in gene expression are dependent on the CGH measurements via a saturation function - as the copy number increases, the gene expression increases, until it saturates to a fixed level, beyond which increasing the copy numbers does not lead to an increase in expression. From a systems biology view point, such a prediction seems reasonable since single CGH mutations will not affect other pathways that are required to be activated for large changes in gene expression.

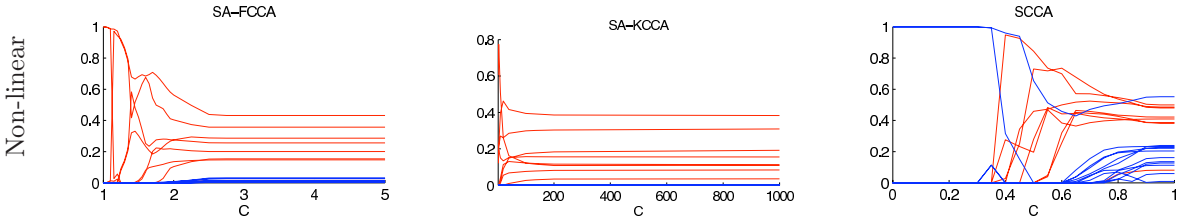

Figure 3. Regularization paths for non-linear correlations in the data, for SA-FCCA, SA-KCCA and SCCA resp. The paths for the relevant variables (in X and Y) are shown in red, the irrelevant variables are shown in blue.

7. Discussion

In this paper we introduced two proposals for nonparametric CCA and demonstrated their effectiveness both in theory and practice. Several interesting questions and extensions remain. CCA is often run on more than two data sets, and one is often interested in more than just the *principal* canonical direction. Chen & Liu (2012) have proposed group sparse linear CCA for situations when a grouping of the covariates is known. These extensions all have natural nonparametric analogues which would be interesting to explore. As in the case of regression (Koltchinskii & Yuan, 2010), the KCCA formulation considered in this paper can also be generalized to involve multiple kernels and kernels over groups of variables in a straightforward way.

While thresholding marginal correlations one can imagine exploiting the structure in the correlations. In particular, in the $(p_1 \times p_2)$ marginal correlations matrix we are looking for a *bicluster* of high entries in the matrix. Leveraging this structure could potentially allow us to detect weaker marginal correlations. Finally, an important application of kernel CCA is as a contrast function in independence testing. The additive formulations we have proposed allow for independence testing over more restricted alternatives but can be used to construct *interpretable* tests of independence.

Acknowledgements

Research supported in part by NSF grant IIS-1116730, AFOSR contract FA9550-09-1-0373, and NIH grant R01-GM093156-03.

References

Bach, Francis R. and Jordan, Michael I. Kernel independent component analysis. *JMLR*, 3:1–48, March 2003.

Bartlett, P. L. and Mendelson, S. Rademacher and Gaussian complexities: Risk bounds and structural results. *JMLR*, 3:463–482, 2002.

Breiman, Leo and Friedman, Jerome H. Estimating optimal transformations for multiple regression and correlation. *JASA*, 80(391):pp. 580–598, 1985.

Chen, Xi and Liu, Han. An efficient optimization algorithm for structured sparse cca, with applications to eqtl mapping. *Statistics in Biosciences*, pp. 1–24, 2012.

Fan, Jianqing and Song, Rui. Sure independence screening in generalized linear models with NP-dimensionality. *Ann. Statist.*, 38(6):3567–3604, 2010.

Greenshtein, Eitan and Ritov, Ya'acov. Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, 10(6):971–988, 2004.

Gretton, A., Smola, A., Bousquet, O., Herbrich, R., , Schoelkopf, B., and Logothetis, N. Behaviour and convergence of the constrained covariance. Technical Report 130, MPI for Biological Cybernetics, 2004.

Hastie, Trevor and Tibshirani, Robert. Generalized additive models. *Statistical Science*, 1(3):pp. 297–310, 1986.

Hotelling, Harold. Relations between two sets of variates. *Biometrika*, 28(3/4):pp. 321–377, 1936.

Johnstone, Iain M. and Lu, Arthur Yu. On consistency and sparsity for principal components analysis in high dimensions. *JASA*, 104(486):682–693, 2009.

Juditsky, Anatoli and Nemirovski, Arkadii. Functional aggregation for nonparametric regression. *Ann. Statist.*, 28 (3):681–712, 2000.

Koltchinskii, Vladimir and Yuan, Ming. Sparsity in multiple kernel learning. *Ann. Statist.*, 38(6):3660–3695, 2010.

Lenz, G. et. al. Molecular subtypes of diffuse large B-cell lymphoma arise by distinct genetic pathways. *Proc. Natl. Acad. Sci. U.S.A.*, 105:13520–13525, Sep 2008.

Meier, Lukas, van de Geer, Sara, and Bühlmann, Peter. High-dimensional additive modeling. *Ann. Statist.*, 37 (6B):3779–3821, 2009.

Parkhomenko, E, Tritchler, D, and Beyene, J. Genome-wide sparse canonical correlation of gene expression with genotypes. *BMC Proc*, 1 Suppl 1, 2007.

Raskutti, Garvesh, Wainwright, Martin J., and Yu, Bin. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *JMLR*, 08 2010.

Ravikumar, Pradeep, Lafferty, John, Liu, Han, and Wasserman, Larry. Sparse additive models. *JRSSB (Statistical Methodology)*, 71(5):1009–1030, 2009.

Witten, D. M. and Tibshirani, R. J. Extensions of sparse canonical correlation analysis with applications to genomic data. *Stat Appl Genet Mol Biol*, 8:Article28, 2009.

Witten, Daniela M., Tibshirani, Robert, and Hastie, Trevor. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534, 2009.

A. Supplementary Material

A.1. Discussion of SA-FCCA v/s SA-KCCA

SA-FCCA and SA-KCCA offer different advantages and disadvantages and neither is completely dominated by the other. The methods are two instances of the same approach, which is to use a nonparametric additive model.

From an optimization perspective, SA-KCCA works over RKHS, leading to an optimization problem over a finite parameter space for which strong convergence guarantees can be made. For SA-FCCA however, we use backfitting, which is typically known to converge only in the population setting. From a statistical perspective, stronger results are known for the kernel version in the regression setting. From a practitioner's perspective, these algorithms perform comparably statistically. Computationally, the SA-FCCA algorithm is considerably more simple - after some pre-computations, the coordinate descent back-fitting algorithm only requires matrix-vector multiplications in each iteration, and typically converges in a small number of iterations. SA-KCCA requires us to optimize a second order cone-program which, although convex, is not amenable to fast coordinate descent algorithms.

There is a clear dichotomy here from a statistical/optimization theory perspective, we would recommend the SA-KCCA formulation but from a practical perspective we would recommend the SA-FCCA formulation.

Computational costs: The computational cost of each inner loop optimization of SA-FCCA when it is done to an accuracy of ϵ is $O(n^2 \max(p_1, p_2)/\epsilon)$ using the algorithm we propose. SA-KCCA using a standard interior point solver has complexity $O(n^3 \max(p_1, p_2)^3 \log(1/\epsilon))$. SA-FCCA also requires a pre-computation of smoother matrices which takes $O(n^3 \max(p_1, p_2))$. These methods typically require a small number of outer-loop iterations to converge.

It is also worth noting that these non-parametric methods are more computationally intensive than both sparse linear CCA which requires $O(n^2/\epsilon)$ for each inner loop iteration, and kernel CCA which requires $O(n^2 \log n)$ in total after computing the Gram matrices.

Notice also that in linear CCA we are learning $p_1 + p_2$ parameters, in kernel CCA we are learning $2n$ parameters, while in SA-KCCA we are learning the much larger $n(p_1 + p_2)$ parameters. A direct comparison of the number of parameters in SA-FCCA is subtle, since at least from a degrees of freedom perspective this depends on the smoothness of the target function.

A.2. A derivation of the backfitting algorithm for FCCA

In this section we derive the biconvex backfitting algorithm for FCCA. In particular, consider the case when g is fixed and let a denote the vector of $(g(Y_1), \ldots, g(Y_n))^T$ in the sample setting, and let it denote the function g in the population setting.

It is instructive to first consider the population setting. The optimization problem becomes

$$\begin{aligned} \max_{f \in \mathcal{F}} \quad & \mathbb{E}[f(X)a] \\ \text{subject to} \quad & \|f\|_2^2 \leq 1 \\ & \|f\|_1 \leq C_f \end{aligned}$$

The norms are defined as $\|f\|_1 = \sum_{j=1}^{p_1} \sqrt{\mathbb{E}(f_j^2(x_j))}$ and $\|f\|_2^2 = \sum_{j=1}^{p_1} \mathbb{E}(f_j^2(x_j))$.

Consider the Lagrange problem,

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

The Frètchet derivative w.r.t. f_j along the direction η gives one of the KKT conditions $\mathbb{E}[(a - 2\lambda f_j - \gamma \nu_j)\eta] = 0$

for all η in the Hilbert space $\mathcal{H}_j$, where $\nu_j = \frac{f_j}{\sqrt{\mathbb{E}(f_j^2)}}$ if $\sqrt{\mathbb{E}(f_j^2)}$ is not 0, and is the set $\{u_j \in \mathcal{H}_j | \mathbb{E}(u_j^2) \leq 1\}$ if $\sqrt{\mathbb{E}(f_j^2)} = 0$.

Using iterated expectations the KKT condition can be written as $\mathbb{E}[(\mathbb{E}(a|X_j) - 2\lambda f_j - \gamma\nu_j)\eta] = 0$. Now, if we denote $E(a|X_j) = P_j$. In particular, if we consider $\eta = \mathbb{E}[(\mathbb{E}(a|X_j) - 2\lambda f_j - \gamma\nu_j]$, we can see that

$$\begin{aligned} \mathbb{E}[(\mathbb{E}(a|X_j) - 2\lambda f_j - \gamma\nu_j)] &= 0 \\ \text{i.e. } \mathbb{E}(a|X_j) - 2\lambda f_j - \gamma\nu_j &= 0 \text{ almost everywhere} \\ P_j - 2\lambda f_j &= \gamma\nu_j \end{aligned}$$

then if $\sqrt{\mathbb{E}(P_j^2)} \leq \gamma$, we have $f_j = 0$, and we arrive at the following

$$f_j \left(2\lambda + \frac{\gamma}{\sqrt{\mathbb{E}(f_j^2)}} \right) = P_j \text{ if } \sqrt{\mathbb{E}(P_j^2)} > \gamma$$

$$f_j = 0 \text{ otherwise}$$

and this gives the following soft threshold update:

$$f_j = \frac{1}{2\lambda} \left[1 - \frac{\gamma}{\sqrt{\mathbb{E}(P_j^2)}} \right]_+ P_j$$

We analyze the Lagrangian in the 3 cases (i.e. all constraints are tight, only the 2-norm constraints are tight, and only the 1-norm constraints are tight).

1. When only the 2-norm constraint is tight, $\gamma = 0$ and λ is selected to make the 2-norm be 1.
2. When only the 1-norm constraint is tight, we use the equation above with $\lambda = 0$ and see that only the f_js with the largest $\sqrt{\mathbb{E}(P_j^2)}$ are non-zero.
3. When both constraints are tight, we use the soft-threshold update with λ and γ selected to make both constraints tight.

Now, we can define the algorithm in the finite sample case, as an analog of the algorithm for the basic problem in the linear case. For a fixed g, FCCA problem can be solved using the following algorithm.

1. Test for case 1 by setting $f_j(X_j) = \frac{S_j a}{\lambda}$ for each j, where $\lambda^2 = \frac{1}{n}\sum_{j=1}^p \|S_j a\|_2^2$. If the solution satisfies $\|f\|_1 \leq c_1$ this is the required f.
2. Test for case 2, in this case we find $\|S_j a\|_2$ for each j, and find all k such that $\|S_k a\|_2 \geq \|S_j a\|_2$ for all j. Denote the cardinality of this set ϕ. Set

   $$f_k(X_k) = \frac{C_f S_k a}{\phi \|S_k a\|}$$

   for all k such that $\|S_k a\|_2 \geq \|S_j a\|_2$ for all j, and all other $f_j = 0$. If $\|f\|_2 \leq 1$ this is the required f.
3. If neither of the above cases are satisfied then in this case $f_j(X_j) = \frac{\mathcal{S}_\gamma(S_j a)}{\lambda}$ where $\lambda^2 = \frac{1}{n}\sum_{j=1}^p \|\mathcal{S}_\gamma(S_j a)\|_2^2$ for each j. where γ is chosen so that $\|f\|_1 = C_f$.

Here S_j is a linear smoother and is used to estimate the conditional expectation of a given X_j, i.e. if $P_j = \mathbb{E}(a|X_j)$ then $\hat{P}_j = S_j a$.

A.3. Uniform bounds

We will first prove Lemma 5.1 and then give a proof sketch for Lemma 5.2.

Proof

We will limit our attention to functions

$$f_j \in B_{\mathcal{H}}(1)$$

since the general case for a constant radius follows by a simple rescaling argument. We have the condition

$$\sup_x |K(x,x)| \leq c < \infty$$

This also implies the uniform boundedness of the univariate functions by a simple argument.

$$\sup_x |f_j(x)| = \sup_x |\langle f_j, K(\cdot, x)\rangle| \leq \sup_x \|f_j\|_{\mathcal{H}} \sqrt{K(x,x)}$$

Thus, we have

$$\sup_x |f_j(x)| \leq C$$

for some absolute constant C.

Recall that we wish to uniformly control

$$\Omega_n = \sup_{f_j, g_k \in \mathcal{C}, j \in \{1,\ldots,p_1\}, k \in \{1,\ldots,p_2\}} \left| \frac{1}{n} \sum_{i=1}^n f_j(X_{ij}) g_k(Y_{ik}) - \mathbb{E}(f_j(X_j) g_k(Y_k)) \right|$$

Let us first analyze

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

which is just Ω_n for a fixed pair j, k. The bound on Ω_n will then follow from a union bound.

Deviation from its expectation is a simple consequence of the boundedness of functions and McDiarmid's inequality, i.e. for some absolute constant C, we have

$$\mathbb{P}(\Theta_n - \mathbb{E}\Theta_n > t) \leq \exp\left(\frac{-nt^2}{C}\right)$$

Now, we need to understand the expectation. A symmetrization argument gives us

$$\mathbb{E}\Theta_n \leq 2\mathcal{R}(\mathcal{C})$$

where

$$\mathcal{R}(\mathcal{C}) = \mathbb{E}_{X,Y,\sigma}\left(\sup_{f_j, g_k \in \mathcal{C}} \frac{1}{n} \sum_{i=1}^n \sigma_i f_j(X_j) g_k(Y_k) \right)$$

A bound on $\mathcal{R}(\mathcal{C})$ is given by Lemma 16 in Gretton et al. (2004). They show,

$$\mathbb{E}_{X,Y,\sigma}\left(\sup_{f_j, g_k \in \mathcal{C}} \frac{1}{n} \sum_{i=1}^n \sigma_i f_j(X_j) g_k(Y_k) \right) \leq \frac{1}{n} \mathbb{E}_{X,Y} \sqrt{\sum_{i=1}^n K(X_{ij}, X_{ij}) K(Y_{ik}, Y_{ik})}$$

This gives us a bound on Θ_n, and to get a bound on Ω_n we just union bound over the $p_1 p_2$ possible choices for j, k.

Defining,

$$\zeta = \max_{j,k} \frac{2}{n} \mathbb{E}_{X \sim x_j, Y \sim y_k} \sqrt{\sum_{i=1}^{n} K(X_{ij}, X_{ij}) K(Y_{ik}, Y_{ik})}$$

We have for some absolute constant C

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

For SA-FCCAwe have a different class of functions. Ravikumar et al. (2009) show the following result for uniformly bounded (by a constant) f and g in a second order Sobolev space, for an absolute constant C,

$$\omega \equiv \mathbb{E}\left(\sup_{f_j, g_k \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^{n} f_j(X_{ij}) g_k(Y_{ik}) - \mathbb{E}(f_j(X_j) g_k(Y_k)) \right| \right) \leq \frac{C}{\sqrt{n}}$$

Since, the functions are uniformly bounded we can now use McDiarmid's inequality to get for some C'

$$\mathbb{P}\left(\sup_{f_j, g_k \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^{n} f_j(X_{ij}) g_k(Y_{ik}) - \mathbb{E}(f_j(X_j) g_k(Y_k)) \right| - \omega \geq t \right) \leq \exp\left(\frac{-t^2 n}{C'} \right)$$

Now, applying the union bound over j and k we get the lemma from the main paper. Again, defining

$$\Omega_n = \sup_{f_j, g_k \in \mathcal{C}, j \in \{1,\ldots,p_1\}, k \in \{1,\ldots,p_2\}} \left| \frac{1}{n} \sum_{i=1}^{n} f_j(X_{ij}) g_k(Y_{ik}) - \mathbb{E}(f_j(X_j) g_k(Y_k)) \right|$$

$$\mathbb{P}\left(\Omega_n \geq \frac{C_1}{\sqrt{n}} + C_2 \sqrt{\frac{\log\left((p_1 p_2)/\delta\right)}{n}} \right) \leq \delta$$

A.4. Marginal thresholding

In this section we prove the following result:

Theorem A.1 *Given*

$$\mathbb{P}\left(\sup_{f_j, g_k \in \mathcal{C}, j \in \{1,\ldots,p_1\}, k \in \{1,\ldots,p_2\}} \left| \frac{1}{n} \sum_{i=1}^{n} f_j(X_{ij}) g_k(Y_{ik}) - \mathbb{E}(f_j(X_j) g_k(Y_k)) \right| \geq \epsilon \right) \leq \delta$$

With probability at least $1 - \delta$, marginal thresholding at ϵ has no false inclusions. Further, if we have that α_j or $\beta_k \geq 2\epsilon$ then under the same $1 - \delta$ probability event marginal thresholding at ϵ correctly includes the relevant covariate X_j or Y_k.

Proof

The first part is straightforward. In particular, we know for any irrelevant X_j for any Y_k and $f_j, g_k \in \mathcal{C}$, $\mathbb{E} f_j(X_j) g_k(Y_k) = 0$, and in the at least $1 - \delta$ probability event we have

$$\max_{f_j, g_k \in \mathcal{C}} \frac{1}{n} \sum_{i=1}^{n} f_j(X_{ij}) g_k(Y_{ik}) < \epsilon$$

For the second part, consider a particular relevant covariate X_j, denote

$$\theta^* = \max_{f_j, g_k \in \mathcal{C}} \frac{1}{n} \sum_{i=1}^{n} f_j(X_{ij}) g_k(Y_{ik})$$

It suffices to show that if $\alpha_j \geq 2\epsilon \implies \theta^* \geq \epsilon$.

Denote, $(f_j^*, g_k^*) = \arg\sup_{f_k, g_k \in \mathcal{C}} \mathbb{E}(f_j(X_j)g_k(Y_k))$. Then in the at least $1-\delta$ probability event we have,

$$\theta^* \geq \frac{1}{n}\sum_{i=1}^{n} f_j^*(X_{ij})g_k^*(Y_{ik}) \geq \mathbb{E}(f_j^*(X_j)g_k^*(Y_k)) - \epsilon \geq \epsilon$$

A.5. Persistence

We will show the high dimensional persistence of the global optimizers of the SA-FCCAand SA-KCCAobjectives.

We will prove the result for SA-FCCAand give a proof sketch for SA-KCCA.

Let us assume that the SA-FCCAestimator is chosen to maximize the objective

$$\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=1}^{p_1}\mu_j f_j(X_{ij})\right]\left[\sum_{k=1}^{p_2}\gamma_k g_k(Y_{ik})\right]$$

over the classes

$$\mathcal{F} = \left\{ f : f(x) = \sum_{j=1}^{p_1}\mu_j f_j(x_j), \mathbb{E}f_j = 0, \quad \mathbb{E}f_j^2 = 1, \|\mu\|_1 \leq C_f, \|\mu\|_2^2 \leq 1 \right\}$$
$$\mathcal{G} = \left\{ g : g(x) = \sum_{k=1}^{p_2}\gamma_k g_k(y_k), \mathbb{E}g_k = 0, \quad \mathbb{E}g_k^2 = 1, \|\gamma\|_1 \leq C_g, \|\gamma\|_2^2 \leq 1 \right\}$$

An analogous role to risk in classification/regression problems is played by the (negative) covariance.

$$\mathrm{cov}(f,g) = \mathbb{E}\left[\sum_{j=1}^{p_1}\mu_j f_j(X_j)\right]\left[\sum_{k=1}^{p_2}\gamma_j g_k(Y_k)\right]$$

Theorem A.2 *If $p_1 p_2 \leq e^{n^\xi}$ for some $\xi < 1$. Then,*

$$\mathrm{cov}(f_n^*, g_n^*) - \mathrm{cov}(\hat{f}_n, \hat{g}_n) = O_P\left(\frac{C_f C_g}{n^{(1-\xi)/2}}\right) \tag{8}$$

If $C_f C_g = o(n^{(1-\xi)/2})$ the FCCA procedure described is persistent, i.e. $\mathrm{cov}(f_n^*, g_n^*) - \mathrm{cov}(\hat{f}_n, \hat{g}_n) \to 0$ as $n \to \infty$.

Proof

We can write

$$\mathrm{cov}(f,g) = \sum_{j=1}^{p_1}\sum_{k=1}^{p_2}\mu_j\gamma_k\mathbb{E}[f_j(X_j)g_k(Y_k)] \tag{9}$$

and

$$\hat{C}(f,g) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p_1}\sum_{k=1}^{p_2}\mu_j\gamma_k f_j(X_{ij})g_k(Y_{ik}) \tag{10}$$

Now, we have (using Holder's inequality)

$$|\hat{C}(f,g) - \mathrm{cov}(f,g)| \leq \|\mu\|_1\|\gamma\|_1 \max_{jk}\left[\frac{1}{n}\sum_{i=1}^{n} f_j(X_{ij})g_k(Y_{ik})\right] - \mathbb{E}(f_j(X_j)g_k(Y_k)) \tag{11}$$

p	5	10	25	50	75	100	150
Test correlation	0.9999	1.0000	1.0000	0.6846	0.9079	0.4967	0.2918
Precision	1.0000	1.0000	1.0000	0.7000	0.9000	0.5000	0.3000
Recall	1.0000	1.0000	1.0000	0.7000	0.9000	0.5000	0.3000

Table 3. Results for SCCAon linear data $Y_1 = X_1 + \mathcal{N}(0,1)$ with $n = 100$ samples. As p increases, the performance of the model decreases.

p	5	10	25	50	75	100	150
Test correlation	0.9672	0.9717	0.6178	0.2564	0.2040	0.0294	0.0959
Precision	1.0000	1.0000	0.6000	0.4000	0.2000	0	0.2000
Recall	1.0000	1.0000	0.6000	0.4000	0.2000	0	0.2000

Table 4. Results for SA-FCCA(without marginal thresholding) on quadratic data $Y_1 = X_1^2 + \mathcal{N}(0,1)$ with $n = 100$ samples. As p increases, the performance of the model decreases.

Now, we are almost done. Using Lemma 5.2 we know that we can uniformly bound

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

over all f_j, g_k in our function class and over all $j \in \{1,\dots,p_1\}, k \in \{1,\dots,p_2\}$. In particular, this quantity is $O_P\left(\sqrt{\frac{\log(p_1 p_2)}{n}}\right)$.

Now, this gives us that

$$|\hat{C}(f,g) - \text{cov}(f,g)| = O_P\left(C_f C_g \sqrt{\frac{\log\,(p_1 p_2)}{n}}\right) = O_P\left(\frac{C_f C_g}{n^{(1-\xi)/2}}\right) \tag{12}$$

Using this we have,

$$\text{cov}(\hat{f}_n, \hat{g}_n) \geq \hat{C}(\hat{f}_n, \hat{g}_n) - O_P\left(\frac{C_f C_g}{n^{(1-\xi)/2}}\right) \geq \hat{C}(f_n^*, g_n^*) - O_P\left(\frac{C_f C_g}{n^{(1-\xi)/2}}\right) \geq \text{cov}(f_n^*, g_n^*) - O_P\left(\frac{C_f C_g}{n^{(1-\xi)/2}}\right) \tag{13}$$

and the result follows.

The proof for the persistence of SA-KCCAfollows an almost identical argument. We make two minor modifications. As described in the main text we bound the Rademacher term as $O(1/\sqrt{n})$ by only using the boundedness of the kernel. We can then follow the proof of this theorem exactly, replacing the use of Lemma 5.2 with Lemma 5.1.

A.6. Marginal thresholding is needed to get high accuracy in high dimensions

We show that for both linear SCCA(Table 3) and non-linear SA-FCCA(Table 4) models to measure correlation, the models do not have good performance when $p \sim n$. Hence, using a screening procedure to extract variables of interest before running CCA is essential.

A.7. Simulation Details

This section describes how the simulated data was generated for the experiments in Section 6.1. The algorithm requires a function $f(x)$ that defines the relationship between X and Y. Four different functions were used, as defined in the results (Figure 1).

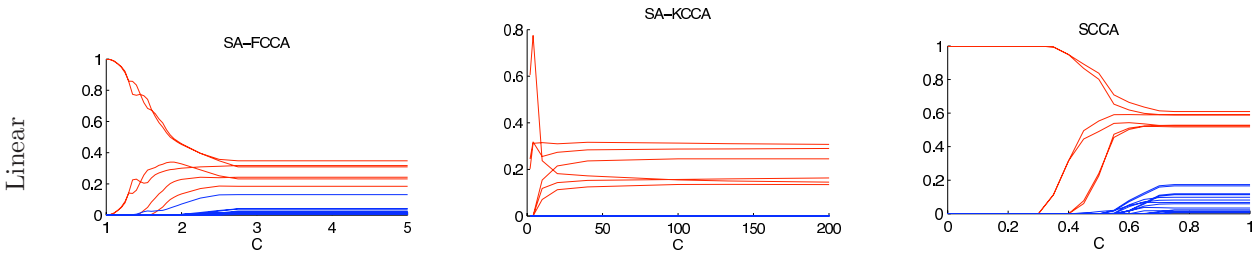

Figure 4. Regularization paths for linear correlations in the data, for SA-FCCA, SA-KCCA and SCCA resp. The paths for the relevant variables (in X and Y) are shown in red, the irrelevant variables are shown in blue.

Algorithm 2 Generate simulated data

input n, p_1, p_2, function $f(x)$.

1. Pick relevant feature r_x and r_y of X and Y randomly from $\{1, \cdots, p_1\}$ and $\{1, \cdots, p_2\}$ resp.
2. For $j = 1 \ldots p_1$
   For $i = 1 \ldots n$
   $X(i, j) = \mathcal{N}(0, 1)$;
3. For $j = 1 \ldots p_2$
   For $i = 1 \ldots n$
   $if (j == r_y)$
   $Y(i, r_y) = f(X(i, r_x)) + \mathcal{N}(0, 0.1^2)$;
   $else$
   $Y(i, j) = \mathcal{N}(0, 1)$;

output X, Y

A.8. Comparison of regularization paths

As the sparsity parameter is varied, different number of features are selected. We plot the regularization paths obtained by varying the sparsity parameter for linear data (Figure 4). The linear data was selected in a similar manner to Section 6.2 with $n = 100$, $p_1 = p_2 = 12$, so that X and Y have 4 relevant variables each.

For SA-FCCA and SA-KCCA, the norm of each function is plotted, and for sparse linear CCA, the absolute values of u and v are shown, as a function of the sparsity parameter. Figure 4 shows that when the true relationship between the variables is linear, all three models separate the relevant and irrelevant variables. Note that the bandwidth of SA-FCCA and SA-KCCA were not tuned in this problem, so both models are capable of extracting the correct linear relationships without adjusting the bandwidth heavily.

A.9. Contrast SA-FCCA with SA-KCCA on DLBCL data

We ran SA-KCCA on the DLBCL data, on which SA-FCCA results were reported in Section 6.3. We observed that the same co-variates were picked as relevant by both SA-FCCA and SA-KCCA. The functions extracted by SA-KCCA are shown in Figure 5. Note that the functions appear to be mirror-images of the ones extracted by SA-FCCA in Figure 2. Since a mirror image of the function still preserves the non-linear correlations, we conclude that SA-FCCA and SA-KCCA work comparably in such predictions.

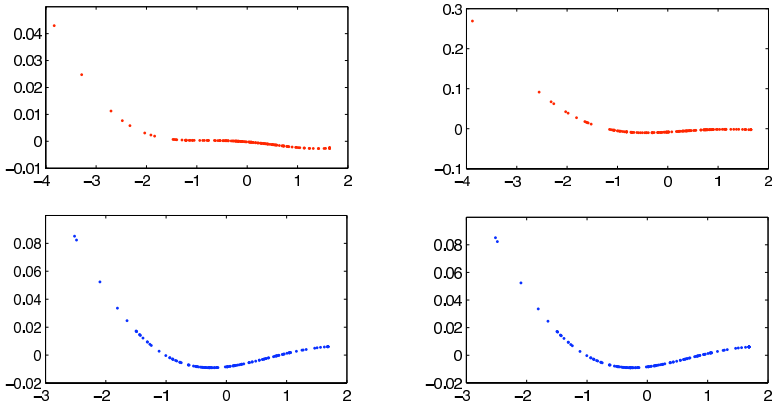

Figure 5. KCCA output on DLBCL data : The top row shows two of the functions $f_i(X_i)$ v/s X_i with non-zero norms for X in red, and the bottom row shows two functions $g_j(Y_j)$ v/s $Y_j$with non-zero norms for Y in blue.