Efficient Dimensionality Reduction for Canonical Correlation Analysis

Haim Avron HAIMAV@US.IBM.COM
Christos Boutsidis CBOUTSI@US.IBM.COM
IBM T.J. Watson Research Center

Sivan Toledo STOLEDO@TAU.AC.IL
Tel-Aviv University

Anastasios Zouzias ZOUZIAS@CS.TORONTO.EDU
University of Toronto

Abstract

We present a fast algorithm for approximate Canonical Correlation Analysis (CCA). Given a pair of tall-and-thin matrices, the proposed algorithm first employs a randomized dimensionality reduction transform to reduce the size of the input matrices, and then applies any standard CCA algorithm to the new pair of matrices. The algorithm computes an approximate CCA to the original pair of matrices with provable guarantees, while requiring asymptotically less operations than the state-of-the-art exact algorithms.

1. Introduction

Canonical Correlation Analysis (CCA), originally due to Hotelling (1936), is an important technique in statistics, data analysis, and data mining. CCA has been successfully applied in many machine learning applications, e.g. dimensionality reduction (Sun et al., 2010), clustering (Chaudhuri et al., 2009), learning of word embeddings (Dhillon et al., 2011), sentiment classification (Dhillon et al., 2012), discriminant learning (Su et al., 2012), and object recognition (Kim et al., 2007). In many ways CCA is analogous to Principal Component Analysis (PCA), but instead of analyzing a single dataset (in matrix form), the goal of CCA is to analyze the relation between a pair of datasets (each in matrix form). From a statistical point of view, PCA extracts the maximum covariance directions between elements in a single matrix, whereas CCA finds the direction of maximal correlation between a pair of matrices. From a linear algebraic point of view, CCA measures the similarities between two subspaces (those spanned by the columns of each of the two matrices analyzed). From a geometric point of view, CCA computes the cosine of the *principle* angles between the two subspaces.

There are different ways to define the canonical correlations of a pair of matrices, and all these ways are equivalent (Golub & Zha, 1995). The linear algebraic formulation of Golub & Zha (1995), which we present shortly, serves our algorithmic point of view best.

Definition 1. *Let* $\mathbf{A} \in \mathbb{R}^{m\times n}$ *and* $\mathbf{B} \in \mathbb{R}^{m\times \ell}$ *, and assume that* $p = \mathrm{rank}(\mathbf{A}) \geq \mathrm{rank}(\mathbf{B}) = q$*. The* canonical correlations $\sigma_1(\mathbf{A},\mathbf{B}) \geq \sigma_2(\mathbf{A},\mathbf{B}) \geq \cdots \geq \sigma_q(\mathbf{A},\mathbf{B})$ *of the matrix pair* $(\mathbf{A},\mathbf{B})$ *are defined recursively by the following formula* $(i = 1,\ldots,q)$*:*

$$\sigma_i(\mathbf{A},\mathbf{B}) = \max_{\mathbf{x}\in\mathcal{A}_i,\mathbf{y}\in\mathcal{B}_i} \sigma(\mathbf{A}\mathbf{x},\mathbf{B}\mathbf{y}) =: \sigma(\mathbf{A}\mathbf{x}_i,\mathbf{B}\mathbf{y}_i)$$

where

- $\sigma(\mathbf{u},\mathbf{v}) = |\mathbf{u}^{\mathrm{T}}\mathbf{v}|/(\|\mathbf{u}\|_2\|\mathbf{v}\|_2)$,
- $\mathcal{A}_i = \{\mathbf{x} : \mathbf{A}\mathbf{x} \neq \mathbf{0}, \mathbf{A}\mathbf{x} \perp \{\mathbf{A}\mathbf{x_1},\ldots,\mathbf{A}\mathbf{x_{i-1}}\}\}$,
- $\mathcal{B}_i = \{\mathbf{y} : \mathbf{B}\mathbf{y} \neq \mathbf{0}, \mathbf{B}\mathbf{y} \perp \{\mathbf{B}\mathbf{y_1},\ldots,\mathbf{B}\mathbf{y_{i-1}}\}\}$.

The unit vectors $\mathbf{A}\mathbf{x}_1/\|\mathbf{A}\mathbf{x}_1\|_2,\ldots,\mathbf{A}\mathbf{x}_q/\|\mathbf{A}\mathbf{x}_q\|_2$, $\mathbf{B}\mathbf{y}_1/\|\mathbf{B}\mathbf{y}_1\|_2,\ldots,\mathbf{B}\mathbf{y}_q/\|\mathbf{B}\mathbf{y}_q\|_2$ *are called the* canonical *or* principal vectors. *The vectors* $\mathbf{x}_1/\|\mathbf{A}\mathbf{x}_1\|_2,\ldots,\mathbf{x}_q/\|\mathbf{A}\mathbf{x}_q\|_2$, $\mathbf{y}_1/\|\mathbf{B}\mathbf{y}_1\|_2,\ldots,\mathbf{y}_q/\|\mathbf{B}\mathbf{y}_q\|_2$ *are called* canonical weights *(or* projection vectors*). Note that the canonical weights and the canonical vectors are* not *uniquely defined.*

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

1.1. Main Result

The main contribution of this paper (see Theorem 12) is a fast algorithm to compute an approximate CCA. The algorithm computes an additive-error approximation to all the canonical correlations. It also computes a set of approximate canonical weights with provable guarantees. We show that the proposed algorithm is asymptotically faster compared to the standard method of Björck & Golub (1973). To the best of our knowledge, this is the first sub-cubic time algorithm for approximate CCA that has provable guarantees.

The proposed algorithm is based on *dimensionality reduction*: given a pair of matrices $(\mathbf{A}, \mathbf{B})$, we transform the pair to a new pair $(\hat{\mathbf{A}}, \hat{\mathbf{B}})$ that has much fewer rows, and then compute the canonical correlations of the new pair exactly, alongside a set of canonical weights, e.g. using the Björck and Golub algorithm. We prove that with high probability the canonical correlations of $(\hat{\mathbf{A}}, \hat{\mathbf{B}})$ are close to the canonical correlations of $(\mathbf{A}, \mathbf{B})$, and that any set of canonical weights of $(\hat{\mathbf{A}}, \hat{\mathbf{B}})$ can be used to construct a set of approximately orthogonal canonical vectors of $(\hat{\mathbf{A}}, \hat{\mathbf{B}})$. The transformation of $(\mathbf{A}, \mathbf{B})$ into $(\hat{\mathbf{A}}, \hat{\mathbf{B}})$ is done in two steps. First, we apply the *Randomized Walsh-Hadamard Transform (RHT)* to both $\mathbf{A}$ and $\mathbf{B}$. This is a unitary transformation, so the canonical correlations are preserved exactly. On the other hand, we show that with high probability, the transformed matrices have their "information" equally spread among all the input rows, so now the transformed matrices are amenable to uniform sampling. In the second step, we uniformly sample (without replacement) a sufficiently large set of rows and rescale them to form $(\hat{\mathbf{A}}, \hat{\mathbf{B}})$. The combination of RHT and uniform sampling is often called *Subsampled Randomized Walsh-Hadamard Transform (SRHT)* in the literature (Tropp, 2011). Note that other variants of dimensionality reduction (Sarlós, 2006) might be appropriate as well, but for concreteness we focus on the SRHT.

Our dimensionality reduction scheme is particularly effective when the matrices are tall-and-thin, that is they have much more rows than columns. Targeting such matrices is natural: in typical CCA applications, columns typically correspond to features or labels and rows correspond to samples or training data. By computing the CCA on as many instances as possible (as much training data as possible), we get the most reliable estimates of application-relevant quantities. However in current algorithms adding instances (rows) is expensive, e.g. in Björck and Golub algorithm we pay $O(n^2 + \ell^2)$ for each row. Our algorithm allows practitioners to run CCA on huge data sets because we reduce the cost of an extra row, making it not much more expensive than $O(n + \ell)$.

1.2. Related Work

Dimensionality reduction has been the driving force behind many recent algorithms for accelerating key machine learning and linear algebraic tasks. A representative example is linear regression, i.e., solve the least squares problem $\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2$, where $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{b} \in \mathbb{R}^m$. If $m \gg n$, then one can use the SRHT to reduce the dimensions of $\mathbf{A}$ and $\mathbf{b}$, to form $\hat{\mathbf{A}}$ and $\hat{\mathbf{b}}$, and then solve the small problem $\min_{\mathbf{x}} \|\hat{\mathbf{A}}\mathbf{x} - \hat{\mathbf{b}}\|_2$. This process will return an approximate solution to the original problem (Sarlós, 2006; Boutsidis & Drineas, 2009; Drineas et al., 2011). Alternatively, one can observe that $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ and $\hat{\mathbf{A}}^{\mathrm{T}}\hat{\mathbf{A}}$ are spectrally close, so $\hat{\mathbf{A}}$ is an effective preconditioner for $\mathbf{A}$ (Rokhlin & Tygert, 2008; Avron et al., 2010). Other problems that can be accelerated using dimensionality reduction include: (i) approximate PCA (via low-rank matrix approximation) (Halko et al., 2011); (ii) matrix multiplication (Sarlós, 2006); (iii) K-means clustering (Boutsidis et al., 2010); (iv) approximation of matrix coherence and statistical leverage (Drineas et al., 2012); to name only a few.

Our approach uses similar techniques as the algorithms mentioned above. For example, Lemma 4 in our article plays a central role in these algorithms as well. However, our analysis requires the use of advanced ideas from matrix perturbation theory and it leads to two new technical lemmas that might be of independent interest: Lemmas 7 and 8 provide bounds for the singular values of the product of two *different* sampled orthonormal matrices. Previous work only provides bounds for products of the *same* matrix (Lemma 4; see also Sarlós (2006, Corollary 11)).

Dimensionality reduction techniques for accelerating CCA have been suggested or used in the past. One common technique is to simply use less samples by uniformly sampling the rows. While this technique might work reasonably well in many instances, it may fail for others unless all rows are sampled. In fact, Theorem 10 analyzes uniform sampling, and establishes bounds on the required sample size.

Sun et al. (2010) suggest a two-stage approach which involves first solving a least-squares problem, and then using the solution to reduce the problem size. However, their technique involves explicitly factoring one of the two matrices, which takes cubic time. Therefore, their method is especially effective when one of the two

matrices has significantly less columns than the other. When the two matrices have about the same number of columns, there is no asymptotic performance gain. In contrast, our method is sub-cubic in any case.

Finally, it is worth noting that CCA itself has been used for dimensionality reduction (Sun et al., 2008; Chaudhuri et al., 2009; Sun et al., 2010). This is not the focus of this paper; we suggest a dimensionality reduction technique to accelerate the computation of CCA.

2. Preliminaries

We use $i:j$ to denote the set $\{i,\ldots,j\}$, and $[n]=1:n$. We use $\mathbf{A},\mathbf{B},\ldots$ to denote matrices and $\mathbf{a},\mathbf{b},\ldots$ to denote column vectors. $\mathbf{I}_n$ is the $n\times n$ identity matrix; $\mathbf{0}_{m\times n}$ is the $m\times n$ matrix of zeros. We denote by $\mathcal{R}(\cdot)$ the column space of its argument matrix. We denote by $[\mathbf{A};\mathbf{B}]$ the matrix obtained by concatenating the columns of $\mathbf{B}$ next to the columns of $\mathbf{A}$. Given a subset of indices $T\subseteq[m]$, the corresponding sampling matrix $\mathbf{S}$ is the $|T|\times m$ matrix obtained by discarding from $\mathbf{I}_m$ the rows whose index is not in T. Note that $\mathbf{SA}$ is the matrix obtained by keeping only the rows in $\mathbf{A}$ whose index *appears* in T. A symmetric matrix $\mathbf{A}$ is positive semi-definite (PSD), denoted by $0\preceq\mathbf{A}$, if $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}\geq 0$ for every vector $\mathbf{x}$. For any two symmetric matrices $\mathbf{X}$ and $\mathbf{Y}$ of the same size, $\mathbf{X}\preceq\mathbf{Y}$ denotes that $\mathbf{Y}-\mathbf{X}$ is a PSD matrix.

We denote the *compact* (or *thin*) SVD of a matrix $\mathbf{A}\in\mathbb{R}^{m\times n}$ of rank p by $\mathbf{A}=\mathbf{U}_{\mathbf{A}}\mathbf{\Sigma}_{\mathbf{A}}\mathbf{V}_{\mathbf{A}}^{\mathrm{T}}$, with $\mathbf{U}_{\mathbf{A}}\in\mathbb{R}^{m\times p}$, $\mathbf{\Sigma}_{\mathbf{A}}\in\mathbb{R}^{p\times p}$, and $\mathbf{V}_{\mathbf{A}}^{\mathrm{T}}\in\mathbb{R}^{p\times n}$. The Moore-Penrose pseudo-inverse of $\mathbf{A}$ is $\mathbf{A}^{+}=\mathbf{V}_{\mathbf{A}}\mathbf{\Sigma}_{\mathbf{A}}^{-1}\mathbf{U}_{\mathbf{A}}^{\mathrm{T}}\in\mathbb{R}^{n\times m}$. We denote the singular values of $\mathbf{A}$ by $\sigma_1(\mathbf{A})\geq\sigma_2(\mathbf{A})\geq\cdots\geq\sigma_p(\mathbf{A})$.

2.1. The Björck and Golub Algorithm

There are quite a few algorithms to compute the canonical correlations (Golub & Zha, 1995). One of the most popular methods is due to Björck & Golub (1973). It is based on the following observation.

Theorem 2 (Björck & Golub (1973))**.** *Assume that the columns of $\mathbf{Q}\in\mathbb{R}^{m\times p}$ ($m\geq p$) and $\mathbf{W}\in\mathbb{R}^{m\times q}$ ($m\geq q$) form an orthonormal basis for the range of $\mathbf{A}$ and $\mathbf{B}$ (respectively). Let $\mathbf{Q}^{\mathrm{T}}\mathbf{W}=\mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}$ be its compact SVD. The diagonal elements of $\mathbf{\Sigma}$ are the canonical correlations of $(\mathbf{A},\mathbf{B})$. The canonical vectors are given by the first q columns of $\mathbf{QU}$ (for $\mathbf{A}$) and $\mathbf{WV}$ (for $\mathbf{B}$).*

Theorem 2 implies that once we have a pair of matrices $\mathbf{Q}$ and $\mathbf{W}$ with orthonormal columns whose column space spans the same column space of $\mathbf{A}$ and $\mathbf{B}$, respectively, then all we need is to compute the singular value decomposition of $\mathbf{Q}^{\mathrm{T}}\mathbf{W}$. Björck and Golub suggest the use of QR decompositions, but $\mathbf{U}_{\mathbf{A}}$ and $\mathbf{U}_{\mathbf{B}}$ will serve as well. Both options require $O\left(m\left(n^2+\ell^2\right)\right)$ time.

Corollary 3. *Frame Definition 1. Let $\mathbf{U}_{\mathbf{A}}^{\mathrm{T}}\mathbf{U}_{\mathbf{B}}=\mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}$ be its compact SVD. Then, for $i\in[q]$: $\sigma_i(\mathbf{A},\mathbf{B})=\mathbf{\Sigma}_{ii}$. The canonical weights are given by the columns of $\mathbf{V}_{\mathbf{A}}\mathbf{\Sigma}_{\mathbf{A}}^{-1}\mathbf{U}$ (for $\mathbf{A}$) and $\mathbf{V}_{\mathbf{B}}\mathbf{\Sigma}_{\mathbf{B}}^{-1}\mathbf{V}$ (for $\mathbf{B}$).*

2.2. Matrix Coherence and Sampling from an Orthonormal Matrix

Matrix coherence is a fundamental concept in the analysis of matrix sampling algorithms (e.g. Talwalkar & Rostamizadeh (2010)). There a quite a few similar but different ways to define the coherence, however in this paper we use the following definition. Given a matrix $\mathbf{A}$ with m rows, the *coherence* of $\mathbf{A}$ is defined as $\mu(\mathbf{A})=\max_{i\in[m]}\|\mathbf{e}_i^{\top}\mathbf{U}_{\mathbf{A}}\|_2^2$, where $\mathbf{e}_i$ is the i-th standard basis (column) vector of $\mathbb{R}^m$. Note that the coherence of $\mathbf{A}$ is a property of the column space of $\mathbf{A}$, and does not depend on the actual choice of $\mathbf{A}$. Therefore, if $\mathcal{R}(\mathbf{A})=\mathcal{R}(\mathbf{B})$ then $\mu(\mathbf{A})=\mu(\mathbf{B})$. Furthermore, it is easy to verify that if $\mathcal{R}(\mathbf{A})\subseteq\mathcal{R}(\mathbf{B})$ then $\mu(\mathbf{A})\leq\mu(\mathbf{B})$. Finally, we mention that for every matrix $\mathbf{A}$ with m rows:

$$\operatorname{rank}(\mathbf{A})/m\leq\mu(\mathbf{A})\leq 1\,.$$

We focus on tall-and-thin matrices, i.e. matrices with (much) more rows than columns. We are interested in dimensionality reduction techniques that (approximately) preserve the singular values of the original matrix. The simplest idea to do dimensionality reduction in tall-and-thin matrices is uniform sampling of the rows of the matrix. Coherence measures how susceptible the matrix is to uniform sampling; the following lemma shows that not too many samples are required when the coherence is small. The bound is almost tight (Tropp, 2011, Section 3.3).

Lemma 4 (Sampling from Orthonormal Matrix, Tropp (2011) Corollary to Lemma 3.4)**.** *Let $\mathbf{Q}\in\mathbb{R}^{m\times d}$ have orthonormal columns. Let $0<\epsilon<1$ and $0<\delta<1$. Let r be an integer such that*

$$6\epsilon^{-2}m\mu(\mathbf{Q})\log(3d/\delta)\leq r\leq m\,.$$

Let T be a random subset of $[m]$ of cardinality r, drawn from a uniform distribution over such subsets, and let $\mathbf{S}$ be the $|T|\times m$ sampling matrix corresponding to T rescaled by $\sqrt{m/r}$. Then, with probability of at least $1-\delta$, for $i\in[d]$:

$$\sqrt{1-\epsilon}\leq\sigma_i(\mathbf{SQ})\leq\sqrt{1+\epsilon}.$$

Proof. Apply Lemma 3.4 from Tropp (2011) with the following choice of parameters: $\ell = \alpha M \log(k/\delta)$, $\alpha = 6/\epsilon^2$, and $\delta_{tropp} = \eta = \epsilon$. Here, ℓ, α, M, k, η are the parameters of Lemma 3.4 from Tropp (2011); also δ_{tropp} plays the role of δ, an error parameter, of Lemma 3.4 from Tropp (2011). ϵ and δ are from our Lemma. ■

In the above lemma, T is obtained by sampling coordinates from $[m]$ *without* replacement. Similar results can be shown for sampling with replacement, or using Bernoulli variables (Ipsen & Wentworth, 2012).

2.3. Randomized Walsh-Hadamard Transform

Matrices with high coherence pose a problem for algorithms based on uniform row sampling. One way to circumvent this problem is to use a coherence-reducing transformation. One popular coherence-reducing transformation is the Randomized Walsh-Hadamard Transform (RHT) matrix. We start with the definition of the deterministic Walsh-Hadamard Transform matrix.

Fix an integer $m = 2^h$, for $h = 1, 2, 3, \ldots$. The (non-normalized) $m \times m$ matrix of the Walsh-Hadamard Transform (WHT) is defined recursively as,

$$\mathbf{H}_m = \begin{bmatrix} \mathbf{H}_{m/2} & \mathbf{H}_{m/2} \\ \mathbf{H}_{m/2} & -\mathbf{H}_{m/2} \end{bmatrix}, \text{ with } \mathbf{H}_2 = \begin{bmatrix} +1 & +1 \\ +1 & -1 \end{bmatrix}.$$

The $m \times m$ normalized matrix of the Walsh-Hadamard transform is $\mathbf{H} = m^{-\frac{1}{2}} \mathbf{H}_m$.

The recursive nature of the WHT allows us to compute $\mathbf{HX}$ for an $m \times n$ matrix $\mathbf{X}$ in time $O(mn \log(m))$. However, in our case we are interested in $\mathbf{SHX}$ where $\mathbf{S}$ is a r-row sampling matrix. To compute $\mathbf{SHX}$ only $O(mn \log(r))$ operations suffice (Ailon & Liberty, 2008, Theorem 2.1).

Definition 5 (Randomized Walsh-Hadamard Transform (RHT))**.** *Let $m = 2^h$ for some positive integer h. A* Randomized Walsh-Hadamard Transform (RHT) *is an $m \times m$ matrix of the form*

$$\mathbf{\Theta} = \mathbf{HD}$$

where $\mathbf{D}$ is a random diagonal matrix of size m whose entries are independent random signs, and $\mathbf{H}$ is a normalized Walsh-Hadamard matrix of size m.

Lemma 6 (RHT bounds Coherence, Tropp (2011) Lemma 3.3)**.** *Let $\mathbf{A}$ be an $m \times n$ ($m \geq n$, $m = 2^h$ for some positive integer h) matrix, and let $\mathbf{\Theta}$ be an RHT. Then, with probability of at least $1 - \delta$,*

$$\mu(\mathbf{\Theta A}) \leq \frac{1}{m} \left(\sqrt{n} + \sqrt{8 \log(m/\delta)} \right)^2 .$$

3. Perturbation Bounds for Matrix Products

This section states three new technical lemmas which analyze the perturbation of the singular values of the product of a pair of matrices after dimensionality reduction. The proofs appear in the full version of the present article (Avron et al., 2012).

Lemma 7. *Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ ($m \geq n$) and $\mathbf{B} \in \mathbb{R}^{m \times \ell}$ ($m \geq \ell$). Define $\mathbf{C} := [\mathbf{A}; \mathbf{B}] \in \mathbb{R}^{m \times (n+\ell)}$, and suppose $\mathbf{C}$ has rank ω, so $\mathbf{U_C} \in \mathbb{R}^{m \times \omega}$. Let $\mathbf{S} \in \mathbb{R}^{r \times m}$ be any matrix such that $\sqrt{1-\epsilon} \leq \sigma_\omega(\mathbf{SU_C}) \leq \sigma_1(\mathbf{SU_C}) \leq \sqrt{1+\epsilon}$, for some $0 < \epsilon < 1$. Then, for $i = 1, \ldots, \min(n, \ell)$,*

$$|\sigma_i\left(\mathbf{A}^{\mathrm{T}}\mathbf{B}\right) - \sigma_i\left(\mathbf{A}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{SB}\right)| \leq \epsilon \cdot \|\mathbf{A}\|_2 \cdot \|\mathbf{B}\|_2 .$$

Lemma 8. *Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ ($m \geq n$) and $\mathbf{B} \in \mathbb{R}^{m \times \ell}$ ($m \geq \ell$). Let $\mathbf{S} \in \mathbb{R}^{r \times m}$ be any matrix such that $\mathrm{rank}(\mathbf{SA}) = \mathrm{rank}(\mathbf{A})$ and $\mathrm{rank}(\mathbf{SB}) = \mathrm{rank}(\mathbf{B})$, and all singular values of $\mathbf{SU_A}$ and $\mathbf{SU_B}$ are inside $[\sqrt{1-\epsilon}, \sqrt{1+\epsilon}]$ for some $0 < \epsilon < 1/2$. Then, for $i = 1, \ldots, \min(n, \ell)$,*

$$|\sigma_i\left(\mathbf{U_A^{\mathrm{T}}}\mathbf{S}^{\mathrm{T}}\mathbf{SU_B}\right) - \sigma_i\left(\mathbf{U_{SA}^{\mathrm{T}}U_{SB}}\right)| \leq 2\epsilon\,(1+\epsilon) .$$

Lemma 9. *Repeat the conditions of Lemma 7. Then, for all $\mathbf{w} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^\ell$, we have*

$$\left|\mathbf{w}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{By} - \mathbf{w}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{SBy}\right| \leq \epsilon \cdot \|\mathbf{Aw}\|_2 \cdot \|\mathbf{By}\|_2.$$

4. CCA of Row Sampled Pairs

Given $\mathbf{A}$ and $\mathbf{B}$, one straightforward way to accelerate CCA is to sample rows uniformly from both matrices, and to compute the CCA of the smaller matrices. In this section we show that if we sample enough rows, then the canonical correlations of the sampled pair are close to the canonical correlations of the original pair. Furthermore, the canonical weights of the sampled pair can be used to find approximate canonical vectors. Not surprisingly, the sample size depends on the coherence. More specifically, it depends on the coherence of $[\mathbf{A}; \mathbf{B}]$.

Theorem 10. *Suppose $\mathbf{A} \in \mathbb{R}^{m \times n}$ ($m \geq n$) has rank p and $\mathbf{B} \in \mathbb{R}^{m \times \ell}$ ($m \geq \ell$) has rank $q \leq p$. Let $0 < \epsilon < 1/2$ be an accuracy parameter and $0 < \delta < 1$ be a failure probability parameter. Let $\omega = \mathrm{rank}([\mathbf{A}; \mathbf{B}]) \leq p + q$. Let r be an integer such that*

$$54\epsilon^{-2} m \mu([\mathbf{A}; \mathbf{B}]) \log(12\omega/\delta) \leq r \leq m .$$

Let T be a random subset of $[m]$ of cardinality r, drawn from a uniform distribution over such subsets, and let

$\mathbf{S} \in \mathbb{R}^{r \times m}$ *be the sampling matrix corresponding to* T *rescaled by* $\sqrt{m/r}$. *Denote* $\hat{\mathbf{A}} = \mathbf{SA}$ *and* $\hat{\mathbf{B}} = \mathbf{SB}$.

Let $\hat{\sigma}_1, \ldots, \hat{\sigma}_q$ *be the exact canonical correlations of* $(\hat{\mathbf{A}}, \hat{\mathbf{B}})$, *and let*

$$\mathbf{w}_1 = \hat{\mathbf{x}}_1 / \|\hat{\mathbf{A}}\hat{\mathbf{x}}_1\|_2, \ldots, \mathbf{w}_q = \hat{\mathbf{x}}_q / \|\hat{\mathbf{A}}\hat{\mathbf{x}}_q\|_2 \,,$$

and

$$\mathbf{p}_1 = \hat{\mathbf{y}}_1 / \|\hat{\mathbf{B}}\hat{\mathbf{y}}_1\|_2, \ldots, p_q = \hat{\mathbf{y}}_q / \|\hat{\mathbf{B}}\hat{\mathbf{y}}_q\|_2$$

be the exact canonical weights of $(\hat{\mathbf{A}}, \hat{\mathbf{B}})$. *With probability of at least* $1 - \delta$ *all the following hold simultaneously:*

(a) (Approximation of Canonical Correlations) For every $i = 1, 2, \ldots, q$: $|\sigma_i(\mathbf{A}, \mathbf{B}) - \sigma_i(\mathbf{SA}, \mathbf{SB})| \leq \epsilon + 2\epsilon^2/9 = O(\epsilon)$.

(b) (Approximate Orthonormal Bases) The vectors $\{\mathbf{Aw}_i\}_{i \in [q]}$ *form an approximately orthonormal basis. That is, for any* $c \in [q]$,

$$\frac{1}{1+\epsilon/3} \leq \|\mathbf{Aw}_c\|_2^2 \leq \frac{1}{1-\epsilon/3} \,,$$

and for any $i \neq j$,

$$|\langle \mathbf{Aw}_i,\ \mathbf{Aw}_j \rangle| \leq \frac{\epsilon}{3-\epsilon}.$$

Similarly, for the set of $\{\mathbf{Bp}_i\}_{i \in [q]}$.

(c) (Approximate Correlation) For every $i = 1, 2, \ldots, q$:

$$\frac{\sigma_i(\mathbf{A}, \mathbf{B})}{1+\epsilon/3} - \frac{\epsilon/3}{1-\epsilon/9} \leq \sigma(\mathbf{Aw}_i, \mathbf{Bp}_i) \leq \frac{\sigma_i(\mathbf{A}, \mathbf{B})}{1-\epsilon/3} + \frac{\epsilon/3}{(1-\epsilon/3)^2} \,.$$

Proof. Let $\mathbf{C} := [\mathbf{U_A}; \mathbf{U_B}]$. Lemma 4 implies that each of the following three assertions hold with probability of at least $1 - \delta/3$, hence all three hold simultaneously with probability of at least $1 - \delta$:

- For every $r \in [p]$:
$$\sqrt{1-\epsilon/3} \leq \sigma_r(\mathbf{SU_A}) \leq \sqrt{1+\epsilon/3} \,.$$
- For every $k \in [q]$:
$$\sqrt{1-\epsilon/3} \leq \sigma_k(\mathbf{SU_B}) \leq \sqrt{1+\epsilon/3} \,.$$
- For every $h \in [\omega]$:
$$\sqrt{1-\epsilon/3} \leq \sigma_h(\mathbf{SU_C}) \leq \sqrt{1+\epsilon/3} \,.$$

We now show that if indeed all three hold, then (a)-(c) hold as well.

Proof of (a). Corollary 3 implies that $\sigma_i(\mathbf{A}, \mathbf{B}) = \sigma_i(\mathbf{U_A^T U_B})$ and $\sigma_i(\mathbf{SA}, \mathbf{SB}) = \sigma_i(\mathbf{U_{SA}^T U_{SB}})$. We now use the triangle inequality to get,

$$\begin{aligned} |\sigma_i(\mathbf{A}, \mathbf{B}) &- \sigma_i(\mathbf{SA}, \mathbf{SB})| \\ &= |\sigma_i\left(\mathbf{U_A^T U_B}\right) - \sigma_i\left(\mathbf{U_{SA}^T U_{SB}}\right)| \\ &\leq |\sigma_i\left(\mathbf{U_A^T U_B}\right) - \sigma_i\left(\mathbf{U_A^T S^T S U_B}\right)| \\ &+ |\sigma_i\left(\mathbf{U_A^T S^T S U_B}\right) - \sigma_i\left(\mathbf{U_{SA}^T U_{SB}}\right)|. \end{aligned}$$

To conclude the proof, use Lemma 7 and Lemma 8 to bound these two terms, respectively.

Proof of (b). For any $c \in [q]$,

$$\|\mathbf{Aw}_c\|_2 = \|\mathbf{Aw}_c\|_2 / \|\hat{\mathbf{A}}\mathbf{w}_c\|_2$$

since $\|\hat{\mathbf{A}}\mathbf{w}_c\|_2 = 1$. Now Lemma 9 implies the first inequality.

For any $i \neq j$

$$\begin{aligned} |\langle \mathbf{Aw}_i,\ \mathbf{Aw}_j \rangle| &\leq |\mathbf{w}_i^T \hat{\mathbf{A}}^T \hat{\mathbf{A}} \mathbf{w}_j| + |\mathbf{w}_i^T (\hat{\mathbf{A}}^T \hat{\mathbf{A}} - \mathbf{A}^T \mathbf{A}) \mathbf{w}_j| \\ &= |\mathbf{w}_i^T (\hat{\mathbf{A}}^T \hat{\mathbf{A}} - \mathbf{A}^T \mathbf{A}) \mathbf{w}_j| \\ &\leq \frac{\epsilon}{3} \|\mathbf{Aw}_i\|_2 \|\mathbf{Aw}_j\|_2 \\ &\leq \frac{\epsilon/3}{1-\epsilon/3} \|\hat{\mathbf{A}}\mathbf{w}_i\|_2 \|\hat{\mathbf{A}}\mathbf{w}_j\|_2 \\ &= \frac{\epsilon}{3-\epsilon}. \end{aligned}$$

In the above, we used the triangle inequality, the fact that the $\mathbf{w}_i$'s are the canonical weights of $\hat{\mathbf{A}}$, and Lemma 9.

Proof of (c). We only prove the upper bound. The lower bound is similar, and we omit it.

$$\begin{aligned} \sigma(\mathbf{Aw}_i, \mathbf{Bp}_i) &= \\ &= \frac{\langle \mathbf{Aw}_i,\ \mathbf{Bp}_i \rangle}{\|\mathbf{Aw}_i\|_2 \|\mathbf{Bp}_i\|_2} \\ &\leq \frac{1}{1-\epsilon/3} \cdot \langle \mathbf{Aw}_i,\ \mathbf{Bp}_i \rangle \\ &= \frac{1}{1-\epsilon/3} \cdot \left(\left\langle \hat{\mathbf{A}}\mathbf{w}_i,\ \hat{\mathbf{B}}\mathbf{p}_i \right\rangle + \mathbf{w}_i^T \left(\mathbf{A}^T\mathbf{B} - \hat{\mathbf{A}}^T\hat{\mathbf{B}} \right) \mathbf{p}_i \right) \\ &\leq \frac{\sigma\left(\hat{\mathbf{A}}\mathbf{x}_i, \hat{\mathbf{B}}\mathbf{y}_i\right)}{1-\epsilon/3} + \frac{\epsilon/3}{1-\epsilon/3} \cdot \|\mathbf{Aw}_i\|_2 \cdot \|\mathbf{Bp}_i\|_2 \\ &\leq \frac{\sigma\left(\hat{\mathbf{A}}\mathbf{w}_i, \hat{\mathbf{B}}\mathbf{p}_i\right)}{1-\epsilon/3} + \frac{\epsilon/3}{(1-\epsilon/3)^2} \end{aligned}$$

In the above, the first equality follows by the definition of $\sigma(\cdot, \cdot)$, the first inequality by using $1 = \|\hat{\mathbf{A}}\mathbf{w}_i\|_2^2 \leq (1+\epsilon)\|\mathbf{Aw}_i\|_2^2$ (same holds for $\mathbf{Bp}_i$), the second inequality from Lemma 9, the third inequality by using $(1-\epsilon)\|\mathbf{Aw}_i\|_2^2 \leq \|\hat{\mathbf{A}}\mathbf{w}_i\|_2^2 = 1$ (same holds for $\mathbf{Bp}_i$), and the last inequality by (a). ∎

5. Fast Approximate CCA

First, we define what we mean by approximate CCA.

Definition 11 (Approximate CCA)**.** *For $0 \leq \eta \leq 1$, an η-*approximate CCA of $(\mathbf{A}, \mathbf{B})$*, is a set of positive numbers $\hat{\sigma}_1, \ldots, \hat{\sigma}_q$ together with a set of vectors $\mathbf{w}_1, \ldots, \mathbf{w}_q$ for $\mathbf{A}$ and a set of vectors $\mathbf{p}_1, \ldots, \mathbf{p}_q$ for $\mathbf{B}$, such that*

(a) For every $i \in [q]$,

$$|\sigma_i(\mathbf{A}, \mathbf{B}) - \hat{\sigma}_i| \leq \eta \,.$$

(b) For every $i \in [q]$,

$$|\|\mathbf{A}\mathbf{w}_i\|_2^2 - 1| \leq \eta \,,$$

and for $i \neq j$,

$$|\langle \mathbf{A}\mathbf{w}_i,\ \mathbf{A}\mathbf{w}_j \rangle| \leq \eta \,.$$

Similarly, for the set of $\{\mathbf{B}\mathbf{p}_i\}_{i \in [q]}$.

(c) For every $i \in [q]$,

$$|\sigma_i(\mathbf{A}, \mathbf{B}) - \sigma(\mathbf{A}\mathbf{w}_i, \mathbf{B}\mathbf{p}_i)| \leq \eta \,.$$

We are now ready to present our fast algorithm for approximate CCA of a pair of tall-and-thin matrices. Algorithm 1 gives the pseudo-code description of our algorithm.

The analysis in the previous section (Theorem 10) shows that if we sample enough rows, the canonical correlations and weights of the sampled matrices are an $O(\epsilon)$-approximate CCA of $(\mathbf{A}, \mathbf{B})$. However, to turn this observation into a concrete algorithm we need an upper bound on the coherence of $[\mathbf{A}; \mathbf{B}]$. It is conceivable that in certain scenarios such an upper bound might be known in advance, or that it can be computed quickly (Drineas et al., 2012). However, even if we know the coherence, it might be as large as one, which will imply that sampling the entire matrix is needed.

To circumvent this problem, our algorithm uses the RHT to reduce the coherence of the matrix pair before sampling rows from it. That is, instead of sampling rows from $(\mathbf{A}, \mathbf{B})$ we sample rows from $(\mathbf{\Theta A}, \mathbf{\Theta B})$, where $\mathbf{\Theta}$ is a RHT matrix (Definition 5). This unitary transformation bounds the coherence with high probability, so we can use Theorem 10 to compute the number of rows required for an $O(\epsilon)$-approximate CCA. We now sample the transformed pair $(\mathbf{\Theta A}, \mathbf{\Theta B})$ to obtain $(\hat{\mathbf{A}}, \hat{\mathbf{B}})$. Now the canonical correlations and weights of $(\hat{\mathbf{A}}, \hat{\mathbf{B}})$ are computed and returned.

Algorithm 1 Fast Approximate CCA

1: **Input:** $\mathbf{A} \in \mathbb{R}^{m \times n}$ of rank p, $\mathbf{B} \in \mathbb{R}^{m \times \ell}$ of rank q, $0 < \epsilon < 1/2$, and δ $(n \geq l,\ p \geq q)$.
2: $r \longleftarrow \min(54\epsilon^{-2}\left[\sqrt{n+\ell} + \sqrt{8\log(12m/\delta)}\right]^2 \log(3(n+\ell)/\delta), m)$
3: Let $\mathbf{S}$ be the sampling matrix of a random subset of $[m]$ of cardinality r (uniform distribution).
4: Draw a random diagonal matrix $\mathbf{D}$ of size m with ± 1 on its diagonal with equal probability.
5: $\hat{\mathbf{A}} \longleftarrow \mathbf{SH}\cdot(\mathbf{DA})$ using fast subsampled WHT (see Section 2.3).
6: $\hat{\mathbf{B}} \longleftarrow \mathbf{SH}\cdot(\mathbf{DB})$ using fast subsampled WHT (see Section 2.3).
7: Compute and return the canonical correlations and the canonical weights of $(\hat{\mathbf{A}}, \hat{\mathbf{B}})$ (e.g. using Björck and Golub's algorithm).

Theorem 12. *With probability of at least $1 - \delta$, Algorithm 1 returns an $O(\epsilon)$-approximate CCA of $(\mathbf{A}, \mathbf{B})$. Assuming Björck and Golub's algorithm is used in line 7, Algorithm 1 runs in time*

$$O\left(mn\log m + \epsilon^{-2}\left[\sqrt{n} + \sqrt{\log(m/\delta)}\right]^2 \log(n/\delta)n^2\right) .$$

Proof. Lemma 6 ensures that with probability of at least $1 - \delta/2$,

$$\mu([\mathbf{\Theta A}; \mathbf{\Theta B}]) \leq \frac{1}{m}\left(\sqrt{n+\ell} + \sqrt{8\log(3m/\delta)}\right)^2 .$$

Assuming that the last inequality holds, Theorem 10 ensures that with probability of at least $1 - \delta/2$, the canonical correlations and weights of $(\hat{\mathbf{A}}, \hat{\mathbf{B}})$ form an $O(\epsilon)$-approximate CCA of $(\mathbf{\Theta A}, \mathbf{\Theta B})$. By the union bound, both events hold together with probability of at least $1 - \delta$. The RHT transforms applied to $\mathbf{A}$ and $\mathbf{B}$ are unitary, so for every η, an η-approximate CCA of $(\mathbf{\Theta A}, \mathbf{\Theta B})$ is also an η-approximate CCA of $(\mathbf{A}, \mathbf{B})$ (and vice versa).

Running time analysis. Step 2 takes $O(1)$ operations. Step 3 requires $O(r)$ operations. Step 4 requires $O(m)$ operations. Step 5 involves the multiplication of $\mathbf{A}$ with $\mathbf{SHD}$ from the left. Computing $\mathbf{DA}$ requires $O(mn)$ time. Multiplying $\mathbf{SH}$ by $\mathbf{DA}$ using fast subsampled WHT requires $O(mn\log r)$ time, as explained in Section 2.3. Similarly, step 6 requires $O(m\ell\log r)$ operations. Finally, step 7 takes $O(rn\ell + r(n^2 + \ell^2))$ time. Assuming that $n \geq \ell$, the total running time is $O(rn^2 + mn\log(r))$. Plugging the value for r, and using the fact that $r \leq m$, established our running time bound. ∎

From a practical point of view, our algorithm is useful for measuring the size of the correlated subspace, and obtaining the principal vectors of it. A reasonable value for ϵ is 0.1, or perhaps 0.01. So for reasonably high correlations, say above 0.2, we get some useful information. However, for lower correlations we get no information at all. Furthermore, it is too expensive to compute all the principal vectors, but once we know the size of the correlated subspace we can use the approximate weights to compute the vectors for that subspace.

6. Relative vs. Additive Error

Now, we demonstrate that, unless $r \approx m$, it is not possible to replace the additive error guarantees of Theorem 12 with relative error guarantees.

Lemma 13. *Assume that given any matrix pair ($\mathbf{A}$, $\mathbf{B}$) and any constant $0 < \epsilon < 1$, Algorithm 1 computes a pair $(\hat{\mathbf{A}}, \hat{\mathbf{B}})$ by setting a sufficient large value for r in Step 2 so that the canonical correlations are relatively preserved with constant probability, i.e., with constant probability ($i = 1, \ldots, q$):*

$$(1-\epsilon)\sigma_i(\mathbf{A},\mathbf{B}) \leq \sigma_i(\hat{\mathbf{A}},\hat{\mathbf{B}}) \leq (1+\epsilon)\sigma_i(\mathbf{A},\mathbf{B})$$

Then, it follows that $r = \Omega(m/\log(m))$.

The proof of this lemma appears in the full version of the present article (Avron et al., 2012).

7. Experiments

We now report the results of a few small-scale experiments. Our experiments are not meant to be exhaustive; however, they do show that our algorithm can be modified slightly to achieve very good performance in practice while still producing acceptable errors.

Our implementation of Algorithm 1 differs from the pseudo-code description in two ways. First, we use

$$r \longleftarrow \min(\epsilon^{-2}\left[\sqrt{n+\ell}+\sqrt{\log(m/\delta)}\right]^2 \log(n+\ell)/\delta), m)$$

for setting the sample size, i.e. we keep the same asymptotic behavior, but drop the constants. The constants in Algorithm 1 are rather large, so they preclude the possibility of beating Björck and Golub's algorithm for reasonable matrix sizes. Our implementation also differs in the choice of underlying mixing matrix. Algorithm 1, and the analysis, uses the WHT. However, it is possible to show that other Fourier-type transforms will work as well (the bounds remain unchanged), and that some of these alternative transforms have certain advantages that make them better suited for an actual implementation (Avron et al., 2010). Specifically, we use the implementation of randomized Discrete Hartley Transform in the Blendenpik library[1].

We report the results of three experiments. In each experiment we run our code five times on a fixed pair of pair of matrices (datasets) $\mathbf{A}$ and $\mathbf{B}$, and compared the different outputs to the true canonical correlations. The first two experiments involved synthetic data-sets, for which we set $\epsilon = 0.25$ and $\delta = 0.05$. The last experiment was conducted on a real-life dataset, and we used $\epsilon = 0.5$ and $\delta = 0.2$. All experiments were conducted in a 64-bit version of MATLAB 7.8. We used a two quad-core Intel E5410 computer running at 2.33 GHz, with 32GB DDR2 800 MHz RAM, running Linux 2.6, but we use a single core only.

Synthetic Experiment 1. In this experiment we first draw five random matrices: three matrices $\mathbf{G}, \mathbf{W}, \mathbf{Z} \in \mathbb{R}^{m\times n}$ with independent entries from the normal distribution, and two matrices $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{n\times n}$ with independent entries from the uniform distribution on $[0, 1]$. We now set $\mathbf{A} = \mathbf{G}\mathbf{X} + 0.1 \cdot \mathbf{W}$ and $\mathbf{B} = \mathbf{G}\mathbf{Y} + 0.1 \cdot \mathbf{Z}$. We use the sizes $m = 120,000$ and $n = 60$. Conceptually, we first take a random basis (the columns of $\mathbf{G}$), and linearly transform it in two different ways (by multiplying by $\mathbf{X}$ and $\mathbf{Y}$). The transformation does not change the space spanned by the bases. We now add to each base some random noise ($0.1 \cdot \mathbf{W}$ and $0.1 \cdot \mathbf{Z}$). Since both $\mathbf{A}$ and $\mathbf{B}$ essentially span the same column space, only polluted by different noise, we expect $(\mathbf{A}, \mathbf{B})$ to have mostly large canonical correlations (close to 1), but also a few small ones. Indeed, Figure 1(a), which plots the canonical correlations of this pair, shows that this is the case.

Figure 2(a) shows the (signed) error in approximating the correlations, in five different runs. The actual error is always an order of magnitude smaller than the input ϵ; the maximum absolute error is only 0.011. For large canonical correlations the error is much smaller, and the approximated value is very accurate. For smaller correlations, the error starts to get larger, but it is still an order of magnitude smaller than the actual value for the smallest correlation. As for the running time, the proposed algorithm takes about 40% less time than Björck and Golub's algorithm (3 sec vs. 5 sec).

Synthetic Experiment 2. In this experiment we first draw three random matrices. The first matrix, $\mathbf{X} \in \mathbb{R}^{m\times n}$ has independent entries from the normal distribution. The second matrix $\mathbf{Y} \in \mathbb{R}^{m\times k}$ has independent entries which take value ± 1 with equal probability, and the third matrix $\mathbf{Z} \in \mathbb{R}^{k\times n}$ has independent

[1] Available at http://www.mathworks.com/matlabcentral/fileexchange/25241-blendenpik.

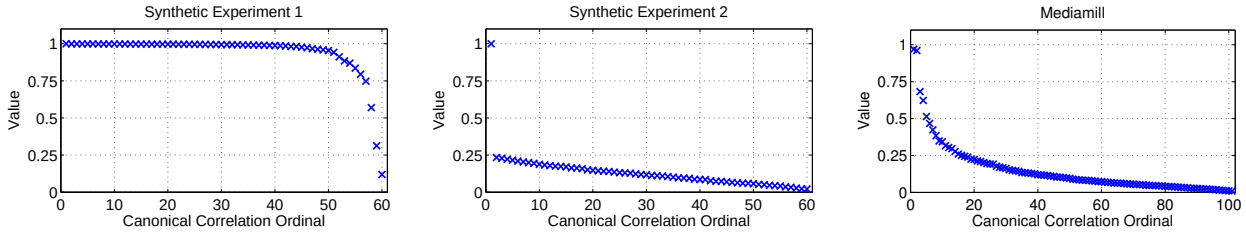

Figure 1. The exact canonical correlations.

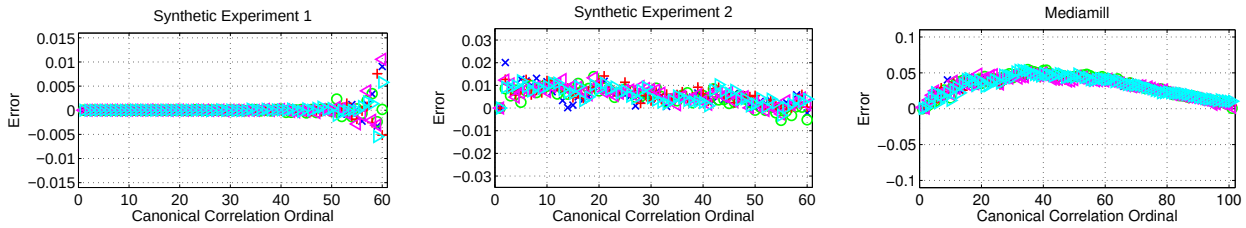

Figure 2. Error in approximation of the canonical correlations.

entries from the uniform distribution on $[0, 1]$. We now set $\mathbf{A} = \mathbf{X} + 0.1 \cdot \mathbf{Y} \cdot (\mathbf{1}_{k\times n} + \mathbf{Z})$ and $\mathbf{B} = \mathbf{Y}$, where $\mathbf{1}_{k\times n}$ is the $k \times n$ all-ones matrix. We use the sizes $m = 80,000$, $n = 80$ and $k = 60$. Here we basically have noise ($\mathbf{B}$) and a matrix polluted with that noise ($\mathbf{A}$). So there is some correlation, but really the two subspaces are different; there is one large correlation (almost 1) and all the rest are small (Figure 1(b)).

Figure 2(b) shows the (signed) error in approximating the correlations, in five different runs. The actual error is an order of magnitude smaller than the target ϵ; the maximum absolute error is only 0.02. Again, for the largest canonical correlation (which is close to 1) the result is very accurate, with tiny errors. For the other correlations it is larger. For tiny correlations the error is about the same magnitude as the actual value. Interestingly, we observe a bias towards over-estimating the correlations. As for the running time, the proposed algorithm takes about 30% less time than Björck and Golub's algorithm (3.1 sec vs. 4.5 sec).

Real-life dataset: Mediamill. We also tested the proposed algorithm on the annotated video dataset from the Mediamill Challenge (Snoek et al., 2006)[2]. Combining the training set and the challenge set, 43907 images are provided, each image is a representative keyframe image of a video shot. The dataset provides 120 features for each image, and the set is annotated with 101 labels. Figure 1(c) shows the exact canonical correlations. We see there is a few high correlations, with very strong decay afterwards.

[2]The dataset is publicly available at http://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets /multilabel.html##mediamill.

Figure 2(c) shows the (signed) error in approximating the correlations, in five different runs. The maximum absolute error is rather small (only 0.055). For the large correlations, which are the more interesting ones in this context, the error is much smaller, so we have a relatively high accuracy approximation. Again, there is an interesting bias towards over-estimating the correlations. As for the running time, the proposed algorithm is considerably faster than Björck and Golub's algorithm (2.05 sec vs. 5.84 sec).

Summary. The experiments are not exhaustive, but they do suggest the following. First, it appears that the sampling size bounds are rather loose. The algorithm achieves much better approximation errors. Second, there seems to be a connection between the canonical correlation value and the error: for larger correlations the error is smaller. Our bounds fail to capture these phenomena. Finally, the experiments show that the proposed is faster than Björck and Golub's algorithm *in practice* on both synthetic and real-life datasets, even if they are fairly small.

Acknoledgements

Haim Avron and Christos Boutsidis acknowledge the support from XDATA program of the Defense Advanced Research Projects Agency (DARPA), administered through Air Force Research Laboratory contract FA8750-12-C-0323. Sivan Toledo was supported by grant 1045/09 from the Israel Science Foundation (founded by the Israel Academy of Sciences and Humanities) and by grant 2010231 from the US-Israel Binational Science Foundation.

References

Ailon, N. and Liberty, E. Fast dimension reduction using Rademacher series on dual BCH codes. In *SODA*, 2008.

Avron, H., Maymounkov, P., and Toledo, S. Blendenpik: Supercharging LAPACK's least-squares solver. *SIAM Journal on Scientific Computing*, 32(3):1217–1236, 2010.

Avron, Haim, Boutsidis, Christos, Toledo, Sivan, and Zouzias, Anastasios. Efficient dimensionality reduction for canonical correlation analysis. *CoRR*, abs/1209.2185, 2012.

Björck, A. and Golub, G.H. Numerical methods for computing angles between linear subspaces. *Mathematics of computation*, 27(123):579–594, 1973.

Boutsidis, C. and Drineas, P. Random projections for the nonnegative least-squares problem. *Linear Algebra and its Applications*, 431(5-7):760–771, 2009.

Boutsidis, C., Zouzias, A., and Drineas, P. Random projections for k-means clustering. In *NIPS*, 2010.

Chaudhuri, K., Kakade, S. M., Livescu, K., and Sridharan, K. Multi-view clustering via canonical correlation analysis. In *ICML*, pp. 129–136, 2009.

Dhillon, P., Rodu, J., Foster, D., and Ungar, L. Using CCA to improve CCA: A new spectral method for estimating vector models of words. In *ICML*, 2012.

Dhillon, P. S., Foster, D., and Ungar, L. Multi-view learning of word embeddings via CCA. In *NIPS*, 2011.

Drineas, P., Mahoney, M.W., Muthukrishnan, S., and Sarlós, T. Faster least squares approximation. *Numerische Mathematik*, 117(2):217–249, 2011.

Drineas, P., Magdon-Ismail, M., Mahoney, M. W., and Woodruff, D. P. Fast approximation of matrix coherence and statistical leverage. In *ICML*, 2012.

Golub, G.H. and Zha, H. The canonical correlations of matrix pairs and their numerical computation. *IMA Volumes in Mathematics and its Applications*, 69:27–27, 1995.

Halko, N., Martinsson, P.G., and Tropp, J.A. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, 2011.

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

Ipsen, I. and Wentworth, T.. The effect of coherence on sampling from matrices with orthonormal columns, and preconditioned least squares problems. *Arxiv preprint arXiv:1203.4809*, 2012.

Kim, T.-K., Kittler, J., and Cipolla, R. Discriminative learning and recognition of image set classes using canonical correlations. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(6):1005–1018, 2007.

Rokhlin, V. and Tygert, M. A fast randomized algorithm for overdetermined linear least-squares regression. *Proceedings of the National Academy of Sciences*, 105(36):13212, 2008.

Sarlós, T. Improved approximation algorithms for large matrices via random projections. In *FOCS*, 2006.

Snoek, C. G. M., Worring, M., van Gemert, J. C., Geusebroek, J. M., and Smeulders, A. W. M. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *Proceedings of the ACM international conference on Multimedia*, pp. 421–430, 2006.

Su, Y., Fu, Y., Gao, X., and Tian, Q. Discriminant learning through multiple principal angles for visual recognition. *Image Processing, IEEE Transactions on*, 21(3):1381 –1390, March 2012.

Sun, L., Ji, S., and Ye, J. A least squares formulation for canonical correlation analysis. In *ICML*, pp. 1024–1031, 2008.

Sun, L., Ceran, B., and Ye, J. A scalable two-stage approach for a class of dimensionality reduction techniques. In *KDD*, pp. 313–322, 2010.

Talwalkar, Ameet and Rostamizadeh, Afshin. Matrix coherence and the nystrom method. In *UAI*, pp. 572–579, 2010.

Tropp, J. Improved analysis of the subsampled randomized Hadamard transform. *Adv. Adapt. Data Anal., special issue, "Sparse Representation of Data and Images"*, 2011.