Sufficient Canonical Correlation Analysis

Yiwen Guo, Xiaoqing Ding, *Fellow, IEEE*, Changsong Liu, *Member, IEEE*, and Jing-Hao Xue

***Abstract*—Canonical correlation analysis (CCA) is an effective way to find two appropriate subspaces in which Pearson's correlation coefficients are maximized between projected random vectors. Due to its well-established theoretical support and relatively efficient computation, CCA is widely used as a joint dimension reduction tool and has been successfully applied to many image processing and computer vision tasks. However, as reported, the traditional CCA suffers from overfitting in many practical cases. In this paper, we propose sufficient CCA (S-CCA) to relieve CCA's overfitting problem, which is inspired by the theory of sufficient dimension reduction. The effectiveness of S-CCA is verified both theoretically and experimentally. Experimental results also demonstrate that our S-CCA outperforms some of CCA's popular extensions during the prediction phase, especially when severe overfitting occurs.**

***Index Terms*—Canonical correlation analysis, multi-class classification, multi-view learning, generalization ability, overfitting, sufficient dimension reduction.**

I. Introduction

In multivariate statistics, canonical correlation analysis (CCA) [1] is a classical and popular technique of analyzing the linear relationship between a pair of multidimensional random vectors. It seeks two orthogonal matrices to project the original pair of random vectors into their subspaces [1] of maximal Pearson correlation coefficients and can be applied to many image processing and computer vision tasks [2]–[5]. Being able to efficiently explore connections between random vectors, CCA has been proven a very useful tool of dimension reduction in various learning tasks, such as multi-label classification [6], in which the two random vectors represent features and class labels respectively, and multi-view learning [7]–[9], in which both random vectors represent features.

Given N pairs of samples $\{(\mathbf{x}_1,\mathbf{y}_1),...,(\mathbf{x}_N,\mathbf{y}_N)\}$ of two different zero-mean random vectors $\mathbf{x}\in\mathbb{R}^p$ and $\mathbf{y}\in\mathbb{R}^q$, we can form two sample matrices $X\in\mathbb{R}^{N\times p}$ and $Y\in\mathbb{R}^{N\times q}$. Aiming at seeking two projection matrices to project X and Y into subspaces with common dimension r, CCA obtains the first pair of projection directions by solving the following optimization problem:

$$\max_{v_1,v_2} \frac{v_1^T \widehat{\mathrm{Cov}}(\mathbf{x},\mathbf{y}) v_2}{\sqrt{v_1^T \widehat{\mathrm{Var}}(\mathbf{x}) v_1}\sqrt{v_2^T \widehat{\mathrm{Var}}(\mathbf{y}) v_2}} , \tag{1}$$

in which $\widehat{\mathrm{Cov}}(\mathbf{x},\mathbf{y}) = X^TY$, $\widehat{\mathrm{Var}}(\mathbf{x}) = X^TX$ and $\widehat{\mathrm{Var}}(\mathbf{y}) = Y^TY$ are sample cross-covariance and covariance matrices, respectively. The second and following pairs of projection vectors are derived from maximizing the same objective function as (1) with additional constraints of orthogonality to the previous directions. To solve these sequential optimization problems, a classical SVD algorithm can be applied [10]. With less than $(\max\{p,q\})^2N$ times of multiplication to calculate projection directions, CCA can be very efficient for many practical tasks. Moreover, it has been proven in [11] that the CCA directions can be interpreted as multipliers of the maximum likelihood estimators under some linear and Gaussian assumptions. The corresponding latent variable model is illustrated in Figure 1a, where $\mathbf{z}$ is the latent vector.

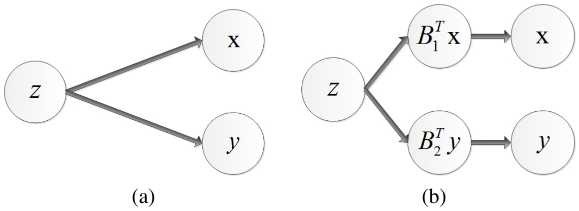


Fig. 1. The graphical models for (a) CCA and (b) sufficient CCA.

Despite the above merits, the original CCA has been reported to suffer from overfitting [12]–[14]. That is, generally, because the number of parameters to be learned, $(p+q)r$, are relatively large when compared with the number of observations, N, the cross-correlation coefficients of projected samples dramatically decrease when a trained CCA model is applied to new test data.

With the development of computer vision, overfitting caused by the curse of dimensionality has become a significant issue in practical learning problems. A number of methods have been developed to cope with the combination of large dimensionality and relatively small data size. Existing methods for CCA are mainly divided into three categories: generative methods based on Bayesian models, ensemble methods and dimension reduction methods. Hence, we shall describe the related methods of different categories separately.

In 2005, Bach and Jordan [11] interpreted CCA from a probabilistic perspective. Following their work up, Klami et al. [15] and Wang [16] proposed Bayesian extensions of CCA. Several other variants of Bayesian CCA [17]–[20] have been successively proposed in the following years. Despite the popularity in the machine learning community, Bayesian methods usually suffer from their model and computational complexity, which makes it not a preferred option in some computer vision applications.

Alternatively, ensemble methods are considered to enhance

This work was partially supported by the National Basic Research Program of China (973 program) under Grant No. 2013CB329403.

Y. Guo, X. Ding and C. Liu are with the State Key Laboratory of Intelligent Technology and Systems, Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing 100084, P. R. China (email: guoyw11@mails.tsinghua.edu.cn; dingxq@tsinghua.edu.cn; lcs@tsinghua.edu.cn).

J.-H. Xue is with the Department of Statistical Science, University College London, London WC1E 6BT, UK (email: jinghao.xue@ucl.ac.uk).

[1] The subspaces may also be referred to as a common subspace in some literature.

the generalization ability of CCA by some researchers [21], [22]. Similarly with Bayesian methods, a major concern is also with its computational complexity. Since quite a few pairs of CCA projection matrices need to be learned, it may not be suitable to the practical applications with high data dimension. Besides, it requires plenty of time to do matrix multiplication even in the test phase.

Another category of approaches to enhancing CCA is through dimension reduction. Due to their effectiveness and expansibility, dimension reduction techniques are popularly used in practical learning cases as a building block. To our knowledge, it is very common to apply principal component analysis (PCA) to each random vector before using CCA, especially when the original data dimension is high. However, few other dimension reduction methods are specifically studied or proposed to tackle the overfitting problem of CCA.

In this paper, we follow the line of dimension reduction and propose a novel enhancement to the original CCA, termed sufficient CCA (S-CCA), in order to mitigate the overfitting problem. Under certain assumptions, we theoretically prove that our method generalizes better on new test data without degrading the optimal value of the canonical correlation. Statistically speaking, we first extract some statistical moments essential to the optimal solutions of CCA and then restrict the hypothesis space further through sufficiently reducing the variable dimensions. The seminal work of using sufficient statistics for dimension reduction was proposed in [23] and called sufficient dimension reduction (SDR) afterwards [24].

The proposed framework is straightforward and effective to enhance the generalization ability of the original CCA. Moreover, because of the recent advancement of SDR, our S-CCA can be efficiently solved in limited time. Hence, by utilizing S-CCA, we can gain even more excellent prediction performance than those of the original CCA and several other extensions of CCA.

The rest of this paper are structured as follows. We present in Section II our S-CCA method and highlight our intuition of S-CCA. In Section III and Section IV, we theoretically and experimentally analyze S-CCA. Section V draws some conclusions.

II. Sufficient canonical correlation analysis

Our ultimate goal is to find explicit linear functions to map the original data into certain spaces with lower dimension and, simultaneously, to maximize Pearson's population correlation coefficients between projected random variables as much as possible. That is to say, we hope to optimize

$$\max_{v_1, v_2} \frac{v_1^T \mathrm{Cov}(\mathbf{x}, \mathbf{y}) v_2}{\sqrt{v_1^T \mathrm{Var}(\mathbf{x}) v_1} \sqrt{v_2^T \mathrm{Var}(\mathbf{y}) v_2}} \tag{2}$$

instead of (1) if possible, in which $\mathrm{Cov}(\cdot, \cdot)$ and $\mathrm{Var}(\cdot)$ are the population moments which are usually unknown in practice. However, the error due to the estimation of the population moments by the sample statistics leads to overfitting. With a fixed amount of data, the variance of each element of estimators $\widehat{\mathrm{Var}}(\mathbf{x})$, $\widehat{\mathrm{Var}}(\mathbf{y})$ and $\widehat{\mathrm{Cov}}(\mathbf{x}, \mathbf{y})$ is only relevant to the probability distribution $p(\mathbf{x}, \mathbf{y})$ [25]. In particular, the total estimation error increases with the number of the elements. Therefore, it would be reasonable to improve the generalization ability of CCA if we can propose a proper protocol to reduce the number of the elements of population moments, i.e., to reduce the original data dimension, beforehand. Following this intuition, we come up with S-CCA, a proper sufficient dimension reduction (SDR) approach for CCA.

S-CCA can be represented as a latent variable model illustrated in Figure 1b. Compared with the original CCA in Figure 1a, S-CCA assumes that there exist rectangular matrices B_1 and B_2 that fulfill

$$E_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x}) = E_{\mathbf{y}|B_1^T\mathbf{x}}(\mathbf{y}|B_1^T\mathbf{x}) \ , \tag{3}$$

$$E_{\mathbf{x}|\mathbf{y}}(B_1^T\mathbf{x}|\mathbf{y}) = E_{\mathbf{x}|B_2^T\mathbf{y}}(B_1^T\mathbf{x}|B_2^T\mathbf{y}) \ , \tag{4}$$

in which the linear projection matrices $B_1 \in \mathbb{R}^{p \times d_1}$, $(r \leq d_1 < p)$, and $B_2 \in \mathbb{R}^{q \times d_2}$, $(r \leq d_2 < q)$, satisfy $B_1^T B_1 = I_{d_1}$ and $B_2^T B_2 = I_{d_2}$. Equations (3) and (4) are chosen because they are considered "sufficient" for the optimization problem (2), which means that the optimal correlation value will not decrease in the new CCA problem for the projected data $B_1^T\mathbf{x}$ and $B_2^T\mathbf{y}$. A theorem with more details about this is provided in section III. See Figure 2 for a schematic illustration.

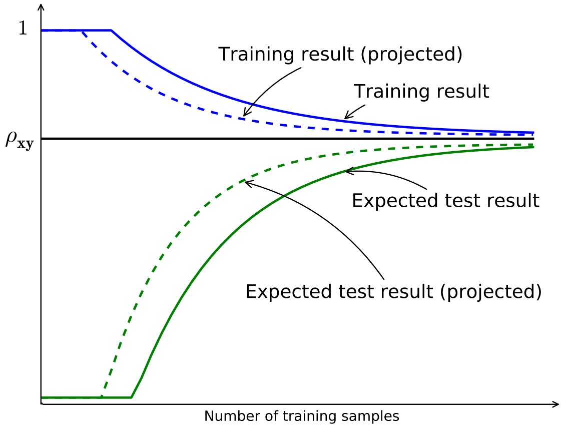


Fig. 2. Let $\rho_{\mathbf{xy}}$ denote the optimal value of the objective function of (2). Suppose training and test canonical correlation results change with the number of training samples in a trend as the solid curves, then the corresponding results for the projected samples pairs $\{(B_1^T\mathbf{x}_i, B_2^T\mathbf{y}_i)\}$ should be like the dashed curves.

The existence of B_1 and B_2 is easy to be verified when we consider the similar linear and Gaussian case as with [11]. That is to say, assume the joint distribution of $\mathbf{x}$ and $\mathbf{y}$ is a zero-mean multivariate Gaussian distribution, $(\mathbf{x}, \mathbf{y}) \sim \mathcal{N}(\mathbf{0}, \Sigma)$, where

$$\Sigma = \begin{bmatrix} W_1 W_1^T + I_p & W_1 W_2^T \\ W_2 W_1^T & W_2 W_2^T + I_q \end{bmatrix} \tag{5}$$

in which W_1 and W_2 are two mapping matrices from latent vectors to observations with unit orthogonal columns, and I_p and I_q are the covariance matrices of the white noises, which are assumed to be identity matrices. As we all know from the joint and conditional Gaussian probabilities,

$$\begin{aligned} E_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x}) &= \mathrm{Cov}(\mathbf{y},\mathbf{x})\mathrm{Var}^{-1}(\mathbf{x})\mathbf{x} \\ &= W_2 W_1^T (W_1 W_1^T + I_p)^{-1}\mathbf{x} \ . \end{aligned} \tag{6}$$

Similarly, we have

$$E_{\mathbf{y}|B_1^T\mathbf{x}}(\mathbf{y}|B_1^T\mathbf{x}) = W_2 W_1^T B_1 (B_1^T (W_1 W_1^T + I_p) B_1)^{-1} B_1^T \mathbf{x}.$$

It is easy to find out that $E_{\mathbf{y}|B_1^T\mathbf{x}}(\mathbf{y}|B_1^T\mathbf{x}) = E_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})$ is satisfied if we define matrix B_1 as W_1 [2]. The existence of B_2 can be similarly verified. For nonlinear cases, we cannot ensure the existence of B_1 and B_2. However, sometimes it is reasonable to assume that a low-dimensional manifold is embedded in the original data space and the mapping function can be approximately represented by a linear projection [27], which also implies the existence of B_1 and B_2.

In our S-CCA, the original optimization problem (2) becomes

$$\begin{aligned} \max_{v_1, v_2} \quad & \frac{v_1^T B_1^T \mathrm{Cov}(\mathbf{x},\mathbf{y}) B_2 v_2}{\sqrt{v_1^T B_1^T \mathrm{Var}(\mathbf{x}) B_1 v_1}\sqrt{v_2^T B_2^T \mathrm{Var}(\mathbf{y}) B_2 v_2}} \\ \text{s.t.} \quad & E_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x}) = E_{\mathbf{y}|B_1^T\mathbf{x}}(\mathbf{y}|B_1^T\mathbf{x}), \\ & E_{\mathbf{x}|\mathbf{y}}(B_1^T\mathbf{x}|\mathbf{y}) = E_{\mathbf{x}|B_2^T\mathbf{y}}(B_1^T\mathbf{x}|B_2^T\mathbf{y}). \end{aligned} \tag{7}$$

To solve the above problem, we shall take advantage of the recent SDR algorithms such as [28] and [29] to get B_1 and B_2 first. Then, problem (7) will degenerate to the original CCA problem with lower data dimension.

III. Theoretical analysis of S-CCA

In this section, we theoretically justify the intuitive use of (3) and (4) for CCA, which is accomplished in two steps. Firstly, we will demonstrate in Theorem 1 that the optimal correlation value learned by CCA remains unchanged with data pre-projection by matrices B_1 and B_2 satisfying (3) and (4). Consequently, Theorem 1 also indicates that S-CCA seeks the same projection directions with the original CCA in the ideal case. Then we try to verify the intuition that applying dimension reduction before performing CCA will help to relieve the overfitting problem of CCA, which corresponds to Theorem 2.

Suppose that the unit optimal solutions to problem (2) are $\tilde{v}_1$ and $\tilde{v}_2$. Without loss of generality, we still assume $E(\mathbf{x}) = 0$ and $E(\mathbf{y}) = 0$. Let us find out the connection between the CCA optimal solutions and the conditional expectations in (3) and (4).

Theorem 1. *S-CCA achieves the same optimal correlation value with CCA and seeks the same projection directions $B_1 v_1$ and $B_2 v_2$ in (7) as v_1 and v_2 in (2), if*

1) there exist matrices B_1 and B_2 fulfilling equations (3) and (4), and

2) $E_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})$ and $E_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})$ are linear functions of $\mathbf{x}$ and $\mathbf{y}$, respectively.

Proof. By applying the Cauchy-Schwarz inequality to (2), we know that the optimal value $\rho_{\mathbf{xy}}$ of its objective function satisfies

$$\rho_{\mathbf{xy}} = \max_{v_2} \frac{\sqrt{v_2^T \mathrm{Cov}(\mathbf{y},\mathbf{x})\mathrm{Var}^{-1}(\mathbf{x})\mathrm{Cov}(\mathbf{x},\mathbf{y}) v_2}}{\sqrt{v_2^T \mathrm{Var}(\mathbf{y}) v_2}} \ , \tag{8}$$

and the relationship between $\tilde{v}_1$ and $\tilde{v}_2$ is

$$\lambda \tilde{v}_1^T \mathbf{x} = \tilde{v}_2^T \mathrm{Cov}(\mathbf{y},\mathbf{x})\mathrm{Var}^{-1}(\mathbf{x})\mathbf{x} \ , \tag{9}$$

in which λ is a scale factor to make sure $\|\tilde{v}_1\| = \|\tilde{v}_2\| = 1$.

For the linear case above, we can easily get

$$\mathrm{Cov}(\mathbf{y},\mathbf{x})\mathrm{Var}^{-1}(\mathbf{x})\mathrm{Cov}(\mathbf{x},\mathbf{y}) = E_{\mathbf{x}}(E_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})E_{\mathbf{y}|\mathbf{x}}(\mathbf{y}^T|\mathbf{x})), \tag{10}$$

by multiplying (6) with $p(\mathbf{x},\mathbf{y})\mathbf{y}^T$ and integrating with respect to $(\mathbf{x},\mathbf{y})$, and

$$\begin{aligned} & E_{\mathbf{x}}(E_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})E_{\mathbf{y}|\mathbf{x}}(\mathbf{y}^T|\mathbf{x})) \\ =& E_{\mathbf{x}}\left(E_{\mathbf{y}|B_1^T\mathbf{x}}(\mathbf{y}|B_1^T\mathbf{x})E_{\mathbf{y}|B_1^T\mathbf{x}}(\mathbf{y}^T|B_1^T\mathbf{x})\right) \end{aligned} \tag{11}$$

by plugging (3) in. Equations (10) and (11) imply

$$\begin{aligned} & \mathrm{Cov}(\mathbf{y},\mathbf{x})\mathrm{Var}^{-1}(\mathbf{x})\mathrm{Cov}(\mathbf{x},\mathbf{y}) \\ =& \mathrm{Cov}(\mathbf{y}, B_1^T\mathbf{x})\mathrm{Var}^{-1}(B_1^T\mathbf{x})\mathrm{Cov}(B_1^T\mathbf{x},\mathbf{y}) \ , \end{aligned}$$

which means that we can still reach the optimal value $\rho_{\mathbf{xy}}$ and the solution of problem (8) holds to be $\tilde{v}_2$ after simply projecting $\mathbf{x}_1$ by B_1.

From (9) we can get $\lambda \tilde{v}_1^T \mathbf{x} = \tilde{v}_2^T E_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})$ and from (3) we can further obtain $\lambda \tilde{v}_1^T \mathbf{x} = \tilde{v}_2^T E_{\mathbf{y}|B_1^T\mathbf{x}}(\mathbf{y}|B_1^T\mathbf{x})$. Therefore, equation (3) implies that the optimal projection $\lambda \tilde{v}_1^T \mathbf{x}$ can also be achieved by using a function of $B_1^T\mathbf{x}$, which in such a linear case can be expressed as

$$\lambda \tilde{v}_1^T \mathbf{x} = \gamma \tilde{\tilde{v}}_1^T B_1^T \mathbf{x} \ , \tag{12}$$

in which $\tilde{\tilde{v}}_1$ represents a unit optimal solution to (7).

Similarly, we can also explain the effectiveness of equation (4) as $\tilde{\tilde{v}}_1^T B_1^T$ holds optimal and

$$\lambda' \tilde{v}_2^T \mathbf{y} = \gamma' \tilde{\tilde{v}}_2^T B_2^T \mathbf{y} \ , \tag{13}$$

in which $\tilde{\tilde{v}}_2$ represents the other unit optimal solution to (7). □

Theorem 1 means that once (3) and (4) are fulfilled, we can turn to an optimization problem equivalent to problem (2) but with fewer parameters (i.e., $(d_1 + d_2)r$ rather than $(p+q)r$ parameters) to estimate.

Since S-CCA is able to reach the optimal correlation coefficients in the end and it does not matter whether we first project $\mathbf{x}$ or $\mathbf{y}$, we can get the following corollary.

Corollary 1. *Reciprocally, the S-CCA method based on equations (3) and (4) is equivalent to the S-CCA method based on*

$$E_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y}) = E_{\mathbf{x}|B_2^T\mathbf{y}}(\mathbf{x}|B_2^T\mathbf{y}) \ , \tag{14}$$

$$E_{\mathbf{y}|\mathbf{x}}(B_2^T\mathbf{y}|\mathbf{x}) = E_{\mathbf{y}|B_1^T\mathbf{x}}(B_2^T\mathbf{y}|B_1^T\mathbf{x}) \ . \tag{15}$$

However, as approximation always occurs during both the sufficient dimension reduction (SDR) and CCA phases, we would usually obtain different results in practical cases when

[2] In the linear and Gaussian case, we have $E_{\mathbf{y}|B_1^T\mathbf{x}}(\mathbf{y}|B_1^T\mathbf{x}) = W_2(I_r + W_1^T W_1)^{-1} W_1^T \mathbf{x}$ with $B_1 = W_1$. According to Lemma 2 in [26], we can further get $(I_r + W_1^T W_1)^{-1} W_1^T = W_1^T (I_p + W_1 W_1^T)^{-1}$. These lead to $E_{\mathbf{y}|B_1^T\mathbf{x}}(\mathbf{y}|B_1^T\mathbf{x}) = E_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})$.

we apply (3)-(4) or (14)-(15). Thus, it might be wise to compare and choose between (3)-(4) and (14)-(15) in practice.

Although it is well known that appropriate dimension reduction can relieve the overfitting problem in regression and classification, we would like to specifically verify it for CCA. Moreover, we are particularly interested in how the data dimension will affect the generalization ability of the original CCA.

Let us define the estimation error of (1) as

$$\Delta = \frac{\left|\widehat{\mathrm{Cov}}\left(\hat{v}_1^T \mathbf{x}, \hat{v}_2^T \mathbf{y}\right) - \mathrm{Cov}\left(\hat{v}_1^T \mathbf{x}, \hat{v}_2^T \mathbf{y}\right)\right|}{\sqrt{\mathrm{Var}\left(\hat{v}_1^T \mathbf{x}\right)}\sqrt{\mathrm{Var}\left(\hat{v}_2^T \mathbf{y}\right)}} \ , \tag{16}$$

in which $\hat{v}_1$ and $\hat{v}_2$ are the solutions to (1) [3] . That is, the error Δ is a function of both the test data $\mathbf{x}$ and $\mathbf{y}$ and the training data X and Y. A decrease in Δ implies the better ability of generalization.

Lemma 1. *Let $\xi_{\mathbf{xy}} = \widehat{Cov}(\mathbf{x}, \mathbf{y}) - Cov(\mathbf{x}, \mathbf{y})$, $Q_1 = Var^{-1}(\mathbf{x})$ and $Q_2 = Var^{-1}(\mathbf{y})$. Then the estimation error Δ can be upper-bounded as*

$$\sup_{\hat{v}_1, \hat{v}_2} \Delta \leq \left[tr\left(Q_2^{1/2} \xi_{\mathbf{xy}}^T Q_1 \xi_{\mathbf{xy}} Q_2^{1/2}\right)\right]^{1/2} . \tag{17}$$

Proof. As we know, given vectors u and v and a positive definite matrix Q, $|u^T v| \leq (u^T Q u)^{1/2} (v^T Q^{-1} v)^{1/2}$. It follows that

$$\begin{aligned}
\Delta &= \frac{\left|tr\left[\xi_{\mathbf{xy}} \hat{v}_2 \hat{v}_1^T\right]\right|}{\left(\hat{v}_1^T Q_1^{-1} \hat{v}_1\right)^{1/2} \left(\hat{v}_2^T Q_2^{-1} \hat{v}_2\right)^{1/2}} \\
&\leq \frac{\left(\hat{v}_2^T Q_2^{-1/2} Q_2^{1/2} \xi_{\mathbf{xy}}^T Q_1 \xi_{\mathbf{xy}} Q_2^{1/2} Q_2^{-1/2} \hat{v}_2\right)^{1/2}}{\left(\hat{v}_2^T Q_2^{-1} \hat{v}_2\right)^{1/2}} \\
&= \frac{\left[tr\left(Q_2^{1/2} \xi_{\mathbf{xy}}^T Q_1 \xi_{\mathbf{xy}} Q_2^{1/2} Q_2^{-1/2} \hat{v}_2 \hat{v}_2^T Q_2^{-1/2}\right)\right]^{1/2}}{\left(\hat{v}_2^T Q_2^{-1} \hat{v}_2\right)^{1/2}} \\
&\leq \left[tr\left(Q_2^{1/2} \xi_{\mathbf{xy}}^T Q_1 \xi_{\mathbf{xy}} Q_2^{1/2}\right)\right]^{1/2} .
\end{aligned}$$

□

Define $\|\xi_{\mathbf{xy}}\|_{Q_1 Q_2} = [tr(Q_2^{1/2} \xi_{\mathbf{xy}}^T Q_1 \xi_{\mathbf{xy}} Q_2^{1/2})]^{1/2}$. Since $\|\xi_{\mathbf{xy}}\|_{Q_1 Q_2} = \|Q_2^{1/2} \xi_{\mathbf{xy}}^T Q_1^{1/2}\|_F$, it is not difficult to verify that $\|\cdot\|_{Q_1 Q_2}$ is a norm operator, which means

$$\sup_{\hat{v}_1, \hat{v}_2} \Delta \leq \left\|\widehat{\mathrm{Cov}}(\mathbf{x}, \mathbf{y}) - \mathrm{Cov}(\mathbf{x}, \mathbf{y})\right\|_{Q_1 Q_2} . \tag{18}$$

Furthermore, we shall follow [30] to use the following Lemma 2 to bound $\left\|\widehat{\mathrm{Cov}}(\mathbf{x}, \mathbf{y}) - \mathrm{Cov}(\mathbf{x}, \mathbf{y})\right\|_{Q_1 Q_2}$ and propose an inequality for the learning bound in Theorem 2.

Lemma 2. *Let ϵ_i be zero-mean independent random vectors in a Hilbert space. If there exist B, $M > 0$ such that for all natural numbers $l \geq 2$ the moment condition $\frac{1}{n}\sum_{i=1}^{n} E\|\epsilon_i\|_H^l \leq \frac{B^2}{2} l! M^{l-2}$ is satisfied, then for all $\delta > 0$: $P(\|\frac{1}{n}\sum_i \epsilon_i\|_H \geq \delta) \leq 2\exp(-\frac{n}{2}\delta^2/(B^2 + \delta M))$ [31] [30].*

Theorem 2. *In order to simplify the formulation, define $\alpha_{\mathbf{xy}} = \|\mathrm{Var}^{-1/2}(\mathbf{y})\mathrm{Cov}(\mathbf{y}, \mathbf{x})\mathrm{Var}^{-1/2}(\mathbf{x})\|_F$ and $\beta_{\mathbf{xy}} = \sup_{\mathbf{x}_0, \mathbf{y}_0} \|\mathrm{Var}^{-1/2}(\mathbf{y})\mathbf{y}_0 \mathbf{x}_0^T \mathrm{Var}^{-1/2}(\mathbf{x})\|_F$. Then the upper bound of the CCA estimation error Δ, the generalization bound, is as follows: with probability of at least $1 - 2\exp(-t)$,*

$$\sup_{\hat{v}_1, \hat{v}_2} \Delta \leq \left(\frac{2t\beta_{\mathbf{xy}}\alpha_{\mathbf{xy}}}{N}\right)^{1/2} + \frac{2t\beta_{\mathbf{xy}}}{N} \ . \tag{19}$$

Proof. In order to make use of Lemma 2, we first construct proper B and M to satisfy the moment condition. Let $\epsilon_i = \mathbf{x}_i \mathbf{y}_i^T - E(\mathbf{x}\mathbf{y}^T)$, then

$$\begin{aligned}
E(\|\epsilon_i\|_{Q_1 Q_2}^2) &= E\left(\left\|\mathbf{x}_i \mathbf{y}_i^T\right\|_{Q_1 Q_2}^2\right) - \left\|E(\mathbf{x}\mathbf{y}^T)\right\|_{Q_1 Q_2} \\
&\leq E\left(\left\|Q_2^{1/2} \mathbf{y}_i \mathbf{x}_i^T Q_1^{1/2}\right\|_F^2\right) \\
&\leq \beta_{\mathbf{xy}} \alpha_{\mathbf{xy}}
\end{aligned}$$

and

$$\|\epsilon_i\|_{Q_1 Q_2} \leq \left\|\mathbf{x}_i \mathbf{y}_i^T\right\|_{Q_1 Q_2} + \left\|E(\mathbf{x}\mathbf{y}^T)\right\|_{Q_1 Q_2} \leq 2\beta_{\mathbf{xy}} \ .$$

In our case, $\|\epsilon_i\|_{Q_1 Q_2}$ can be also treated as the Euclidean norm of $vec(Q_2^{1/2} \epsilon_i Q_1^{1/2})$, in which $vec(\cdot)$ represents the operator to unfold matrices into vectors. Therefore, Lemma 2 can be properly applied.

Let $B = \sqrt{\beta_{\mathbf{xy}} \alpha_{\mathbf{xy}}}$ and $M = \beta_{\mathbf{xy}}$, then the moment condition in Lemma 2 will be fulfilled. That is to say,

$$P\left(\left\|\frac{1}{N}\sum_i \epsilon_i\right\|_{Q_1 Q_2} \geq \delta\right) \leq 2\mathrm{exp}\left(-\frac{N\delta^2}{2B^2 + 2\delta M}\right) .$$

Equivalently,

$$P\left(\left\|\frac{1}{N}\sum_i \epsilon_i\right\|_{Q_1 Q_2} \geq \sqrt{\frac{2t}{N}} B + \frac{2tM}{N}\right) \leq 2\exp\left(-t\right) .$$

In consequence,

$$\left\|\widehat{\mathrm{Cov}}(\mathbf{x}, \mathbf{y}) - \mathrm{Cov}(\mathbf{x}, \mathbf{y})\right\|_{Q_1 Q_2} \leq \sqrt{\frac{2t\beta_{\mathbf{xy}}\alpha_{\mathbf{xy}}}{N}} + \frac{2t\beta_{\mathbf{xy}}}{N}$$

with probability of at least $1 - 2\exp(-t)$. □

Since the elements of $\mathrm{Var}^{-1/2}(\mathbf{y})\mathrm{Cov}(\mathbf{y}, \mathbf{x})\mathrm{Var}^{-1/2}(\mathbf{x})$ are in the range $[-1, +1]$, $\alpha_{\mathbf{xy}} \leq \sqrt{pq}$. Besides, $\alpha_{\mathbf{xy}}$ is a lower bound of $\beta_{\mathbf{xy}}$. Thus, reducing the data dimension can be beneficial to tighten the generalization bound of Δ, which is consistent with our intuition.

IV. Experiments

We apply our S-CCA to both synthetic and real multi-view data to illustrate and verify the effectiveness of S-CCA. First of all, we introduce the methods to be evaluated and compared in the experiments:

P-CCA: As a popular data analysis and processing technique, principal component analysis (PCA) can be used for dimension reduction as well. As we know, PCA has the ability to minimize the squared reconstruction error while reducing the data dimension. Besides, it has a probabilistic

[3] In fact, a better option of estimation error should be the absolute difference between the estimated optimal canonical correlation and $\rho_{\mathbf{xy}}$. Here the definition of Δ is given on account of the deduction simplicity.

interpretation as the maximum likelihood estimator of the Gaussian latent variable model and it is closely connected to CCA. Thus, it is straightforward to combine PCA with CCA and we call it P-CCA in this paper. More specifically, P-CCA applies PCA to X and Y seperately before the CCA calculation. That is to say, it seeks canonical directions from the linear combinations of the first k_{PCA} principal components of X and Y, respectively.

B-CCA: Despite the fact that PCA is popular in practical learning systems, it ignores the relationship between the two random vectors studied in CCA. B-CCA is the combination of the bilinear model (BLM) [32] and CCA. It applies CCA after projecting the original data with the first k_{BLM} eigenvectors of symmetric matrix $[X\ Y]^T[X\ Y]$, which can be treated as the PCA decomposition of the stacked data matrix $[X\ Y]$. Taking advantage of the mutual information of different random vectors, BLM is also a popular method for dimension reduction for multi-view data. Since the calculation of BLM only involves the eigendecomposition of covariance matrices, BLM would inherit the merit of PCA and be considered as an appropriate choice for pre-projecting high-dimensional data before applying CCA.

S-CCA: As described and analyzed in Section II and Section III respectively, our S-CCA tries to maintain the optimal correlation coefficients while reducing the data dimension. In order to efficiently solve equations (3) and (4), we make use of Fukumizu and Leng's gradient-based sufficient dimension reduction method [28]. Since it is a kernel-based method to learn B_1 and B_2, the bandwidth of Gram matrices $G_{\mathbf{x}}$ and $G_{\mathbf{y}}$ should be set in advance. We complete this by using the average of scaled pairwise distances $c_{\mathbf{x}} \sum_{i \neq j} \|\mathbf{x}_i - \mathbf{x}_j\|^2/N/(N-1)$ and $c_{\mathbf{y}} \sum_{i \neq j} \|\mathbf{y}_i - \mathbf{y}_j\|^2/N/(N-1)$, in which $c_{\mathbf{x}} \in [0.01, 100]$ and $c_{\mathbf{y}} \in [0.01, 100]$ are the scale factors for $G_{\mathbf{x}}$ and $G_{\mathbf{x}}$, respectively. Note that in normal cases when the data dimensions $d_{\mathbf{x}}$ and $d_{\mathbf{y}}$ are similar, we can simply set $c_{\mathbf{x}} = c_{\mathbf{y}}$ to simplify the tuning process. The regularization factors of the Gram matrix inversions are chosen from $\{10^{-3}, 10^{-5}, 10^{-7}\}$.

A. Synthetic data

We have theoretically shown in Theorem 2 that a preprocessing of dimension reduction or feature selection can be beneficial to tighten the upper bound of Δ. However, it is still unclear in practice how much P-CCA, B-CCA and S-CCA will prevail over CCA. On the basis of Theorem 1, S-CCA should be superior, but this also need to be confirmed by experiments. In this subsection we compare the performance of these methods by using simulated samples and analyze S-CCA's advantages over CCA, P-CCA and B-CCA.

Here we conduct experiments on two types of synthetic data corresponding to two cases unable to be handled well by analyzing principal data component. First we consider random vectors comprising both multivariate Gaussian and non-Gaussian elements:

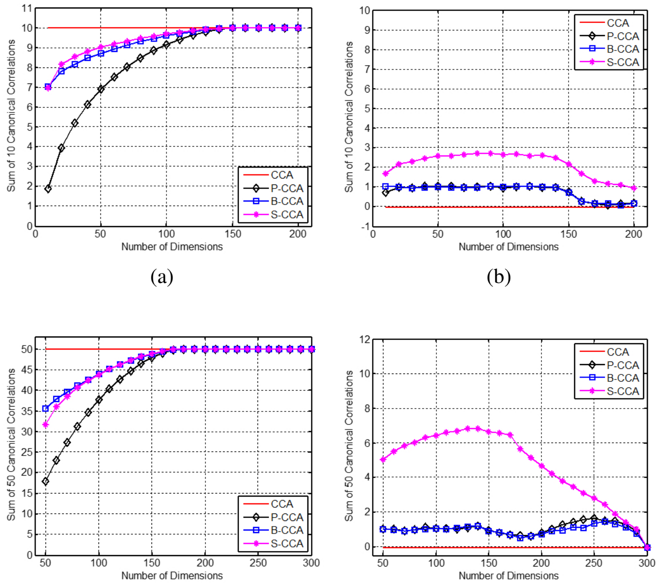
$$\text{Simulation 1)}: \quad \begin{aligned} \mathbf{x} &= \left[\mathbf{z}^T W_1^T \quad \exp(\eta_{\mathbf{x}})^T\right]^T , \\ \mathbf{y} &= \left[\mathbf{z}^T W_2^T \quad (\eta_{\mathbf{y}} \odot \eta_{\mathbf{y}})^T\right]^T , \end{aligned} \tag{20}$$

in which $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, I_{100})$, W_1 and W_2 are two different 200×100 mapping matrices, $\eta_{\mathbf{x}}$, $\eta_{\mathbf{y}} \sim \mathcal{N}(\mathbf{0}, 2 \times I_{200})$ are random noise vectors which are independent of $\mathbf{z}$, and $\odot$ indicates the Hadamard product. That is to say, each of the views consists of 400 features. A hundred repetitions of 500 samples were generated for the experiments, in which 300 of them were used as the training samples and the rest 200 as the test samples. Due to the impact of noise elements, $\mathbf{x}$ and $\mathbf{y}$ are highly uncorrelated. Besides, there are non-Gaussian elements. Hence we anticipate that P-CCA and B-CCA may not perform well in this case.

Then we consider the linear case with additive Gaussian noise:

$$\text{Simulation 2)}: \quad \begin{aligned} \mathbf{x} &= W\mathbf{z} + \eta_{\mathbf{x}} , \\ \mathbf{y} &= V^T W\mathbf{z} + \eta_{\mathbf{y}} , \end{aligned} \tag{21}$$

in which $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, I_{300})$, W is a 300×300 mapping matrix, $\eta_{\mathbf{x}} \sim \mathcal{N}(\mathbf{0}, 10^{-4} \times I_{300})$ and $\eta_{\mathbf{y}} \sim \mathcal{N}(\mathbf{0}, I_{150})$ are two random noise vectors which are independent of $\mathbf{z}$, and V is a 300×150 orthogonal matrix whose rows correspond to the eigenvectors of WW^T with relatively small eigenvalues. That is to say, one view consists of 300 features and the other view consists of 150 features. The same number of samples are generated as in Simulation 1).



(a)

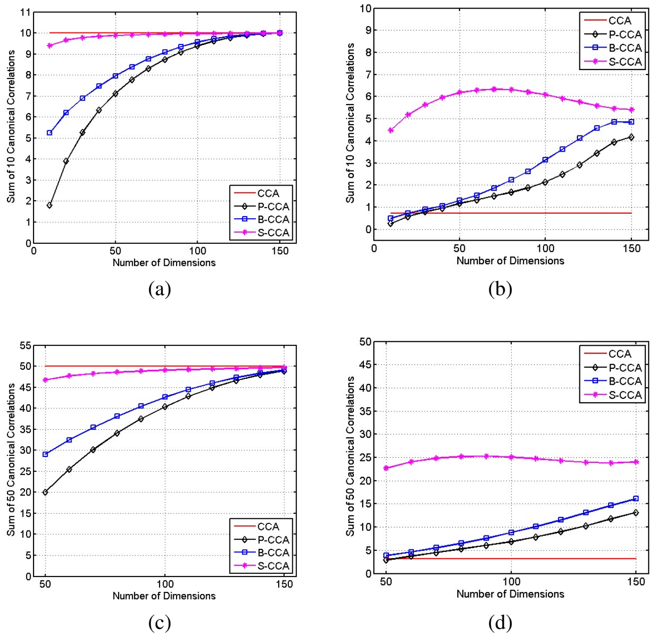
(b)

(c)

(d)

Fig. 3. For Simulation 1), the sum of the top-10 or top-50 canonical correlations in the training or test phase versus the pre-projection dimension: (a) top-10, training; (b) top-10, test; (c) top-50, training; (d) top-50, test. S-CCA performs the best on the test set, indicating its superior ability of generalization.

Following [22], we use the sum of canonical correlations to evaluate the performance of different methods. From Figure 3 and Figure 4, we can observe clearly that P-CCA, B-CCA and S-CCA all prevail over CCA, which is severely overfit to the training data. This confirms our theoretical result in Section III that a proper pre-projection of the original data can improve



(a) (b) (c) (d)

Fig. 4. For Simulation 2), the sum of the top-10 or top-50 canonical correlations in the training or test phase versus the pre-projection dimension: (a) top-10, training; (b) top-10, test; (c) top-50, training; (d) top-50, test. Similarly to Simulation 1), S-CCA performs markedly the best on the test set, which indicates its superior ability of generalization.

the generalization ability of CCA. It is also apparent that S-CCA significantly outperforms P-CCA and B-CCA in the test phase. The inferiority of P-CCA to S-CCA is due to its neglect of cross-view relationship. Although both B-CCA and S-CCA take the cross-view variance into consideration, the intra-view variances may interfere with B-CCA. More specifically, BLM searches for the vectors v_1 and v_2 that maximize $v_1^T\widehat{\mathrm{Var}}(\mathbf{x})v_1+ v_2^T\widehat{\mathrm{Var}}(\mathbf{y})v_2 + 2v_1^T\widehat{\mathrm{Cov}}(\mathbf{x},\mathbf{y})v_2$, which means that v_1 and v_2 are with high possibility to be chosen if their projected intra-view variances are relatively large. These drawbacks degrade P-CCA and B-CCA in many practical cases, as we shall see further in the next three subsections of experiments on real data.

Furthermore, although we have theoretically proven in Theorem 1 that the optimal value $\rho_{\mathbf{xy}}$ in (8) holds after the pre-projection of S-CCA, we are also interested in how much it will degrade in practical cases, because a high $\rho_{\mathbf{xy}}$ implies the potential to get a better result on the test set. In order to study this, we can exploit the advantage of data simulation, by further generating a sufficiently large number of training samples to investigate how well S-CCA, P-CCA and B-CCA can approximate the $\rho_{\mathbf{xy}}$ of CCA. The top-10 canonical correlations are used as the evaluation criterion and summarized in Table I, where we use as many training samples as no further noticeable improvement in the result can be obtained.

As shown in Table I, S-CCA approximates the best to CCA in terms of $\rho_{\mathbf{xy}}$. This demonstrate that the optimal canonical correlation obtained by S-CCA degrades the least when compared with those of P-CCA and B-CCA.

B. The Dexter dataset

'Dexter' is a public dataset for text classification and it is one of the datasets for the NIPS 2003 feature selection challenge [33]. In this dataset, there are 1300 samples that each is represented by 20000 integer features, in which only 9947 features are effective. The remaining 10053 distractor features are manually added to make the dataset more challenging. Similarly to Lu and Foster [34], we split the features into two parts as different views.

Since the features are extremely sparse, many features are always zeros in the training samples. That is to say, the corresponding words do not occur in any of the text and thus have no contribution to the training procedure, so we ignore these features. There are 7751 features left and considered to be adequate, among which the first 500 features are assigned to the first view and the rest are for the second view. According to the competition, the Dexter dataset is divided into a training set of 300 samples, a validation set of 300 samples and a test set of 1000 samples. We follow this division. The sum of the top-10 canonical correlations obtained by CCA, P-CCA, B-CCA and S-CCA are summarized in Table II.

TABLE II
DEXTER: THE SUM OF THE TOP-10 CANONICAL CORRELATIONS ON THE TEST AND TRAINING SETS, AND THE NUMBER OF PRE-PROJECTION BASES OBTAINED BY CROSS-VALIDATION.

Method	Test corr.	Training corr.	Pre-proj dim.
CCA	0.1473	10.0000	-
P-CCA	3.2783	8.4521	80
B-CCA	3.1689	8.6841	60
S-CCA	**3.5575**	9.7865	30

As with the simulated experiments, S-CCA prevails over P-CCA and B-CCA and significantly over CCA on the test set. Note that this time P-CCA shows superior performance when compared with B-CCA, which suggests that a separate PCA projection may mitigate the negative effect of the distractors.

C. PASCAL VOC 2012

The PASCAL Visual Object Classes Challenge 2012 (PASCAL VOC 2012) [35] is an object recognition competition that still draws great attention of the computer vision community. The dataset consists of 17125 images for several different tasks, in which 11540 images are annotated for training and validation of image classification task. Each image can be classified into one or several object classes including 'person', 'bird', 'cat', 'aeroplane', 'bicycle', 'bottle' and so on. Typical samples from the dataset are illustrated in Figure 5. Since multi-class classification is a popular application of CCA, we propose to use CCA, P-CCA, B-CCA and S-CCA for PASCAL VOC 2012 and compare their performance.

In the original configuration of the competition, 5717 images are used for training and 5823 images for validation while the test is done by the evaluation server. We treat the validation sample set as the test set and use two-fold cross-validation within the training set to roughly tune the parameters, which include the bandwidth scale factors of S-CCA and the numbers of pre-projection bases. The 128-dimensional SIFT descriptors

TABLE I
VALUES OF THE TOP-10 CANONICAL CORRELATIONS WHICH CAN BE REACHED AND THEIR SUMMATIONS.

Method	Pre-proj. dim.	corr. 1	corr. 2	corr. 3	corr. 4	corr. 5	corr. 6	corr. 7	corr. 8	corr. 9	corr. 10	sum
P-CCA	10	0.7000	0.0351	0.0298	0.0249	0.0165	0.0132	0.0109	0.0094	0.0065	0.0017	0.8482
	50	0.9834	0.0865	0.0755	0.0740	0.0714	0.0681	0.0672	0.0647	0.0619	0.0579	1.6106
	100	0.9930	0.1130	0.1113	0.1074	0.1051	0.1012	0.1006	0.0998	0.0985	0.0952	1.9251
	200	0.9989	0.5269	0.2434	0.1799	0.1679	0.1589	0.1577	0.1549	0.1547	0.1521	2.8952
B-CCA	10	0.9341	0.0686	0.0510	0.0375	0.0328	0.0291	0.0246	0.0121	0.0049	0.0018	1.1967
	50	0.9854	0.0892	0.0776	0.0735	0.0731	0.0703	0.0700	0.0660	0.0653	0.0621	1.6325
	100	0.9934	0.1128	0.1096	0.1081	0.1054	0.1020	0.1009	0.0982	0.0979	0.0968	1.9250
	200	0.9990	0.5702	0.3903	0.3062	0.2609	0.2366	0.1751	0.1704	0.1610	0.1589	3.4286
S-CCA	10	0.9842	0.1439	0.1340	0.1033	0.0965	0.0837	0.0630	0.0439	0.0306	0.0118	1.6948
	50	0.9979	0.4285	0.3654	0.3440	0.3291	0.2842	0.2805	0.2682	0.2386	0.2266	3.7631
	100	0.9989	0.5954	0.5391	0.5312	0.5266	0.4893	0.4776	0.4749	0.4439	0.4356	5.5124
	200	0.9995	0.9039	0.8803	0.8654	0.8643	0.8422	0.8383	0.8337	0.8178	0.7870	8.6324
CCA	-	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	10.0000

(a) Results on Simulation 1).

Method	Pre-proj. dim.	corr. 1	corr. 2	corr. 3	corr. 4	corr. 5	corr. 6	corr. 7	corr. 8	corr. 9	corr. 10	sum
P-CCA	10	0.0848	0.0560	0.0463	0.0418	0.0385	0.0302	0.0208	0.0123	0.0026	0.0005	0.3337
	50	0.2235	0.2056	0.1762	0.1709	0.1585	0.1567	0.1493	0.1388	0.1372	0.1312	1.6479
	100	0.5051	0.4773	0.4309	0.4083	0.3720	0.3653	0.3491	0.3462	0.3055	0.3010	3.8609
	150	0.9601	0.9566	0.9498	0.9393	0.9283	0.9087	0.8681	0.8530	0.8229	0.7798	8.9665
B-CCA	10	0.1192	0.0918	0.0827	0.0750	0.0683	0.0622	0.0531	0.0500	0.0288	0.0033	0.6343
	50	0.3091	0.2894	0.2343	0.2165	0.2037	0.1912	0.1861	0.1699	0.1672	0.1603	2.1275
	100	0.7476	0.6528	0.6402	0.6002	0.5533	0.5466	0.5159	0.4712	0.4302	0.4120	5.5702
	150	0.9598	0.9595	0.9589	0.9507	0.9449	0.9404	0.9280	0.9070	0.8979	0.8811	9.3281
S-CCA	10	0.6343	0.5607	0.5437	0.5273	0.4916	0.4184	0.3955	0.3839	0.2978	0.2239	4.4770
	50	0.8612	0.8554	0.8425	0.8168	0.8129	0.8069	0.8009	0.7852	0.7796	0.7589	8.1202
	100	0.9346	0.9308	0.9272	0.9214	0.9199	0.9164	0.9121	0.9092	0.9018	0.8992	9.1727
	150	0.9633	0.9606	0.9596	0.9591	0.9588	0.9574	0.9561	0.9548	0.9534	0.9524	9.5754
CCA	-	0.9716	0.9712	0.9710	0.9701	0.9699	0.9694	0.9685	0.9681	0.9678	0.9675	9.6950

(b) Results on Simulation 2).

Fig. 5. Samples from the PASCAL VOC 2012 dataset. The label of these images are ‘aeroplane’, ‘train’, ‘diningtable+chair’, ‘sofa+pottedplant+chair’, ‘person’, ‘person+chair+pottedplant’, ‘person+motorbike+pottedplant’, ‘car+person’, ‘boat’ and ’person’.

densely extracted from the grayscale image are adopted as the original features and encoded by a Fisher coding [36] procedure. The SIFT features are extracted with a step of 8 and the codebook size is chosen to be 20, making the samples represented by 5120-dimensional vectors. There are 20 object classes, so the label view will be represented by 20-dimensional vectors. Similarly, we first evaluate the sum of the top-20 canonical correlations and the results are summarized in Table III. We note that, since the pre-projection dimension is set to be larger than 20, we only pre-project the 5120-dimensional feature vectors.

Furthermore, to get rid of the classifier’s influence, we simply use the Euclidean distance between the projected feature vectors and label vectors as the output score, and then calculate the average precision (AP) [35] and ‘area under

TABLE III
PASCAL VOC 2012: THE SUM OF THE TOP-20 CANONICAL CORRELATIONS ON THE TEST AND TRAINING SETS, AND THE NUMBER OF PRE-PROJECTION BASES.

Method	Test corr.	Training corr.	Pre-proj dim.
CCA	1.9509	18.5535	-
P-CCA	7.0071	9.1307	310
B-CCA	7.0088	9.1618	310
S-CCA	**7.0753**	9.0190	250

TABLE IV
PASCAL VOC 2012: AVERAGE AP ON THE TEST AND TRAINING SETS, WITH THE STANDARD DEVIATION PRESENTED IN PARENTHESES.

Method	Test AP	Training AP
CCA	0.1863 (0.0787)	0.9990 (0.0020)
P-CCA	0.4039 (0.1410)	0.5595 (0.1089)
B-CCA	0.4040 (0.1409)	0.5619 (0.1089)
S-CCA	**0.4105** (0.1397)	0.5551 (0.1085)

TABLE V
PASCAL VOC 2012: AVERAGE AUC ON THE TEST AND TRAINING SETS, WITH THE STANDARD DEVIATION PRESENTED IN PARENTHESES.

Method	Test AUC	Training AUC
CCA	0.1114 (0.0898)	0.9999 (0.0004)
P-CCA	0.3759 (0.1626)	0.5554 (0.1228)
B-CCA	0.3764 (0.1629)	0.5579 (0.1231)
S-CCA	**0.3825** (0.1607)	0.5498 (0.1239)

the curve' (AUC) to measure the classification performance. The comparisons of the results are summarized in Table IV, Table V and Figure 6.

It can be clearly observed from Tables III, IV and V that, as with the previous experiments, P-CCA, B-CCA and S-CCA all perform remarkably better than CCA, and S-CCA keeps performing the best in terms of the sum of canonical correlations, average AUC and average AP. Figure 6 also demonstrates that S-CCA obtains better classification accuracy than P-CCA and B-CCA in most of the object cases.

D. Sketch-to-photo face recognition

As one of the principal problems of heterogeneous face recognition (or multi-view face recognition), sketch-to-photo recognition has attracted great attention among face recognition researchers. In order to evaluate and compare the performance of CCA, P-CCA, B-CCA and S-CCA in tackling this problem, we choose the CUHK Face Sketch database (CUFS) [37] for experiments. The CUFS database consists of 606 corresponding sketch-photo pairs, in which 188 faces come from the Chinese University of Hong Kong (CUHK) student database and the rest are collected from other public databases. The sketch images form the probe set and the photo images form the gallery set.

TABLE VI
CUFS: THE RANK-1 ACCURACY (%) AND THE CORRESPONDING SUM OF CANONICAL CORRELATIONS.

Method	Test set	Accuracy	Training corr.	Test corr.
CCA	100	3.0	30	1.8454
Eigenface [38]	100	31	-	-
Tang [38]	100	71	-	-
Liu [39]	300	87.67	-	-
FaceVACS [40] [41]	300	90.37	-	-
B-CCA	100	92.2	29.2670	22.7274
PLS [42]	100	93.6	-	-
P-CCA	100	94.2	28.1587	23.2108
S-CCA	100	95.6	29.3592	23.9529
Wang [37]	300	96.3	-	-
Klare [41]	300	99.47	-	-

We also compare S-CCA with some popular approaches to sketch-to-photo recognition. As shown in Table VI, all the results, except for CCA, P-CCA, B-CCA and S-CCA, are directly cited from the original papers, in which the test sets contain 100 or 300 faces. We follow the test protocol of [42], which consists of five times of partition of the dataset and utilizes the average rank-1 accuracy as the performance measure. For P-CCA, B-CCA and SCCA, the first 30 canonical correlations are chosen as output and a simple nearest neighbor classifier is applied. Same as the previous experiments, the sum of these canonical correlations are illustrated in Table VI. In fact all the CCA-based methods achieve 100% accuracy on the training set, and thus we only list in the table the accuracy on the test set.

Apparently, CCA suffers from a severe overfitting problem. Although it has the excellent performance on the training set, it performs even worse than the classical Eigenface method [38] on the test set. Keeping the training accuracy unchanged, P-CCA, B-CCA and S-CCA all achieve better test accuracy than CCA (see Table VI). As a specific solution to overcoming overfitting, S-CCA shows significant superiority for this problem when compared with P-CCA and B-CCA. In addition, we can observe that S-CCA, with no need for a complicated feature extraction procedure, can beat many methods specially designed for sketch-to-photo face recognition. We also note that, for this task, Wang [37] and Klare [41] methods performed the best, as they either consist of complicated classifier or utilize score fusion technique. Sophisticated feature extraction like the one applied by [41] also helps to reach a better performance, which indicates the importance of image preprocessing and feature learning.

V. CONCLUSIONS

In this paper, we have presented sufficient CCA (S-CCA) to mitigate to overfitting of CCA, inspired by sufficient dimension reduction. We have theoretically proven that S-CCA is able to reach the optimal correlation coefficient even after projecting the original random vectors into subspaces. Besides, we have also quantitatively investigated the effectiveness of such pre-projection by deriving a generalization bound of CCA. S-CCA can be solved by existing methods and has demonstrated superior empirical performance to CCA or established pre-projection methods based on other mechanisms.

REFERENCES

[1] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3/4, pp. 321–337, 1936.

[2] A. A. Nielsen, "Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data," *IEEE Transactions on Image Processing*, vol. 11, no. 3, pp. 293–305, 2002.

[3] R. Donner, M. Reiter, G. Langs, P. Peloschek, and H. Bischof, "Fast active appearance model search using canonical correlation analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 10, pp. 1690–1694, 2006.

[4] T.-K. Kim and R. Cipolla, "Canonical correlation analysis of video volume tensors for action categorization and detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 8, pp. 1415–1428, 2009.

[5] Y.-R. Yeh, C.-H. Huang, and Y.-C. F. Wang, "Heterogeneous domain adaptation and classification by exploiting the correlation subspace," *IEEE Transactions on Image Processing*, vol. 23, no. 5, pp. 2009–2018, 2014.

[6] L. Sun, S. Ji, and J. Ye, "Canonical correlation analysis for multilabel classification: A least-squares formulation, extensions, and analysis," *IEEE Transaction of Pattern Analysis and Machine Intelligence*, vol. 33, no. 1, pp. 194–200, 2011.

[7] S. M. Kakade and D. P. Foster, "Multi-view regression via canonical correlation analysis," in *Annual Conference on Learning Theory (COLT)*, 2007.

[8] K. Chaudhuri, S. M. Kakade, K. Livescu, and K. Sridharan, "Multi-view clustering via canonical correlation analysis," in *International Conference on Machine Learning (ICML)*, 2009.

[9] A. Li, S. Shan, X. Chen, and W. Gao, "Maximizing intra-individual correlations for face recognition across pose differences," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[10] A. Björck and G. H. Golub, "Numerical methods for computing angles between linear subspaces," *Mathematics of Computation*, vol. 27, no. 123, pp. 579–594, 1973.

[11] F. R. Bach and M. I. Jordan, "A probabilistic interpretation of canonical correlation analysis," Technical Report 688, Department of Statistics, University of California, Berkely, Tech. Rep., April 2005.

[12] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[13] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. Cambridge university press, 2004.

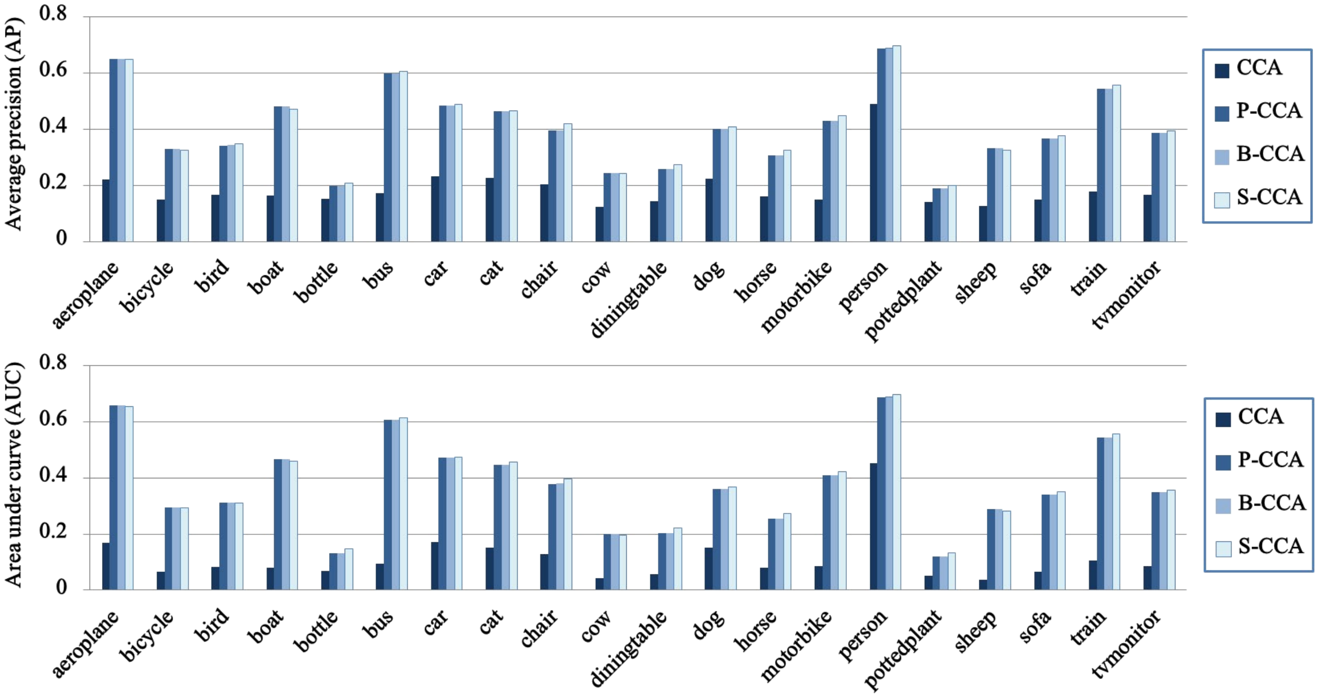


Fig. 6. The results on the PASCAL VOC 2012 dataset, in terms of AP (upper panel) and AUC (lower panel), respectively.

[14] L. Lahti, S. Myllykangas, S. Knuutila, and S. Kaski, "Dependency detection with similarity constraints," in *IEEE International Workshop on Machine Learning for Signal Processing*, 2009, pp. 1–6.

[15] A. Klami and S. Kaski, "Local dependent components," in *International Conference on Machine Learning(ICML)*, 2007.

[16] C. Wang, "Variational Bayesian approach to canonical correlation analysis," *IEEE Transactions on Neural Networks*, vol. 18, no. 3, pp. 905–910, 2007.

[17] C. Archambeau and F. R. Bach, "Sparse probabilistic projections," in *Advances in neural information processing systems(NIPS)*, 2009.

[18] P. Rai and H. D. III, "Multi-label prediction via sparse infinite CCA," in *Advances in Neural Information Processing Systems(NIPS)*, 2009.

[19] S. Virtanen, A. Klami, and S. Kaski, "Bayesian CCA via group sparsity," in *International Conference on Machine Learning (ICML)*, 2011.

[20] A. Klami, S. Virtanen, and S. Kaski, "Bayesian canonical correlation analysis," *Journal of Machine Learning Research*, vol. 14, pp. 965–1003, 2013.

[21] J. Zhang and D. Zhang, "A novel ensemble construction method for multi-view data using random cross-view correlation between within-class examples," *Pattern Recognition*, vol. 44, no. 6, pp. 1162–1171, 2011.

[22] C. O. Sakar, O. Kursun, and F. Gurgen, "Ensemble canonical correlation analysis," *Applied Intelligence*, vol. 40, no. 2, pp. 291–304, 2014.

[23] K.-C. Li, "Sliced inverse regression for dimension reduction," *Journal of the American Statistical Association*, vol. 86, no. 414, pp. 316–327, 1991.

[24] R. D. Cook, *Regression Graphics: Ideas for Studying Regressions through Graphics*. Wiley, 2009.

[25] K. Fukunaga, *Introduction to Statistical Pattern Recognition (Second Edition)*. Academic Press, 1990.

[26] M. Welling, "The Kalman filter," *Lecture Note*, 2010.

[27] X. He and P. Niyogi, "Locality preserving projections," in *Advances in Neural Information Processing Systems(NIPS)*, 2003.

[28] K. Fukumizu and C. Leng, "Gradient-based kernel dimension reduction for regression," *Journal of the American Statistical Association*, vol. 109, no. 505, pp. 359–370, 2014.

[29] K.-Y. Lee, B. Li, and F. Chiaromonte, "A general theory for nonlinear sufficient dimension reduction: Formulation and estimation," *The Annals of Statistics*, vol. 41, no. 1, pp. 221–249, 2013.

[30] Z. Tong, "Learning bounds for kernel regression using effective data dimensionality," *Neural Computation*, vol. 17, no. 9, pp. 2077–2098, 2005.

[31] V. Yurinsky, *Sums and Gaussian vectors*. Springer-Verlag, 1995.

[32] J. B. Tenenbaum and W. T. Freeman, "Separating style and content with bilinear models," *Neural Computation*, vol. 12, no. 6, pp. 1247–1283, 2000.

[33] I. Guyon, S. Gunn, A. B. Hur, and G. Dror, "Result analysis of the NIPS 2003 feature selection challenge," in *Advances in Neural Information Processing Systems (NIPS)*, 2004.

[34] Y. Lu and D. P. Foster, "Large scale canonical correlation analysis with iterative least squares," in *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[35] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes challenge 2012 (VOC2012) results," http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[36] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[37] X. Wang and X. Tang, "Face photo-sketch synthesis and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 11, pp. 1955–1967, 2009.

[38] X. Tang and X. Wang, "Face sketch recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 1, pp. 50–57, 2004.

[39] Q. Liu, X. Tang, H. Jin, H. Lu, and S. Ma, "Nonlinear approach for face sketch synthesis and recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[40] "FaceVACS Software Developer Kit, Cognitec Systems GmbH," http://www.cognitec-systems.de, 2010.

[41] B. F. Klare, Z. Li, and A. K. Jain, "Matching forensic sketches to mug shot photos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 3, pp. 639–646, 2011.

[42] A. Sharma and D. W. Jacobs, "Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

PLACE
PHOTO
HERE

Yiwen Guo received the B.E. degree from Wuhan University, Wuhan, China, in 2011. He is currently working toward the Ph.D. degree with the Department of Electronic Engineering in Tsinghua University, Beijing, China. His current research interests include computer vision, machine learning, pattern recognition and face recognition.

PLACE
PHOTO
HERE

Xiaoqing Ding received the B.E. degree from Tsinghua University, Beijing, China, in 1962. She is currently a Professor and a Ph.D. Supervisor with the Department of Electronic Engineering, Tsinghua University. Her research interests include computer vision, pattern recognition, machine learning, image processing, face recognition and character recognition, etc. Prof. Ding is an IEEE Fellow and IAPR Fellow. She was a recipient of a series of achievements on Chinese/multilanguage character recognition, face recognition, etc. She was a recipient of the most prestigious National Scientific and Technical Progress Awards in China in 1992, 1998, 2003, and 2008.

PLACE
PHOTO
HERE

Changsong Liu is an Associate Professor in the Department of Electronic Engineering at Tsinghua University. He received the B.S. degree in both Mechanics Engineering and Electronic Engineering in 1992, the Master degree in Electronic Engineering in 1995, the Ph.D. degree in 2007 from Tsinghua University, China. His fields of interests include image processing, pattern recognition, and nature language processing. He has published more than 80 papers.

PLACE
PHOTO
HERE

Jing-Hao Xue received the B.Eng. degree in telecommunication and information systems in 1993 and the Dr.Eng. degree in signal and information processing in 1998, both from Tsinghua University, the M.Sc. degree in medical imaging and the M.Sc. degree in statistics, both from Katholieke Universiteit Leuven in 2004, and the degree of Ph.D. in statistics from the University of Glasgow in 2008. Since 2008, he has worked in the Department of Statistical Science at University College London as a Lecturer and Senior Lecturer. His current research interests include statistical classification, high-dimensional data analysis, computer vision and pattern recognition.