Sparse Discrimination based Multiset Canonical Correlation Analysis for Multi-Feature Fusion and Recognition

Hongkun Ji
jihongkunnust@126.com
Xiaobo Shen
njust.shenxiaobo@gmail.com
Quansen Sun
qssun@126.com
Zexuan Ji
jizexuan840820@hotmail.com

School of Computer Science and Engineering
Nanjing University of Science and Technology
Nanjing 210094, China

Abstract

Multiset canonical correlation analysis is a powerful technique for analyzing linear correlations among multiple representation data. However, it usually fails to discover the intrinsic sparse reconstructive relationship and discriminating structure of multiple data spaces in real-world applications. In this paper, by taking discriminative information of within-class and between-class sparse reconstruction into account, we propose a novel algorithm, called sparse discrimination based multiset canonical correlations (SDbMCCs), to explicitly consider both discriminative structure and sparse reconstructive relationship in multiple representation data. In addition to maximizing between-set cumulative correlations, SDbMCC minimizes within-class sparse reconstructive distances and maximizes between-class sparse reconstructive distances, simultaneously. The feasibility and effectiveness of the proposed method is verified on four popular databases (CMU PIE, ETH-80, AR and Extended Yale-B) with promising results.

1 Introduction

Canonical correlation analysis (CCA) [7] is a powerful tool for finding the correlation between two sets of multidimensional variables. It investigates the linear correlations between two sets of random variables and linearly projects two sets of random variables into a lower-dimensional space in which they are maximally correlated. In order to handle supervision information, nonlinear relationships and singularity, several extensions based on CCA have been proposed [1, 4, 5, 6, 13, 16, 17, 18, 19, 20, 21].

However, most of CCA-based methods are not efficient for multiple (more than two) feature representations classification tasks [10]. As a generalized extension of CCA, multiset canonical correlation analysis (MCCA) [11, 14] was proposed to solve this problem. Hou et al. [8] proposed a multiple component analysis (MCA) by utilizing a higher-order covariance tensor for joint feature extraction. MCA can obtain orthogonal subspaces corresponding to

Pages 141.1-141.9
DOI: https://dx.doi.org/10.5244/C.29.141

each feature set by higher-order singular value decomposition (HOSVD) [12]. Subsequently, Yuan et al. [23] imported the generalized correlation coefficient and presented a novel multiset integrated canonical correlation analysis (MICCA) framework. MICCA projects multiple high-dimensional representations in parallel into respective low-dimensional subspaces and then fuses multiple sets of feature vectors by given strategies to form discriminative feature vectors for recognition tasks.

In recent years, with the development of sparse representation [2, 3, 9, 22], sparse representation based classification (SRC) has been proved to be robust for feature selection and efficient in handling occlusion and corruption. Sparsity preserving projections (SPP) [15] aims to preserve the sparse reconstructive relationship of the data. Extensive experiments on some common datasets show that the projections obtained by SPP are invariant to rotations, rescalings and translations of the data. Motivated by recent progress in correlation analysis and sparse representation, in this paper, we propose a novel algorithm, called sparse discrimination based multiset canonical correlations (SDbMCCs). Some aspects of the proposed SDbMCC method are worth being highlighted: 1) SDbMCC explicitly considers both discriminative structure and sparse reconstructive relationship in multiple representation data; 2) By taking the discriminative information into account, SDbMCC can minimize within-class sparse reconstructive distances and maximize between-class sparse reconstructive distances, simultaneously; 3) The extracted features by SDbMCC are proven to be more effective and robust to occlusion and corruption. The proposed algorithm has been compared with state-of-the-art methods on four popular databases (PIE, ETH-80, AR and Extended Yale-B) to demonstrate the superior performances.

2 Sparse discrimination based multiset canonical correlations

2.1 Motivation

This work is motivated by the following three aspects: first of all, as we mentioned before, MCCA is efficient for multiple feature representations classification tasks, but it only considers the cumulative correlation information and ignores the intrinsic sparse reconstructive relationship and discriminating structure in multiple representation data. Secondly, SR-based methods have been proved to be invariant to rotations, rescalings and translations of the data. Therefore, it can significantly improve the model's robustness to various noise. Finally, traditional SR-based methods are essentially unsupervised. In this paper, we take the discriminative information into account which can minimize within-class sparse reconstructive residual and maximize between-class sparse reconstructive residual, simultaneously. As a result, it can significantly improve the discriminative ability of the extracted low dimensional features. A natural question is whether we can introduce this kind of sparse reconstructive relationship and discriminating information into MCCA to improve its performance for joint feature extraction. In this paper, we give a positive answer to this question.

2.2 Construct within-class and between-class scatters

In this paper, we characterize the within-class and between-class scatters, i.e., minimizing the average within-class distance and maximizing the average between-class distance, simultaneously.

Given m feature sets $\{X^{(i)} = [x_1^{(i)}, x_2^{(i)}, \cdots, x_n^{(i)}] \in \Re^{d_i \times n}\}_{i=1}^{m}$ extracted from n samples of c classes. Let $\Phi : \Re^n \to \Re^n$ be the characteristic function that selects the coefficients of $s_j^{(i)}$ associated with the same class of the sample $x_j^{(i)}$, where $s_j^{(i)}$ is the sparse weight vector of $x_j^{(i)}$ as in SPP. For $s_j^{(i)} \in \Re^n$, with the assumption that $x_j^{(i)}$ belongs to the kth class, $\Phi(s_j^{(i)})$ is a vector whose only nonzero entries are the entries in $s_j^{(i)}$ that are associated with class k. For each feature set $X^{(i)}$, after the projections of $\{x_1^{(i)}, x_2^{(i)}, \cdots, x_n^{(i)}\}$ onto the projection axis $\alpha^{(i)}$, we can get the within-class scatter defined by

$$\begin{aligned} J_w^{(i)} &= \sum_{i=1}^{n} \| \alpha^{(i)T} x_j^{(i)} - \alpha^{(i)T} X^{(i)} \Phi(s_j^{(i)}) \|^2 \\ &= \| \alpha^{(i)T} X^{(i)} - \alpha^{(i)T} X^{(i)} \Phi(S^{(i)}) \|^2 \\ &= \alpha^{(i)T} X^{(i)} S_\Phi^{(i)} X^{(i)T} \alpha^{(i)} \end{aligned} \tag{1}$$

where $S^{(i)} = [s_1^{(i)}, s_2^{(i)}, \cdots, s_n^{(i)}]$ and $S_\Phi^{(i)} = [I - \Phi(S^{(i)})] \cdot [I - \Phi(S^{(i)})]^T$. Then we can construct the total within-class scatter for all m feature sets as follows:

$$J_w = \sum_{i=1}^{m} J_w^{(i)} = \sum_{i=1}^{m} \alpha^{(i)T} X^{(i)} S_\Phi^{(i)} X^{(i)T} \alpha^{(i)} \tag{2}$$

Obviously, minimizing the total within-class scatter J_w can lead to a better sparse reconstructive relationship of the samples belonging to the same class.

Similarly to the within-class scatter, let $\Psi : \Re^n \to \Re^n$ be the characteristic function that selects the coefficients of $s_j^{(i)}$ associated with the different classes of the sample $x_j^{(i)}$. We can construct the total between-class scatter for all m feature sets as follows:

$$J_b = \sum_{i=1}^{m} J_b^{(i)} = \sum_{i=1}^{m} \alpha^{(i)T} X^{(i)} S_\Psi^{(i)} X^{(i)T} \alpha^{(i)} \tag{3}$$

where $S_\Psi^{(i)} = [I - \Psi(S^{(i)})] \cdot [I - \Psi(S^{(i)})]^T$. Maximizing the total between-class scatter can significantly enhance the discriminative capability.

2.3 Model of SDbMCC

By considering both discriminative and sparse reconstructive relationship in multiple representation data, we tend to minimize the total within-class sample distance and maximize the total between-class sample distance in the low-dimensional embedding subspace, as well as maximizing the correlations. Combining with the original MCCA objective function, we can construct the model of SDbMCC as follows:

$$\begin{aligned} &max J(\alpha) = \Sigma_{i=1}^{m} \Sigma_{j=1}^{m} \alpha^{(i)T} S_{ij} \alpha^{(j)} - \tau[\eta J_w - (1-\eta) J_b] \\ &s.t.\ \ \alpha^{(i)T} S_{ii} \alpha^{(i)} = 1, \ \ i = 1, 2, \cdots, m. \end{aligned} \tag{4}$$

where $\alpha^T = [\alpha^{(1)T}, \alpha^{(2)T}, \cdots, \alpha^{(m)T}]$, S_{ii} is the within-set covariance matrix of feature set $X^{(i)}$, and S_{ij} is the between-set covariance matrix between feature sets $X^{(i)}$ and $X^{(j)}$. Because the ratio between within-class and between-class sparse representation weights suffers

from a variety of noises, we further hope that the tradeoff between between-class scatters and within-class scatters, as well as the tradeoff between correlations and the sparse reconstructive relationship, can be tuned. Therefore, we construct the corresponding regularizations with two tradeoff parameters τ and η. It is generally difficult to obtain exact solutions by solving above optimization problem. Thus, we couple the constraints to obtain a relaxed version with a single constraint as

$$\begin{aligned} max J(\alpha) &= \Sigma_{i=1}^{m}\Sigma_{j=1}^{m}\alpha^{(i)T}S_{ij}\alpha^{(j)} - \tau[\eta J_w - (1-\eta)J_b] \\ s.t.\ \ &\Sigma_{i=1}^{m}\alpha^{(i)T}S_{ii}\alpha^{(i)} = 1. \end{aligned} \tag{5}$$

By utilizing Lagrange multiplier to solve the optimization problem in Eq. (5), we can obtain the following equation:

$$\begin{aligned} &\begin{pmatrix} \widetilde{S}_{11} & S_{12} & \cdots & S_{1m} \\ S_{21} & \widetilde{S}_{22} & \cdots & S_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ S_{m1} & S_{m2} & \cdots & \widetilde{S}_{mm} \end{pmatrix} \begin{pmatrix} \alpha^{(1)} \\ \alpha^{(2)} \\ \vdots \\ \alpha^{(m)} \end{pmatrix} \\ &= \lambda \begin{pmatrix} S_{11} & & & \\ & S_{22} & & \\ & & \ddots & \\ & & & S_{mm} \end{pmatrix} \begin{pmatrix} \alpha^{(1)} \\ \alpha^{(2)} \\ \vdots \\ \alpha^{(m)} \end{pmatrix} \end{aligned} \tag{6}$$

where λ is the Lagrange multiplier and $\{\widetilde{S}_{ii} = S_{ii} - \tau \cdot X^{(i)}[\eta S_{\Phi}^{(i)} - (1-\eta)S_{\Psi}^{(i)}]X^{(i)T}\}_{i=1}^{m}$. So the solutions of Eq. (5) are the eigenvectors of the generalized eigenvalue problem in Eq. (6) corresponding to the top d largest eigenvalues, where d denotes the number of the projection axes.

3 Experiments

In order to evaluate the proposed method, we compared the performance of the proposed method with several state-of-the-art methods, including MCCA, MCA and MICCA, on four widely used benchmark datasets in both face and object classification tasks, i.e., AR, CMU PIE, Extended Yale-B and ETH-80. The statistics of each data set are briefly listed in Table 1. In this paper, the original images were considered as the first set of features. Since orthonormal wavelet transforms can keep the important information of original images, and the low-frequency sub-images include more shape information in contrast with high-frequency sub-images, we performed Daubechies and Coiflets orthonormal wavelet transforms to obtain the second and the third sets of low-frequency sub-images. To enhance the robustness and avoid the singularity problem in the transformed spaces, we applied the K-L transform to reduce the dimensions of the three feature sets to $d_1 = d_2 = d_3$ corresponding to more than 90% of data energy, respectively. In each experiments, the regularization parameter τ was selected from $\{10^{-4}, 10^{-3}, \cdots, 10^{5}\}$ and η was selected from $\{0.1, 0.2, \cdots, 1\}$. Ten independent tests were performed to get the average recognition rates and the NN classifier with cosine distance measure was employed for recognition.

Dataset	Yale-B	AR	PIE	ETH-80
Number of classes (c)	38	120	68	8
Number per Class	64	26	49	410
Feature Dimensionality	70	150	150	150

Table 1: Brief description of the data sets for classification.

3.1 Face recognition

In order to verify the effectiveness of SDbMCC in face recognition, three experiments on the Extended Yale-B, AR and CMU PIE face datasets were carried out. In the experiment on the Yale-B dataset, $N = \{16, 24, 32\}$ images per individual were randomly chosen for training, while the remaining $\{64 - N\}$ images were used for testing. And the numbers of training images per individual corresponding to the AR and PIE datasets were $N = \{7, 10, 13\}$ and $N = \{5, 10, 15\}$, respectively. Table 2, Table 3 and Table 4 show the maximal average recognition accuracies (%) across ten runs of each method together with their corresponding standard deviations and dimensions on the Yale-B, AR and PIE datasets, respectively. Fig. 1 reveals the recognition rates of each method versus the variation of the dimension on all four datasets.

Method	16 Train	24 Train	32 Train
MCCA	85.0±0.8(70)	88.0±1.0(70)	89.5±0.9(70)
MCA	40.1±4.1(70)	43.3±3.6(70)	49.1±1.7(70)
MICCA	82.3±1.9(70)	86.9±1.1(70)	88.8±1.0(70)
SDbMCC	89.7±0.9(59)	93.1±0.5(70)	94.8±0.6(61)

Table 2: Recognition accuracies (%) on the Yale-B database.

Method	7 Train	10 Train	13 Train
MCCA	78.8±1.0(150)	84.4±0.7(149)	88.5±0.8(148)
MCA	30.5±1.2(148)	36.6±1.5(150)	42.7±1.7(148)
MICCA	62.9±1.0(150)	73.1±0.6(150)	79.0±1.5(150)
SDbMCC	86.4±1.1(82)	92.0±0.6(79)	95.0±0.5(81)

Table 3: Recognition accuracies (%) on the AR database.

Method	5 Train	10 Train	15 Train
MCCA	91.8±0.9(116)	95.5±0.4(129)	96.8±0.4(141)
MCA	50.9±2.2(150)	68.1±1.5(150)	78.7±1.5(149)
MICCA	84.9±1.3(150)	93.6±0.4(149)	95.3±0.5(149)
SDbMCC	94.3±0.4(136)	97.0±0.4(97)	97.8±0.2(64)

Table 4: Recognition accuracies (%) on the PIE database.

From Table 2, Table 3 and Table 4, we can see three main points. Firstly, SDbMCC outperformed MCCA, MCA and MICCA in terms of recognition accuracy, no matter how many training samples per individual were selected. Secondly, when obtaining the best recognition

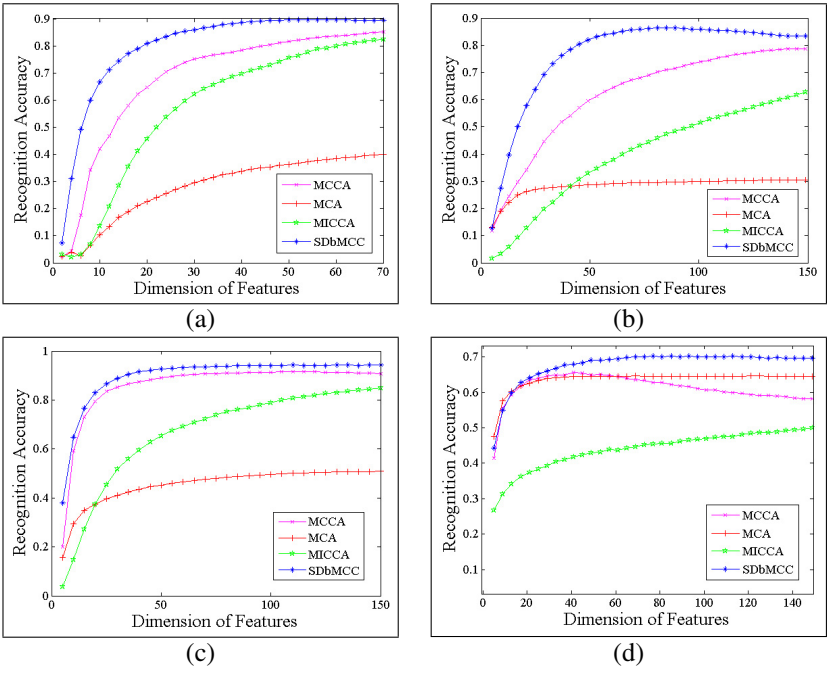


Figure 1: The recognition accuracy of MCCA, MCA, MICCA and SDbMCC with cosine distance metric versus the dimension on (a) the Yale-B database (16 images per individual) (b) the AR database (7 images per individual) (c) the PIE database (5 images per individual) (d) the ETH-80 database.

accuracies, SDbMCC and MCCA always achieved comparable standard deviations which were much smaller than those of MCA and MICCA. This suggests that SDbMCC is relatively stable and robust. Thirdly, the Yale-B and AR datasets contain more occlusion and corruption (glasses and scarves) than PIE whose images are only under various different poses, illumination conditions and facial expressions. By compare the recognition accuracies in Table 2, Table 3 and Table 4, we can discover that SDbMCC reflect more superiority in contrast with the other methods when the datasets contain more occlusion and corruption, which demonstrates that SDbMCC is much more robust to noise.

From Fig. 1(a), Fig. 1(b) and Fig. 1(c), we can see that SDbMCC still outperformed MCCA, MCA and MICCA in terms of recognition accuracy versus the variation of dimension. This indicates that the SDbMCC algorithm is very suitable and robust for joint feature extraction in the face recognition task.

3.2 Object recognition

The ETH-80 object database consists of 8 different categories and each category has 10 objects which consist of 41 images, respectively. In this experiment, three objects per category were randomly chosen for training, while the remaining seven objects were used for testing. Therefore, the total number of training samples was $8 \times 3 \times 41 = 984$, and the total number of testing samples was $8 \times 7 \times 41 = 2296$. Table 5 shows the maximal average recognition

Method	MCCA	MCA	MICCA	SDbMCC
Accuracy	65.6±2.7	64.7±4.1	49.9±2.5	70.2±2.4
Dimension	41	70	149	150

Table 5: Recognition accuracies (%) on the ETH-80 database.

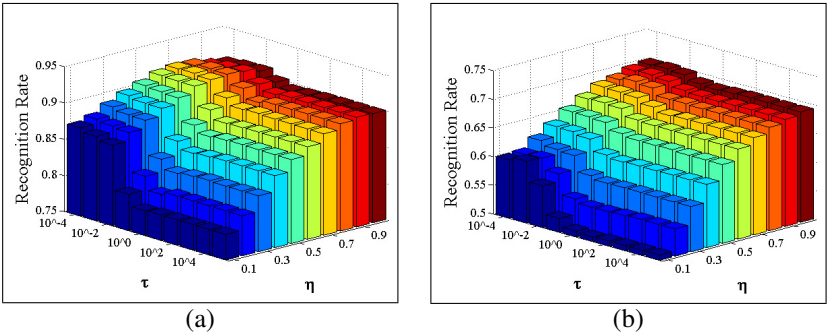
accuracies (%) across ten runs of each method together with their corresponding standard deviations and dimensions on the ETH-80 dataset.

Table 5 and Fig. 1(d) demonstrate that SDbMCC outperforms MCCA, MCA and MICCA in terms of recognition accuracy in object category recognition task. The recognition rate of SDbMCC exceeded those of MCCA 4.6%, MCA 5.5% and MICCA 20.3%, respectively.

3.3 Parameters selection

Our model has two regularization parameters η and τ. Firstly, we can restrict $\eta \in [0,1]$ without strain, and we traversed it from $\{0.1, 0.2, \cdots, 1\}$ in our experiments. Secondly, τ is a tradeoff between correlation and sparse reconstructive relationship, and it was unclear how to determine the optimal parameters at first. Thus, we set a wide range for it as $\{10^{-4}, 10^{-3}, \cdots, 10^{5}\}$ in all experiments.

In this section, we evaluate how SDbMCC performs with different parameter values. For this test, we adopted the AR (10 training samples per individual) and ETH-80 datasets on $10 \times 10 = 100$ pairwise parameter combinations. Fig. 2 shows the maximal recognition accuracy with different η and τ. We can see that the performance of our model has some big fluctuations with respect to η and τ. And setting $\eta \in [0.7, 1]$ and $\tau \in \{10^{-3}, 10^{-2}\}$ for a real world application may be a good choice.



(a) (b)

Figure 2: The recognition accuracy of SDbMCC with cosine distance metric versus the parameters τ and η on the (a) the AR database (10 images per individual) (b) the ETH-80 database.

4 Conclusions

In this paper, we have developed a new technique for joint dimensionality reduction or subspace learning of high dimensional data, called sparse discrimination based multiset canonical correlations (SDbMCCs). As we can see, SDbMCC has more discriminating abilities

than the MCCA based methods in hand. On the other hand, SDbMCC can minimize within-class sparse reconstructive residual and maximize between-class sparse reconstructive residual simultaneously. Therefore, SDbMCCs is more robust to noise, occlusion and corruption. The proposed method has been evaluated on multiple recognition tasks with several popular databases. The experimental results demonstrate that our algorithm facilitates the effective learning of multiset feature fusion and exhibits impressive classification accuracy.

Our model is based on the assumption that the sparse reconstructive relationship of the data in the original high-dimensional space will be preserved in the embedding low-dimensional space. But due to the noise or somehow causes, this assumption may not always set up. Especially in multiple feature representations classification tasks, the characteristics of different feature sets may differ widely with each other. Therefore, it is very worthwhile to build a more robust model to deal with this problem. We are currently exploring these problems both in theory and practice.

References

[1] D.L. Chu, L.Z. Liao, Michael K. Ng, and X.W. Zhang. Sparse canonical correlation analysis: New formulation and algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35:3050–3065, 2013.

[2] M. Davenport, M. Duarte, M. Wakin, D. Takhar, K. Kelly, and R. Baraniuk. The smashed filter for compressive classification and target recognition. In *In Proc. IS&T/SPIE Symposiumon Electronic Imaging: Computational Imaging*, 2006.

[3] M. Davenport, M. Wakin, and R. Baraniuk. Detection and estimation with compressive measurements. In *Technical Report*, 2007.

[4] D.R. Hardoon and J.R. Shawe-Tayler. Sparse canonical correlation analysis. *Machine Learning Journal*, 83:331–353, 2011.

[5] D.R. Hardoon and J.R. Shawe-Taylor. Sparse canonical correlation analysis. In *Tech. Rep.* Department of Computer Science,University College London, 2007.

[6] D.R. Hardoon, S.R. Szedmak, and J.R. Shawe-Taylor. Canonical correlation analysis: an overview with application to learning methods. *Neural Computation*, 16:2639–2664, 2004.

[7] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28:321–377, December 1936.

[8] S.D. Hou, Q.S. Sun, and D.S. Xia. Feature fusion using multiple component analysis. *Neural Process Letters*, 34:259–275, 2011.

[9] K. Huang and S. Aviyente. Sparse representation for signal classification. In *Advances in Neural Information Processing Systems (NIPS)*, 2006.

[10] M. Kan, S.G. Shan, H.H. Zhang, S.H. Lao, and X.L. Chen. Multi-view discriminant analysis. In *European Conference on Computer Vision (ECCV)*, volume 7572, pages 808–821, 2012.

[11] J.R. Kettenring. Canonical analysis of several sets of variables. *Biometrika*, 58:433–451, 1971.

[12] L.D. Lathauwer, B.D. Moor, and J. Vandewalle. A multilinear singular value decomposition. *SIAM J. Matrix Anal*, 21:1253–1278, 2000.

[13] T. Melzer, M. Reiter, and H. Bischof. Appearance models based on kernel canonical correlation analysis. *Pattern Recognition*, 36:1961–1971, 2003.

[14] A.A. Nielsen. Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data. *IEEE Transactions on image processing*, 11:293–305, 2002.

[15] L.S. Qiao, S.C. Chen, and X.Y. Tan. Sparsity preserving projections with applications to face recognition. *Pattern Recognition*, 43:331–341, 2010.

[16] Q.S. Sun, Z.D. Liu, P.A. Heng, and D.S. Xia. A theorem on the generalized canonical projective vectors. *Pattern Recognition*, 38:449–452, 2005.

[17] T.K. Sun and S.C. Chen. Locality preserving cca with applications to data visualization and pose estimation. *Image and Vision Computing*, 25:531–543, 2007.

[18] T.K. Sun, S.C. Chen, J.Y. Yang, and P.F. Shi. A supervised combined feature extraction method for recognition. In *IEEE International Conference on Data Mining*, pages 1043–1048, 2008.

[19] S. Waaijenborg, P.C.V. de Witt Hamer, and A.H. Zwinderman. Quantifying the association between gene expressions and dna-markers by penalized canonical correlation analysis. *Statistical Applications in Genetics and Molecular Biology*, 7:Article 3, 2008.

[20] D.M. Witten and R. Tibshirani. Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical Applications in Genetics and Molecular Biology*, 8:Article 28, 2009.

[21] D.M. Witten, R. Tibshirani, and T. Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10:515–534, 2009.

[22] J. Wright, A. Yang, S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31:210–227, 2009.

[23] Y.H. Yuan, Q.S. Sun, Q. Zhou, and D.S. Xia. A novel multiset integrated canonical correlation analysis framework and its application in feature fusion. *Pattern Recognition*, 44:1031–1040, 2011.