Hypergraph Spectra for Unsupervised Feature Selection

Zhihong Zhang and Edwin R. Hancock*

Department of Computer Science,
University of York, UK

Abstract. Most existing feature selection methods focus on ranking individual features based on a utility criterion, and select the optimal feature set in a greedy manner. However, the feature combinations found in this way do not give optimal classification performance, since they neglect the correlations among features. In an attempt to overcome this problem, we develop a novel unsupervised feature selection technique by using hypergraph spectral embedding, where the projection matrix is constrained to be a selection matrix designed to select the optimal feature subset. Specifically, by incorporating multidimensional interaction information (MII) for higher order similarities measure, we establish a novel hypergraph framework which is used for characterizing the multiple relationships within a set of samples. Thus, the structural information latent in the data can be more effectively modeled. Secondly, we derive a hypergraph embedding view of feature selection which casting the feature discriminant analysis into a regression framework that considers the correlations among features. As a result, we can evaluate joint feature combinations, rather than being confined to consider them individually, and are thus able to handle feature redundancy. Experimental results demonstrate the effectiveness of our feature selection method on a number of standard datasets.

Keywords: Hypergraph representation, Hypergraph subspace learning.

1 Introduction

In order to render the analysis of high-dimensional data tractable, it is crucial to identify a smaller subset of features that are informative for classification and clustering. Dimensionality reduction aims to reduce the number of variables under consideration, and the process can be divided into feature extraction and feature selection. Feature extraction usually projects the features onto a low-dimensional and distinct feature space, e.g., kernel PCA [1], Locality preserving Projection (LPP) [2] and Laplacian eigenmap [3]. Unlike feature extraction, feature selection identifies the optimal feature subset in the original feature space. By maintaining the original features, feature selection improves the interpretability of the data, which is preferred in many real world applications, such as face recognition and text mining. Feature selection algorithms can be roughly classified into two groups, namely a) supervised feature selection and b) unsupervised feature selection.

* Edwin Hancock is supported by a Royal Society Wolfson Research Merit Award.

G.L. Gimel' farb et al. (Eds.): SSPR & SPR 2012, LNCS 7626, pp. 620–628, 2012.

While the labeled data required by supervised feature selection can be scarce, there is usually no shortage of unlabeled data. Hence, there are obvious attractions in developing unsupervised feature selection algorithms which can utilize this data. The typical examples in unsupervised learning are graph-based spectral learning algorithms. Examples include the Laplacian score [6], SPEC [5], Multi-Cluster Feature Selection (MCFS) [8] and Unsupervised Discriminative Feature Selection (UDFS) [10]. Given d features, and a similarity matrix S for the samples, the idea of spectral feature selection algorithms is to identify features that align well with the leading eigenvectors of S. The leading eigenvectors of S contain information of concerning the structure of the sample distribution and group similar samples into compact clusters. Consequently, features that align closely to them will better preserve sample similarity [5]. For example, the Laplacian score [6] uses a nearest neighbor graph to model the local geometric structure of the data, using the pairwise similarities between features are calculated using the heat kernel. In this framework, the features are evaluated individually and are selected one by one. The SPEC [5] algorithm is an extension of the Laplacian score that render it more robust to noise. The method selects the features most consistent with the graph structure. Note that SPEC also evaluates features independently.

However, there are two limitations to the above graph-based spectral feature selection methods. Firstly, they evaluate features individually, and hence cannot handle redundant features. Redundant features increase the dimensionality unnecessarily, and worsen learning performance when faced with a shortage of data. It is also shown empirically that removing redundant features can result in significant performance improvement. The second weakness is that in many situations the graph representation for relational patterns can lead to substantial loss of information. This is because in real-world problems objects and their features tend to exhibit multiple relationships rather than simple pairwise ones. For example, consider the problem of classifying faces which are under different lighting conditions [7]. Therefore, the higher order relations cannot be suitably characterized by pairwise similarity measures.

A natural way for remedying the misleading representation described above is to represent the dataset as a hypergraph instead of a graph. Hypergraph representations allow vertices to be multiply connected by hyperedges and can hence capture multiple or higher order relationships between features. Due to their effectiveness in representing multiple relationships, for the task of feature selection addressed in this paper, we introduce a hypergraph embedding view of feature selection by subspace learning. The method jointly evaluates the utility sets of features rather than individual feature. There are three novel ingredients. The first is that by incorporating hypergraph representation into feature selection, we can be more effective capture the higher order relations among samples. Secondly, inspired from the recent works on mutual information [16], we determine the weight of a hyperedge using an information measure referred to as multidimensional interaction information (MII) which precisely preserves the higher order relations captured by the hypergraph. The advantage of MII is that it is sensitive to the relations between sample combinations, and as a result can be used to seek third or even higher order dependencies among the relevant samples. Thus, the structural information latent in the data can be more effectively modeled. Finally, we describe a new feature selection strategy through hypergraph embedding, which casts the feature

discriminant analysis into a regression framework that considers the correlations among features. As a result, we can evaluate joint feature combinations, rather than being confined to consider them individually, thus it is able to handle feature redundancy.

2 Hypergraph Construction

In this section, we establish a novel hypergraph framework which is used for characterizing the multiple relationships within a set of samples. To this end, we commence by introducing a new method for measuring higher order similarities among samples based on information theory. According to Shannon's study, the uncertainty of a random variable X can be measured by the entropy $H(X)$. For two random variables X and Y, the conditional entropy $H(Y|X)$ measures the remaining uncertainty about Y when X is known. The mutual information $I(X;Y)$ of X and Y quantifies the information gain about Y provided by X. The relationship between $H(Y)$, $H(Y|X)$ and $I(X;Y)$ is $I(X;Y) = H(Y) - H(Y|X)$. As defined by Shannon, the initial uncertainty for X is $H(X) = -\sum_{x\in Y} P(x)\log P(x)$, where $P(x)$ is the prior probability density function over $x \in X$. The remaining uncertainty for Y if X is known is defined by the conditional entropy $H(Y|X) = -\int_x p(x)\{\sum_{y\in Y} p(y|x)\log p(y|x)\}dx$, where $p(y|x)$ denotes the posterior probability for $y \in Y$ given $x \in X$. After observing x, the amount of additional information gain is given by the mutual information

$$I(X;Y) = \sum_{y\in Y}\int_x p(y,x)log\frac{p(y,x)}{p(y)p(x)}dx\ . \tag{1}$$

The mutual information (1) quantifies the information which is shared by X and Y. When the $I(X;Y)$ is large, it implies that x and y are closely related. Otherwise, when $I(X;Y)$ is equal to 0, it means that two variables are totally unrelated. Analogically, the conditional mutual information of X and Y given Z, denoted as $I(X;Y|Z) = H(X|Z) - H(X|Y,Z)$, represents the quantity of information shared by X and Y when Z is known. The conditioning on a third random variable may either increase or decrease the original mutual information. In this context, the Interaction Information $I(X;Y;Z)$ is defined as the difference between the conditional mutual information and the simple mutual information, i.e.

$$I(X;Y;Z) = I(X;Y|Z) - I(X;Y)\ . \tag{2}$$

The interaction information $I(X;Y;Z)$ measures the influence of the variable Z on the amount of information shared between variables X and Y. Its value can be positive, negative, or zero. Zero valued Interaction Information $I(X;Y;Z)$ implies that the relation between X and Y entirely depends on Z. A positive value of $I(X;Y;Z)$ implies that X and Y are independent of each other themselves, but are correlated with each other when combined with Z. A negative value of $I(X;Y;Z)$ indicates that Z can account for or explain the correlation between X and Y. The generalization of Interaction Information to K variables is defined recursively as follow

$$I(\{X_1,\cdots,X_K\}) = I(\{X_2,\cdots,X_K\}|X_1) - I(\{X_2,\cdots,X_K\})\ . \tag{3}$$

Based on the higher order similarity measure, we establish a hypergraph framework for characterizing a set of high dimensional samples. A hypergraph is defined as a triplet $H = (V, E, w)$. Here V denotes the vertex set, E denotes the hyperedge set in which each hyperedge $e \in E$ represents a subset of V, and w is a weight function which assigns a real value $w(e)$ to each hyperedge $e \in E$. We only consider K-uniform hypergraphs (i.e. those for which the hyperedges have identical cardinality K) in our work. Given a set of high dimensional samples $\mathbf{X} = [x_1, \cdots x_N]^T$ where $x_i \in \mathbb{R}^d$, we establish a K-uniform hypergraph, with each hypergraph vertex representing an individual sample and each hyperedge representing the Kth order relations among a K-tuple of participating samples. A K-uniform hypergraph can be represented in terms of Kth order matrix, i.e. a tensor $\mathcal{W}$ of order K, whose element $W_{i_1,\cdots,i_K}$ is the hyperedge weight associated with the K-tuple of participating vertices $\{v_{i_1}, \cdots, v_{i_K}\}$. In our work, the hyperedge weight associating with $\{x_{i_1}, x_{i_2}, \cdots, x_{i_K}\}$ is computed as follows

$$W_{i_1,\cdots,i_K} = K \frac{I(x_{i_1}, x_{i_2}, \cdots, x_{i_K})}{H(x_{i_1}) + H(x_{i_2}) + \cdots H(x_{i_K})} \,. \tag{4}$$

It is clear that $W_{i_1,\cdots,i_K}$ is a normalized version of Kth order Interaction Information. The greater the value of $W_{i_1,\cdots,i_K}$ is, the more relevant the K samples are. On the other hand, if $W_{i_1,\cdots,i_K} = 0$, the K samples are totally unrelated.

3 Hypergraph Representation

Unlike matrix eigen-decomposition, there has not yet been a widely accepted method for spanning a rationale eigen-space for a tensor [13]. Therefore, it is hard to directly embed a hypergraph into a feature space spanned by its tensor representation through eigen-decomposition. In our work, we consider the transformation of a K-uniform hypergraph into a graph. Accordingly, the associated hypergraph tensor $\mathcal{W}$ is transformed to a graph adjacency matrix $\boldsymbol{A}$, and the higher order information exhibited in the original hypergraph can be encoded in an embedding space spanned by the related matrix representation. In this scenario, one straightforward way for the transformation is marginalization which computes the arithmetical average over all the hyperedge weights $W_{i_1,\cdots,i_{K-2},i,j}$ associated with the edge weight $A_{i,j}$

$$\tilde{A}_{i,j} = \sum_{i_1=1}^{|V|} \cdots \sum_{i_{K-2}=1}^{|V|} W_{i_1,\cdots,i_{K-2},i,j} \tag{5}$$

The edge weight $\tilde{A}_{i,j}$ for edge ij is generated by a uniformly weighted sum of hyperedge weights $W_{i_1,\cdots,i_{K-2},i,j}$. However,the form appearing in (5) behaves as a low pass filter, and thus results in information loss through marginalization.

To make the process of marginalization more comprehensive, we use marginalization to constrain the sum of edge weights and then estimate their values through solving an over-constrained system of linear equations. Our idea is motivated by the so called *clique average* introduced in the higher order clustering literature [11]. We characterize the relationships between $\boldsymbol{A}$ and $\mathcal{W}$ as follows

$$W_{i_1,\cdots,i_K} = \sum_{\{i,j\}\subseteq\{i_1,\cdots,i_K\}} A_{i,j} \tag{6}$$

There are $\binom{|V|}{2}$ variables and $\binom{|V|}{K}$ equations in the system of equations described in (5). When $K > 2$, the linear system (5) is over-determined and cannot be solved analytically. We thus approximate the solution to (5) by minimizing the least squares error

$$\widehat{\boldsymbol{A}} = \arg\max_{\boldsymbol{A}} \sum_{i_1,\cdots,i_K} \left(\sum_{\{i,j\}\subseteq\{i_1,\cdots,i_K\}} A_{i,j} - W_{i_1,\cdots,i_K} \right)^2 \tag{7}$$

In practical computation, we normalize the compatibility tensor $\mathcal{W}$ by using the extended Sinkhorn normalization scheme [14], and constrain the element of $\boldsymbol{A}$ to be in the interval $[0, 1]$ to avoid unexpected infinities. Effective iterative numerical methods are used to compute the approximated solutions [15].

The adjacency matrix $\boldsymbol{A}$ computed through (7) is one effective representation for a K-uniform hypergraph, because it naturally avoids the operation of arithmetic average and thus to a certain degree overcomes the low pass information loss arising in (5). Furthermore, the Laplacian matrix $\boldsymbol{L}$ for a hypergraph can be defined as $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}$, where $\boldsymbol{D}$ is the diagonal matrix with its ith diagonal element being $A_{ii} = \sum_j A_{ij}$. In this context, a hypergraph can be easily embedded into a feature space spanned by its Laplacian matrix, which will be explained in detail in the next Section.

4 Feature Selection through Hypergraph Embedding

In this section, we formulate the procedure of feature extraction on a basis of hypergraph spectral embedding. One goal of spectral embedding is to represent the high dimensional data $\mathbf{X} \in \mathbb{R}^{N\times d}$ by a low dimensional representation $\mathbf{Y} \in \mathbb{R}^{N\times C}$ $(C \ll d)$ in the low dimensional feature space such that the structural characteristics of the high dimensional data are well preserved or are more "obvious". Here we use the representations $\mathbf{X} = [x_1, \cdots x_N]^T$ and $\mathbf{Y} = [y_1, \cdots, y_k, \cdots, y_C]$, where y_k is a N-dimensional vector and its N elements represent the N samples $x_1, \cdots x_N$ separately in the kth dimension of the low dimensional feature space.

Based on the hypergraph transformation described in Section 3 and the scheme of Laplacian eigen-decomposition [3], the hypergraph spectral embedding can be easily conducted as follows

$$\boldsymbol{D}^{-1}\boldsymbol{L}\mathbf{Y} = \lambda\mathbf{Y} \ . \tag{8}$$

The hypergraph embedding procedure can be viewed as feature extraction, and can be expressed as $\mathbf{Y} = \mathbf{X}\Phi$ where $\Phi \in \mathbb{R}^{d\times C}$ is a column-full-rank projection matrix. However, unlike feature extraction, feature selection attempts to select the optimal feature subset in the original feature space. Therefore, for the task of feature selection, the projection matrix $\Phi = [\Phi_1, \ldots \Phi_C]$ can be constrained to be a selection matrix which contains the combination coefficients for different features in approximating $\mathbf{Y} = [y_1, \ldots, y_C]$. That is, given the kth column of $\mathbf{Y}$, i.e y_k, we aim to find a subset

of features, such that their linear span is close to y_k. This idea can be formulated as the minimization problem

$$\widehat{\Phi} = \underset{\Phi}{\operatorname{argmin}} \sum_{k=1}^{C} \|y_k - X\Phi_k\|^2 . \tag{9}$$

where $\Phi = [\Phi_1, \cdots, \Phi_k, \cdots, \Phi_C]$ and Φ_k is a d dimensional vector that contains the combination coefficients required to compute for different features in approximating y_k. However, feature selection requires to locate a optimal subset of features that are close to y_k. This is a combinatorial problem which is NP-hard. Thus we approximate the problem in (9) subject to the constraint

$$|\Phi_k| \leq \gamma \tag{10}$$

where $|\Phi_k|$ is the ℓ_1-norm and $|\Phi_k| = \sum_{j=1}^{d} |\Phi_{j,k}|$. When applied in regression, the ℓ_1-norm constraint is equivalent to applying a Laplace prior on Φ_k. This tends to force some entries in Φ_k to be zero, resulting in a sparse solution. Therefore, the representation $\mathbf{Y}$ is generated by using only a small set of selected features in $\mathbf{X}$.

In order to efficiently solve the optimization problem in Equations (9) and (10), we use the Least Angle Regression (LARs) algorithm [9]. Instead of setting the parameter γ, LARs allow us to control the sparseness of Φ_k. This is done by specifying the cardinality of the number of nonzero subset of Φ_k, which is particularly convenient for feature selection.

We consider selecting m features from the d feature candidates. For a dataset containing C clusters, we can compute C selection vectors $\{\Phi_k\}_{k=1}^{C} \in R^d$. The cardinality of each Φ_k is m and each entry in Φ_k corresponds to a feature. Here, we use the following computationally effective method for selecting exactly m features based on the C selection vectors. For every feature j, we define the HG score for the feature as

$$HGscore(j) = \max_{k} |\Phi_{j,k}| . \tag{11}$$

where $\Phi_{j,k}$ is the jth element of vector Φ_k. We then sort the features in descending order according to their HG scores, and then select the top m features.

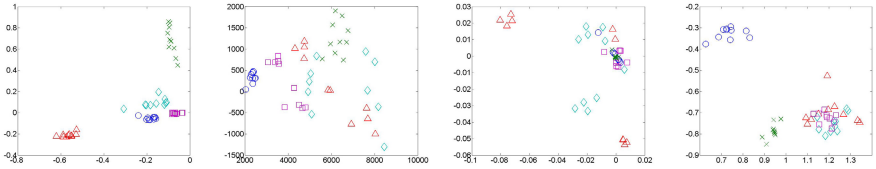
5 Experiments and Comparisons

We test the performance of our proposed algorithm on one publicly available face database (ORL) and one handwritten digit databases (MNIST). Table. 1 summarizes the coverage and properties of the two benchmark datasets.

Data Transformation: We compare the data transformation performance of our proposed method using hypergraph embedding (HG embedding) with alternative methods, including kernel PCA [1], the Laplacian eigenmap [3] and LPP [2]. In order to visualize the results, we have used five randomly selected subjects from each dataset, and these are shown in Fig. 1 and Fig. 2. In each figure, we have shown the projections onto the leading two most significant eigenmodes from different spectral embedding methods,

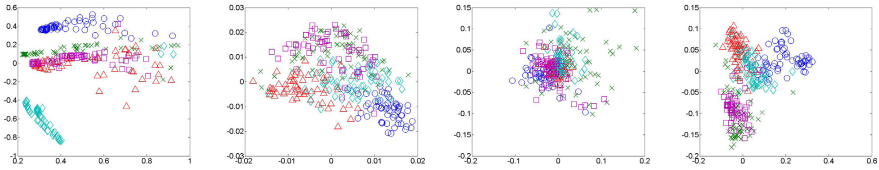
Table 1. Summary of benchmark datasets

Dataset	Examples	Features	Classes
ORL	400	1024	40
MNIST	4000	784	10



(a) HG embedding (b) kernel PCA (c) Laplacian eigenmaps (d) LPP

Fig. 1. Distribution of samples of five subjects in ORL dataset

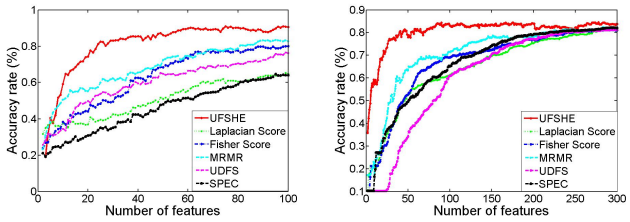


(a) HG embedding (b) kernel PCA (c) Laplacian eigenmaps (d) LPP

Fig. 2. Distribution of samples of five subjects in MNIST dataset

ordered according to their eigenvalues. This provides a low-dimensional representation for the images. From the above figures, it is clear that our hypergraph spectral embedding method demonstrates much clearer cluster structure than alternative spectral clustering methods. This implies that the hypergraph representation is more appropriate and more complete in describing feature relations and structures existing in these datasets.

Classification Accuracy: In order to explore the discriminative capabilities of the information captured by our method, we use the selected features for further classification. We compare the classification results from our proposed method (UFSHE) with five alternative feature selection algorithms. For unsupervised learning, three alternative feature selection algorithms are selected as baselines. These methods are the Laplacian score [6], SPEC [5] and UDFS [10]. We also compare our results with two state-of-the-art supervised feature selection methods, namely a) the Fisher score [4] and b) the MRMR algorithm [12]. We use 5-fold cross-validation for the SVM classifier on the feature subsets obtained by the feature selection algorithms to verify their classification performance. Here we use the linear SVM with LIBSVM.



(a) ORL dataset (b) MNIST dataset

Fig. 3. Accuracy rate vs. the number of selected features on two benchmark image datasets

Table 2. The best result of all methods and their corresponding size of selected feature subset on two benchmark image datasets

Dataset	MRMR	Fisher Score	Laplacian Score	SPEC	UDFS	UFSHE
ORL	83.5%(95)	80%(99)	65.25%(99)	64.5%(95)	76.5%(99)	$\mathbf{91\%(75)}$
MNIST	82.5%(284)	81.25%(293)	82.05%(291)	82.1%(292)	81.3%(293)	$\mathbf{84.33\%(90)}$

The classification accuracies obtained with different feature subsets are shown in Fig. 3. From the figure, it is clear that our proposed method UFSHE is, by and large, superior to the alternative feature selection methods. Specifically, it selects both a smaller and better performing (in terms of classification accuracy) set of discriminative features on both datasets. Moreover, UFSHE rapidly converges, with typically around 30 features. Each of the alternative unsupervised methods, usually require more than 100 features to achieve a comparable result. The reason for this improvement is that the hypergraph representation is effective in capturing the higher order relations among samples and thus the structural information latent in the data can be effectively preserved. Additionally, our hypergraph based feature selection method casts the feature discriminant analysis into a regression framework which suitably characterizes the correlations among features. As a result, the optimal feature combinations can be located so as to remove redundant features.

The best result for each method together with the corresponding size of the selected feature subset are shown in Table. 2. In this table, the classification accuracy is shown first and the optimal number of features selected is reported in brackets. Overall, UFSHE achieves the highest degree of dimensionality reduction, i.e. it selects a smaller feature subset compared with those obtained by the alternative methods. For example, in the MNIST dataset, the best result obtained by the alternative feature selection methods is 82.5% with the MRMR algorithm and 284 features. However, our proposed method (UFSHE) gives a better accuracy of 84.33% when only 90 features are used. The results further verify that our feature selection method can guarantee the optimal size of the feature subset, as it not only achieves a higher degree of dimensionality reduction but it also gives better discriminability.

6 Conclusion

In this paper, we have presented an unsupervised feature selection method based on hypergraph embedding. The proposed feature selection method offers two major advantages. The first is that by incorporating MII for higher order similarities measure, we establish a novel hypergraph framework which is used for characterizing the multiple relationships within a set of samples. Thus, the structural information latent in the data can be more effectively modeled. Secondly, we derive a hypergraph embedding view of feature selection which casting the feature discriminant analysis into a regression framework that considers the correlations among features. As a result, we can evaluate joint feature combinations, rather than being confined to consider them individually. These properties enable our method to be able to handle feature redundancy effectively.

References

1. Scholkopf, B., Smola, A., Muller, K.R.: Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation 10(5), 1299–1319 (1998)
2. He, X., Niyogi, P.: Locality preserving projections (LPP). In: Proc. NIPS (2004)
3. Belkin, M., Niyogi, P.: Laplacian eigenmaps and spectral techniques for embedding and clustering. In: Proc. NIPS, pp. 585–592 (2002)
4. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. John Wiley & Sons, New York (2001)
5. Zhao, Z., Liu, H.: Spectral feature selection for supervised and unsupervised learning. In: Proc. ICML, pp. 1151–1157 (2007)
6. He, X., Cai, D., Niyogi, P.: Laplacian score for feature selection. In: Proc. NIPS (2005)
7. Belhumeur, P.N., Kriegman, D.J.: What is the set of images of an object under all possible illumination conditions? IJCV 28(3), 245–260 (1998)
8. Cai, D., Zhang, C., He, X.: Unsupervised feature selection for multi-cluster data. In: Proc. ACM SIGKDD, pp. 333–342 (2010)
9. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least angle regression. The Annals of Statistics 32(2), 407–499 (2004)
10. Yang, Y., Shen, H.T., Ma, Z., Huang, Z., Zhou, X.: L21-norm regularized discriminative feature selection for unsupervised learning. In: Proc. IJCAI, pp. 1589–1594 (2011)
11. Agarwal, S., Lim, J., Zelnik-Manor, L., Perona, P., Kriegman, D., Belongie, S.: Beyond pairwise clustering. In: Proc. CVPR, pp. 838–845 (2005)
12. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. on PAMI 27(8), 1226–1238 (2005)
13. Kolda, T.G., Bader, B.W.: Tensor decompositions and applications. SIAM Review 51(3), 455–500 (2009)
14. Shashua, A., Zass, R., Hazan, T.: Multi-way Clustering Using Super-Symmetric Non-negative Tensor Factorization. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3954, pp. 595–608. Springer, Heidelberg (2006)
15. Björck, A.: Numberical methods for least squares problems. In: Proc. SIAM (1996)
16. Zhang, Z., Hancock, E.R.: Hypergraph based Information-theoretic Feature Selection. Pattern Recognition Letters (2012)