A Comparison of Linear ICA and Local Linear ICA for Mutual Information Based Feature Ranking

Tian Lan[1], Yonghong Huang[2], Deniz Erdogmus[1,2]

[1] BME Department, OGI, Oregon Health & Science University, Portland, OR, USA.
lantian@bme.ogi.edu
[2] CSEE Department, OGI, Oregon Health & Science University, Portland, OR, USA.
{huang,deniz}@csee.ogi.edu

Abstract. Feature selection and dimensionality reduction is important for high dimensional signal processing and pattern recognition problems. Feature selection can be achieved by filter approach, in which certain criteria must be optimized. By using mutual information (MI) between feature vectors and class labels as the criterion, we proposed an ICA-MI framework for feature selection. In this paper, we will compare the linear ICA and local linear ICA for the accuracy of MI estimation, and study the bias-variance trade-off on feature projections and ranking.

1 Introduction

Recent trends in multi-sensor signal processing coupled with multidimensional statistical feature extraction techniques for pattern recognition leads to extremely high dimensional classification problems, EEG-based pattern recognition problems being one such scenario. Dimensionality reduction and feature selection, therefore, becomes crucial for accurate and robust classifier design. Techniques based on mutual information maximization between features and class labels has attracted increasing attention, because this approach can find out the most relevant features, therefore (i) reduces the computational load in real-time system; (ii) can eliminate irrelevant or noisy features, hence increases the robustness of the system; (iii) is a filter approach, which is independent of the design of classifier, and is more flexible.

The MI based method for feature selection is motivated by lower and upper bounds in information theory [1,2]. The average probability of error has been shown to be related to MI between the feature vectors and the class labels. Fano's and Hellman & Raviv's bounds demonstrate that probability of error is bounded from below and above by quantities that depend on the Shannon MI between these variables. Specifically, Hellman & Raviv showed that the upper bound on Bayes error is given by $(H_S(C)-I_S(\mathbf{Y},C))/2$, where $H_S(C)$ is the Shannon entropy of the a priori probabilities of the classes and $I_S(\mathbf{Y},C)$ is the Shannon MI between the continuous-valued feature vector and the discrete-valued class label. Maximizing this MI reduces the upper bound as well as Fano's lower bound, therefore, forces the probability of error to decrease.

Estimating MI requires the knowledge of the joint pdf of the data in the feature space. This is an especially data consuming estimation problem, and if possible must be avoided. Utilizing individual mutual information of the features with the class labels will surely lead to suboptimal selections, since features are generally mutually dependent and information redundancies cannot be captured with such an approach. Several MI-based methods have been developed for feature selection in the past years [3-8]. Unfortunately, all of these methods failed to solve the particularly difficult high dimensional situation – partly because of the curse of dimensionality that is particularly severe for MI estimation.

In practice, MI must be estimated non-parametrically from the training samples. Although this is a challenging problem for multiple continuous-valued random variables, in classification, the discrete-valued class labels simplify the problem to estimating joint entropy of continuous random vectors. Further simplification is possible if the components of the random vectors are independent or made independent – then the joint entropy becomes the sum of marginal entropies, which are easier to estimate. Thus, the joint mutual information of a feature vector with the class labels is equal to the sum of marginal mutual information of each individual feature with the class labels, provided that the features are independent. In previous work, we exploited this fact and proposed a framework using ICA transformation and sample-spacing estimator to estimate the mutual information between features and class labels [9]. This framework is superior because it is open to diverse algorithms, i.e. each component, including ICA transformation and entropy estimator can be replaced by any qualified algorithm/alternative. Applying linear ICA to an arbitrary feature vector has the drawback that in nonlinear classification problems, the linear ICA model possibly fails, thus estimated MI values are inaccurate. For such situations, nonlinear ICA methods become necessary, and we focus particularly on local linear ICA for this purpose.

In this paper, we will investigate the use of linear and local linear ICA for mutual information estimation. We will compute the estimation bias arising from the possibility that linear ICA might not achieve perfect independence, and study the bias-variance trade-off on feature projections and ranking.

2 Problem Formulation and Asymptotic Analysis

Consider a group of nonlinearly distributed, n-dimensional feature vectors: $\mathbf{x}=[x_1,x_2,...,x_n]^T$. Dimensionality reduction on such a feature vector has to be done to improve the generalization capability of the following classifier without compromising accuracy. The information inequalities mentioned above indicate that the subspace projection should be carried out in a manner that maintains as much mutual information with the class labels as possible. The subspace projection can be achieved by linear/nonlinear projections, as well as feature selection (the latter is a special case of linear projections with binary matrix entries – 0 or 1).

Projection approach: The goal is to determine linear or nonlinear projections that jointly maximize their mutual information with the class labels. Specifically, if $\mathbf{y}=\mathbf{g}(\mathbf{x})$, then we must determine $\mathbf{g}(.)$ such that $I_S(\mathbf{Y};C)$ is maximal. If $\mathbf{g}$ is a solution to

the nonlinear ICA problem given mixture $\mathbf{x}$, then the best m-dimensional nonlinear projection for this NICA solution is the subset $y_1,\ldots,y_m$ such that $I_S(Y_1;C)>I_S(Y_2;C)>\ldots>I_S(Y_n;C)$. Since there are infinitely many solutions to the NICA problem, additional constraints on the form of $\mathbf{g}$ must be imposed. These constraints are typically imposed as model order limitations for parametric nonlinear projections (such as a neural network) or simply as the utilization of a linear projection. For further discussions, we will focus on feature selection for simplicity.

Feature selection: Given a high dimensional feature vector $\mathbf{x}$, our goal is to find the best m dimensional subset of features (in terms of maximum MI with C). This is a combinatorial search problem, and often m is not defined *a priori*. An alternative strategy is to rank the features and pick the top m features from this ranking. Given previously ranked d-1 features $x_{(1)},\ldots, x_{(d-1)}$ the d^{th} feature is the one that maximizes the joint MI: $I_S(x_{(1)},\ldots, x_{(d-1)},x_{(?)};C)$. The joint mutual information takes into account any redundancies in the new feature with the previously ranked d-1 features. This ranking procedure requires the repeated evaluation of d-dimensional MI values. The following procedure is utilized for this purpose.

We first apply a suitable clustering algorithm to segment the data into p partitions: $\mathbf{x}^{(1)}$, $\mathbf{x}^{(2)}$, …, $\mathbf{x}^{(p)}$. We assume that within each partition $\mathbf{x}^{(i)}$, the data is d dimensional, and distributed in accordance with the linear ICA model. We apply the linear ICA transformation on each partition C+1 times to get feature vectors: $\mathbf{y}^{(i|c)}$ and $\mathbf{y}^{(i)}$ for each partition, where c denotes class labels and $\mathbf{y}^{(i|c)}$ are the independent components of data in cluster i from class c only, $\mathbf{y}^{(i)}$ are the independent components of data in cluster i regardless of class labels. As a result of the linear ICA transformation, we have:

$$\begin{aligned} H_S(\mathbf{x}^{(i)}) &= H_S(\mathbf{y}^{(i)}) - \log\,|\,\mathbf{W}^{i}\,| \\ H_S(\mathbf{x}^{(i|c)}) &= H_S(\mathbf{y}^{(i|c)}) - \log\,|\,\mathbf{W}^{i|c}\,| \end{aligned} \tag{1}$$

where $i = 1,\ldots,p$, and $\mathbf{W}^{i}$ and $\mathbf{W}^{i|c}$ are the corresponding ICA separation matrices. If linear ICA works perfectly, then the joint entropy of $\mathbf{y}^{(i|c)}$ and $\mathbf{y}^{(i)}$ reduces to the sum of marginal entropies. However, this is not guaranteed, therefore, the residual mutual information will remain as an estimation bias. In practice, we have an imperfect ICA solution and

$$\begin{aligned} H_S(\mathbf{x}^{(i)}) &= \sum_{l=1}^{d} H_S(y_l^{(i)}) - \log\,|\,\mathbf{W}^{i}\,| - I_S(\mathbf{y}^{(i)}) \\ H_S(\mathbf{x}^{(i|c)}) &= \sum_{l=1}^{d} H_S(y_l^{(i|c)}) - \log\,|\,\mathbf{W}^{i|c}\,| - I_S(\mathbf{y}^{(i|c)}) \end{aligned} \tag{2}$$

Mutual information satisfies the following additivity property for any partition (q_i denoting the probability mass of the corresponding partition):

$$I_S(\mathbf{x};C) = \sum_i q_i I_S(\mathbf{x}^{(i)};C) \tag{3}$$

The mutual information within each partition can be expressed as a linear combination of entropy values as follows:

$$I_S(\mathbf{x}^{(i)};C) = H_S(\mathbf{x}^{(i)}) - \sum_c p_{ic} H_S(\mathbf{x}^{(i)}\,|\,c) \tag{4}$$

where p_{ic} denotes the probability mass of class c in partition i. Substituting (2) in (4)

$$I_S(\mathbf{x}^{(i)};C) = \left(\sum_{l=1}^{d} H_S(y_l^{(i)}) - \sum_c p_{ic} \sum_{l=1}^{d} H_S(y_l^{(i|c)}) \right) - \left(\log|\mathbf{W}^i| - \sum_c p_{ic} \log|\mathbf{W}^{i|c}| \right) - \left(I_S(\mathbf{y}^{(i)}) - \sum_c p_{ic} I_S(\mathbf{y}^{(i|c)}) \right) \quad (5)$$

The last parenthesis in (5) shows the estimation bias one makes when estimating the MI within each partition if it is assumed that the local linear ICA solution in that partition achieved perfect separation. Over all partitions, the total estimation bias (estimated MI minus the actual MI) is averaged as follows:

$$Bias = \sum_i q_i \left(I_S(\mathbf{y}^{(i)}) - \sum_c p_{ic} I_S(\mathbf{y}^{(i|c)}) \right) \quad (6)$$

Note that asymptotically as the number of partitions approach infinity, one could utilize a grid partitioning structure within which the probability distributions would be uniform, thus local linear ICA would achieve perfect separation within each infinitesimal hypercube. However, in practice, one cannot utilize infinitely many partitions given a finite number of samples. Note that the analysis above also holds for the case where linear ICA is employed directly on the whole dataset without any partitions.

3 Empirical Study

We have employed the feature ranking method described above to benchmark datasets. Partitions are identified via K-means clustering, local linear ICA solutions are determined using joint diagonalization of second and fourth order cumulants [10], and marginal entropies are estimated using sample spacing estimators [11].

3.1 Experiments on a Synthetic Dataset

This dataset consists of four dimensional feature vectors: x_i (i=1,…,4), where x_1 and x_2 are nonlinearly related (Fig. 1 - left), x_3 and x_4 are independent from the first two features and are Gaussian distributed with different mean and variance (Fig. 1 - right). There are two classes in this dataset (represented as blue/red or different gray-scale levels in print). These two classes are separable in the x_1 and x_2 plane, but overlapping in the x_3 and x_4 plane. It is clear that this dataset can be well classified only using x_1 and x_2, while x_3 and x_4 provides redundant and insufficient information for perfect classification. From Fig. 1 we can see that x_2 has less overlap compared with x_1, while x_3 has less overlap than x_4. So ideally, the feature ranking in descending order of importance in terms of classification rate should be x_2, x_1, x_3, x_4. In our experiments, we choose the sample size as 1000 used 20 partitions. The '+' in Fig.1 represents the partition centers. We also apply linear ICA without any partitioning.

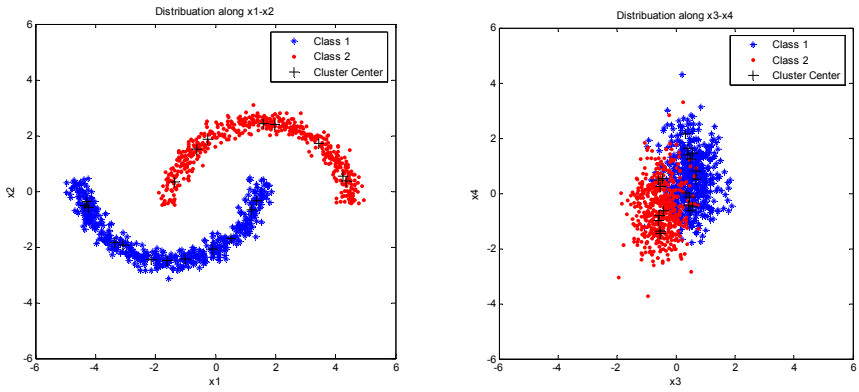

Figure 1 Four-dimensional Synthetic dataset and corresponding cluster centers. Left: distribution of x_1 and x_2; Right: distribution of x_3 and x_4.

Table 1 Feature ranking frequencies on the Iris dataset.

Methods	Ranking indices				
Linear ICA	4	3	2	1	(10)
Local linear ICA	4	1	2	3	(5)
	4	2	3	1	(3)
	4	2	1	3	(2)

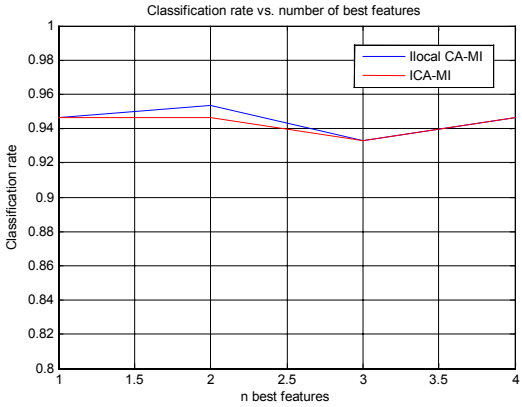

Figure 2 Classification accuracy for Iris data by linear ICA-MI and Local linear ICA-MI methods. The classification accuracy is the average over 10 Monte Carlo simulations.

The linear ICA approach finds the ranking to be x_2, x_1, x_4, x_3, while the local linear ICA approach with 20 partitions finds the expected *correct* ranking.

3.2 Experiments on the Iris Dataset

In this experiment, we applied linear and local linear ICA (with 2 partitions) approaches to the ranking of the features for the Iris dataset from the UCI database [12].

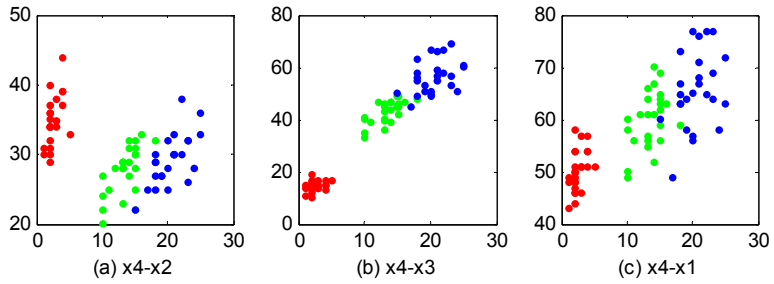

Figure 3 Combinational distribution of 2 feature vectors of Iris dataset. Left: distribution of x_4 and x_2; Middle: distribution of x_4 and x_3; Right: distribution of x_4 and x_1.

Table 2. Feature Ranking results on Wisconsin Breast Cancer dataset for different ICA-MI methods in 10 Monte Carlo simulations. The frequency of different ranking of 10 Monte Carlo simulations are shown inside the bracket.

Methods	Ranking indices
Linear ICA	3 2 9 4 5 6 7 8 1 (9)
	3 2 9 4 5 8 7 6 1 (1)
Local linear ICA	3 1 2 4 5 6 7 8 9 (4)
	3 4 6 8 7 1 9 2 5 (3)
	3 1 4 5 9 6 8 2 7 (3)

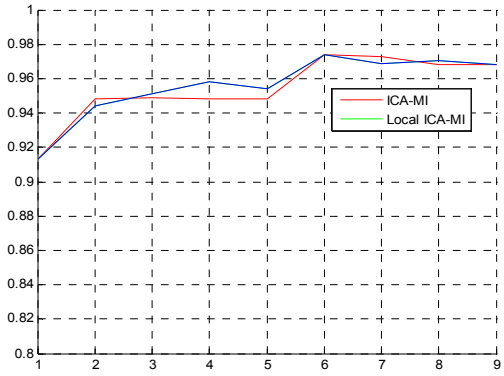

Figure 4 Classification accuracy for Wisconsin Breast Cancer data by different ICA-MI methods. The classification accuracy is the average over 10 Monte Carlo simulations.

Due to the small sample size, 10 Monte Carlo rankings with randomly selected training (used for ranking) and test sets are utilized, each consisting of 50% of the available samples. For each ranked subset, a Gaussian Mixture Model (GMM) based Bayesian classifier is employed. The frequency of rankings and classification accuracy are shown in Table 1. and Fig. 2. Since both methods agree on the fourth feature as the top one, pairwise scatter plots of this feature with the remaining features are shown in Fig. 3 for visual comparison. Feature 3 seems to yield a more compact class

distribution, while features 1 and 4 seem to have less overlapping samples. Still, it is difficult to judge and we rely on the GMM performances on the testing set for the final comparison. The classification accuracy in Fig. 2. shows that local linear ICA yields better performance than linear ICA in Iris data.

3.3 Experiments on the Wisconsin Breast Cancer Dataset

The two methods are applied to this benchmark dataset, which has higher dimensionality than the previous two case studies. Local linear ICA approach uses 2 partitions and the Monte Carlo ranking approach is employed as before. The ranking and classification accuracy are shown in Table 2. and Fig. 4. Local linear ICA also exhibit better performance than linear ICA. Consider the number of data samples and the dimensions, if we partition data into more segments, the performance degrades due to the lack of data for reliable linear ICA transformation within each partition.

4 Conclusions

Feature projections and feature selection are important preprocessing procedures in contemporary pattern recognition problems with extremely high dimensional feature vectors. Mutual information maximization provides a suitable *filter* methodology with proven optimality properties regarding the minimization of bounds for the probability of error one would attain when features selected based on this criteria are utilized.

In this paper, we analyzed the finite sample bias of a local linear ICA based mutual information estimation scheme that can be conveniently used for ranking features for subset selection. Experimental evaluation of the proposed method using 1 and more partitions in localization have revealed that as expected, more accurate results are obtained when large sample sets are available for MI evaluation. The sample size must increase appropriately with increasing data dimensionality; otherwise, the estimates are prone to breaking down at higher dimensional estimations, yielding unreliable rankings after a few dimensions. In very high dimensional and small data size situations, simply assuming a single partition and employing linear ICA rather than local linear ICA might lead to more robust ranking and selection results, though will be based on more biased MI estimates. The bias-variance trade-off will be the determining factor in the choice of the number of partitions for local linear ICA.

Acknowledgments

This work was supported by DARPA under contract DAAD-16-03-C-0054 and by NSF under grant ECS-0524835.

References

1. R. M. Fano, *Transmission of Information: A Statistical Theory of Communications*. Wiley, New York, 1961.
2. M.E. Hellman, J. Raviv, "Probability of Error, Equivocation and the Chernoff Bound," IEEE Transactions on Information Theory, vol. 16, pp. 368-372, 1970.
3 K. Torkkola, "Feature Extraction by Non-Parametric Mutual Information Maximization," Journal of Machine Learning Research, vol. 3, pp. 1415-1438, 2003.
4 R. Battiti, "Using Mutual Information for Selecting Features in Supervised Neural Networks learning," IEEE Trans. Neural Networks, vol. 5, no. 4, pp. 537-550, 1994.
5 A. Ai-ani, M. Deriche, "An Optimal Feature Selection Technique Using the Concept of Mutual Information," Proceedings of ISSPA, pp. 477-480, 2001.
6 N. Kwak, C-H. Choi, "Input Feature Selection for Classification Problems," IEEE Transactions on Neural Networks, vol. 13, no. 1, pp. 143-159, 2002.
7 H.H. Yang, J. Moody, "Feature Selection Based on Joint Mutual Information," in Advances in Intelligent Data Analysis and Computational Intelligent Methods and Application, 1999.
8 H.H. Yang, J. Moody, "Data Visualization and Feature Selection: New Algorithms for Nongaussian Data," Advances in NIPS, pp. 687-693, 2000.
9 T. Lan, D. Erdogmus, A. Adami, M. Pavel, "Feature Selection by Independent Component Analysis and Mutual Information Maximization in EEG Signal Classification," Proceedings of IJCNN'05, Montreal, Canada, pp. 3011-3016, Aug. 2005.
10 L. Parra, P. Sajda, "Blind Source Separation via Generalized Eigenvalue Decomposition," Journal of Machine Learning Research, vol. 4, pp. 1261-1269, 2003.
11 E.G. Learned-Miller, J.W. Fisher III, "ICA Using Spacings Estimates of Entropy," Journal of Machine Learning Research, vol. 4, pp. 1271-1295, 2003.
12 http://www.ics.uci.edu/~mlearn/MLRepository.html