ICA with Sparse Connections: Revisited

Kun Zhang[1], Heng Peng[2], Laiwan Chan[3], and Aapo Hyvärinen[1,4]

[1] Dept of Computer Science & HIIT, University of Helsinki, Finland
[2] Dept of Mathematics, Hong Kong Baptist University, Hong Kong
[3] Dept of Computer Science and Engineering, Chinese University of Hong Kong
[4] Dept of Mathematics and Statistics, University of Helsinki, Finland

Abstract. When applying independent component analysis (ICA), sometimes we expect the connections between the observed mixtures and the recovered independent components (or the original sources) to be sparse, to make the interpretation easier or to reduce the random effect in the results. In this paper we propose two methods to tackle this problem. One is based on adaptive Lasso, which exploits the L_1 penalty with data-adaptive weights. We show the relationship between this method and the classic information criteria such as BIC and AIC. The other is based on optimal brain surgeon, and we show how its stopping criterion is related to the information criteria. This method produces the solution path of the transformation matrix, with different number of zero entries. These methods involve low computational loads. Moreover, in each method, the parameter controlling the sparsity level of the transformation matrix has clear interpretations. By setting such parameters to certain values, the results of the proposed methods are consistent with those produced by classic information criteria.

1 Introduction

Independent component analysis (ICA) aims at recovering latent independent sources from their observable linear mixtures [4]. Denote by $\mathbf{x} = (x_1, ..., x_n)^T$ the vector of observable signals. $\mathbf{x}$ is assumed to be generated by $\mathbf{x} = \mathbf{A}\mathbf{s}$, where $\mathbf{s} = (s_1, ..., s_n)^T$ has mutually independent components. For simplicity we assume the number of observed signals is equal to that of the independent sources. Under certain conditions on the mixing matrix $\mathbf{A}$ and the distributions of s_i, ICA applies a linear transformation on $\mathbf{x}$, i.e., $\mathbf{y} = \mathbf{W}\mathbf{x}$, and tunes the de-mixing matrix $\mathbf{W}$ to make the components of $\mathbf{y} = (y_1, ..., y_n)^T$ mutually as independent as possible; finally y_i provide an estimate of the original sources s_i.

We sometimes prefer the transformation matrix (the de-mixing matrix $\mathbf{W}$ or mixing matrix $\mathbf{A}$) to be sparse, under the condition that y_i are independent, for reliable parameter estimation, or for an easier interpretation purpose [5,11]. For example, when performing LiNGAM (short for linear, non-Gaussian, acyclic models) causality analysis based on ICA [8], we prefer $\mathbf{W}$ to be sparse, since the LiNGAM analysis requires that $\mathbf{W}$ can be permuted to lower triangularity.

Generally speaking, sparsity of the transformation matrix can be easily achieved. One can simply resort to the hard thresholding (which sets small coefficients to

T. Adali et al. (Eds.): ICA 2009, LNCS 5441, pp. 195–202, 2009.

zero), sparse priors [5], the SCAD penalty [11] (which corresponds to an improper prior). Wald test can also be used to set insignificant connections to zero [8]. The problem with these methods is how to determine the free parameter in these methods which controls the level of sparsity. Moreover, for most of them, it is unclear if the subset of the non-zero entries of the transforation matrix could be found consistently (e.g., the estimated subset converges to the correct one in probability when the data follow the model and the sample size grows infinite). On the other hand, one may exploit traditional information criteria, such as BIC [7] and AIC [1], to find the subset of non-zero coefficients in the transformation matrix. The properties of model selection based on information criteria have been well studied. For example, BIC can select the true model consistently, while the model selected by AIC has good prediction performance. Unfortunately, this model selection approach requires exhaustive search over all possible models, which usually involves two stages (training all models followed by comparison of the criteria) and is computationally intensive. Generally speaking, in ICA, the transformation matrix has many entries, and the space of candidate models is too large. Consequently this approach is not practical.

We propose two methods to do ICA with sparse connections which combine the strengths of the two model selection approaches mentioned above. The first one is based on adaptive Lasso [14], which exploits modified L_1 penalties and was recently proposed for variable selection in linear regression. We relate adaptive Lasso with the traditional information criteria, and show how to select the penalization parameter in adaptive Lasso to make its model selection results consistent with those based on information criteria. As L_1 penalties are not differentiable at zero, optimization involving such penalties based on gradients is generally troublesome. We further propose a very simple, yet effective scheme to solve this problem. The second method is based on optimal brain surgeon (OBS) [3] for network pruning. We also show the relationship between this approach and model selection based on traditional information criteria.

2 ICA Based on Maximum Likelihood

Since we will develop ICA with sparse connections by maximizing the penalized likelihood, in this section we briefly review the derivation of ICA algorithms from a maximum likelihood point of view [6]. Denote by f_i the density functions of s_i. The log likelihood of the observed data $\mathbf{x}$ is

$$L_T = \sum_{t=1}^{T} \sum_{i=1}^{n} \log f_i(y_{i,t}) + T \log |\det \mathbf{W}|, \tag{1}$$

where T denotes the sample size. Note that if f_i are not given (say, if it is estimated from data), the scale of y_i and $\mathbf{W}$ estimated by maximizing the above likelihood is arbitrary, due to the scaling indeterminacy in ICA solutions. This can be avoided by constraining the variances of y_i or by keeping certain entries of $\mathbf{W}$ (or $\mathbf{A}$) constant. One scheme is to maximize the likelihood using gradient (or natural gradient) based methods: $\frac{1}{T} \cdot \frac{\partial L_T}{\partial \mathbf{W}} = -E\{\boldsymbol{\psi}(\mathbf{y})\mathbf{x}^T\} + [\mathbf{W}^T]^{-1}$, or, $\frac{1}{T} \cdot$

$\frac{\partial L_T}{\partial \mathbf{A}} = [\mathbf{A}^T]^{-1} \cdot [E\{\boldsymbol{\psi}(\mathbf{y})\mathbf{y}^T\} - \mathbf{I}]$, where $\boldsymbol{\psi}(\mathbf{y}) = (\psi_1(y_1), \cdots, \psi_n(y_n))^T$ with $\psi_1(y_1) = -\frac{f_i'(y_i)}{f(y_i)}$, and in each iteration the variance of each y_i is normalized.

Alternatively, one may incorporate the constraint $E\{y_i^2\} = 1$using the regularization technique. The objective function to be maximized then becomes

$$J_T = \sum_{t=1}^{T}\sum_{i=1}^{n}\{\log f_i(y_{i,t})\} + T\log|\det \mathbf{W}| - \beta\sum_{j=1}^{n}(E\{y_i^2\} - 1)^2, \qquad (2)$$

where β is a regularization parameter. In our experiments we used $\beta = 1$. The gradients of the above function w.r.t. $\mathbf{W}$ and $\mathbf{A}$ can be easily derived.

3 ICA with Sparse Connections Based on Adaptive Lasso

We first propose to achieve the sparsity of the transformation matrix by penalized maximum likelihood. The penalty term we adopt is based on adaptive Lasso [14]. We will show that the result of this penalization method is consistent with that based on traditional information criteria, by setting the penalization parameter to certain given values.

3.1 Idea of Adaptive Lasso

Here we assume that the model under consideration satisfies some regularity conditions including identification conditions for the parameters $\boldsymbol{\theta}$, the consistency of the estimate $\hat{\boldsymbol{\theta}}$ when the sample size T tends to infinity, and the asymptotical normality of $\hat{\boldsymbol{\theta}}$. The penalized likelihood can be written as

$$pL(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) - \lambda p_\lambda(\boldsymbol{\theta}), \qquad (3)$$

where $L(\boldsymbol{\theta})$ is the log likelihood, $\boldsymbol{\theta}$ contains the parameters (which are not redundant), and $p_\lambda(\boldsymbol{\theta}) = \sum_i p_\lambda(\theta_i)$ is the penalty.

The L_1 penalty is well known for producing sparse and continuous estimates [10]. However, it also causes bias in the estimate of significant parameters, and more importantly, it could select the true model consistently only when the data satisfy certain conditions [13]. Adaptive Lasso [14] was proposed to overcome the disadvantage of the L_1 penalty. In adaptive Lasso, $p_\lambda(\boldsymbol{\theta}) = \sum_i \hat{c}_i|\theta_i|$, with $\hat{\boldsymbol{c}} = 1/|\hat{\boldsymbol{\theta}}|^\gamma$, where $\gamma > 0$ and $\hat{\boldsymbol{\theta}}$ is a consistent estimator to $\boldsymbol{\theta}$. In this way, the strength for penalizing different parameters may be different, depending on the magnitude of their estimate. It was shown that under some regularity conditions and the condition $\lambda_T/\sqrt{T} \to 0$ and $\lambda_T T^{(\gamma-1)/2} \to \infty$ (the subscript T in λ_T is used to indicate the dependence of λ on T), the adaptive Lasso estimate is consistent in model selection.

3.2 Relating Adaptive Lasso to Information Criteria

Let us focus on the case $\gamma = 1$ of adaptive Lasso, meaning that

$$p_\lambda(\theta_i) = \hat{c}_i|\theta_i| = |\theta_i|/|\hat{\theta}_i|, \qquad (4)$$

where $\hat{\boldsymbol{\theta}}$ can be any consistent estimator, e.g., the maximum likelihood estimator. After the convergence of the adaptive Lasso procedure, insignificant parameters become zero, and $p_\lambda(\theta_i) = 0$ for such parameters. On the other hand, the "oracle property" [2] holds for adaptive Lasso with suitable λ, meaning that the pointwise asymptotic distribution of the estimators is the same as if the true underlying model were given in advance. Significant parameters are then changed very little by the penalty, when the sample size is not small. Consequently, at convergence, $p_\lambda(\theta_i) = |\hat{\theta}_{i,ALasso}|/|\hat{\theta}_i| \approx 1$ for non-zero parameters, where $\hat{\theta}_{i,ALasso}$ denotes the adaptive Lasso estimator. In other words, the penalty $p_\lambda(\theta_i)$ *indicates whether the parameter* θ_i *is active or not.* Suppose the parameters considered are not redundant. $\sum_i p_\lambda(\theta_i)$ is then an approximator of the number of free parameters, denoted by D, in the resulting model. Recall that the traditional information criteria for model selection can be written as

$$\mathrm{IC}_D = -L(\hat{\boldsymbol{\theta}}_{D,ML}) + \lambda_{IC} D \tag{5}$$

The BIC [7] and AIC [1] criteria are obtained by setting the value of λ_{IC} to

$$\lambda_{BIC} = 1/2 \cdot \log T, \quad \text{and} \quad \lambda_{AIC} = 1, \tag{6}$$

respectively. Relating Eq. 5 to the penalized likelihood Eq. 3, one can see that by setting λ in adaptive Lasso considered here to λ_{IC} in Eq. 5 (which may be λ_{BIC}, λ_{AIC}, etc.), the model selection result of adaptive Lasso would be consistent with that obtained by minimizing the information criterion corresponding to λ_{IC}.

We give the following remarks for model selection based on adaptive Lasso. First, when the initialized model is very large (i.e., it involves very many parameters), $\hat{\boldsymbol{\theta}}$ may be too rough due to finite sample effects, and it is useful to update $\hat{\boldsymbol{\theta}}$ using a consistent estimator when a smaller model is derived. Second, in practice, especially when the sample size is not large, adaptive Lasso still causes bias in the estimate of significant parameters: usually the adaptive Lasso estimator still gives $p_\lambda(\theta_i) = |\hat{\theta}_{i,ALasso}|/|\hat{\theta}_i| < 1$ for significant parameters. Therefore, at convergence, the penalty $p_\lambda(\boldsymbol{\theta}) = \sum_i p_\lambda(\theta_i)$ is expected to be a little smaller than the number of parameters that are set to zero. To achieve that, we should give a heavier weight for the penalization term. That is, λ_T should be a little larger than the recommended values given above. (Or equivalently, $\hat{c}_i$ should be a little larger than $1/|\hat{\theta}_i|$.) In our experiments, we set $\lambda = 1.5\lambda_{BIC} = \frac{1.5}{2}\log T$ to achieve the BIC-like model selection.

3.3 ICA with Sparse Connections Based on Adaptive Lasso

It is obvious that without specifying the variance of y_i or specifying certain entries of $\mathbf{W}$ (or $\mathbf{A}$), applying adaptive Lasso will make the involved parameters smaller and smaller. To avoid that, we can either normalize the variance of y_i in each iteration or enforce $E\{y_i^2\} = 1$ by using Eq. 2 as the objective function. Here we adopt the former scheme. Consequently, the objective function for ICA with a sparse de-mixing matrix is the penalized likelihood:

$$pL_T = \sum_{t=1}^{T}\sum_{i=1}^{n} \log f_i(y_{i,t}) + T\log|\det \mathbf{W}| - \lambda \sum_{i,j=1}^{n} |w_{ij}|/|\hat{w}_{ij}|, \tag{7}$$

where $\hat{w}_{ij}$ are entries of $\hat{\mathbf{W}}$, which is an estimate of $\mathbf{W}$ obtained by conventional ICA. Similarly, replacing $|w_{ij}|/|\hat{w}_{ij}|$ in Eq. 7 with $|a_{ij}|/|\hat{a}_{ij}|$ will produce ICA with a sparse mixing matrix. Note that unlike other methods, here λ is easily determined, based on the relationship between adaptive Lasso and information criteria discussed in Subsection 3.2

Now we aim to maximize the above penalized likelihood. Since the L_1 function is not differentiable at 0, gradient-based methods could not be directly applied for optimization involving L_1 penalties. Most existing methods for such optimization are not easy to implement or could not set insignificant parameters to 0 exactly. We propose a very simple but effective way for this problem.

3.4 A Simple Approach for Optimization Involving L_1 Penalties

The difficulties in optimization involving L_1 penalties are caused by the "sudden change" of the L_1 function. We can then consider such penalties as ravines that are parallel to some axes. The so-called adaptive step size technique [9], which was originally proposed for accelerating the optimization procedure in neural networks learning, can then be exploited for optimization involving such penalties. Note that for a ravine in the objective function parallel to an axis, use of an appropriate individual step size is equivalent to re-scaling the ravine. Moreover, if two successive updates of a given parameter are performed in the same/opposite directions, the step size should be increased/decreased. Consequently, the parameters that should be shrunk to 0 by L_1 penalties will gradually stop oscillation and converge to 0, due to the diminishing step size.

Suppose we aim to maximize the objective function J (Eq. 7, in this case). With an adaptive step size, the change of the parameter θ_i in the kth iteration is given by $\triangle\theta_i^{(k)} = \eta_i^{(k)}(\frac{\partial J}{\partial \theta_i})^{(k)}$, where the step size for parameter θ_i depends on the successive signs of the gradient: $\eta_i^{(k)} = \eta_i^{(k-1)}u$, if $(\frac{\partial J}{\partial \theta_i})^{(k)} \cdot (\frac{\partial J}{\partial \theta_i})^{(k-1)} > 0$, and $\eta_i^{(k)} = \eta_i^{(k-1)}d$, if $(\frac{\partial J}{\partial \theta_i})^{(k)} \cdot (\frac{\partial J}{\partial \theta_i})^{(k-1)} < 0$, with $u > 1$ and $d < 1$. We used $u = 1.1$ and $d = 0.5$ in experiments, and found that they work quite well.

4 ICA with Sparse Connections Based on Optimal Brain Surgeon

Sometimes we may want to obtain the solution path of ICA with sparse connections, which gives all possible solutions with different sparsity levels we are interested in. This can be achieved by using optimal brain surgeon (OBS) [3] for network pruning. We further show the relationship between the stopping criterion of OBS and traditional information criteria, and show how to make OBS produce similar results as information criteria do.

4.1 Optimal Brain Surgeon

Suppose we aim to maximize the objective function J. Assuming that the change of J around its (local) optimum is nearly quadratic in the perturbation of its parameters, i.e., $\delta J = -\frac{1}{2}\delta\boldsymbol{\theta}^T\mathbf{H}\delta\boldsymbol{\theta}$, where $\delta\boldsymbol{\theta}$ denotes the perturbation of the parameters and $-\mathbf{H}$ is the Hessian matrix (containing all second order derivatives). We are looking for a set of parameters whose deletion causes the least change in the value of J.

Mathematically, the least change in J caused by eliminating θ_q can be written as $\min_{\delta\boldsymbol{\theta}}\{\frac{1}{2}\delta\boldsymbol{\theta}^T\mathbf{H}\delta\boldsymbol{\theta}\}$, subject to $\mathbf{e}_q^T\delta\boldsymbol{\theta} + \theta_q = 0$, where $\mathbf{e}_q$ is the unit vector with only the qth element being 1. Using the Lagrangian multiplier, one can find that the optimal weight change and the resulting change in J are

$$\delta\boldsymbol{\theta} = -\frac{\theta_q}{[\mathbf{H}^{-1}]_{qq}}\mathbf{H}^{-1} \cdot \mathbf{w}_q, \quad \text{and} \quad S_q = \frac{1}{2}\frac{\theta_q^2}{[\mathbf{H}^{-1}]_{qq}}, \tag{8}$$

where S_q is called the saliency of the θ_q. OBS finds the q that gives the smallest saliency and prunes it. If J is not very close to quadratic, one needs to adjust the remaining parameters to maximize J, after pruning a parameter and recalculating other parameters according to Eq. 8. We repeat the above pruning procedure ultil the smallest saliency of remaining parameters is larger than T_h, a threshold whose determination is discussed below. One advantage of OBS is that it does not cause any bias in the estimate of the remaining parameters.

4.2 Relating Stopping Criterion of OBS to Information Criteria

Suppose the objective function J is the log-likelihood of the data. We make the following assumptions. *1.* The information criterion Eq. 5 for model selection has no local minimum. *2.* For the OBS procedure, J is well approximated by a quadratic form, and no parameter pruned earlier becomes significant in a smaller model. Under assumption 1, the model selected by minimizing the information criterion has D^* free parameters if $IC_{D^*} > IC_{D^*-1}$ and $IC_{D^*} > IC_{D^*+1}$. According to Eq. 5, this gives $L(\hat{\boldsymbol{\theta}}_{D^*+1,ML}) - L(\hat{\boldsymbol{\theta}}_{D^*,ML}) < \lambda_{IC}$ while $L(\hat{\boldsymbol{\theta}}_{D^*,ML}) - L(\hat{\boldsymbol{\theta}}_{D^*-1,ML}) > \lambda_{IC}$. Assumption 2 implies that $L(\hat{\boldsymbol{\theta}}_{D+1,ML}) - L(\hat{\boldsymbol{\theta}}_{D,ML})$ is actually the smallest saliency S_q when we eliminate a parameter among all the $D+1$ parameters. One can then see that *by setting the threshold T_h for stopping the OBS procedure to λ_{IC} in the information criterion, OBS gives the same model selection result as the corresponding information criterion does.*

4.3 ICA with Entries Pruned by OBS

Entries of the ICA transformation matrix can be pruned by OBS, with Eq. 2 as the objective function. Note that Eq. 2 has incorporated the constraint on the scale of y_i. Due to space limitation, the calculation of the Hessian matrix, as well as how to avoid the heavy computational load in calculating its inverse, is not given here. Note that the quadratic approximation of the objective function may not be very accurate. Consequently, after pruning a parameter and updating others according to Eq. 8, one needs to update remaining parameters to reach the (local) optimum, by making use of the gradient of Eq. 2; alternatively, the Newton-Raphson method may be adopted, as the Hessian matrix has been calculated in the OBS stage.

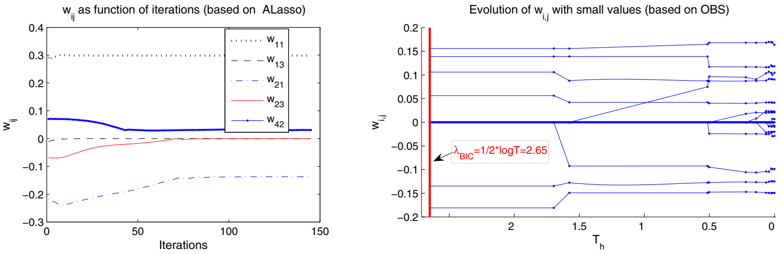
5 Experiments

We first illustrate the performance of the proposed methods by simulation studies. These mothods can carry out both ICA with a sparse de-mixing matrix and ICA with a sparse mixing matrix. Here only the former is demonstrated. We randomly generated a 5 × 5 lower-triangular matrix $\mathbf{W}$. The magnitude of its non-zero entries is uniformly distributed between 0.1 and 1. The mixing matrix was constructed as $\mathbf{A} = \mathbf{W}^{-1}$. The five sources s_i were obtained by passing independent Gaussian i.i.d. signals through power nonlinearities with the exponent between 1.5 and 2. The variances of the sources are randomly chosen between 0.2 and 1. The observations were generated by $\mathbf{x} = \mathbf{A}\mathbf{s}$. We examined two cases in which the sample size is 200 and 500, respectively. ICA with a sparse de-mixing matrix based on adaptive Lasso and that based on OBS, proposed above, were used to separate such mixtures. To make their results similar to that given by BIC, for the former method, we set the penalization parameter $\lambda = 1.5\lambda_{BIC} = \frac{1.5}{2}\log T$, and for the latter one, we set the threshold $T_h = \lambda_{BIC} = \frac{\log T}{2}$.

We repeated the simulation for 40 trials. The percentages of lost connections (non-zero connections that were wrongly set to 0) and spurious connections (zero entries that were not set to 0) are summarized in Table 1. One can see that there are very few entries of $\mathbf{W}$ wrongly identified. As the sample size increases, the error rate diminishes. This coincides with the fact the BIC is consistent in model selection. Fig. 1(a) gives some of w_{ij} in the training process of ICA with sparse $\mathbf{W}$ based on adaptive Lasso in a typical run, while (b) plots the solution path of the OBS-based method for small parameters in a typical run (it shows the

Table 1. Percentages of lost non-zero entries and spurious connections (40 trials)

Sample size T	200		500	
Method	ALasso-based	OBS-based	ALasso-based	OBS-based
Lost connections (%)	2.83%	5%	1%	1.17%
Spurious connections (%)	3%	2%	0.33%	0.25%



(a) (b)

Fig. 1. (a) Some of w_{ij} in the learning process of ICA with sparse $\mathbf{W}$ based on adaptive Lasso (T=200). (b) A typical solution path of ICA with sparse $\mathbf{W}$ based on OBS (T=200). For clarity, only small weigts are shown.

solution of w_{ij} for each possible T_h between 0 and λ_{BIC}). Clearly the pruning result does not depend solely on the magnitudes of the parameters.

We also applied the proposed methods for ICA with sparse $\mathbf{W}$ to separate the 14-dimensional financial returns used in [12]. To obtain BIC-like model selection results, we used the same settings as in the simulations above. The resulting $\mathbf{W}$ could not be permuted to lower triangularity, meaning that LiNGAM [8] does not hold for this data set. This is consistent with the claim in [12].

6 Conclusion and Discussions

We have proposed two methods to perform ICA with a sparse transformation matrix (the mixing matrix $\mathbf{A}$ or de-mixing matrix $\mathbf{W}$). The methods are based on the adaptive Lasso penalty and the optimal brain surgeon technique, respectively. We have shown how to relate the proposed methods to model selection based on traditional information criteria (e.g., BIC and AIC). The proposed methods involve comparatively light computational load, and most importantly, one can easily determine the parameters that control the level of sparsity to make the model selection results consistent with those based on information criteria.

References

1. Akaike, H.: Information theory and an extension of the maximum likelihood principle. In: Proc. 2nd Int. Symp. on Information Theory, pp. 267–281 (1973)
2. Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. J. Amer. Statist. Assoc. 96, 1348–1360 (2001)
3. Hassibi, B., Stork, D.G.: Second order derivatives for network pruning: Optimal brain surgeon. In: NIPS 5, pp. 164–171. Morgan Kaufmann, San Francisco (1993)
4. Hyvärinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. John Wiley & Sons, Inc., Chichester (2001)
5. Hyvärinen, A., Karthikesh, R.: Imposing sparsity on the mixing matrix in independent component analysis. Neurocomputing 49, 151–162 (2002)
6. Pham, D.T., Garat, P.: Blind separation of mixture of independent sources through a quasi-maximum likelihood approach. IEEE Trans. on Signal Processing 45(7), 1712–1725 (1997)
7. Schwarz, G.: Estimating the dimension of a model. The Annals of Statistics 6, 461–464 (1978)
8. Shimizu, S., Hoyer, P.O., Hyvärinen, A., Kerminen, A.J.: A linear non-Gaussian acyclic model for causal discovery. JMLR 7, 2003–2030 (2006)
9. Silva, F.M., Almeida, L.B.: Acceleration techniques for the backpropagation algorithm. In: Neural Networks, pp. 110–119. Springer, Heidelberg (1990)
10. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society 58(1), 267–288 (1996)
11. Zhang, K., Chan, L.-W.: ICA with sparse connections. In: Corchado, E., Yin, H., Botti, V., Fyfe, C. (eds.) IDEAL 2006. LNCS, vol. 4224, pp. 530–537. Springer, Heidelberg (2006)
12. Zhang, K., Chan, L.: Minimal nonlinear distortion principle for nonlinear independent component analysis. JMLR 9, 2455–2487 (2008)
13. Zhao, P., Yu, B.: On model selection consistency of lasso. JMLR 7, 2541–2563 (2006)
14. Zou, H.: The adaptive lasso and its oracle properties. Journal of the American Statistical Association 101(476), 1417–1429 (2006)