Structure Learning by Pruning in Independent Component Analysis

Andreas Brinch Nielsen and Lars Kai Hansen

Informatics and Mathematical Modelling,
Technical University of Denmark,
DK-2800 Kgs. Lyngby, Denmark

Abstract

We discuss pruning as a means of structure learning in independent component analysis. Sparse models are attractive in both signal processing and in analysis of abstract data, they can assist model interpretation, generalizability and reduce computation. We derive the relevant saliency expressions and compare with magnitude based pruning and Bayesian sparsification. We show in simulations that pruning is able to identify underlying sparse structures without prior knowledge on the degree of sparsity. We find that for ICA magnitude based pruning is as efficient as saliency based methods and Bayesian methods, for both small and large samples. The Bayesian information criterion (BIC) seems to outperform both AIC and test sets as tools for determining the optimal degree of sparsity.

Key words: Independent component analysis, ICA, pruning, sparsity, automatic relevance determination, OBD, OBS

Independent component analysis is a topic of significant current interest in many scientific areas, see, e.g., (Hyvärinen, 2001). ICA is used in a variety of signal processing applications such as reconstruction of sound sources from multi-microphone recordings, image processing, as well as for understanding EEG signals from multiple independent sources in the human brain. ICA has found numerous uses in data analysis ranging from text mining to bioinformatics and modelling of music, see, e.g., the recent conference proceedings (Rosca et al., 2006).

Structure learning is of interest in independent component analysis of abstract data and in causal inference where it can assist *model interpretation*, see, e.g., (Tenenbaum and Griffiths, 2000; Hansen et al., 2001; Shimizu et al., 2005; Hoyer et al., 2006). Furthermore, structure learning is relevant to classical blind signal separation applications. In ICA of complex multi-agent visual or auditory scenes sources

Email address: abn,lkh@imm.dtu.dk (Andreas Brinch Nielsen and Lars Kai Hansen).

can be present only in a subset of the recorded mixtures. For example in video where objects can be obscured (Bronstein et al., 2005). In blind separation of audio mixtures of sources in motion a source can be blocked by a physical barrier so that it practically disappears from the scope of a microphone.

In the recent work on ICA based causal modeling (Shimizu et al., 2005; Hoyer et al., 2006) a sparse Markovian structure is sought, causing the separation matrix to become lower triangular or a row/column permutation of a lower triangular matrix.

In general, structure learning reduces complexity and can reduce overfit in short data sequences, hence improves generalizability by Occam's principle. ICA complexity control at the level of estimation of the number of sources has earlier been shown to improve generalizability, see, e.g., (Hansen et al., 2001). Finally, we note that structurally optimized sparse models may reduce the computational burden.

A number of schemes have been proposed for learning sparse representations in ICA, most are based on using sparse priors for the mixing coefficients, see, e.g., (Knuth, 1999; Højen-Sørensen et al., 2002; He and Cichocki, 2006; Park et al., 2006) . However, a more direct approach to sparse models is provided by parameter pruning. In (Shimizu et al., 2005; Hoyer et al., 2006) a scheme was proposed for causal structure discovery, which essentially amounts to weight magnitude based pruning.

Pruning approaches to sparsification are widely used in other machine learning contexts, often based on estimating the individual parameter saliency, see, e.g., (Bishop, 1996) for a comprehensive review. Pruning typically proceeds by eliminating the least salient parameters, retraining the reduced model, and repeating the procedure until a minimal configuration is obtained. Selecting the optimal model within the derived nested sequence of pruned models can be based on cross-validation, or on criteria such as BIC, AIC or MacKay's Evidence (Bishop, 1996). Here we will investigate these schemes for ICA. In particular we are interested in testing whether magnitude based pruning is viable.

The paper is organized as follows. First we outline the basic ICA model in section 1. In section 2 pruning based on estimated saliency will be reviewed. We run a simulation experiment in 3 to illustrate the ranking quality. In section 4 we will test the basic model selection criteria. Finally, in section 5, we present a more elaborate simulation experiment to demonstrate the efficiency of pruning based structure learning in auditory scene analysis.

1 Independent Component Analysis

The basic independent component model involves a linear mixture of L sources into L observations, $\mathbf{X} = \mathbf{AS}$, where $\mathbf{S} = \{\mathbf{s}^1, ..., \mathbf{s}^T\}$ is the source matrix, $\mathbf{X} = \{\mathbf{x}^1, ..., \mathbf{x}^T\}$ the observation matrix, $\mathbf{s}^t$ and $\mathbf{x}^t$ the source and observation L-vectors at time t. $\mathbf{A}$ is a square $L \times L$ mixing matrix. Assuming that the source signals are i.i.d., $p(\mathbf{s}^t) = \prod_i p(s_i^t)$, we obtain the likelihood function,

$$\begin{aligned} p(\mathbf{X}|\mathbf{A}) &= \prod_t p(\mathbf{x}^t|\mathbf{A}), \\ &= \prod_t \int p(\mathbf{x}^t|\mathbf{s}^t, \mathbf{A}) \prod_i p(s_i^t) \mathrm{d}\mathbf{s}^t. \end{aligned}$$

Introducing the separation matrix $\mathbf{W} = \mathbf{A}^{-1}$ the likelihood can be written,

$$\begin{aligned} p(\mathbf{x}^t|\mathbf{s}^t, \mathbf{A}) &= \delta(\mathbf{x}^t - \mathbf{A}\mathbf{s}^t), \\ p(\mathbf{x}^t|\mathbf{W}) &= \int \delta(\mathbf{x}^t - \mathbf{W}^{-1}\mathbf{s}^t) \prod_i p(s_i^t) \mathrm{d}\mathbf{s}^t, \\ &= |\det \mathbf{W}| \prod_i p(\textstyle\sum_j W_{ij} x_j^t). \end{aligned}$$

We will assume a simple parameter free source distribution for computational convenience, $p(s) = \frac{1}{\pi \cosh(s)}$. The negative log likelihood can thus be written in terms of either the mixing or the separation matrix,

$$\begin{aligned} \mathcal{L} &= -T \log |\det \mathbf{W}| + \sum_t \sum_i \log \cosh(\textstyle\sum_j W_{ij} x_j^t) + TL \log \pi, \\ &= M(\mathbf{W}), \\ \mathcal{L} &= T \log |\det \mathbf{A}| + \sum_t \sum_i \log \cosh(\textstyle\sum_j [\mathbf{A}^{-1}]_{ij} x_j^t) + TL \log \pi, \\ &= M(\mathbf{A}). \end{aligned}$$

We base estimation of the mixing parameters on maximum likelihood leading to Infomax-like algorithms (Bell and Sejnowski, 1995; Cardoso, 1997). As discussed in Pedersen et al. (2005) optimization of the mixing or separation matrices do not necessarily produce equivalent results. In the context of sparsity there is further the issue of whether we seek sparsity of the mixing $\mathbf{A}$ or separation $\mathbf{W}$ matrices. Here we will discuss both options and in appendix A we provide the necessary gradients and Hessian expressions.

2 Sparsity by pruning

Pruning has been widely used for simplification of neural network architectures (Le Cun et al., 1990; Thodberg, 1991; Hassibi and Stork, 1993; Gorodkin et al., 1993; Pedersen et al., 1996; Thodberg, 1996; Ragg et al., 1997; Gorodkin et al., 1997; Kaashoek and van Dijk, 2002). Pruning typically proceeds from some over-parametrized model which is simplified in a sequence of pruning/retraining steps leading to some minimum null configuration. The sequence of nested architectures can then be inspected for performance to obtain a given objective, e.g., cross-validation performance. The pruning process can be seen as an efficient heuristic substituting for a complete search through all possible architecture subsets of the initial configuration.

The first and simplest approach to importance ranking is by parameter magnitude (Le Cun et al., 1990) and it has recently been proposed for ICA in Shimizu et al. (2005). In (Le Cun et al., 1990) however, it was shown that for optimization of neural network architectures for generalizability significantly better results can be obtained using the saliency concept. The key idea is to compute the increase in the cost function - or equivalently the decrease in likelihood - incurred by setting a given parameter to zero. One way to achieve this is by actually setting the parameter to zero and recompute the error, this has been referred to as 'brute force' pruning (Thodberg, 1991). To reduce the computational burden, ranking based on *estimated* saliency has been proposed.

We note that if model selection is based on either probability using the Bayesian information criterion or estimated by Akaike's information criterion, these criteria take the form of the training set log-likelihood plus a penalty term that is only a function of the *number* of parameters, hence is optimal if we remove the parameter that incurs the least decrease in the training set log-likelihood. Hence, the saliency should in these cases be based on the likelihood cost.

In the so-called optimal brain damage scheme (OBD) (Le Cun et al., 1990) the saliency estimator is based precisely on an expansion of the cost function which can be the negative log-likelihood,

$$\delta E_i = \frac{\partial E}{\partial \theta_i}\delta\theta_i + \frac{1}{2}H_{ii}(\delta\theta_i)^2 + \mathcal{O}(\delta\theta_i^3).$$

$\mathbf{H}$ is the Hessian matrix of second derivatives. If the set of weights used for expansion are obtained by cost minimization, the first term can be assumed zero. If we further neglect third order terms the price of setting a single weight to zero can be estimated by the *OBD saliency*,

$$\delta E_i \sim S_i \equiv \frac{1}{2} H_{ii} \delta\theta_i^2 = \frac{1}{2} H_{ii} \theta_i^2.$$

Later Hassibi and Stork (1993) noted that within the second order approximation it is possible to estimate the result of both deleting a parameter and *retrain* the remaining parameters. This leads to the so-called optimal brain surgeon (OBS) saliency,

$$S_i = \frac{1}{2} \frac{\theta_i^2}{[\mathbf{H}^{-1}]_{ii}}.$$

Sparsification based on parameter priors can be cast in terms so-called *pruning priors* (Goutte, 1996; Goutte and Hansen, 1997) which is equivalent to shrinkage regression with an L1 norm regularization, see, e.g., (Tibshirani, 1996). Alternatively, hierarchical priors with adaptive hyper-parameters can lead to pruning (Hansen and Rasmussen, 1994). This mechanism was used in (Tipping, 2001) to obtain sparse Bayesian models. The basic mechanism is also referred to as automatic relevance determination (ARD), see, e.g., (MacKay, 1995). We here review the basic results, see also (Bishop, 1996) for additional detail.

We consider the joint pdf of data and parameters obtained by introducing a prior on the parameters,

$$p(\mathbf{X}, \boldsymbol{\theta}|\boldsymbol{\alpha}) = p(\mathbf{X}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{\alpha}),$$

where $\boldsymbol{\theta}$ can be either $\mathbf{A}$ or $\mathbf{W}$ and $\boldsymbol{\alpha}$ is the regularization parameter vector. We will use the conventional Gaussian prior with the hyper-parameters playing the role of precisions $p(\boldsymbol{\theta}|\boldsymbol{\alpha}) \propto \exp(-\frac{1}{2}\sum_i \alpha_i \theta_i)$.

In the so-called *maximum likelihood II* approach the hyper-parameter vector $\boldsymbol{\alpha}$ is estimated using the so-called evidence, which is the marginal likelihood obtained by integrating out the parameters. In ICA models it is not possible to perform the marginalization in closed form, but we can invoke the Laplace approximation which is based on a Taylor expansion of the negative log-likelihood in the vicinity of the maximum posterior value of $\boldsymbol{\theta}$,

$$p(\boldsymbol{\theta}|\boldsymbol{\alpha}, \mathbf{X}) \cong e^{M(\boldsymbol{\theta}_{MP}, \boldsymbol{\alpha})} \left| \frac{\mathbf{H}}{2\pi} \right|^{-\frac{1}{2}} e^{-M(\boldsymbol{\theta}_{MP}, \boldsymbol{\alpha}) - \frac{1}{2}\Delta\boldsymbol{\theta}^T \mathbf{H} \Delta\boldsymbol{\theta}},$$

where $\mathbf{H}$ is the Hessian computed at the ($\boldsymbol{\alpha}$-dependent) current values of $\boldsymbol{\theta}_{MP}$. The negative log-likelihood of the hyper-parameters is estimated by integrating the parameters out using the approximate posterior,

$$-\log p(\mathbf{X}|\boldsymbol{\alpha}) \cong M(\boldsymbol{\theta}_{MP}, \boldsymbol{\alpha}) - \frac{1}{2}\log|\mathbf{H}| + \frac{1}{2} L^2 \log 2\pi.$$

This expression is minimized with respect to the hyper-parameters in an EM-like iterative scheme in which first, the maximum posterior weights are estimated for fixed hyper-parameters, and next, we re-estimate the hyper-parameters for fixed weights using gradient descent based on,

$$\begin{aligned}\frac{\partial(-\log p(\mathbf{X}|\boldsymbol{\alpha}))}{\partial\alpha_{ij}} &= \frac{\partial M(\boldsymbol{\theta}_{MP},\boldsymbol{\alpha})}{\partial\alpha_{ij}} + \frac{1}{2}\frac{\partial\log|\mathbf{H}|}{\partial\alpha_{ij}},\\ &= \frac{1}{2}(\boldsymbol{\theta}_{MP})_{ij}^2 - \frac{1}{2\alpha_{ij}} + \frac{1}{2}[\mathbf{H}^{-1}]_{(ij)(ij)} = 0 \Leftrightarrow \qquad (1)\\ \alpha_{ij}^{new} &= \frac{1-\alpha_{ij}^{old}[\mathbf{H}^{-1}]_{(ij)(ij)}}{\theta_{MP,ij}^2}.\end{aligned}$$

The derivative of $\log|\mathbf{H}|$ is,

$$\begin{aligned}\frac{\log|\mathbf{H}|}{\partial\alpha_{ij}} &= \mathrm{Tr}(\mathbf{H}^{-1}\frac{\partial\mathbf{H}}{\partial\alpha_{ij}}),\\ \mathbf{H} &= \nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}}M(\boldsymbol{\theta}) + \mathrm{diag}(\boldsymbol{\alpha}).\end{aligned}$$

Equation (1) could be 'solved' and the optimal hyper-parameters found in one step, however practise has shown that this is too greedy. Therefore hyper-parameters and parameters are updated iteratively (Bishop, 1996).

For a Gaussian weight prior the individual hyper-parameters can either converge to some finite value, hence producing finite weights or they can escape to infinity, implying a delta-like posterior distribution for the corresponding parameters, i.e., effectively pruning the associated parameter away (Hansen and Rasmussen, 1994; MacKay, 1995).

3 Quality of ranking

In this section experiments will evaluate the quality of the pruning algorithms in a simulated example. We will initially assume that the true number of non-zero parameters is known. We will consider pruning based on magnitude (MB), ARD, OBD and OBS for a synthetic data set.

A random mixing or separation matrix is generated with a number of zeros and is used to generate the observations. The non-zero entries are generated using $W = r + \frac{1}{2}\mathrm{sign}(r)$, where $r \sim \mathcal{N}(0,1)$. In the experiment the matrix is chosen quadratic with $L = 5$ sources. To gauge the dependence on sample size we vary the number of observations $T \in [0; 1000]$.

In ARD based pruning the procedure is carried out once (iterating weights and hyper-parameter to convergence) and the resulting hyper-parameters are used to define the ranking order for deletion towards the desired number of non-zero elements. For MB, OBD and OBS we prune one weight at a time until the desired number is achieved, including intermediate re-training steps of each partially pruned model. The experiment is repeated 1000 times with different random sources and mixing/separation matrices. In this study we have created 'true' models that are lower triangular. Since a lower triangular matrix has a lower triangular inverse this corresponds to the Markovian structures investigated in (Hoyer et al., 2006). We report in figure 1 the miss-classification rate as a function of the number of samples in the training data. Both for estimating mixing and separation matrices we find that performance is strongly dependent on the amount of data provided. The maybe most surprising result is that in both cases the magnitude based method outperforms the more involved estimators with OBD being significantly poorer in the range of larger sample size where the performance of the three other methods are more or less at par.

To understand the properties of the Hessian we can pass to the large sample limit, by replacing sums over t with expectations over the assumed distribution of the source signals, the Hessian (parameterized using the separation matrix) then becomes,

$$\widehat{H}_{i'j',i''j''} = T(A_{j'i''}A_{j''i'} + \delta_{i',i''}\sum_i A_{j'i}A_{j''i}\mathcal{E}_t[(s_i^t)^2\left(1-\tanh^2 s_{i'}^t\right)]).$$

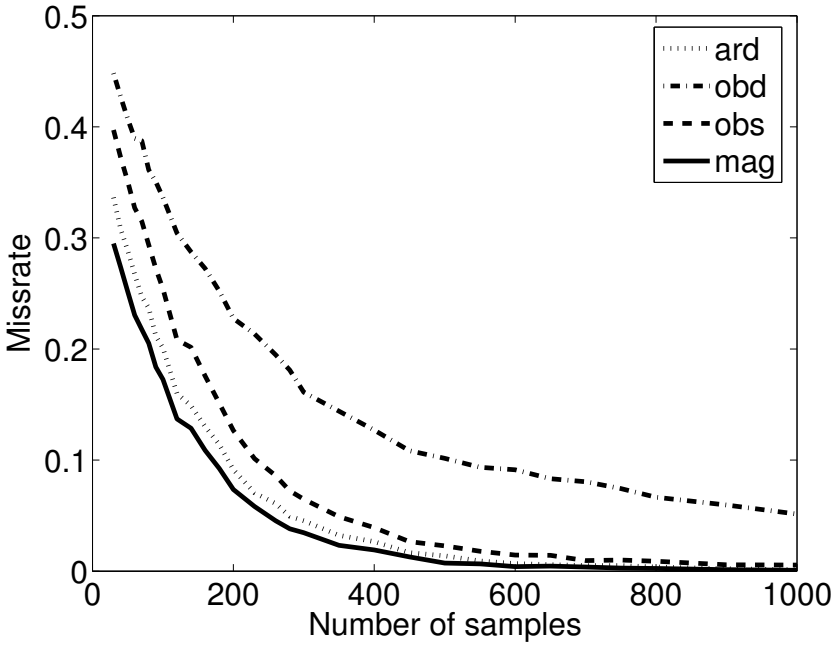
The expectation value is given by $\mathcal{E}_t[(s_i^t)^2(1-\tanh^2 s_{i'}^t)] = \frac{\pi^2}{8} - \delta_{ii'}$. Hence, we obtain,

$$\widehat{H}_{i'j',i''j''} = T((1-\delta_{i',i''})A_{j'i''}A_{j''i'} + \delta_{i',i''}\frac{\pi^2}{8}\sum_i A_{j'i}A_{j''i}). \quad (2)$$

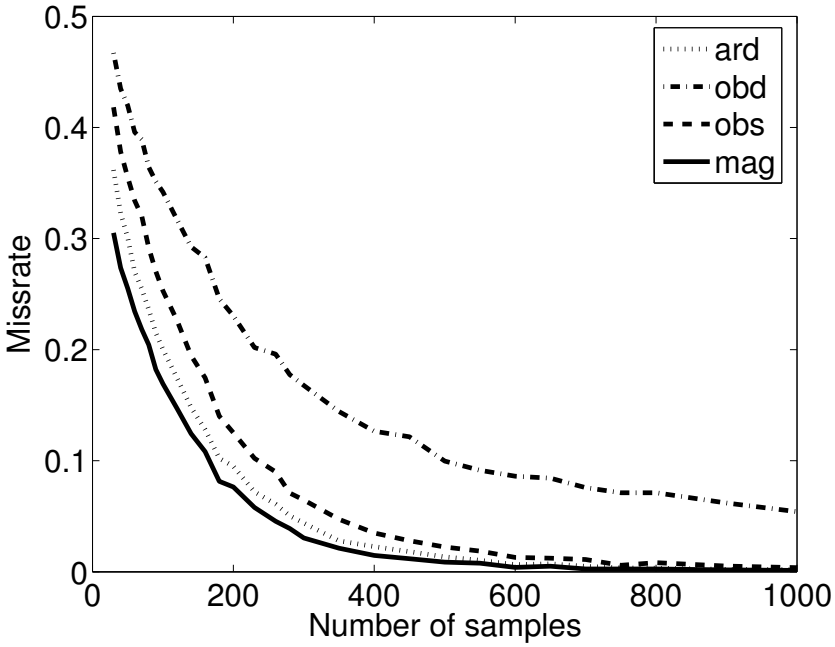
Now, for OBD only the diagonal elements of the Hessian are of interest and the saliency becomes,

$$S_{i'j'} = \frac{1}{2}\widehat{H}_{i'j',i'j'}W_{i'j'}^2 = \frac{\pi^2}{16}TW_{i'j'}^2\sum_i A_{j'i}^2.$$

The Hessian elements in equation (2) are seen to depend only weakly on the actual location of the weight, i.e., the second term takes a constant value for all elements in a row of the mixing matrix. This may explain in part why the saliency is not significantly better than mere MB pruning. Overall, the fact the estimated saliencies are less efficient may indicate that the second order expansion is not always valid. It is well-known that the ICA likelihood has singularities due to co-linearity in the mixing matrix, see, e.g., (Pedersen et al., 2005). Clearly, the radius of convergence for Taylor expansions do not extend beyond such singularities. To further probe



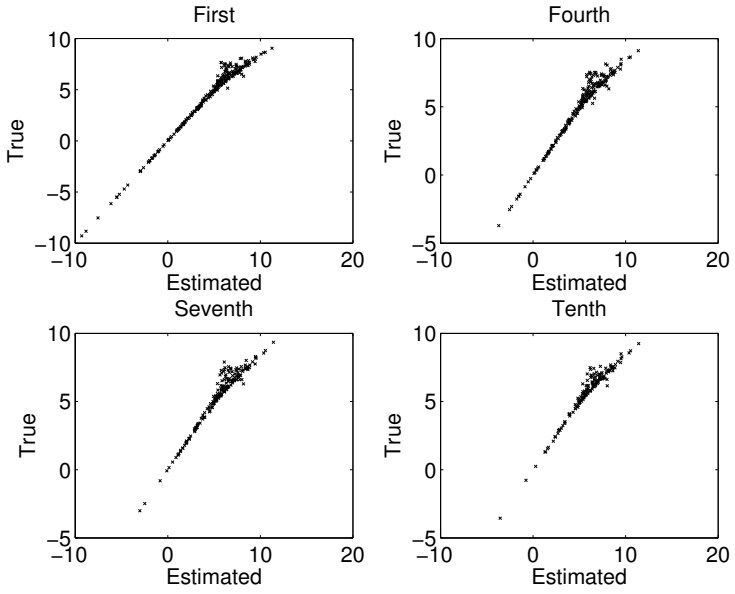
(A)



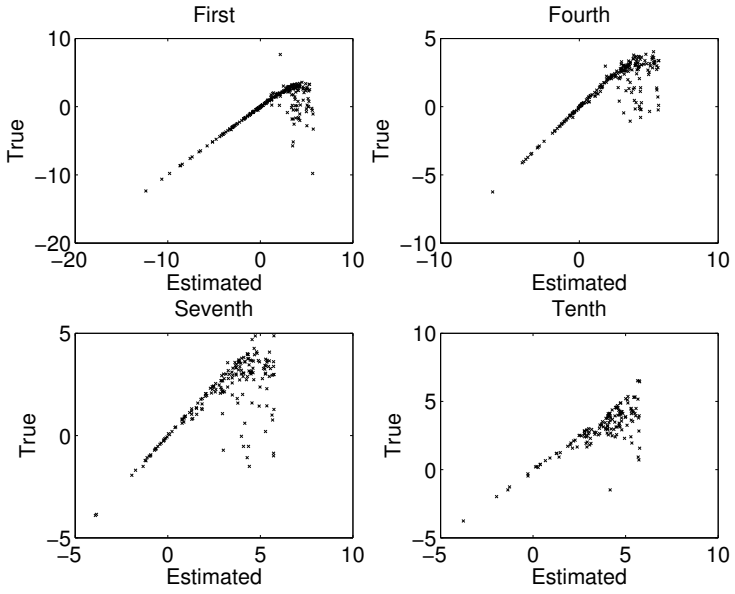
(B)

Fig. 1. Mis-classification rate in 1000 pruning sequences. The rate is reported versus sample size. The ICA problem is square with D=5, the true mixing matrix is lower triangular, hence having 10 zeros. In (A) we show performance for pruning based on the separation matrix while in (B) the result is based on the mixing matrix. Performance is strongly dependent on sample size. While performances of magnitude based, ARD and OBS are comparable, OBD seems to fall short even after learning in relatively large data sets.

the validity of the estimated saliency, we have setup an experiment similar to an evaluation carried out in Gorodkin et al. (1993). At four positions along a pruning



(A)



(B)

Fig. 2. Estimated vs. true saliency. A brute force measurement of the cost in likelihood of deleting parameters at four positions along a pruning sequence (when pruning the first parameter, the fourth, the seventh and the tenth). Log-log axes are used to accomodate the significant dynamic range. (A) is OBD and (B) is OBS.

sequence (when pruning the first parameter, the fourth, the seventh and the tenth) we measure the 'real saliency' by brute force elimination of all parameters in turn, for the OBS saliency we also retrain after every tentative deletion. These measures are compared to the estimated saliency as seen in figure 2. There is a comforting agreement, with a tendency towards over estimation of saliency for OBS and under-estimation for OBD. The largest deviations are found for OBS indicating that the

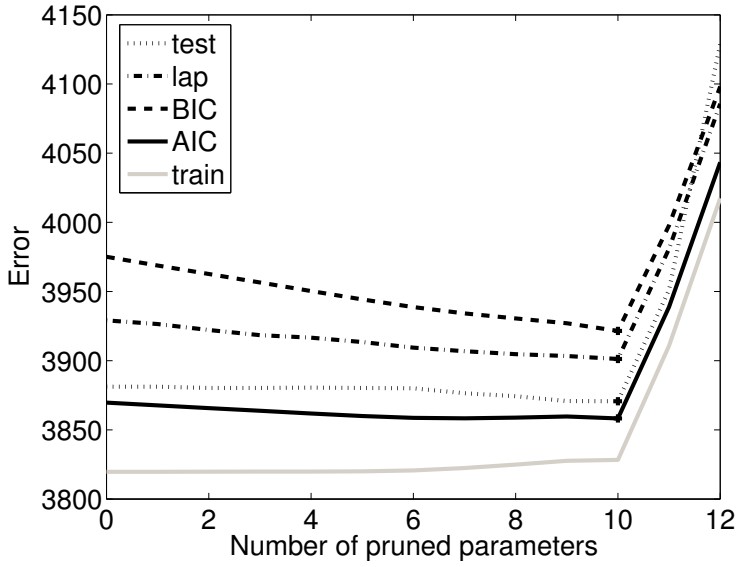
second order expansion is not quite reliable for estimating the retraining configuration, while the cost of deletion is fine for most weights. This however does not translate into OBD being better for pruning, because the importance is not directly related to the absolute value of the weight, hence, not an indication of whether a weight is zero in the 'true' configuration.

4 Model selection

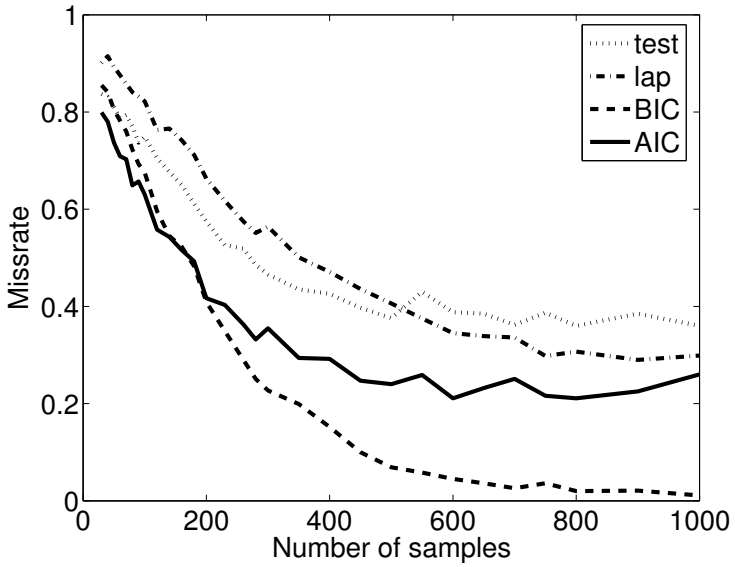
As we have seen the pruning procedure is able to produce acceptable results if informed about the number of non-zero elements, hence, the optimal structure is in the nested sequence, and we just need to find it. We propose to follow the procedures developed for artificial neural networks and base model selection on either expected generalizability or posterior probability. The posterior probability is aimed at finding the correct architecture, i.e., the placement of zero and non-zero elements. However, for finite training samples the correct architecture need not be the one that has the highest expected log-likelihood, since the non-zero parameters may have a wrong value. As the training sample size increases the two measures will converge to make identical decisions. The posterior probability can be estimated either using the Bayes information criterion (BIC) or the higher order approximation denoted the 'Laplace approximation' (LAP) (Bishop, 1996). The expected log-likelihood can be estimated using either a test set (cross-validation) or by Akaike's information criterion (AIC). In figure 3 (A) these different measures are compared for a single run. As seen the 'training error' (negative log-likelihood on the training sample) can not be used for model selection as it is steadily increasing during pruning. The other criteria all finds the true number of zero parameters in this example.

In figure 3 (B) we show the miss classification rate of the correct number of zeros over 1000 trials. As before we test the ability to find the correct structure as function of the training set size. In this experiment we used MB pruning. We find that BIC is superior for larger data sets. The BIC and the Laplace approximation are both aimed at locating the correct model, hence, more consistent with our evaluation in terms of miss-rate than the generalizability measures. The Laplace approximation seems to be too sensitive due to the problems with the second order expansion we have mentioned earlier induced by the singularities in the ICA likelihood, hence, fails more often than the simpler BIC. The failure of the test error can be due to two factors. First, it focuses more on getting the density function correct, hence getting the correct parameters, and secondly, it is a quite general finding that the cross-validated test error is a noisy quantity.

We finally evaluate the efficiency of the different pruning schemes in combination with BIC for model selection. The picture in figure 4 confirms our earlier findings: MB, OBS and ARD all perform well for larger data sets. We conclude that the



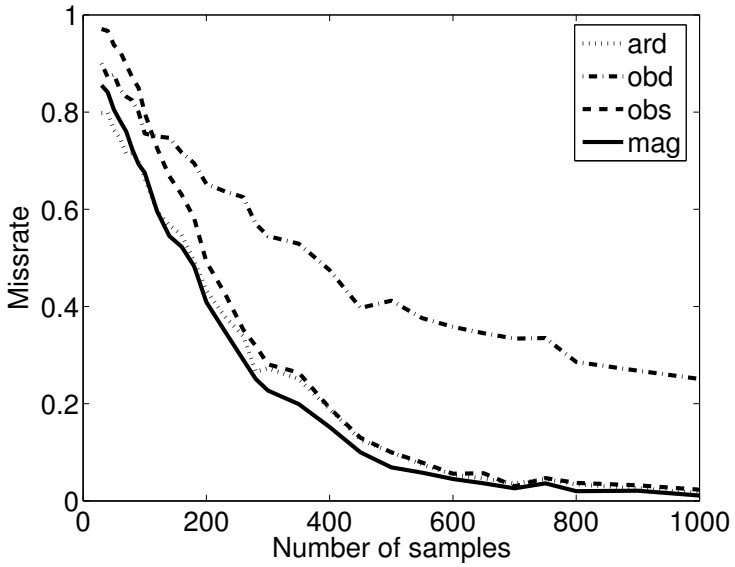
(A)



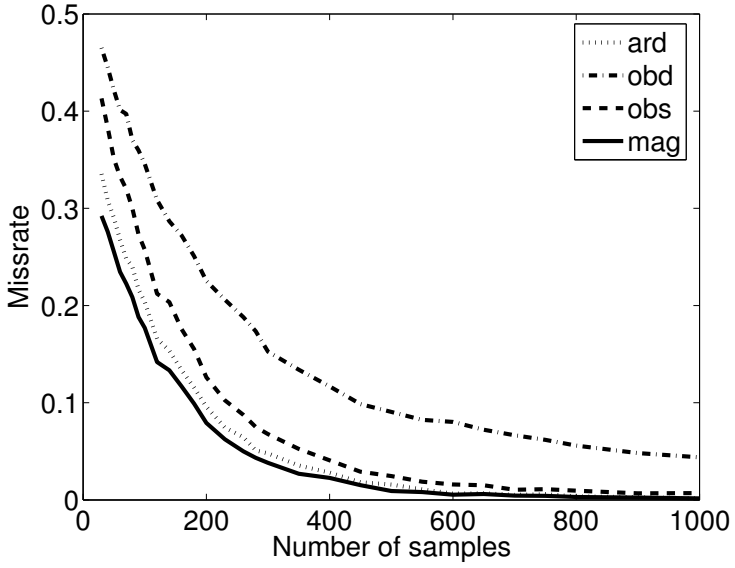
(B)

Fig. 3. Evaluation of standard statistical stop criteria. The 'true' structure is lower triangular as before and has 10 zero parameters out of 25 parameters for the five source model. In figure (A) a single experiment is plotted, each candidate stop criterion locates the correct model. In (B) the miss classification rate of the correct number of zeros over 1000 trials is reported.

magnitude based pruning in combination with the BIC criteria can provide the correct number of zero elements (A) and that they are also in correct position (B) for sufficiently large sample sizes.



(A)



(B)

Fig. 4. In (A) the miss classification rate of finding the *correct number* of zero parameters is reported using BIC with the different pruning ranking schemes. In (B) we report the miss classification rate for the complete process of finding the correct zero parameters using BIC as pruning stop criterion.

5 Auditory scene analysis

In order to demonstrate the viability of pruning based sparsification in audio data we consider a blind signal separation scenario involving again five sources and five observations. The set of sources consists of three speech sources, a 'confounding' music piece and a street noise source. The scene consists of two rooms connected by a hallway. The music and the noise sources are considered stationary in position

corresponding to a stereo and open windows. The speech sources are placed in different locations aimed to mimic movement. The five observations comes from five different microphones placed so that two are located in each room and one in the hallway. The microphones will record from sources in the room where it is placed and from adjacent rooms. This means that the microphones in a given room will record from that room and the hallway but not from the other room. The hallway microphone records everything. Three different placement schemes have been simulated and are illustrated in figure 5.

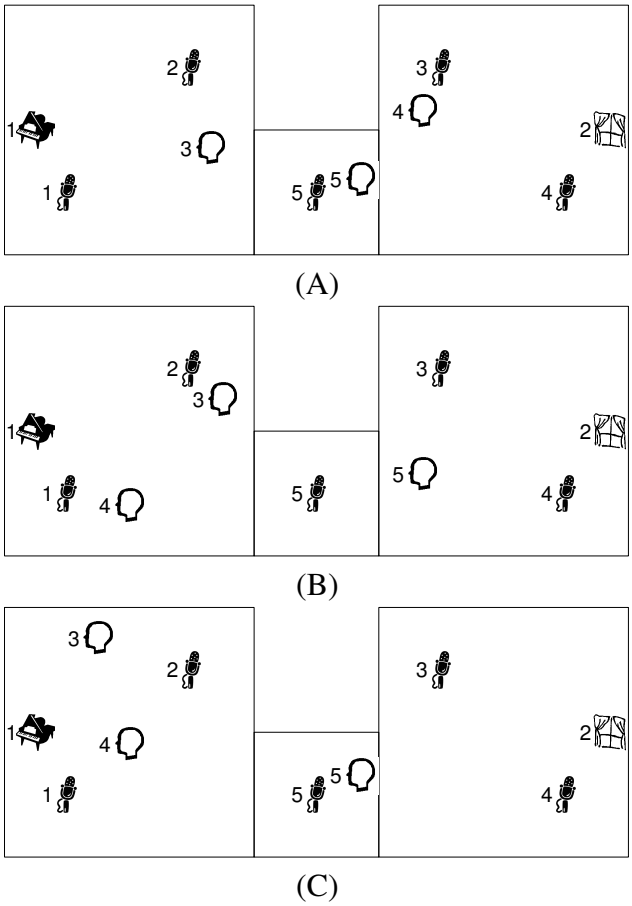


Fig. 5. Auditory scene analysis example. The diagram shows placement of observations and sources in three different situations (A), (B) and (C). Numbers specify the order in the mixing matrix.

Using the setup from above a mixing matrix is generated using the distances between sources and observations. The setup in figure 5 (A) gives a mixing matrix,

$$\mathbf{A}_{(\mathrm{A})} = \begin{bmatrix} 0.89 & 0 & 0.41 & 0 & 0.21 \\ 0.37 & 0 & 0.74 & 0 & 0.30 \\ 0 & 0.32 & 0 & 1.31 & 0.45 \\ 0 & 0.74 & 0 & 0.37 & 0.30 \\ 0.21 & 0.20 & 0.54 & 0.46 & 1.37 \end{bmatrix}$$

As can be seen this is a sparse matrix, and by identifying the zeros it is possible to allocate sources to the different rooms. The sound is divided into segments of fixed length which will assure stationarity of the positions during a given configuration (A,B,C). The ICA algorithm is run and pruning is performed using MB and BIC. Different signal lengths (sample sizes) are considered to compare performance and the resulting miss classification rates are reported in figure 6.

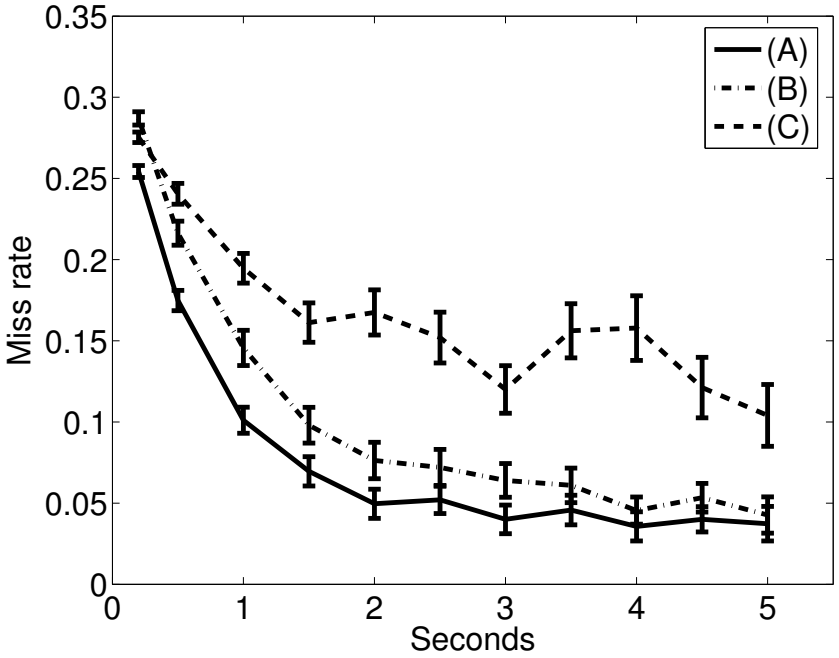


Fig. 6. Auditory scene analysis by sparsified ICA. We report the miss classification rates for the mixing matrix zeros in three different scenarios as a function of the training set size, i.e., at variable window lengths. (A), (B) and (C) refers to the different placement schemes in figure 5.

The performance is as before strongly dependent on sample size. For these real signals this is actually more severe due to non-stationarity. For example, the silent parts of speech signals severely limits performance, clearly dramatically so if they span an entire segment. Furthermore, real signals can be colored and they are not necessarily distributed according to the model assumptions. It can also be seen that the performance is somewhat dependent on the actual placement of the sources. In the first scheme (A) the sources are most evenly placed and this is also the scheme

that gives rise to the best performance. The miss-classification rate increases in the second and is highest in the third scenario (C) which has most sources located in one room.

Despite the higher miss rates, a reasonable performance is achieved for larger samples. If the speakers are not moving too fast it is reasonable to assume stationarity say, within 2-3 seconds then the zero miss classification rate is less than 5 percent in scenarios (A,B).

6 Discussion and conclusion

We have shown that ICA models can be efficiently sparsified by pruning. This is highly relevant for causal models, as well as it can assist interpretation for data analytic applications of ICA.

We have derived expressions for the Hessians of the likelihood function based on both estimation of mixing and separation matrices. We have shown that the sparsity of a generative model can be recovered with high precision at larger sample sizes. We found that simple magnitude based pruning works well, especially for relatively small samples. This is in contrast to published results for neural networks. This discrepancy may be due to the well-known singularities in the ICA likelihood function induced by co-linearities in the mixing matrix.

To determine the optimal degree of sparsity we recommend using the Bayesian information criterion. This approximation works better than both test sets and the nominally higher order Laplace approximation. Test sets are notoriously noisy, while the problems for the Laplace approximation may again be attributed to the limited validity of Taylor expansions.

References

Bell, A., Sejnowski, T., 1995. An information-maximisation approach to blind separation and blind deconvolution. Neural Computation 7 (6), 1129–1159.

Bishop, C. M., 1996. Neural Networks for Pattern Recognition. Oxford University Press, Oxford, UK.

Bronstein, A. M., Bronstein, M. M., Zibulevsky, M., Zeevi, Y. Y., 2005. Sparse ica for blind separation of transmitted and reflected images. International Journal of Imaging Science and Technology (IJIST) 15 (1), 84–91.

Cardoso, J.-F., 1997. Infomax and maximum likelihood for blind source separation. IEEE Signal Processing Letters 4 (4), 112–114.

Gorodkin, J., Hansen, L., Krogh, A., Svarer, C., Winther, O., 1993. A quantita-

tive study of pruning by optimal brain damage. International Journal of Neural Systems 4, 159–169.

Gorodkin, J., Hansen, L. K., Lautrup, B., Solla, S. A., 1997. Universal distribution of saliencies for pruning in layered neural networks. Int. J. Neural Syst. 8 (5-6), 489–498.

Goutte, C., 1996. On the use of a pruning prior for neural networks. pp. 52–61.

Goutte, C., Hansen, L. K., 1997. Regularization with a pruning prior. Neural Networks 10 (6), 1053–1059.

Hansen, L. K., Larsen, J., Kolenda, T., 2001. Blind detection of independent dynamic components. In: IEEE International Conference on Acoustics, Speech, and Signal Processing 2001. Vol. 5. pp. 3197–3200.

Hansen, L. K., Rasmussen, C. E., 1994. Pruning from adaptive regularization. Neural Computation 6 (6), 1222–1231.

Hassibi, B., Stork, D. G., 1993. Second order derivatives for network pruning: Optimal brain surgeon. In: Hanson, S. J., Cowan, J. D., Giles, C. L. (Eds.), Advances in Neural Information Processing Systems. Vol. 5. Morgan Kaufmann, San Mateo, CA, pp. 164–171.

He, Z., Cichocki, A., 2006. K-evd clustering and its applications to sparse component analysis. In: Rosca et al. (2006), pp. 90–97.

Højen-Sørensen, P. A., Winther, O., Hansen, L. K., 2002. Mean field approaches to independent component analysis. Neural Computation 14, 889–918.

Hoyer, P. O., Shimizu, S., Hyvärinen, A., Kano, Y., Kerminen, A. J., 2006. New permutation algorithms for causal discovery using ica. In: Rosca et al. (2006), pp. 115–122.

Hyvärinen, A., 2001. Independent Component Analysis, 1st Edition. Wiley-Interscience.

Kaashoek, J. F., van Dijk, H. K., December 2002. Neural network pruning applied to real exchange rate analysis. Journal of Forecasting 21 (8), 559–77.

Knuth, K. H., 1999. A Bayesian approach to source separation. In: ICA99 Proceedings, Aussois, France, Jan. 1999. pp. 283–288.

Le Cun, Y., Denker, J., Solla, S., 1990. Optimal brain damage. Vol. 2. (Denver 1989), Morgan Kaufmann, San Mateo, pp. 598–605.

MacKay, D. J. C., 1995. Probable networks and plausible predictions - a review of practical Bayesian methods for supervised neural networks. Network: Computation in Neural Systems 6, 469–505.

Park, H.-M., Lee, J.-H., Oh, S.-H., Lee, S.-Y., 2006. Blind deconvolution with sparse priors on the deconvolution filters. In: Rosca et al. (2006), pp. 658–665.

Pedersen, M. S., Larsen, J., Kjems, U., mar 2005. On the difference between updating the mixing matrix and updating the separation matrix. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP'05). Vol. V. Philadelphia, PA, USA, pp. 297–300.

Pedersen, M. W., Hansen, L. K., Larsen, J., 1996. Pruning with generalization based weight saliencies: γOBD, γOBS. In: Touretzky, D. S., Mozer, M. C., Hasselmo, M. E. (Eds.), Advances in Neural Information Processing Systems. Vol. 8. The MIT Press, pp. 521–527.

Ragg, T., Braun, H., Landsberg, H., 1997. A comparative study of neural network optimization techniques. In: Proc. of the ICANNGA 97. Springer-Verlag.

Rosca, J. P., Erdogmus, D., Príncipe, J. C., Haykin, S. (Eds.), 2006. Independent Component Analysis and Blind Signal Separation, 6th International Conference, ICA 2006, Charleston, SC, USA, March 5-8, 2006, Proceedings. Vol. 3889 of Lecture Notes in Computer Science. Springer.

Shimizu, S., Hyvarinen, A., Kano, Y., Hoyer, P. O., 2005. Discovery of non-gaussian linear causal models using ICA. In: Proceedings of the 21th Annual Conference on Uncertainty in Artificial Intelligence (UAI-05). AUAI Press, Arlington, Virginia, pp. 525–553.

Tenenbaum, J. B., Griffiths, T. L., 2000. Structure learning in human causal induction. In: NIPS. pp. 59–65.

Thodberg, H. H., 1991. Improving generalization of neural networks through pruning. Int. J. Neural Syst. 1 (4), 317–326.

Thodberg, H. H., 1996. Review of Bayesian neural networks with an application to near infrared spectroscopy. IEEE Transactions on Neural Networks 7 (1), 56–72.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. Royal. Statist. Soc B. 58 (1), 267–288.

Tipping, M. E., 2001. Sparse bayesian learning and the relevance vector machine. Journal of Machine Learning Research 1, 211–244.

A Gradients and Hessians for ICA

For both saliency based pruning and for the ARD approach we need the Hessian of the negative log-likelihood. We here provide the necessary expressions for parameterization by mixing and separation matrices

$$
\begin{aligned}
M(\mathbf{W}, \boldsymbol{\alpha}) &= -T \log |\det \mathbf{W}| + \sum_t \sum_i \log \cosh(s_i^t) + TL \log \pi + \\
&\quad \frac{1}{2} \sum_{i,j} \alpha_{ij} W_{ij}^2 - \frac{1}{2} \sum_{i,j} \log \alpha_{ij} + \frac{1}{2} L^2 \log 2\pi, \\
\frac{\partial M(\mathbf{W}, \boldsymbol{\alpha})}{\partial W_{i'j'}} &= -T A_{j'i'} + \sum_t x_{j'}^t \tanh s_{i'}^t + \alpha_{i'j'} W_{i'j'}, \\
\frac{\partial^2 M(\mathbf{W}, \boldsymbol{\alpha})}{\partial W_{i'j'} \partial W_{i''j''}} &= T A_{j'i''} A_{j''i'} + \delta_{i',i''} \sum_t x_{j'}^t \left(1 - \tanh^2 s_{i'}^t\right) x_{j''}^t + \\
&\quad \delta_{i',i''} \delta_{j'j''} \alpha_{i'j'}.
\end{aligned} \tag{A.1}
$$

$$
\begin{aligned}
M(\mathbf{A}, \boldsymbol{\alpha}) &= T \log |\det \mathbf{A}| + \sum_t \sum_i \log \cosh(s_i^t) + TL \log \pi + \\
&\quad \frac{1}{2} \sum_{i,j} \alpha_{ij} A_{ij}^2 - \frac{1}{2} \sum_{i,j} \log \alpha_{ij} + \frac{1}{2} L^2 \log 2\pi, \\
\frac{\partial M(\mathbf{A}, \boldsymbol{\alpha})}{\partial A_{i'j'}} &= T W_{j'i'} - \sum_t \sum_i \tanh(s_i^t) W_{ii'} s_{j'}^t + \alpha_{i'j'} W_{i'j'}, \\
\frac{\partial^2 M(\mathbf{A}, \boldsymbol{\alpha})}{\partial A_{i'j'} \partial A_{i''j''}} &= -T W_{j'i''} W_{j''i'} + \sum_t \sum_i \left(1 - \tanh^2 s_i^t\right) W_{ii''} s_{j''}^t W_{ii'} s_{j'}^t + \\
&\quad \sum_t \sum_i \tanh(s_i^t) \left(W_{ii''} W_{j''i'} s_{j'}^t + W_{j'i''} W_{ii'} s_{j''}^t\right) + \\
&\quad \delta_{i'i''} \delta_{j'j''} \alpha_{i''j''}.
\end{aligned} \tag{A.2}
$$