Blind separation of noisy mixtures

Lars Kai Hansen*
Brain Image Analysis Lab (BIAL)
University of California San Diego
La Jolla CA 92093-0949

DRAFT – ver 5.0 – February 1998

Abstract

We analyze blind separation of independent sources in the face of additive noise. The analysis is carried out in a likelihood framework, which allows estimates of the source signals, the mixing matrix and the noise level. In the face of noisy signals, the estimated sources are non-linear functionals of the observed signals, in contrast to the linear noise free case. We also discuss the generalizability of the estimated models and show that an empirical test error estimate may be used to optimize model dimensionality, in particular the optimal number of sources. In the context of a neuroimaging application, we address the "ill-posed" case of short sequences of high dimensional signals driven by a few sources.

1 Introduction

Reconstruction of statistically independent source signals from linear mixtures is relevant to many signal processing contexts [1, 4, 6, 11, 21]. Taking off Principal Component Analysis the problem is often referred to as Independent Component Analysis (ICA)[1].

The source separation problem has a likelihood formulation, see e.g., [5, 22, 23, 14]. The likelihood formulation is attractive for several reasons. First, it allows a principled discussion of the inevitable priors implicit in any separation scheme. The prior distribution of the source signals can take many forms and *factorizes* in the source index expressing the fact that we look for *independent* sources. Secondly, the likelihood approach allows for direct adaptation of the plethora of powerful schemes for parameter optimization, regularization, and evaluation of supervised learning algorithms. Finally, for the case of linear mixtures without noise, the likelihood approach is equivalent to another popular approach based on information maximization [1, 4, 3].

The source separation problem can be analyzed under the assumption that the sources are either independent samples, i.e., "white" sequences, or assuming a more general temporal structure. The separation problem for *autocorrelated* sequences was studied by Molgedey and Schuster [17]. They proposed a source separation scheme based on assumed non-vanishing temporal autocorrelation functions of the independent source sequences evaluated at a specific *time lag*. Their analysis was developed for sources mixed by square, non-singular matrices. Attias and Schreiner derived a likelihood based algorithm for separation of correlated sequences with a frequency domain implementation [2]. Belouchrani and Cardoso

*Permanent Address: CONNECT, Dept. Mathematical Modeling, Technical University of Denmark, DK-2800 Lyngby, DENMARK. email: lkhansen@imm.dtu.dk

[1] There are a number of very useful ICA Web pages providing links to theoretical analyses, implementations and applications see e.g., the page published by Anthony Bell `http://www.cnl.salk.edu/ tony/ica.html`

presented a general likelihood approach allowing for *additive noise* and for non-square mixing matrices. They applied the method to separation of sources taking discrete values [5] estimating the mixing matrix using an Estimate-Maximize (EM) approach with both a determistic and a stochastic formulation. Moulines et al. generalized the EM approach to separation of autocorrelated sequences in presence of noise, and they explored a family of flexible source priors based on Gaussian Mixtures [18]. The difficult problem of noisy, overcomplete source models (more sources than mixture signals) is recently analyzed by Lewicki and Sejnowski within the likelihood framework [13].

In this note we study the likelihood approach focusing on the conceptually simpler white source case. Our aim is to analyze short image sequences as found e.g., in the context of neuroimaging. Hence, we address noisy mixtures and secondly we want to approach highdimensional mixtures where the number of sources is much smaller than the number of sensors (pixels). An additional objective is to find models that *generalize*, i.e., models that perform well on test data.

We give a maximum posterior estimate for the sources which interestingly turns out to be non-linear in the observed signal. The specific model investigated here is a special case of the general framework proposed by Belouchrani and Cardoso[5], however we formulate the parameter estimation problem in terms of the Boltzmann learning rule, which allows for a particular transparent derivation of the mixing matrix estimate.

2 The likelihood

Let the observed mixture signals be denoted X, a matrix of size $D \times N$, where D is the number of sensor channels and N is the number of observations (time points). The noisy mixing model takes the form,

$$X = AS + \Gamma, \tag{1}$$

where S is the source signal matrix (size $M \times N$, M is the number of sources), A is the $D \times M$ mixing matrix, while Γ is a matrix of noise signals with a parameterized distribution.

The properties of the source signals are introduced by a parametrized prior distribution $P(S|\psi)$. The source distribution factorizes in the source index expressing the fact that the sources are independent. In the instantaneous model analyzed by Bell and Sejnowski[4] and by MacKay[14], the distribution of S also factorizes in time, viz., assuming "white" sources.

The likelihood of the parameters of the noise distribution, the parameters of the source distribution and of the mixing matrix is then given by,

$$L(A, \theta, \psi) = P(X|A, \theta, \psi) = \int dS P(X - AS|\theta) P(S|\psi), \tag{2}$$

where $P(.|\theta)$ is the parametrized noise distribution. Note that the source signals are considered "hidden" variables.

For clarity we focus on i.i.d. sources and we will assume that the noise can modeled by i.i.d. Gaussians variables with variance $\theta = \sigma^2$,

$$P(\Gamma|\sigma^2) = \frac{1}{(2\pi\sigma^2)^{PN}} \exp\left(-\frac{1}{2\sigma^2}\mathrm{Tr}_t \Gamma^\top \Gamma\right). \tag{3}$$

Tr_t indicates a trace on the time index.

Finally, we will assume that the source distribution is that studied in [4, 14] (no parameters),

$$P(S) = \frac{1}{\pi^{NM}} \exp\left(-\sum_{t,m} \log\cosh S\right) \tag{4}$$

the summation extending over all sources ($m = 1, ...M$) and times ($t = 1, ..., N$).

3 Estimation of sources for the case of known parameters

Let us first address the problem of estimating the sources if the mixing parameters are known, i.e., for given A, σ^2. We use Bayes formula $P(S|X) \propto P(X|S)P(S)$ to obtain a the posterior distribution of the sources

$$P(S|X, A, \sigma^2) \propto \exp\left(-\frac{1}{2\sigma^2}\mathrm{Tr}_t((X - AS)^\top (X - AS)) - \sum_{t,m} \log\cosh S\right). \tag{5}$$

The *maximum a posteriori* source estimate is found by maximizing this expression w.r.t. S, providing the following non-linear equation to solve for the MAP estimate $\widehat{S}$,

$$-A^\top A\widehat{S} + A^\top X - \sigma^2 \tanh \widehat{S} = 0. \tag{6}$$

There are two problems facing in Eq. (6). First, the equation is non-linear – though only weakly so for low noise levels[2]. Second, the "system matrix", $A^\top A$, may ill-conditioned or even singular. A useful rewriting taking care of potential ill-conditioning of the system matrix is,

$$\widehat{S} = \left(A^\top A + \sigma^2\right)^{-1} \left(A^\top X + \sigma^2 \left(\widehat{S} - \tanh(\widehat{S})\right)\right). \tag{7}$$

This form suggests an approximate solution for low noise levels

$$\begin{aligned} \widehat{S} &= S^0 + \sigma^2 H^{-1}\left(S^0 - \tanh S^0\right) \\ S^0 &= H^{-1}A^\top X, \\ H &= A^\top A + \sigma^2, \end{aligned} \tag{8}$$

exposing the fact that the presence of additive noise turns the otherwise linear separation problem in to a non-linear one. A non-linear source estimate is also found in Lewicki and Sejnowski's analysis of the overcomplete problem [13].

In the numerical experiments reported below we illustrate the difference between the estimate obtained through Eq. (8) and that obtained by the classical ICA solution (equivalent to S^0). We show that the non-linear solution indeed provides an improved estimate of the sources.

4 Joint estimation of sources, mixing matrix, and noise level

Since the likelihood is of the hidden-Gibbs form we can use a generalized Boltzmann learning rule (see [10] and the appendix) to find the gradients of the likelihood of the parameters A, σ^2,

$$\frac{\partial \log L}{\partial A} = \langle \frac{\partial \epsilon}{\partial A} \rangle_{\mathrm{free}} - \langle \frac{\partial \epsilon}{\partial A} \rangle_{\mathrm{clamp}}, \tag{9}$$

$$\frac{\partial \log L}{\partial \sigma^2} = \langle \frac{\partial \epsilon}{\partial \sigma^2} \rangle_{\mathrm{free}} - \langle \frac{\partial \epsilon}{\partial \sigma^2} \rangle_{\mathrm{clamp}}, \tag{10}$$

where ϵ is the "energy function" for the Gibbs distribution,

$$\epsilon = \frac{1}{2\sigma^2}\mathrm{Tr}_t((X - AS)^\top (X - AS)) + \sum_{t,m} \log\cosh S. \tag{11}$$

[2] This expression is the gradient of the exponent of the posterior distribution. A globally convergent iterative solution can be assured if solving by gradient descent $\delta S = -\eta \frac{\partial \epsilon}{\partial S}$, with a sufficiently small η. Here, however, we aim for a fast approximate solution for S

The averages are calculate w.r.t. the joint distribution $P(X,S)$ for the "free" average, and w.r.t. the mixed joint distribution $P_0(X)P(S|X)$ for the "clamped" case, respectively. $P_0(X)$ is the empirical distribution associated with the training set.

The operators to be averaged are given by,

$$\begin{aligned}\frac{\partial\epsilon}{\partial A} &= \frac{1}{\sigma^2}\left(ASS^\top - XS^\top\right) &(12)\\ \frac{\partial\epsilon}{\partial\sigma^2} &= -\frac{1}{2\sigma^4}\mathrm{Tr}_t(X-AS)^\top(X-AS). &(13)\end{aligned}$$

The free average is the derivative of the normalization constant of the joint distribution $P(X,S)$, which in the present case only depends on the noise parameter (σ^2). Hence the free average for the operator conjugate to A vanishes and we need only to calculate expectations, $\langle S\rangle$, $\langle SS^\top\rangle$, for *fixed* observed signals (clamped average). In the general case these expectations are only accessible via simulations. Peterson and Anderson [24] suggested the "Mean Field" approximation (MFA) for the Boltzmann learning problem,

$$\begin{aligned}\langle S\rangle &\approx \widehat{S} &(14)\\ \langle SS^\top\rangle &\approx \widehat{S}\widehat{S}^\top &(15)\end{aligned}$$

Using the MFA we find the estimates $\widehat{A}, \widehat{\sigma^2}$ to be determined by,

$$\begin{aligned}0 &= \frac{1}{\widehat{\sigma^2}}\left(\widehat{A}\widehat{S}\widehat{S}^\top - X\widehat{S}^\top\right) &(16)\\ 0 &= \frac{DN}{\widehat{\sigma^2}} - \frac{1}{2\widehat{\sigma^2}^2}\mathrm{Tr}_t(X-\widehat{A}\widehat{S})^\top(X-\widehat{A}\widehat{S}). &(17)\end{aligned}$$

The approximation in Eq. (17) neglects the covariance contribution which may be crucial if the estimated source matrix becomes ill-conditioned. Hence, we augment the estimate by a regularization constant $\langle SS^\top\rangle \to \widehat{S}\widehat{S}^\top + \beta\mathbf{1}$. β represents a "lumped" effect of the covariance contribution. $\mathbf{1}$ is an $M\times M$ unit matrix. The recursive estimates of A and σ^2 then become,

$$\begin{aligned}\widehat{A} &= X\widehat{S}^\top\left(\widehat{S}\widehat{S}^\top + \beta\mathbf{1}\right)^{-1} &(18)\\ \widehat{\sigma^2} &= \frac{1}{DN}\mathrm{Tr}_t(X-\widehat{A}\widehat{S})^\top(X-\widehat{A}\widehat{S}) &(19)\end{aligned}$$

Fluctuation corrections (hence the magnitude of β) can derived in the low noise limit,based on a Gaussian approximation to the likelihood. The fluctuation expansion to second order (for the neighborhood of the mode $\widehat{S}$) of the energy functional reads $\epsilon \approx \epsilon_0 + \epsilon_1 + \epsilon_2$, with

$$\begin{aligned}\epsilon_0 &= \frac{1}{2\sigma^2}\mathrm{Tr}_t((X-A\widehat{S})^\top(X-A\widehat{S}))\sum_{t,m}\log\cosh\widehat{S} &(20)\\ \epsilon_1 &= -\frac{1}{\sigma^2}\mathrm{Tr}_m(A^\top(X-A\widehat{S})\Delta S^\top) + \sum_{t,m}(\tanh\widehat{S}).*\Delta S &(21)\\ \epsilon_2 &= \frac{1}{2\sigma^2}\mathrm{Tr}_t\Delta S^\top A^\top A\Delta S + \frac{1}{2}\sum_{t,m}(1-\tanh^2\widehat{S}).*\Delta S.*\Delta S, &(22)\end{aligned}$$

where $.*$ is used to indicate element-wise multiplication of matrices, and where $\widehat{S}$ is the estimate found above. The first order term ϵ_1 is zero by definition of $\widehat{S}$.

5 Simulation example

We illustrate first the performance of the proposed ICA scheme on simulated data. The first experiment is designed to study the role of the non-linear source estimate. We study a data matrix X of dimension $D \times N$, with $D = 4$ sensors and $N = 200$ observations. The signal is created as in Eq. (1) with variable noise levels and $M = 3$ sources. The mixing matrix A was chosen randomly with i.i.d. unit variance,

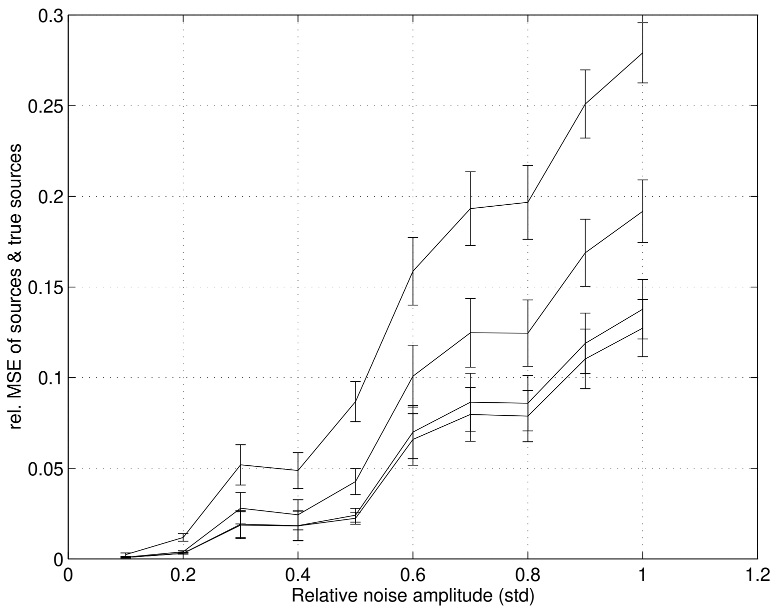

Figure 1: The normalized "Mean Square Error" of the estimator relative to the (assumed) known value of the true source signals as function of noise variance. The error is normalized by the variance of the true source signals. In terms of performance, i.e., from below: S^{30}, S^5, S^1, S^0. We report the average and std. of the average for 50 repetitions.

normal components. The "true" sources S_{true} were created from i.i.d. Gaussian sequences. The two first source sequences are raised to power 2.0 (keeping the sign), while last source was raised to the power 0.5 (keeping the sign). The histograms of the source signals are provided in Figure 2. Next, four different estimates are evaluated using the given instances of A, σ^2. In particular we evaluate the low-noise estimate given in Eq. (8), the "linear" estimate S^0, and finally two estimates obtained by performing $Q = 5$ or $Q = 30$ iterations of the non-linear equation Eq. (7). In Figure 1 we quantify the ability of the solution to estimate the source signals, given that we provide the correct value of both A, σ^2. We find that the (one-time iterated) explicit solution provides a significantly better estimate of the sources than using the linear solution S^0. In Figure 3 a qualitative impression is provided by scatterplots of the various estimators versus the true source signals for an experiment at the highest noise level of Figure 1. In Figure 2 we show the histograms of the true source signals revealing the super-Gaussian nature of the two upper signals, and the sub-Gaussian nature of the third signal.

To illustrate the viability of the A, σ^2 learning algorithm we initialize a mixing matrix as a $D \times M$ random matrix, and the noise variance parameter at 1/100 of the mean signal variance. Ten thousand iterations of Boltzmann learning is then applied to the two parameters A, σ^2 as illustrated in Figures 4-5. The estimates converged and inspection of the estimated sources shows that very close approximations were obtained. The scatterplots in Figure 6 between the three estimated source signals and the three true

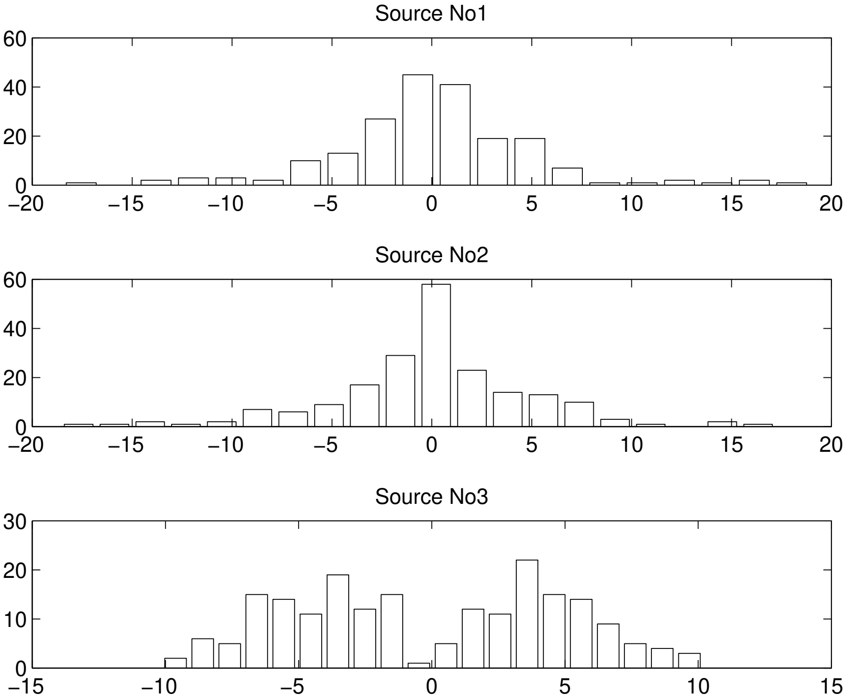

Figure 2: Histograms of the $M = 3$ true source signals used in the numerical experiment. The two upper sources have super-Gaussian distributions, while last source has sub-Gaussian distribution.

source signals show that even for the difficult sub-Gaussian component quite satisfactory separation has been obtained. For reference we show in Figure 7 a similar scatterplot, but now for the three Principal Components that carry the highest variances in a PCA of the same data set. Although these signals are by definition are uncorrelated on the training sample, they provide a rather poor approximation to the separation problem. As a further illustration of the dynamics of the Boltzmann learning process we show an "easier" case with three super-Gaussian sources. These sources are better matched to the model prior. As seen in Figure 11 PCA is capable of finding a reasonable approximation in this case. In Figure 8 we see that initially, the ICA solution is worse (as measured by correlation between estimated and true sources), however from iteration 2000 and on the ICA solution prevails.

6 Generalization and the Bias-Variance Dilemma

The parameters of our blind separation model are estimated from a finite sample and therefore they too are random variables inheriting noise from the data set they were trained on. Hence, like other adaptive schemes the generalization ability is of high interest – how well does the trained model perform on an independent test set? Within the likelihood formulation the generalization error of a specific set of parameters is given by the average negative log-likelihood,

$$\Gamma(A,\theta,\psi) = \int P_*(X)[-\log P(X|A,\theta,\psi)]dX = \int P_*(X)[-\log \int dSP(X-AS|\theta)P(S|\psi)]dX, \qquad (23)$$

cf. Eq. (2). $P_*(X)$ is the true distribution of data. Generalization errors for unsupervised learning schemes –including PCA– have been discussed in[8].

The generalization error is a principled tool for model selection. In the context of blind separation the optimal number of sources retained in the model is of crucial interest. We face a typical bias-variance

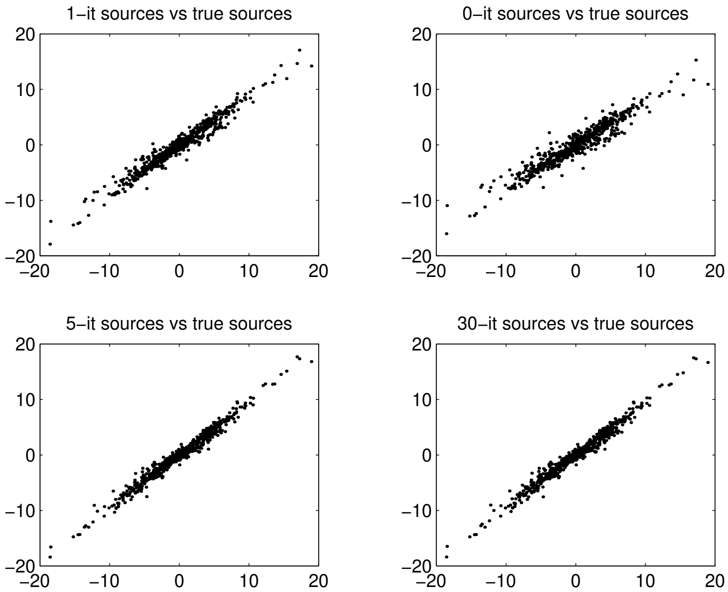

Figure 3: Scatterplots of the four "solutions" (S^1, S^0, S^5, S^{30}, respectively) versus the true sources for the maximal noise level used in Figure 1.

dilemma [7, 20]. If too few components are used, a structured part of the signal will be lumped with the noise, hence leading to a high generalization error because of "lack of fit". On the other hand, if too many sources are used we expect "overfit" since the model will use the additional degrees of freedom to fit non-generic details in the training data.

To illustrate generalizability in the context of ICA we generate a simulated data set along the lines above now with $D = 8$ and a "true" number of super-Gaussian sources $M_{\text{true}} = 3$. Models with $M = 2, ..., 7$ sources were adapted and the generalization error was estimated on a test set of another $N_{\text{test}} = 1000$ point generated using the same procedure as we used for the training set. The variation of the generalization error as model dimension was increased is shown in figure 12. As expected we find a bias-variance trade-off with excess generalization error for large and small models, while the optimal model size is $M = M_{\text{true}}$. In ten repetitions of the experiment nine showed the best generalization performance for $M = M_{\text{true}} = 3$ while one experiment had minimal generalization error at $M = 4$. In general we expect a distribution of "optimal models" which is peaked at the "true" model dimension and with a width that decreaces with increasing sample and test set sizes.

7 Application to neuroimaging

Principal Component Analysis is a popular tool for analysis of short image sequences. The use of PCA in functional neuroimaging, has been pioneered by Moeller and Strother in the so-called SSM model[16]. Independent Component Analysis has been pursued in this context by McKeown et al. [15]. In [15] the fMRI signals were analyzed for independent *spatial patterns*. Here we use our ICA scheme in the form discussed above, looking for independent temporal patterns, and we compare the independent components and their spatial patterns with those obtained by PCA.

The basic tool for PCA is Singular Value Decomposition (SVD). In the SVD of a short image sequence ($N < M$), X is decomposed,

$$X = U\Lambda V^\top, \tag{24}$$

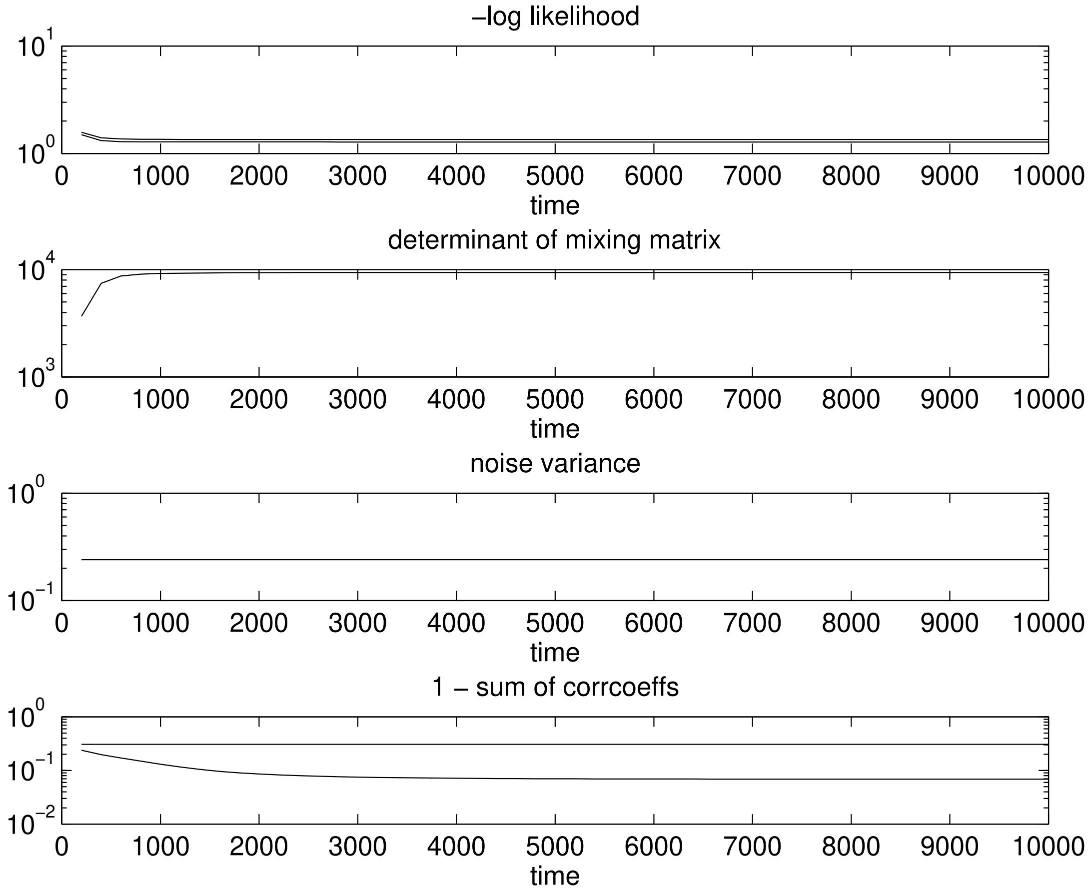

Figure 4: Convergence of Boltzmann Learning for a mixing problem based on the sources of Figure 3. The first panel shows the evolution of the negative log-likelihood on the training sequence and on a test sequence. The second panel shows the evolution of the determinant of the system matrix $A^\top A$. The third panel shows the rapid convergence of the noise variance estimate, while the fourth panel shows the temporal development of the average correlation coefficients between the true sources and the estimated sources (we show $1 - \frac{1}{3}\sum \mathrm{corrcoef}$). For reference we show the average correlation coefficients for the sources estimated by Principal Component Analysis of the training sequence (horizontal line).

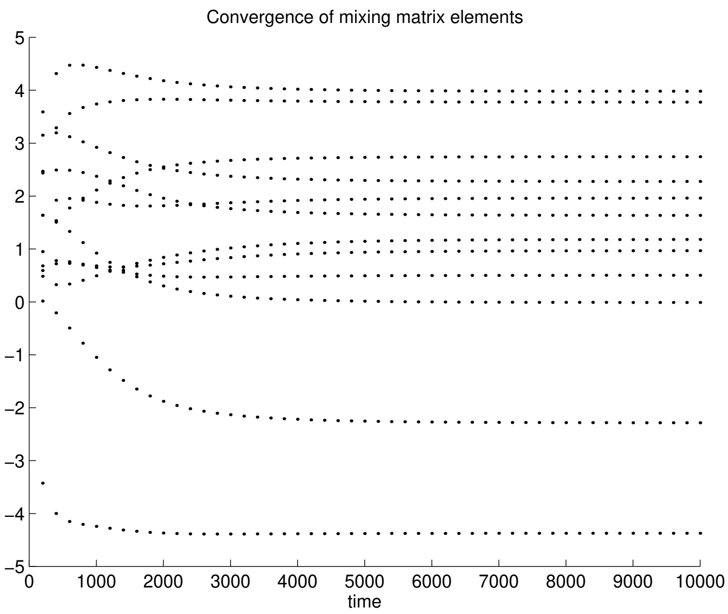

Figure 5: Temporal evolution of the $D \times M = 12$ matrix elements of the estimated mixing matrix.

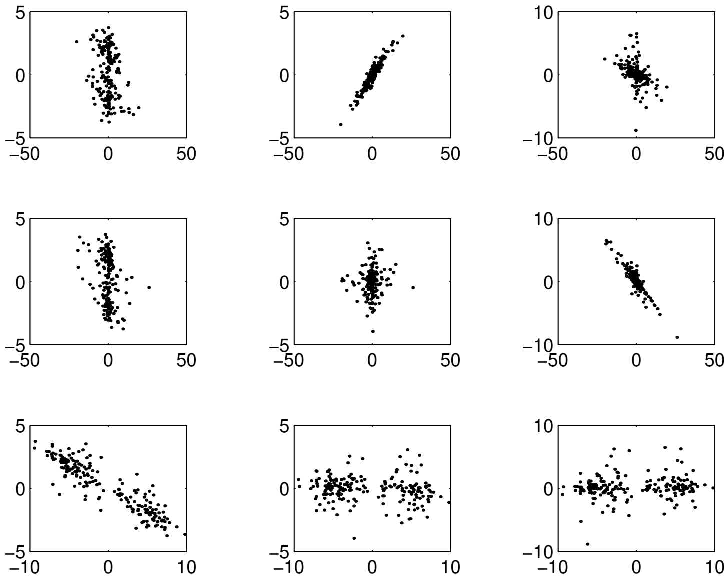

Figure 6: Scatterplots of the ICA estimated sources versus the true sources at the end of the estimation process.

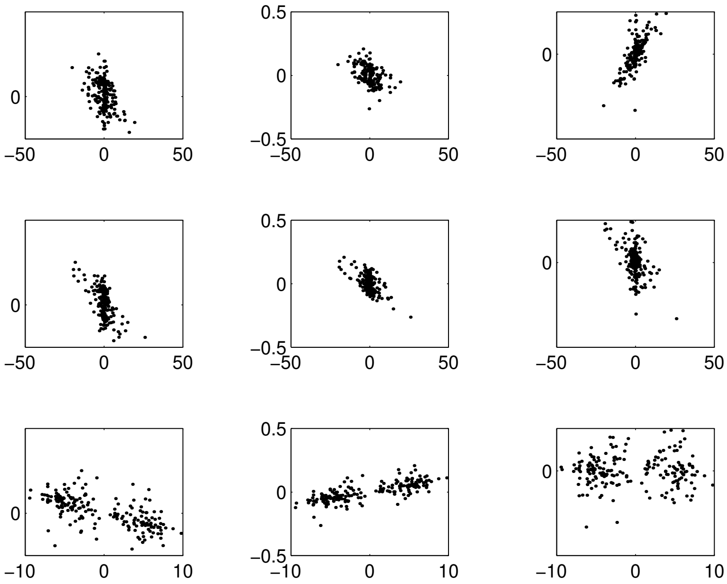

Figure 7: For reference we show the scatterplots of the PCA estimated sources versus the true sources.

where U is a $M \times N$ matrix, while Λ, V are $N \times N$ matrices. Λ is a diagonal matrix with non-negative ordered elements (the largest first)– the singular values. U are eigenvectors corresponding to the N non-zero eigenvalues of $XX^\top$, while V is an orthogonal matrix os the eigenvectors of $X^\top X$,

$$XX^\top = U\Lambda V^\top V\Lambda U^\top = U\Lambda^2 U^\top \tag{25}$$

$$X^\top X = V\Lambda U^\top U^\top \Lambda V^\top = V\Lambda^2 V^\top, \tag{26}$$

hence Λ holds the square roots of the corresponding eigenvalues. Viewed as temporal sequences, $\Lambda V^\top$ are "uncorrelated source signals". The associated spatial patterns of course are subject to an orthogonality constraint being eigenvectors to a symmetric real matrix. However, as above we are often interested in a slightly more general separation of image sources that are independent in time, but not necessarily orthogonal in space, i.e., we would like to be able to perform a more general decomposition of the data matrix, corresponding to the model in Eq. (1). For short image sequences we can can make use of the SVD for simplification of the ICA problem. The approach taken here is similar to the so-called "Cure for Extremely Ill-posed Learning" [12] used to simplify supervised learning in short image sequences.

We first note that the likelihood, considered a function of the columns of A, (spatial patterns) can be split in two parts. A part A_1 orthogonal to the subspace spanned by the N rows of X, and a part A_2 situated in the subspace spanned by the N columns of X. The first is part trivially minimized for any non-zero configuration of sources by putting $A_1 = 0$. It simply does not "couple" to data. The remaining part A_2 can be written as UB in terms of an $N \times N$ matrix B, simplifying the mixing model for the short sequence problem to,

$$Y \equiv U^\top X = BS + \Gamma, \tag{27}$$

with Y, Γ being $N \times N$ matrices. We note that it often may be possible to further limit the dimensionality of the SVD subspace retained in Eq. (27), hence further reducing the "sensor" dimensionality, D, of the remaining problem.

We use the above scheme for the analysis of a functional Magnetic Resonance Imaging sequence, acquired by Dr. Egill Rostrup, Hvidovre Hospital, Denmark, as part of a study involving visual stimulation.

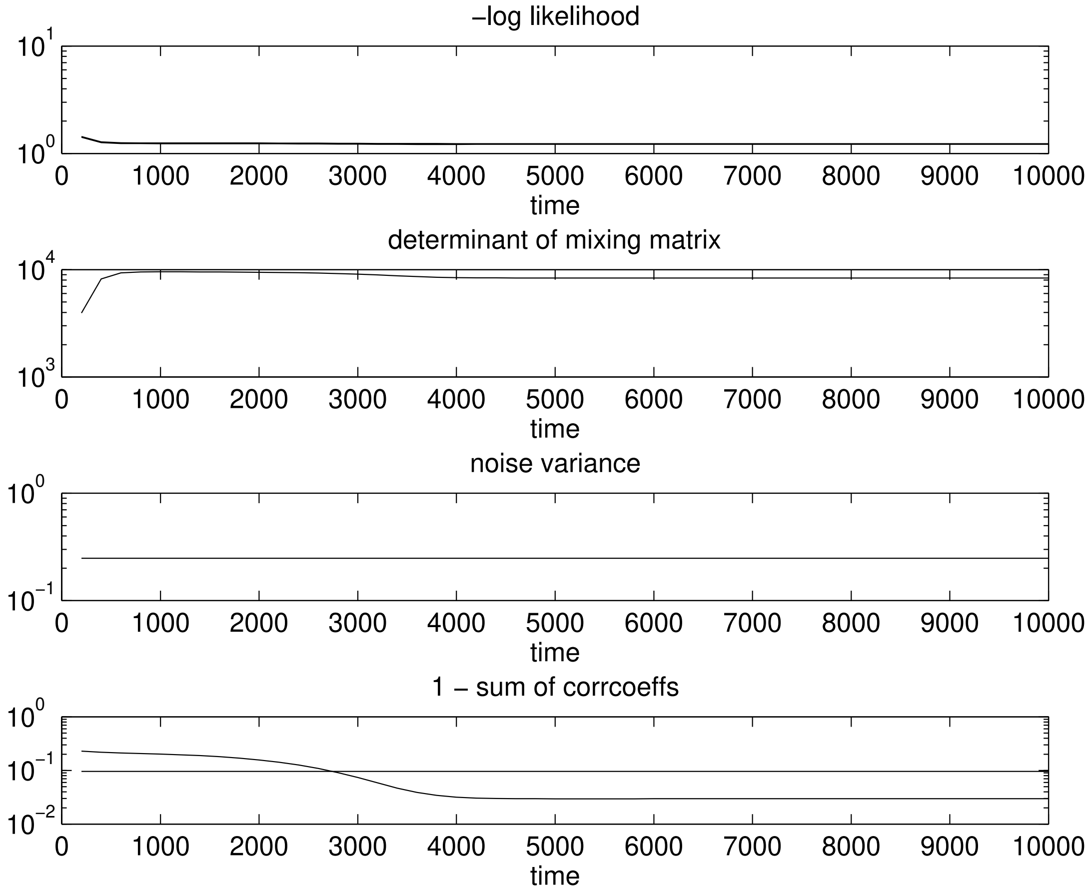

Figure 8: As Figure 4: Convergence of Boltzmann Learning for an "easier" mixing problem based on three super-Gaussian sources. The first panel shows the evolution of the negative log-likelihood on the training sequence and on a test sequence. The second panel shows the evolution of the determinant of the system matrix $A^{\top}A$. The third panel shows the quick convergence of the noise variance estimate, while the fourth panel shows the temporal development of the correlation coefficients between the true sources and the estimated sources. For reference we also show the same expression for the sources estimated by Principal Component Analysis of the training sequence (horizontal line).

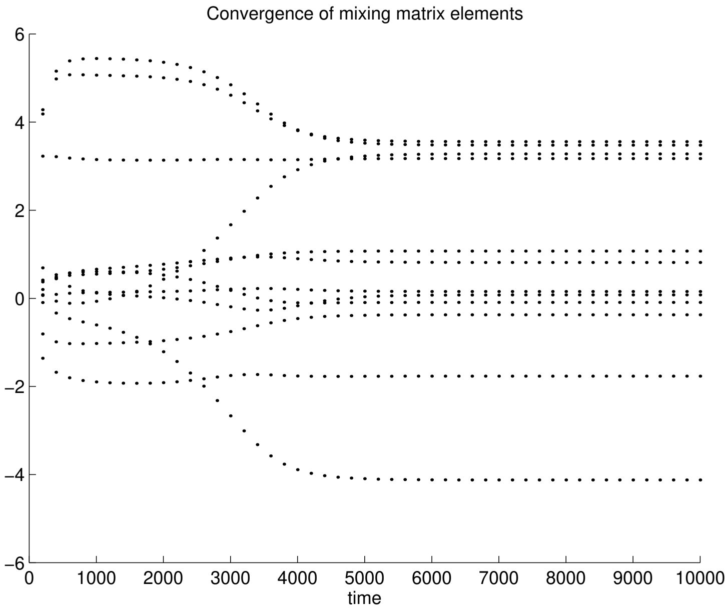

Figure 9: As Figure 5, however for three super-Gaussian sources: Temporal evolution of the $D \times M = 12$ matrix elements of the estimated mixing matrix.

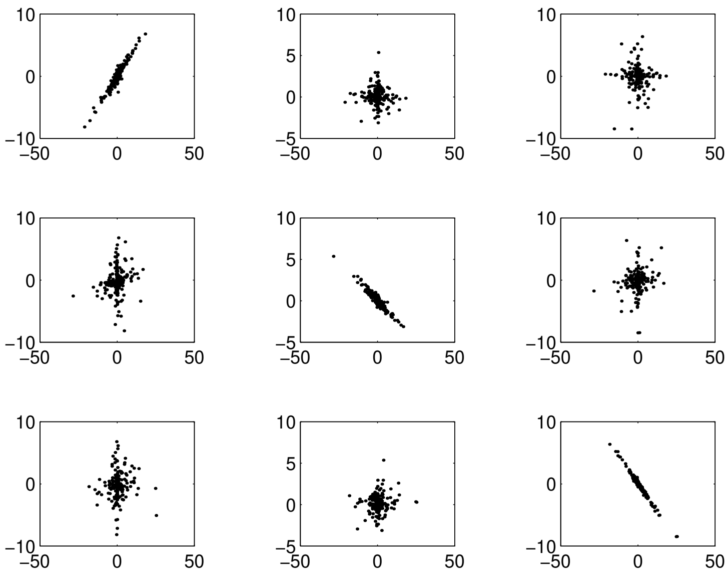

Figure 10: As Figure 6, now for three super-Gaussian sources: Scatterplots of the ICA estimated sources versus the true sources at the end of the estimation process.

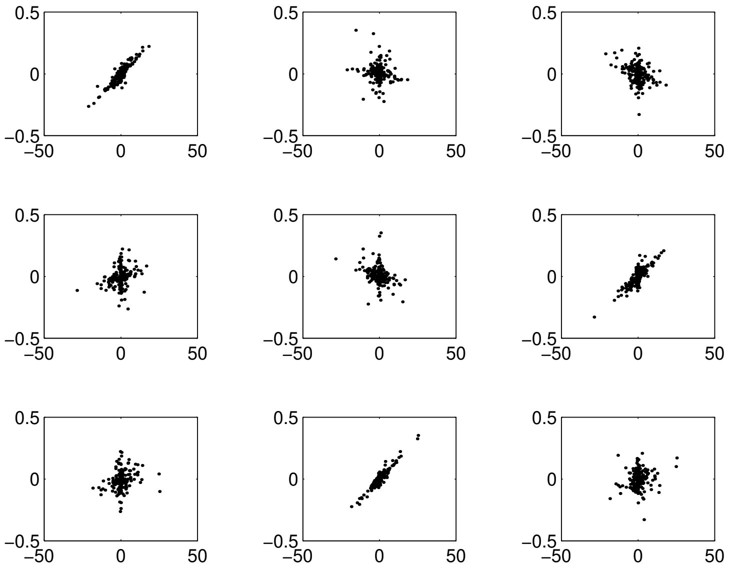

Figure 11: As Figure 7, however for three super-Gaussian sources: For reference we show the scatterplots of the PCA estimated sources versus the true sources.

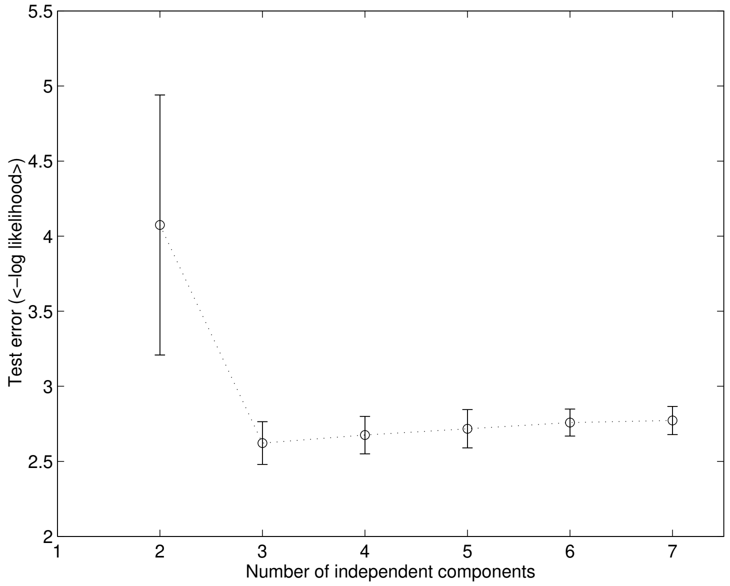

Figure 12: Bias-variance tradeoff for blind separation as model dimensionality is varied. The simulated data set was generated with $M_{\mathrm{true}} = 3$ sources. Smaller or larger models have excess generalization error because of "lack-of-fit" and "overfit" respectively.

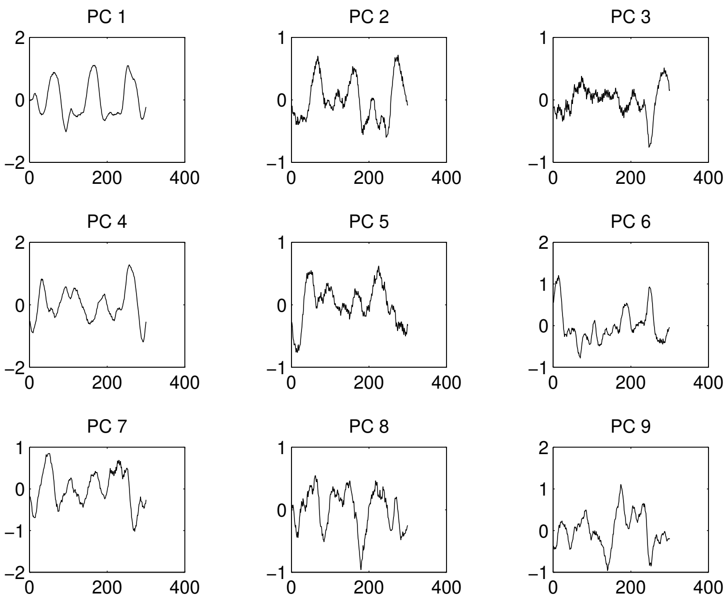

Figure 13: Time courses of the Principal Components corresponding to the nine largest covariance eigenvalues for a sequence of 300 fMRI scans of a visually stimulated subject. Stimulation takes places at scan times $\tau = 25-75$, $\tau = 125-175$, and $\tau = 225-275$ with a time interval between scans of $TR = 333msec$. The time courses have been smoothened for presentation, eliminating noise and high-frequency physiological components. Note that most of the response is captured by the first Principal Components, showing a strong response to all three periods of stimulation.

A single slice, holding 91×71 pixels through visual cortex was acquired with a short time interval between successive scans of TR = 333 msec. Visual stimulation in the form of a flashing annular checkerboard pattern was interleaved with periods of fixation. A "run" consisting of 25 scans of rest, 50 scans of stimulation, and 25 scans of rest was repeated 10 times. For this analysis a contiguous mask was created with 2440 pixels holding the essential parts of the slice including the visual cortex. SVD was performed on a subset of three runs ($N = 300$) providing matrices $U \sim 2440 \times 300$, $\Lambda \sim 300 \times 300$ and $V \sim 300 \times 300$.

In figures 13-14 we show the time course of the nine most significant Principal Components and their associated eigenimages. A distinct activation response is picked up in the first component, and the thresholded eigenimage associated with it is dominated by a cluster of pixel in the visual cortex area as expected. The Principal Components 3,4 show evidence of "non-stationarity" pointing to the third run ($t = 201-300$) being different from the two first runs, however the spatial pattern does not provide much clue to the nature of these differences. In figures 15-16 we show the time courses and spatial patterns of the ICA solution. The differentiation of the first two runs from the third run is very prominent in the estimated source signals. Components 3,7, and 8 are activated in the two first runs, while components 1,5, and 9 are activated in the third run. Comparing the spatial patterns of component 1 (active in the two first runs) and component 8 (active in the third run) we find that there are small areas anterior to the main activation hotspot in visual cortex which are correlated with the activity of the visual cortex (both are found in the positive excursion set) for the two first runs, while anti-correlated in the third run (the main area is in the positive excursion set, while the smaller regions are now colored blue, hence belong to the negative excursion set). Such subtle differentiation of the temporal structure of the fMRI signal is not seen in PCA, presumably due the additional spatial orthogonality constraint.

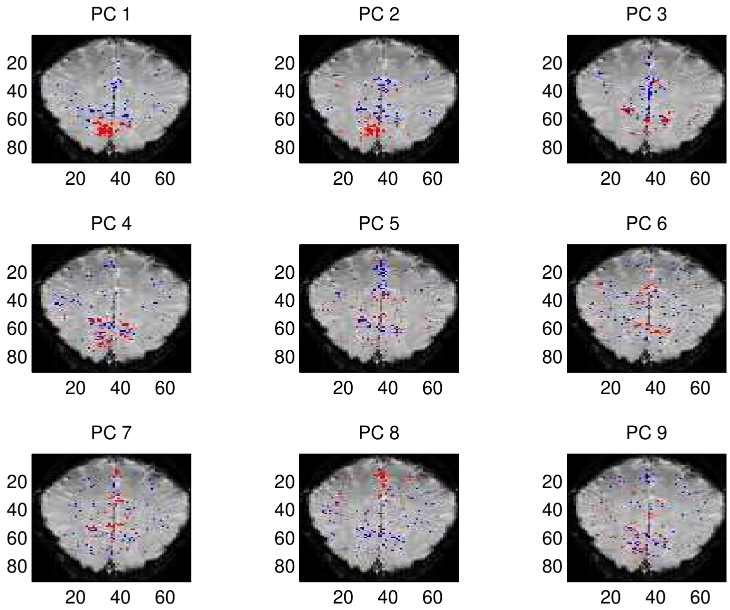

Figure 14: Covariance eigenimages corresponding nine most significant Principal Components. The eigenimage corresponding to the dominating first PC is, as expected, focused in the region of the primary visual cortex. The image is formed by thresholding the eigenimages, so that the 3% fractiles (positive: red, negative: blue) are shown in the context of the averaged fMRI image.

8 Conclusion

A likelihood-MAP formulation of the noisy separation problem has been given. In the face of additive noise, the MAP estimate of the independent sources is a nonlinear functional of the observed signal, in contrast with the linear "classical" solution. A numerical experiment showed that the non-linear MAP estimate indeed provides a closer approximation to the true source signals in the presence of additive white noise.

As the signal model distribution takes the form of a hidden variable-Gibbs distribution, adaption of the mixing matrix and the noise parameters can be derived using the Boltzmann learning rule. Experiments on simulated data showed that the learning rule was capable of finding both the mixing matrix and an estimate of the noise level.

In real-world applications of adaptive systems *generalizability* is of key importance. Generalizability depends on the sample size and on model complexity. A relevant complexity parameter for ICA based on the simple prior distribution in Eq. 4 is the number of sources. We proposed the generalization error – defined as the average negative log-likelihood and estimated on an independent test set– as a means for optimization of source model complexity. An experiment on simulated data showed that the estimated generalization error was able to identify the correct source dimensionality.

In an application to short image sequences we showed how one may simplify the ill-posed modeling problem considerably by use of the SVD. Furthermore, it was shown that the ICA was able to differentiate the three repetitions of the experiment, using its more flexible definition of spatial patterns. The corresponding PCA was not able to capture this apparent non-stationary aspect of the experiment.

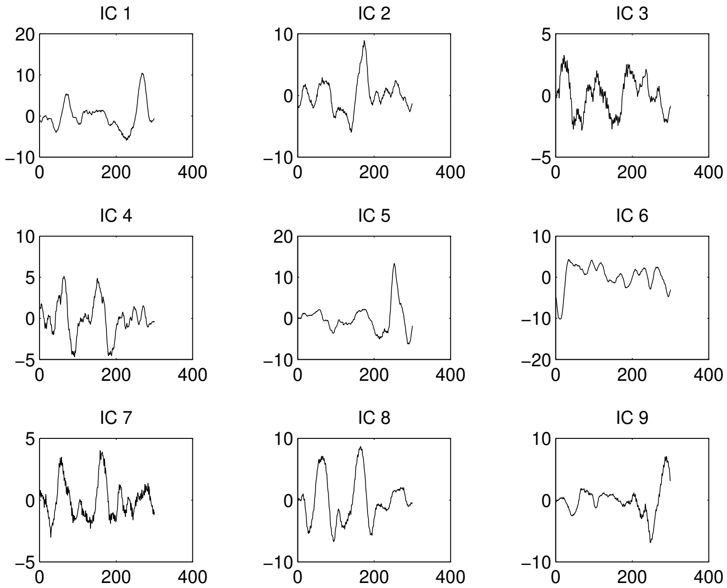

Figure 15: Time courses of the Independent Components of an analysis based on nine sources. Stimulation takes places at scan times $\tau = 25 - 75$, $\tau = 125 - 175$, and $\tau = 225 - 275$ with a time interval between scans of $TR = 333msec$. The time courses have been smoothened for presentation, eliminating noise and high-frequency physiological components. In contrast to the PCA results in figure 12, we find here that the response is distributed across several sources. IC's number 3,7 and 8, have picked up most of the response for the two first runs, while IC's 1,5 and 9, are dominated by response in the last of the three runs. This more detailed decomposition is possible because the spatial patterns are not subject to the orthogonality constraint. Hence, while the spatial patterns corresponding to the third runs are quite similar to those associated with the first two the ICA solution has differentiated them spatially.

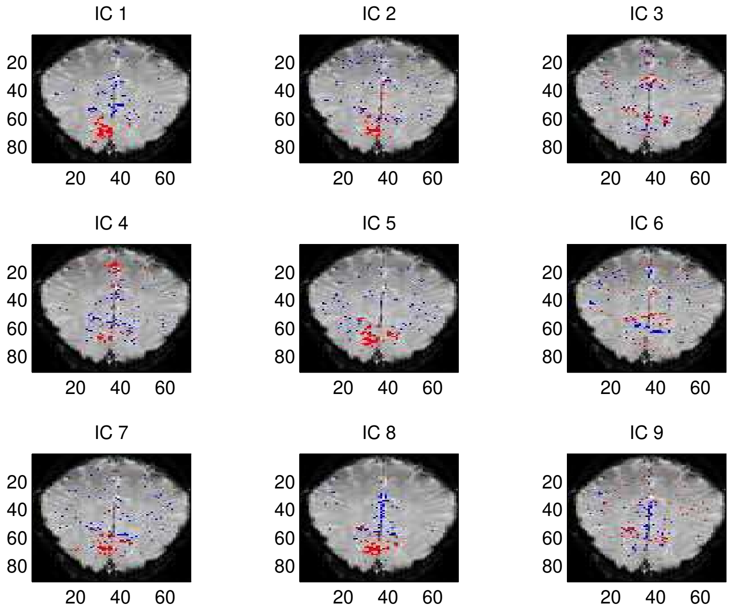

Figure 16: Spatial patterns corresponding to the Independent Components of figure 14. We note the similarities of the patterns corresponding components 3,7 and 8 (holding the response to the first two runs) and components 1,5, and 9 (holding the response for the third and last run). The ICA algorithm can separate temporal structures that are associated with similar, hence overlapping, but different spatial systems. The image is formed by thresholding the eigenimages, so that the 3% fractiles (positive: red, negative: blue) are shown in the context of the averaged fMRI image.

9 Acknowledgment

I thank Thomas Kolenda, Jan Larsen, Te-Won Lee, Michael Lewicki, Martin McKeown, Niels Mørch and Terry Sejnowski for valuable comments and discussions. I thank Egill Rostrup for access to the fMRI data set. The author was supported in part by a McDonnell-Pew Center travel grant. This research was carried out while the author was visiting the Brain Image Analysis Lab (BIAL). I wish to thank Terry Jernigan and the BIAL staff and researchers for creating an enjoyable atmosphere for the visit.

Appendix: The Generalized Boltzmann Learning Rule

The Boltzmann Learning rule was the first learning rule for neural networks with hidden units and was formulated for binary stochastic units [10]. The main virtue from a modern practical point is that it allows learning in Gibbs distributions with hidden variables *without computing the normalization integral*. Here I provide the general gradient calculation. A systematic 2.order (Newton) scheme is also available, but is not necesarry for this context, since we can solve the stationarity equations directly. The generalized Boltzmann rule was earlier used in [9] for parameter estimation in Random Markov Fields.

Let $T = \{x_n | n = 1, ..., N\}$ be a sample of a stochastic variable with distribution $P_*(x)$. Let our model distribution be of the hidden variable-Gibbs form,

$$P(x|\theta) = \int dSP(x, S|\theta), \tag{28}$$

with

$$\begin{aligned} P(x, S|\theta) &= Z^{-1} \exp\left(-\epsilon(x, S|\theta)\right) && (29) \\ Z &= \int dx dS \exp\left(-\epsilon(x, S|\theta)\right). && (30) \end{aligned}$$

The objective function for the parameters θ is the negative log-likelihood function,

$$C = \int dx P_0(x)[-\log P(x|\theta)], \tag{31}$$

where the distribution $P_0(x)$ can be either $P_*(x)$, i.e., C is the generalization error, or

$$P_0(x) = P_{\mathrm{emp}}(x) = \frac{1}{N} \sum_{n=1}^{N} \delta(x - x_n) \tag{32}$$

in which case C is the training error. The objective function can be augmented with a log-prior term leading to a penalized maximum likelihood estimation.

We estimate parameters by minimization of C, hence we look for the gradients $\frac{\partial C}{\partial \theta}$. Computing the derivative we face two separate terms,

$$\frac{\partial C}{\partial \theta} = -\frac{\partial \log Z}{\partial \theta} + \frac{\partial}{\partial \theta} \int dS \exp\left(-\epsilon(x, S|\theta)\right). \tag{33}$$

The first derivative on the left hand side of Eq. (33) is

$$\begin{aligned} -\frac{\partial \log Z}{\partial \theta} &= -\frac{1}{Z} \int dx dS \frac{\partial}{\partial \theta} e^{-\epsilon} && (34) \\ &= \int dx dS \left(\frac{\partial \epsilon}{\partial \theta}\right) P(x, S|\theta) && (35) \\ &\equiv \langle \left(\frac{\partial \epsilon}{\partial \theta}\right) \rangle_{\mathrm{free}} && (36) \end{aligned}$$

The second term on the left hand side of Eq. (33) is

$$\int dx P_0(x) \frac{\partial}{\partial\theta} \log \int dS e^{-\epsilon} = \int dx P_0(x) \frac{1}{ZP(x)} \int dS \frac{\partial}{\partial\theta} e^{-\epsilon} \quad (37)$$

$$= -\int dx P_0(x) \int dS \left(\frac{\partial\epsilon}{\partial\theta}\right) \frac{P(x,S|\theta)}{P(x)} \quad (38)$$

$$= -\int dx P_0(x) \int dS \left(\frac{\partial\epsilon}{\partial\theta}\right) P(x|S,\theta) \quad (39)$$

$$-\equiv \langle\left(\frac{\partial\epsilon}{\partial\theta}\right)\rangle_{\text{clamped}} \quad (40)$$

References

[1] S. Amari, A. Cichocki and H.H. Yang: *A New Learning Algorithm for Blind Signal Separation.* In Advances in Neural Information Processing Systems 8, Eds. D. Touretzky, M. Mozer, and M. Hasselmo, MIT Press, Cambridge MA, 757-63 (1996).

[2] H. Attias and C.E. Schreiner: *Blind Source Separation and Deconvolution by Dynamic Component Analysis.* In Neural Networks for Signal Processing VII: Proceedings of the 1997 IEEE Workshop (1997).

[3] T.-W. Lee, M. Girolami, A.J. Bell and T.J. Sejnowski: *A unifying Information-theoretic framework for Independent Component Analysis.* International Journal on Mathematical and Computer Modeling, in press (1998)

[4] A. Bell & T.J. Sejnowski: *An Information-Maximization Approach to Blind Separation and Blind Deconvolution* Neural Computation **7**, 1129-59 (1995).

[5] A. Belouchrani and J.-F. Cardoso. *Maximum likelihood source separation by the expectation-maximization technique: deterministic and stochastic implementation* In Proc. NOLTA, 49-53 (1995).

[6] P. Comon: *Independent Component Analysis: A New Concept?* Signal Processing, **36**, 287-314 (1994).

[7] S. Geman, E. Bienenstock, and R. Doursat *Neural Networks and the Bias/Variance Dilemma,* Neural Computation, **4**, 1-58 (1992).

[8] L.K. Hansen and J. Larsen *Unsupervised Learning and Generalization.* In Proceedings of the IEEE International Conference on Neural Networks 1996, Washington DC, vol. 1, 25-30 (1996).

[9] L.K. Hansen, L. Nonboe Andersen, U. Kjems, J. Larsen *Revisiting Boltzmann Learning: Parameter Estimation in Markov Random Fields.* In Proceedings of International Conference on Acoustics Speech and Signal Processing, Atlanta, GA, vol. 6, 3395-98 (1996).

[10] G.E. Hinton and T.J. Sejnowski: *Optimal Perceptual Inference* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Washington 1983, 448, (1983).
G.E. Hinton and T.J. Sejnowski: *Learning and Relearning in Boltzmann Machines* In D.E. Rumelhart and J.L. McClelland, Eds. "Parallel Distributed Processing: Explorations in the Microstructure of Cognition" , MIT Press Cambridge Vol. 1, 282 (1986).

[11] C. Jutten and J. Herault: *Blind Separation of Sources: An Adaptive Algorithm Based on Neuromimetic Architecture* Signal Processing **24**, 1-10 (1991).

[12] B. Lautrup, L.K. Hansen I. Law, N. Mørch, C. Svarer, S.C. Strother: *Massive weight sharing: A cure for extremely ill-posed problems.* In H.J. Hermanet al., eds. Supercomputing in Brain Research: ¿From Tomography to Neural Networks. World Scientific Pub. Corp. 137-148 (1995).

[13] M.S. Lewicki and T.J. Sejnowski: *Learning overcomplete representations* Preprint "Draft 2.1" (1998).

[14] D. MacKay: *Maximum Likelihood and Covariant Algorithms for Independent Components Analysis.* "Draft 3.7" (1996).

[15] M.J. McKeown, T.P. Jung, S. Makeig, G. Brown, S.S. Kindermann, T.-W. Lee and T.J. Sejnowski: *Spatially independent activity patterns in functional magnetic resonance imaging data during the stroop color-naming task.* Proc.Natl.Acad.Sci. USA. In press (1998)

[16] J.R. Moeller and S.C. Strother: *A regional covariance approach to the analysis of functional patterns in positron emission tomographic data* J. Cereb. Blood Flow Metab. **11**, A121-A135 (1991). S.C. Strother, A.R. Anderson, K.A. Schaper, J.J. Sidtis, R.P. Woods. D.A. Rottenberg: *Principal Component Analysis and the Scaled Subprofile Model compared to Intersubject Averaging a Statistical Parametric Mapping: I. "Functional Connectivity" of the Human Motor System Studied with [^{15}O]PET.* J. Cereb. Blood Flow Metab. **15**, 738-75 (1995).

[17] L. Molgedey & H. Schuster: *Separation of Independent Signals using Time-Delayed Correlations* Physical Review Letters **72** 3634-37 (1994).

[18] E. Moulines, J.-F. Cardoso, E. Gassiat. *Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models* Proc. ICASSP'97 Munich, vol. 5, pp. 3617-20 (1997).

[19] N, Mørch, U. Kjems, L.K. Hansen, C. Svarer, I. Law, B. Lautrup, S.C. Strother, and K. Rehm. *Visualization of Neural Networks Using Saliency Maps.* In Proceedings of 1995 IEEE International Conference on Neural Networks, 2085-90 (1995).

[20] N. Mørch, L.K. Hansen, S.C. Strother,C. Svarer, D.A. Rottenberg, B. Lautrup, R. Savoy, O.B. Paulson: *Nonlinear versus Linear Models in Functional Neuroimaging: Learning Curves and Generalization Crossover.* In Proceedings of the 15th International Conference on Information Processing in Medical Imaging (1997).

[21] E. Oja: *PCA, ICA, and nonlinear Hebbian learning.* In Proc.Int.Conf. on Artificial Neural Networks ICANN-95, 89-94 (1995).

[22] B.A. Olshausen: *Learning linear, sparse, factorial codes* A.I. Memo 1580, Massachusetts Institute of Technology (1996).

[23] B.A. Pearlmutter and L.C. Parra: *A context-sensitive generalization of ICA.* In Proc. 1996 International Conference on Neural Information Processing. Hong Kong (1996).

[24] C. Peterson & J.R. Anderson: *A Mean Field Theory Learning Algorithm for Neural Networks* Complex Systems **1**, 995-1019 (1987).