Bayesian Independent Component Analysis as Applied to One-Channel Speech Enhancement

Ilyas Potamitis, Nikos Fakotakis, George Kokkinakis,

Wire Communications Lab., Electrical & Computer Engineering Dept.,
University of Patras, Rion-26 500, Patras, Greece
potamitis@wcl.ee.upatras.gr

Abstract. Our work applies a unifying Bayesian-Independent Component Analysis (BICA) framework in the context of speech enhancement and robust Automatic Speech Recognition (ASR). The corrupted speech waveform is reshaped in overlapping speech frames, and is assumed to be composed as a linear sum of the underlying clean speech and noise. Subsequently, a linear sum of latent independent functions is proposed to span each clean frame. Two different techniques are applied following a Bayesian formulation: In the first case the posterior probability of a clean speech frame is formed conditioned on the noisy one on which a maximum *a posteriori* (MAP) approach is applied, leading to Sparse Code Shrinkage (SCS) - a fairly new statistical technique originally presented to applied mathematics and image denoising, but its much promising potential for speech enhancement has not yet been exploited. In the second case, viewed within the Variational Bayes framework, the model for noisy speech generation is stated in a block-based fashion as a noisy, blind source separation problem from which we infer the independent basis functions that span the space of a speech frame and their mixing matrix, thus reconstructing directly the corresponding clean frames.

1 Introduction

The primary objective of noise compensation methods as applied in the context of speech processing is to reduce the effect of any signal which is alien to and disruptive of the message conveyed among participants in a communicative event (whether humans or ASR machines). Depending on the application, speech enhancement methods aim at improving the quality of perception and/or robust speech recognition.

The subtractive and the attenuating type of filters are well-established, one-channel noise compensation methods that predominate in speech enhancement literature [1]. Namely: a) spectral subtraction b) least mean square, adaptive filtering c) filter-based parametric approaches d) model-based, short-time spectral magnitude estimation, e) Hidden Markov Model (HMM)-based speech enhancement techniques. In what follows we foreground the main presuppositions of these techniques from which the BICA framework departs:

Most speech enhancement algorithms are based on a transformation, which facilitates the estimation of the clean speech model parameters such as Discrete

Fourier Transform, Discrete Cosine Transform, Karhunen-Loève Transform. The transformation itself is more or less determined on an ad hoc basis. BICA is based on a data driven transformation kernel adapted to the structure of speech data.

Most methods focus on the distorted short-time amplitude of the speech signal leaving the phase unprocessed. In the BICA framework the speech signal is processed uniformly, meaning that there is no inherent need that compels us to concentrate on amplitude and ignore phase processing.

The noise reduction process in the subtractive types of algorithms introduces a trade-off between distortion of spectral balance of the processed speech signal and noise suppression factor. The BICA framework is not a subtractive technique and musical noise is not perceivable.

Spectral Subtraction and some versions of Wiener/Kalman and HMM-based algorithms require an accurate estimate of corrupting noise statistics. This is acquired during speech pauses through the use of a Voice Activity Detector (VAD). The construction of a robust VAD at low SNRs is a task still open to research. The framework of BICA has no need of a VAD to estimate speech-presence.

A linear modelling approach allows the corrupted speech frame to be decomposed in a linear sum of the underlying speech waveform and noise. Subsequently, the speech waveform is assumed to be composed of a linear sum of independent functions. This generation process can be viewed within a Bayesian Independent Component Analysis framework in two ways.

In the first case we apply ICA to a large ensemble of frames derived from clean, phonetically balanced recordings, revealing their underlying independent component structure based on Bells' seminal work on the higher order structure of images and sounds [2]. ICA is a statistical technique that determines a linear coordinate system, whose axes are defined by all higher moments of the data to which it is applied. Projecting the overlapping frames of a noisy recording on this basis, the time-domain observations are linearly transformed in order to reduce their dependency. Based on Hyvärinen's MAP formulation on the independent bases of images and Maximum Likelihood denoising of non-Gaussian data [3], we apply MAP inference on the posterior pdf of the ICA transformed clean frame conditioned on the noisy one, which leads to a threshold function optimally derived from the statistics of each independent component that effectively reduces white and coloured Gaussian noise. Finally, an inverse transformation from the ICA transformed domain back to time domain, taking into account the frame overlap, reconstructs the enhanced signal.

SCS requires the derivation of the unmixing matrix from clean data prior to applying the method. In the second technique there is no need to preprocess clean speech recordings, as the mixing matrix and the independent functions are directly inferred from the degraded observations. The key idea is to state the one-channel speech enhancement problem as a noisy, blind source separation case. As in SCS, the underlying clean speech frame is proposed to be constructed by the mixing of independent functions. The posterior distribution of the model for the generation of observed noisy frames defines a probability distribution over all sets of parameters. The posterior pdf is expressed as a computationally feasible approximation. The variables of interest, (i.e the mixing matrix and independent sources), are inferred in the process of fitting an approximating posterior to the true posterior, where the accuracy of the fit is assessed by the Kullback-Leibler (KL) divergence measure

between the two distributions. The clean frame is then reconstructed by multiplying the mixing matrix with the estimated independent components.

We support theoretical derivations by extensive experimentation using recorded speech signals and real noise sources from the NOISEX-92 database[1]. Assessment criteria are based on objective and subjective measurements. The former category includes the segmental signal to noise ratio (SNR) measure, Itakura-Saito distortion measurements (allegedly correlated with subjective perception of speech quality), as well as word recognition accuracy of a speech recognition system. The latter category includes visual comparison of speech signals and informal listening tests.

2 Problem Formulation

Consider a clean, time-domain speech signal $\mathbf{x_t}$ corrupted by additive Gaussian noise $\mathbf{n_t}$, where (t) denotes the sample index. We process $\mathbf{x_t}$ in overlapping blocks of N=80 samples with 79 samples overlap (N corresponds to 10 ms frame at 8 kHz sampling rate). The resulting matrices are:

$$\mathbf{Y=X+N} \quad (\mathbf{1})$$

where $\mathbf{Y}=[\mathbf{y}_1,..,\mathbf{y}_F]$, $\mathbf{X}=[\mathbf{x}_1,..,\mathbf{x}_F]$, $\mathbf{N}=[\mathbf{n}_1,..,\mathbf{n}_F]$. Each $\mathbf{x}_i$, $\mathbf{y}_i$, $\mathbf{z}_i$, i=1,..,F, corresponds to a frame and F denotes the number of frames. N has been selected in such a way that the speech stationarity assumption is accurate enough and the window is small enough to capture rapid varying phenomena as transients and stop bursts.

We assume the existence of an underlying process that is responsible for the composition of every frame $\mathbf{x}_i \in \mathbf{X}$. This generative process is summarized as:

$$\mathbf{X=AS} \quad (\mathbf{2})$$

where $\mathbf{A} \in \Re^{NxN}$ is an unknown constant matrix called the mixing matrix and $\mathbf{S} \in \Re^{NxF}$. Eq. 2 suggests that any frame $\mathbf{x}$ is a linear combination of 'templates' (columns of $\mathbf{A}$), weighted by the corresponding coefficients of $\mathbf{s}$, or equivalently; the columns of $\mathbf{A}$ span the space of 10 ms frames and the independent coefficients $\mathbf{s} \in \mathbf{S}$ are responsible for the linear combination of the columns of $\mathbf{A}$ that reconstruct every phonetic segment $\mathbf{x}$. In SCS, in order to derive the unmixing matrix $\mathbf{W}$, which, when applied on any speech frame provides its variables with maximum independency, we make use of 1000 clean, phonetically balanced recordings uttered by an equal number of speakers of both genders at 8 kHz sampling rate. One can observe that a time varying mixing/unmixing matrix is more proper for the time varying case of speech. However, the unmixing matrix aims at reducing the dependency of consecutive speech samples and forming sparser distributions of the resulting pdf's of the ICA transformed speech frames. Sparsity can be achieved with a constant unmixing matrix as well. In the case of Variational Approximation the speech waveform is segmented in small blocks for which a different mixing matrix is derived and the enhanced signal is reconstructed in a block by block basis. Finally, heavy overlapping in both techniques is necessary, in order to result in a smooth, enhanced speech waveform.

[1] Varga A., "NOISEX-92 study," Technical Report, DRA Speech Research Unit, 1992.

3 Speech Enhancement Using the SCS Technique

There is a broad class of algorithmic versions implementing ICA for the instantaneous noiseless mixing case. We made use of a batch version of the Fastica algorithm [4] that derives **W** iteratively, through successive presentations of the frames of each recording separately until convergence in order to avoid processing at once the huge amount of total observational data. After estimating **W** we proceeded in an orthogonalizing transformation:

$$\mathbf{W}_{orth}=\mathbf{W}(\mathbf{W}^T\mathbf{W}) \quad (3)$$

The orthogonalization procedure normalizes the eigenvectors of **W** resulting in a more numerically-stable matrix. **W**orth was found superior to **W** on the grounds of better performance as assessed using the objective and subjective criteria of section 5. By applying $\mathbf{W}_{orth}$ to Eq. 1 we map the time domain matrices to the ICA domain:

$$\mathbf{Z}=\mathbf{W}_{orth}\mathbf{Y}=\mathbf{W}_{orth}\mathbf{X}+\mathbf{W}_{orth}\mathbf{N}=\mathbf{U}+\mathbf{N'} \quad (4)$$

Gausssian distributions are invariant to linear transformations, therefore, $\mathbf{N'}=\mathbf{W}_{orth}\mathbf{N}$ is also zero mean Gaussian. Each $\mathbf{u}\in\mathbf{U}:\mathbf{U}=\mathbf{W}_{orth}\mathbf{X}$, where $\mathbf{U}=[\mathbf{u}_1,..,\mathbf{u}_F]$, corresponds to a clean speech frame in the ICA transform domain. The components composing each **u** can be considered almost independent and can be factorized in terms of pdf marginals. The Bayesian form of the posterior pdf of **u** is (index (i) is dropped):

$$f_{\mathbf{u/z}}(\mathbf{u}/\mathbf{z})=\frac{f_{\mathbf{z/u}}(\mathbf{z/u})f_{\mathbf{u}}(\mathbf{u})}{f_{\mathbf{z}}(\mathbf{z})}=\frac{1}{f_{\mathbf{z}}(\mathbf{z})}f_{\mathbf{n'}}(\mathbf{z}-\mathbf{u})f_{\mathbf{u}}(\mathbf{u}) \quad (5)$$

Assuming that there is no correlation of noise between components (the assumption holds approximately for non white Gaussian noise) and based on the factorization of $f_{\mathbf{u}}(\mathbf{u})$, we obtain the MAP estimate of **u** by minimizing the uniform cost function:

$$\hat{\mathbf{u}}=\underset{u}{\operatorname{argmax}}\left\{\prod_{i=1}^{N}f_{n'}(z_i-u_i)\prod_{i=1}^{N}f_{ui}(u_i)\right\} \quad (6)$$

As $\hat{\mathbf{u}}$ is now factorized, it can be decomposed and the denoising method can be applied to each individual component as soon as a closed-form of the densities $f_{ui}(u_i)$ is derived. Evaluating Eq. 6 for i=1,..,N allows $\hat{\mathbf{u}}_i=[\hat{u}_1,\ldots,\hat{u}_N]^T$ to be constructed and subsequently $\bar{\mathbf{U}}=[\hat{\mathbf{u}}_1,\ldots,\hat{\mathbf{u}}_F]$. By making use of the relation $\bar{\mathbf{U}}=\mathbf{W}_{orth}\mathbf{X}$ and the orthogonality of $\mathbf{W}_{orth}$, we return back to time domain by: $\mathbf{X}=\mathbf{W}^{\mathbf{T}}{}_{orth}\bar{\mathbf{U}}$.

We reconstruct the enhanced waveform by reshaping matrix **X** to vector form, taking into account the overlapping of vectors. The appropriate density representing each independent component $u_i(t)$ i=1,..,N in Eq. 6, can be selected according to the smallest KL divergence between the non-parametric density estimate of each component (histogram method) and a suitable, fitted parametric density (e.g Gaussian mixtures, double Laplacian). However, this selection procedue was found unnecessary. We estimated the normalized kurtosis of each $u_i(t)$ for a large number of recordings and we found with no exception that all components were very sparse. (Let $\mathbf{K}_{rec}$ denote the normalized kurtosis over all recordings, then mean$\mathbf{K}_{rec}$ =15, min$\mathbf{K}_{rec}$=8, max$\mathbf{K}_{rec}$=48 and it is obvious that they diverge significantly from the zero

value of normalized kurtosis of a Gaussian pdf). Therefore, the representative of a family of very sparse super-Gaussian pdfs' is selected in advance as described in [3].

$$f_{ui}(u_i) = \frac{1}{2d}\frac{(a+2)[a(a+1)/2]^{a/2+1}}{[\sqrt{a(a+1)/2} + |u_i/d|]^{a+3}} \quad (7)$$

d denotes standard deviation and $a=(2-k+(k(k+4))^{1/2})/(2k-1)$ where $k=d^2 f_{ui}(0)$ controls the sparsity of the distribution. Substituting Eq. 7 in Eq. 5 and setting the derivative of the log-likelihood to zero results to a non-linearity applied to z_i.

$$\bar{u}_i = \text{sign}(z_i)\max(0, \frac{|z_i| - bd}{2} + \frac{1}{2}\sqrt{(|z_i| + bd)^2 - 4\sigma^2(a+3)}) \quad (8)$$

where i=1,..,N and $b = \sqrt{a(a+1)/2}$. As can be observed from the coefficient $|z_i|$-bd , the non-linearity has a threshold, 'shrinking' effect by setting small values to zero.

4 Speech Enhancement using Variational Bayes Approximation.

Allowing Eq. 1 and Eq. 2 to be combined, the problem of speech enhancement can be re-stated as a problem of noisy, blind signal separation. We segment the noisy speech signal in blocks in order to infer a different mixing matrix **A** and independent sources for each segment, allowing clean speech to be constructed by a simple multiplication of the mixing matrix and the independent components in a block by block basis. For each segment we assign a mixture of Gaussian priors to the sources as in [6], and [5].

$$f(s_m) = \sum_{c=1}^{Nc} p_{mc} N(s_m \mid 0, b_{mc}) \quad (9)$$

We adopt the same zero-mean Gaussian noise model as in Eq. 4. The prior pdfs of the source and noise models and their parameters are stated in conjugate forms (see Appendix). A simple separable formulation of the posterior distribution is selected:

$$Q(\mathbf{s}, p, b, \mathbf{A}, a, \mathbf{L}) = Q(\mathbf{s})Q(p)Q(b)Q(\mathbf{A})Q(a)Q(\mathbf{L}) \quad (10)$$

Applying Jensen's inequality to the log evidence one can obtain an upper bound of the Maximum Likelihood for a model with visible variables the frame observations and hidden variables and their parameters as expressed in Eq. 1, 2, 4, 9. Jensen's inequality leads to a simplified cost function C_{KL} (Eq. 11), which can be minimized with respect to each of the Q distributions (see also [7]). The model parameters are inferred by a free-form functional minimization of the Kullback-Leibler divergence between the approximate distribution of Eq. 10 and the true posterior f(**s**, p, b, **A**, a, **x**, **L**), (refer to [5] for a detailed derivation and implementation).

$$C_{KL} = \left\langle \log \frac{Q(\mathbf{s})Q(p)Q(b)}{f(\mathbf{s}, p, b)} \right\rangle_q + \left\langle \log \frac{Q(\mathbf{A})Q(a)}{f(\mathbf{A}, a)} \right\rangle_q + \left\langle \log \frac{Q(\mathbf{L})}{f(\mathbf{z}, \mathbf{L})} \right\rangle_q \quad (11)$$

The minimization process of the cost function results in set of coupled equations that are updated iteratively until the parameters of interest (i.e **A** and **S**), are inferred.

5 Experimental results

The BICA framework makes the assumption that the distribution of noise is Gaussian. Therefore, we expect some impact on the effectiveness of BICA for noise cases exhibiting divergence from the normality assumption. Nevertheless, from Eq. 4, it becomes obvious that every $\mathbf{z} \in \mathbf{Z}$: $\mathbf{Z}=\mathbf{WX}+\mathbf{N}'$ accumulates a number of noise components. According to the Central Limit Theorem for each $\mathbf{z}$ component the density of noise will be closer to Gaussian than the distribution of each of the noise components. Therefore, we expect small impact on the effectiveness of BICA framework for noise cases exhibiting moderate divergence from the normality assumption. As regards the SNR of the recordings: each noise type is added to 34 clean speech files of 5 sec. mean duration so that the corrupted waveform ranges from –10 to 20 SNR_{dB}. The objective criteria are based on the mean value over all recordings. Let s(t) and $\hat{s}(t)$ be the original undegraded and enhanced speech signals:

Segmental SNR provides simple error estimation over time and frequency and is calculated for (M), non-overlapping frames of 15 ms duration while the result is averaged over all waveform segments. Fig. 1 and Fig. 2 illustrate that best performance is achieved for the Gaussian type of noise. We attribute the extensive denoising capability to the fact that Gaussian noise complies with the assumptions of the algorithm, more than any other type of noise. Good performance is observed for the 'Operations Room' and 'STITEL' noise cases which are slowly varying, though, the algorithm cannot suppress the impulsively occurring components.

$$\mathrm{SNR}_{dB} = \frac{1}{M}\sum_{j=0}^{M-1} 10\log_{10}\frac{\sum {s_M}^2(t)}{\sum (s_M(t) - \hat{s}_M(t))^2} \qquad (12)$$

Itakura-Saito (IS) distortion measure is based on the spectral distance between the AR coefficient (a_c, a_p) of the clean and enhanced speech waveforms over synchronous frames of 15ms duration (g_c, g_p: all-pole gains). (IS) measure is closely associated with speech quality assessment and it is very sensitive to spectrum variations. It confirms our observation based on total SNR measure. (IS) measure reveals that although noise reduction is high at low SNRs the mismatch between the true and the enhanced version of the signal is augmented as the noise variance increases.

$$d_{IS}(a_p, a_c) = \left[\frac{g^2_c}{g^2_p}\right]\left[\frac{a_p R_c a^T_p}{a_c R_c a^T_c}\right] + \log\left[\frac{g^2_p}{g^2_c}\right] - 1 \qquad (13)$$

Word Recognition accuracy was assessed by using a speech recognition module built with HTK Hidden Markov Models toolkit (Young S., "The HTK Book", Cambridge University, 1995). The basic recognition units are tied state context dependent triphones of five states each. Given this set of HMMs, and the corresponding dictionary, the HTK recognition unit produces the best path of the word network using the Frame Synchronous Viterbi Beam Search algorithm. The testing set consists of 100 files, part of the identity card corpus of the SpeechDat database (one speaker for each recording). The results of the tests conducted are depicted in Fig 1 and 2 and demonstrate that BICA enhancement method cooperates well with the HMMs framework, achieving consistently good results in low SNRs.

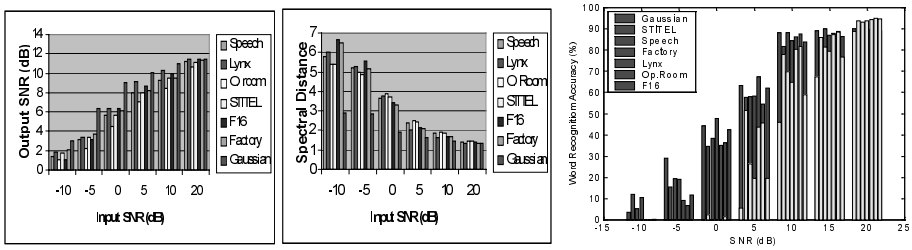

Fig. 1. SCS technique. (From left to right): Seg. SNR, Itakura-Saito, Word Recognition Accuracy (Black: SCS enhancement applied, White: no enhancement applied).

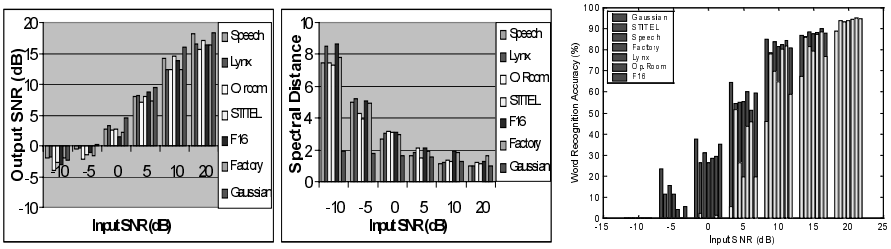

Fig. 2. Variational Bayes. (Left to right): Seg. SNR, Itakura-Saito, Word Recognition Accuracy (Black: SCS enhancement applied, White: no enhancement applied).

Listening tests. What we find most interesting is that at low SNRs as –10 dB, musical noise is hardly perceivable, though, the enhanced signal suffers due to distortions induced on the speech waveform. In such adverse conditions the algorithm shows its best performance in the case of Gaussian noise, since it is the noise that complies the most with the normality assumption (refer also to the time domain plots of noisy and enhanced signals of Fig. 3). At –5 dB the enhanced signal is quite perceptible though speech distortion is still present. The quality of the enhanced speech is quite high at higher SNRs, though, the enhanced coloured-noise speech is less clear. The non-stationary, impulsively occurring components of some noises cannot be suppressed, (e.g. slums of door in factory noise, voices in Operation room noise-case), a fact that we attribute to the violation of the stationarity assumption.

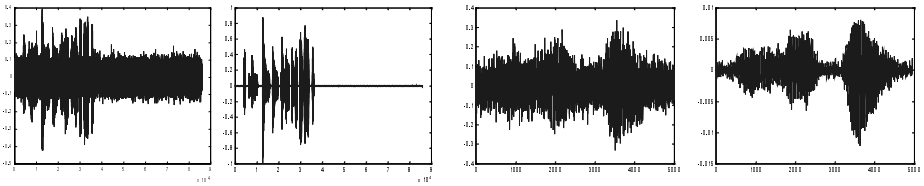

Fig. 3. Time domain plots for noisy and enhanced signals degraded with Gaussian noise at 0 dB input SNR. From Left to Right (a),(b): SCS enhancement, (c), (d): Variational Bayes method.

6 Conclusions

A novel view for the enhancement of signals is applied successfully to speech. It is based on the idea of decomposing clean speech frames corresponding to 10 ms phonetic segments into their independent basis functions. We proposed the application of a unifying BICA framework of two different techniques. In the first case an unmixing projection matrix is derived from a large ensemble of clean frames. This ensemble facilitates the application of MAP formulation that leads to a shrinkage function optimally derived from the statistics of each component. In the second case we demonstrated how a speech-enhancement problem can be stated as a noisy blind source separation problem faced with Bayesian Variational Approximation. Extensive experimentation with these techniques gave excellent results in the case of white Gaussian noise. They proved very effective in coloured types of Gaussian noise that do not diverge significantly from the stationarity assumption, and, most important, they preserve natural sound. As regards SCS, the most time-consuming task of the algorithm is the computation of the projection matrix **W**, which is computed off-line and only once for all subsequent restorations, whereas the Variational Bayes approach, though more computationally demanding, infers the clean frames directly from the noisy observations without the need of a preprocessing stage.

Appendix: Conjugate priors and optimal posterior distributions

The conjugate priors (left side) of the optimal posterior distributions (right side) are:

$f(b_{mc}) \sim G(b_{mc} \mid b^{(b)}, c^{(b)})$	$Q(b_{mc}) \sim G(b_{mc} \mid b_{mc}^{(b)}, c_{mc}^{(b)})$
$f(A_{nm}) \sim N(A_{nm} \mid 0, a_m)$	$Q(\mathbf{s}) \sim G(\mathbf{s} \mid \hat{\mathbf{s}}, \check{\mathbf{s}})$
$f(\{\pi_{mc}\}_{c=1,..,N}) \sim D(\{\pi_{mc}\}_{c=1,..,N} \mid c_{(p)})$	$Q(\{p_{mc}\}_{c=1,..,N}) \sim D(\{p_{mc}\}_{c=1,..,N} \mid c^{(p)}{}_{c=1,..N})$
$f(a_m) \sim G(a_m \mid b^{(a)}, c^{(a)})$	$Q(a_{mc}) \sim G(a_{mc} \mid b_{mc}^{(a)}, c_{mc}^{(a)})$
$f(L_{nn}) \sim G(L_{nn} \mid b^{(L)}, c^{(L)})$	$Q(L_{nm}) \sim G(L_{nn} \mid b^{(L)}, c^{(L)})$

where N, G, D denote Normal, Gamma and Dirichlet pdfs respectively.

References

1. Deller J., Proakis G., Hansen J.: "Discrete Time Processing of Speech Signals," New York, Macmillan Publishing Company, (1993).
2. Bell A., "The Independent Components of Natural Scenes are Edge Filters," Vision Research, (1997), 3327-3338.
3. Hyvärinen A., Hoyer O., Oja E., "Image Denoising by Sparse Code Shrinkage," S. Haykin and B. Kosko (eds), Intelligent Signal Processing, IEEE Press, 2000
4. Hyvärinen A., Oja E., "A Fast Fixed-Point Algorithm for Independent Component Analysis," Neural Computation, 9(7), (1997), 1483-1492.
5. Miskin J., MacKay D.,"Ensemble Learning for Blind Source Separation," Principles and Practice, Cambridge University Press.
6. Attias H., "Independent Factor Analysis," Neural Comput., (11), (1999), 803-851.
7. Ghahramani Z., "On Structured Variational Approx.," Tech. Report CRG-TR-97-1.