Accepted to ICA'99
http://sig.enst.fr/~ica99/
Updated version, September 9th, 1998

ENSEMBLE LEARNING FOR INDEPENDENT COMPONENT ANALYSIS

Harri Lappalainen

Helsinki University of Technology
Neural Networks Research Centre
P.O.Box 2200
FIN-02015 HUT, FINLAND
Harri.Lappalainen@hut.fi
http://www.cis.hut.fi/~harri/

ABSTRACT

In this paper, a recently developed Bayesian method called ensemble learning is applied to independent component analysis (ICA).

Ensemble learning is a computationally efficient approximation for exact Bayesian analysis. In general, the posterior probability density function (pdf) is a complex high dimensional function whose exact treatment is difficult. In ensemble learning, the posterior pdf is approximated by a more simple function and Kullback-Leibler information is used as the criterion for minimising the misfit between the actual posterior pdf and its parametric approximation. In this paper, the posterior pdf is approximated by a diagonal Gaussian pdf.

According to the ICA-model used in this paper, the measurements are generated by a linear mapping from mutually independent source signals whose distributions are mixtures of Gaussians. The measurements are also assumed to have additive Gaussian noise with diagonal covariance.

The model structure and all parameters of the distributions are estimated from the data.

1. INTRODUCTION

Recently there has been a lot of interest in Bayesian methods but few applications for unsupervised learning. One of the most important benefits of Bayesian methods is the possibility for model comparison. In supervised learning, cross validation or other methods can be used. For unsupervised learning this is usually not possible, however, since reconstruction error decreases also for the test set as the complexity of the model increases. The ability to optimise the structure of the model is thus particularly valuable in unsupervised learning.

Both ensemble learning, first used in [1], and independent component analysis using mixture of Gaussians model for sources, first used in [2], are existing techniques, but they have not been combined previously.

The reader is assumed to have basic knowledge about Bayesian probability theory and ICA.

2. ENSEMBLE LEARNING

Assume we would like to make a prediction, decision, etc., based on measurements and some kind of models. From the axioms of Bayesian probability theory it follows that all the models should be used in the process and the models should be weighted according to the posterior probabilities of the models. This averaging over models is the essence of Bayesian analysis.

Usually the models include unknown real values and therefore the posterior probability is expressed by a posterior pdf. Unfortunately the posterior pdf is typically a complex high dimensional function whose exact treatment is difficult. In practice, it has to be approximated in one way or another.

In ensemble learning, a parametric computationally tractable approximation – an ensemble – is chosen for the posterior pdf. Let P denote the exact posterior pdf and Q the ensemble. The misfit between P and Q is measured with Kullback-Leibler information I_{KL} between Q and P.

$$I_{\mathrm{K}L}(Q;P) = E_Q\left\{\ln\frac{Q}{P}\right\}$$

The parameters of the ensemble are optimised to fit the posterior by minimising $I_{\mathrm{K}L}(Q;P)$.

Ensemble learning was first used in [1] where it was applied to a multi-layer perceptron with one hidden layer. Since then it has been used e.g. in [3-9].

2.1. Model selection

An important special case of approximation of the posterior pdf is the model selection. The posterior pdf is typically multimodal, but often almost all of the probability mass is located around the largest peak of the posterior pdf. When there is a lot of data compared to the complexity of the models, this is almost always the case. In our case, approximating the posterior pdf with only one peak is usually reasonably accurate.

Notice that the posterior density itself has no special meaning regarding the averaging over models; only the probability mass matters. A broad peak with low density can be more important than a sharp peak with high density. Over-learning results in high but very narrow peaks. The Kullback-Leibler information automatically takes into account the probability mass and is therefore robust against over-learning.

3. MODEL FOR THE MEASUREMENTS

The measurements vectors $\{x(t)\}$ are assumed to be generated by a linear mapping A from mutually independent source signals $\{s(t)\}$ and additive Gaussian noise $\{v(t)\}$.

$$x(t) = As(t) + v(t)$$

The components $v_i(t)$ of the noise are assumed to have means b_i and variances $e^{2\sigma_i}$. Another way to put this is to say that $x(t)$ has Gaussian distribution with mean $As(t) + b$ and diagonal covariance with components $e^{2\sigma_i}$. Each component A_{ij} of the linear mapping is assumed to have zero mean and unit variance.

The distribution of each source signal is a mixture of Gaussians (MOG).

$$p(s_i(t)|c_i, S_i, \gamma_i) = \frac{\sum_j e^{c_{ij}} \mathcal{G}(s_i(t); S_{ij}, e^{2\gamma_{ij}})}{\sum_j e^{c_{ij}}}$$

The parameters c_{ij} are the logarithms of mixture coefficients, S the means and γ the logarithms of the standard deviations of the Gaussians[1] (here $\mathcal{G}(a; b, c)$ denotes a Gaussian distribution over a with mean b and variance c).

The distributions of parameters c_{ij}, S_{ij}, γ_{ij}, b_i and σ_i are $\mathcal{G}(c_{ij}; 0, e^{2\alpha})$, $\mathcal{G}(S_{ij}; 0, e^{2\epsilon})$, $\mathcal{G}(\gamma_{ij}; \Gamma, e^{2\delta})$, $\mathcal{G}(b_i; B, e^{2\epsilon})$ and $\mathcal{G}(\sigma_i; \Sigma, e^{2\eta})$.

The prior distribution of the hyperparameters α, ϵ, Γ, δ, B, β, Σ and η is assumed to be uniform in the area of reasonable values for the hyperparameters.

[1] The Gaussians are parametrised by the logarithms of the standard deviations in order to make the future assumption of roughly Gaussian posterior pdf valid.

To summarise: the eight hyperparameters are assigned flat prior pdfs. The distributions of other parameters are defined hierarchically from these using Gaussian distributions each parametrised by the mean and the logarithm of the standard deviation. The joint pdf of $\{x(t), s(t), A, b, \sigma, c, S, \gamma, \alpha, \epsilon, \Gamma, \delta, B, \beta, \Sigma, \eta\}$ is simply the product of the independent pdfs.

4. DIAGONAL GAUSSIAN ENSEMBLE

Given the measurements, the unknown variables of the model are the source signals, the mixing matrix, the parameters of the noise and source distributions and the hyperparameters. The posterior P is thus a pdf of all these unknown variables. For notational simplicity, we shall sometimes denote these n variables by $\theta_1, \theta_2, \ldots, \theta_n$.

In order to make the approximation of the posterior pdf computationally tractable, we shall choose the ensemble Q to be a Gaussian pdf with diagonal covariance. The ensemble has twice as many parameters as there are unknown variables in the model because each dimension of the posterior pdf is parametrised by a mean and variance in the ensemble. A hat over a symbol denotes the mean and a tilde the variance of the corresponding variable.

$$Q(\theta_1, \ldots, \theta_n) = \prod_{i=1}^{n} \mathcal{G}(\theta_i; \hat{\theta}_i, \tilde{\theta}_i)$$

The factorised ensemble makes the computation of the Kullback-Leibler information $I_{\mathrm{KL}}(Q; P)$ simple since the logarithm can be split into a sum of terms: the terms $E_{\mathcal{G}(\theta_i)}\{\ln \mathcal{G}(\theta_i)\} = -1/2 \ln 2\pi e \tilde{\theta}_i$ (entropies of Gaussian distributions) and terms $-E_Q\{\ln P_i\}$, where P_i are the factors of the posterior pdf. Notice that the posterior pdf factorises into simple terms; due to the hierarchical structure of the model because the posterior pdf equals to the joint pdf divided by a normalising term.

To see how to compute the terms $-E_Q\{\ln P_i\}$, let $P_i = p(\gamma_{ij}|\Gamma, \epsilon) = \mathcal{G}(\gamma_{ij}; \Gamma, \epsilon)$.

$$-E_Q\{\ln P_i\} = E_Q\{\frac{(\gamma_{ij} - \Gamma)^2 e^{-2\epsilon} + \ln 2\pi}{2} + \epsilon\} \tag{1}$$

It is easy to show that this equals to

$$\frac{[(\hat{\gamma}_{ij} - \hat{\Gamma})^2 + \tilde{\gamma}_{ij} + \tilde{\Gamma}]e^{2(\tilde{\epsilon} - \hat{\epsilon})} + \ln 2\pi}{2} + \hat{\epsilon},$$

since according to the choice of Q, the parameters γ_{ij}, Γ and ϵ have independent Gaussian distributions.

The most difficult terms are the expectations of $\ln p(s_i(t)|c_i, S_i, \gamma_i)$. Approximation for the expectation of this form is given in appendix A.

The normalising term in the posterior pdf only depends on those variables which are given, in this case $\{x(t)\}$, and can therefore be neglected when minimising the Kullback-Leibler information $I_{\mathrm{KL}}(Q; P)$.

5. SIMULATIONS

5.1. Data

Speech data was used for the simulations: 30 s of Finnish speech was digitised with 16 kHz sampling rate and high-pass filtered. Power spectra were computed every 8 ms using short time Fourier transformations with Hamming windows of length 16 ms. This results in 3749 vectors of dimension 128. Energy was computed for 34 channels whose spacing imitates the frequency scale of human ear. Logarithms were taken from the energies after adding small constants. The final data thus consisted of 3749 vectors of dimension 34.

This particular preprocessing was chosen because it is typical in current speech recognition systems.

5.2. Minimisation

The ensemble Q was fitted to the posterior pdf by minimising $I_{\mathrm{KL}}(Q; P)$. The particular technique for minimisation is not important regarding the results; everybody can use their favourite algorithm. In these simulations, a variation of Newton's method was used. The parameters $\hat{s}(t)$ and $\tilde{s}(t)$ were iterated 15 times for each measurement vector $x(t)$. Other parameters were updated after going through all the measurements, and this was iterated 200 times. The number of iterations was chosen conservatively. The details of the minimisation procedure will be given in [10].

5.3. Results

Several different structures for the model were tested. The number of source signals and the number of Gaussians in the mixtures was varied. The number of Gaussians in the mixtures was same for all sources in each network. This is not to say that it would not be perfectly simple to optimise the number of Gaussians for each source separately.

It turned out that the Kullback-Leibler information was minimised by a network with 23 dimensional source signals whose distributions were mixtures of three Gaussians. There were 87 292 unknown variables in the model: 86 227 in $s(t)$; 782 in A; 68 in b and σ; 207 in c, S and γ; and 8 in hyperparameters.

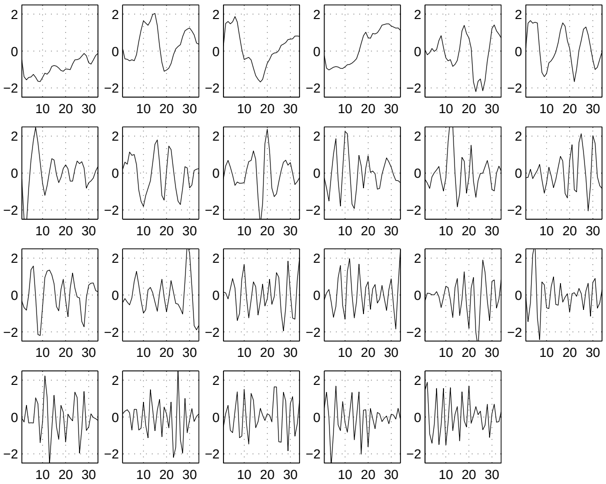

Figure 1: Each row vector of $\hat{A}$ is a 34 dimensional basis vector corresponding to one source. The frequency increaces from left to right in all the subimages.

Figures 1–3 show the basis vectors, histograms and reconstructed distributions of the 23 source signals of the best network. The ordering in all three figures is the same. The basis vectors in figure 1 probably do not seem very interesting for someone who is not working with speech recognition. The basis is fairly close to cosine transformation which is widely used for processing the spectra in speech recognition.

The histograms in figure 2 and the corresponding distributions in figure 3 show that the algorithm works. The model has captures the salient features of the source distributions, some of which are multimodal, skewed or kurtotic.

The second best fit was obtained by a network with two Gaussian in the mixtures, but the probability mass it captured[2] was over 10^{76} times smaller! It is therefore reasonable to approximate the whole posterior pdf of all model structures and parameters by an ensemble with a peak in only one model structure. It is also evident that in this case, any prior information about the model structure has no significance.

For comparison, also models with only one Gaussian in the mixtures were tested. In this case the logarithmic mixture coefficients c_{ij} can be dropped out from the model. The best model with only one Gaussian was found to have over 10^{1238} times less probability mass. This shows that the algorithm agrees with human eye: it is clear that the distributions of at least some of the source signals are far from Gaussian.

[2] In [10] it will be explained how the relative probability masses of different models can be compared using the Kullback-Leibler information computed by the algorithm. See also [9].

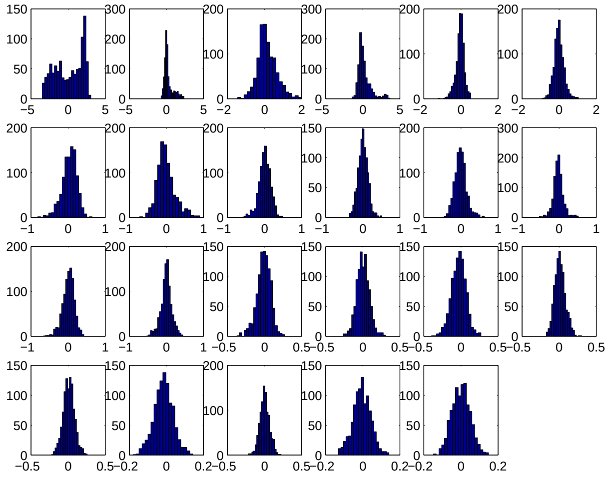

Figure 2: The histograms of the means $\hat{s}_i(t)$ of the sources.

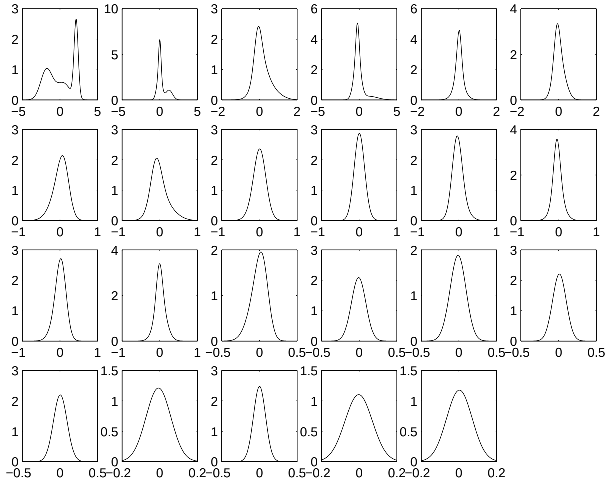

Figure 3: The distributions of the sources reconstructed from $\hat{c}$, $\hat{S}$ and $\hat{\gamma}$.

6. DISCUSSION

6.1. Benefits

Probably the most important benefit of Bayesian analysis regarding unsupervised learning is the ability to compare models. Not only can the different ICA-models be compared to each other, they can also be compared to vector quantisation or any other statistical models, provided that Bayesian analysis is also applied to these other models.

In ensemble learning, the treatment for missing values is very simple. Since they are unknown variables, they are included in the ensemble and therefore their distribution will be estimated similarly as for any other unknown variables.

6.2. Limitations

The method proposed here has limitations due to the simple structure of the ensemble. The posterior pdf of the unknown variables is often close to Gaussian, but there can be significant correlations. On the other hand, as the algorithm tries to fit the ensemble to the posterior pdf, it tries to find a peak in the posterior which would satisfy the diagonality assumption. In the simulations with only one Gaussian in the mixtures, for instance, the linear mapping found by the algorithm will orthogonal because that makes the sources independent in the posterior pdf.

There are two significant cases where the factorial assumption for the ensemble is too strong. If the row vectors of the linear mapping A are far from orthogonal, the source signals are correlated in the posterior pdf. Another case is when the amount of noise in the data is small and there are not very many data samples. In that case the components of the linear mapping are correlated with the source signals in the posterior pdf.

The limitations can be overcome by adding off-diagonal terms to the covariance matrix of the ensemble, but then the formulas for the Kullback-Leibler information become more complicated. Full covariance matrix is, of course, out of the question: the ensemble already had 174 584 parameters and with full covariance matrix the number would have been almost 7.62×10^9. It is more feasible to try to find model structures which make the off-diagonal terms in the covariance matrix of the unknown variables small.

6.3. Relation to previous work

Many current ICA-algorithms do not estimate the noise level from the data. The ability to estimate the noise is not due to the Bayesian analysis, however. It stems from the generative model used here. Similar models have been used for instance in [2, 11]. The treatment in the latter is very similar to ours: a factorial ensemble was fitted to the posterior pdf of the source signals by minimising their Kullback-Leibler information. However, the posterior did not include other parameters than the sources and only the maximum of the posterior pdf was used. As argued in section 2.1, large density does not necessarily imply large mass, and therefore the EM-algorithm used in [11] does not necessarily find a probable model for the data, and can not, in any case, be used for comparing models with different structures.

6.4. Some generalisations

The most obvious generalisation would be to take into account the significant temporal correlations. Also, the linear mapping can be replaced by a nonlinear one. An interesting generalisation is to let the variances of the sources vary in time. According to this model, higher lever source signals modulate the variances of source signals.

Simulations with these and other variants will be published in [10]. All these models have significantly larger probabilities than the simple ICA-models studied in this paper.

7. REFERENCES

[1] Geoffrey Hinton and Drew van Camp. Keeping neural networks simple by minimizing the description length of the weights. In *Proceedings of the COLT'93*, pages 5–13, Santa Cruz, California, 1993.

[2] Éric Moulines, Jean-François Cardoso, and Elisabeth Gassiat. Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models. In *Proceedings of the ICASSP'97*, pages 3617–3620, Munich, Germany, 1997.

[3] David J. C. MacKay. Developments in probabilistic modelling with neural networks—ensemble learning. In *Neural Networks: Artificial Intelligence and Industrial Applications. Proceedings of the 3rd Annual Symposium on Neural Networks, Nijmegen, Netherlands, 14-15 September 1995*, pages 191–198, Berlin, 1995. Springer.

[4] Lawrence K. Saul, Tommi Jaakkola, and Michael I. Jordan. Mean field theory for sigmoid belief networks. *Journal of Artificial Intelligence Research*, 4:61–76, 1996.

[5] David J. C. MacKay. Comparison of approximate methods for handling hyperparameters. *Neural Computation*. Submitted.

[6] David J. C. MacKay. Ensemble learning for hidden Markov models. Available from `http://wol.ra.phy.cam.ac.uk/`, 1997.

[7] David Barber and Christopher M. Bishop. Ensemble learning for multi-layer networks. In M. I. Jordan, M. J. Kearns, and S. A. Solla, editors, *Advances in Neural Information Processing Systems 10*, pages 395–401. MIT Press, 1998.

[8] David Barber and Bernhard Schottky. Radial basis functions: a bayesian treatement. In M. I. Jordan, M. J. Kearns, and S. A. Solla, editors, *Advances in Neural Information Processing Systems 10*, pages 402–408. MIT Press, 1998.

[9] Christopher M. Bishop, Neil Lawrence, Tommi Jaakkola, and Michael I. Jordan. Approximating posterior distributions in belief networks using mixtures. In M. I. Jordan, M. J. Kearns, and S. A. Solla, editors, *Advances in Neural Information Processing Systems 10*, pages 416–422. MIT Press, 1998.

[10] Harri Lappalainen. Ensemble learning for unsupervised neural networks. Technical report, Helsinki University of Technology, Laboratory of Computer and Information Science, 1998. In preparation.

[11] Hagai Attias. Independent factor analysis. *Neural Computation*, 1998. Submitted.

A. DERIVATIONS

An approximation for $-E_Q\{\ln p(s_i(t)|c_i, S_i, \gamma_i)$ is derived here. The derivation makes use of the Jensen's inequality and second order Taylor's series expansion.

Let $g_{ij} = -\ln \mathcal{G}(s_i(t); S_{ij}, \gamma_{ij})$. The expectation to be approximated is then

$$E_Q\left\{\ln \sum_j e^{c_{ij}}\right\} - E_Q\left\{\ln \sum_j e^{c_{ij}-g_{ij}}\right\}.$$

Let us first concider the latter term. The logarithm of the sum is a strictly convex function of g_{ij}. By Jensen's inequality, moving the expectation inside a convex function cannot result in increment. Replacing the latter expectation by

$$E_{s_i(t),c_i}\left\{\ln \sum_j e^{c_{ij}-E_{S_{ij},\gamma_{ij}}\{g_{ij}\}}\right\}$$

can therefore only result in increment in the approximation. This is safe because we are trying to minimise the Kullback-Leibler information and approximating it above is thus conservative.

The expectation $E_{S_{ij},\gamma_{ij}}\{g_{ij}\}$ is similar to equation 1 and equals to

$$\frac{[(s_i(t) - \hat{S}_{ij})^2 + \tilde{S}_{ij}]e^{2(\tilde{\gamma}_{ij}-\hat{\gamma}_{ij})} + \ln 2\pi}{2} + \hat{\gamma}_{ij}.$$

Let us denote this by $\hat{g}_{ij}(s_i(t))$. At this point, the approximation equals to

$$E_{c_i}\left\{\ln\sum_j e^{c_{ij}}\right\} - E_{s_i(t),c_i}\left\{\ln\sum_j e^{c_{ij}-\hat{g}_{ij}(s_i(t))}\right\}.$$

The terms inside the expectations are functions of $s_i(t)$ and c_i.

Next, let us concider a second order Taylor's series expansion about $\hat{s}_i(t)$ and $\hat{c}_i$. Notice that the first order terms and all second order crossterms disappear in the expectations and only the constant and the pure second order terms remain. This is because the variables are independent in the ensemble.

The constant term will be

$$\ln\sum_j e^{\hat{c}_{ij}} - \ln\sum_j e^{\hat{c}_{ij}-\hat{g}_{ij}(\hat{s}_i(t))} \tag{2}$$

and the remaining second order terms of c_{ij}

$$E_{c_i}\left\{\sum_j \frac{(c_{ij}-\hat{c}_{ij})^2}{2}\zeta_{ij}(1-\zeta_{ij})\right\}$$

and

$$-E_{c_i}\left\{\sum_j \frac{(c_{ij}-\hat{c}_{ij})^2}{2}\xi_{ij}(1-\xi_{ij})\right\},$$

where

$$\zeta_{ij} = \frac{e^{\hat{c}_{ij}}}{\sum_k e^{\hat{c}_{ik}}}$$

and

$$\xi_{ij} = \frac{e^{\hat{c}_{ij}-\hat{g}_{ij}(\hat{s}_i(t))}}{\sum_k e^{\hat{c}_{ik}-\hat{g}_{ik}(\hat{s}_i(t))}}.$$

Taking the expectations yields

$$\sum_j \frac{\tilde{c}_{ij}}{2}[\zeta_{ij}(1-\zeta_{ij}) - \xi_{ij}(1-\xi_{ij})]. \tag{3}$$

The second order term of $s_i(t)$ will be

$$E_{s_i(t)}\left\{\frac{(s_i(t)-\hat{s}_i(t))^2}{2}(\phi_i+\chi_i^2-\psi_i)\right\},$$

where

$$\phi_i = \sum_j \xi_{ij}e^{2(\tilde{\gamma}_{ij}-\hat{\gamma}_{ij})},$$

$$\chi_i = \sum_j \xi_{ij}(\hat{s}_i(t)-\hat{S}_{ij})e^{2(\tilde{\gamma}_{ij}-\hat{\gamma}_{ij})}$$

and

$$\psi_i = \sum_j \xi_{ij}\left[(\hat{s}_i(t)-\hat{S}_{ij})e^{2(\tilde{\gamma}_{ij}-\hat{\gamma}_{ij})}\right]^2.$$

Taking the expectation yields

$$\frac{\tilde{s}_i(t)}{2}(\phi_i+\chi_i^2-\psi_i). \tag{4}$$

At this point, the approximation of the original expectation is thus the sum of terms in equations 2–4.

Some care has to be taken with the approximation resulting from the Taylor's series expansion because it utilises only local information about the shape of the posterior pdf. For instance, if the mean $\hat{s}_i(t)$ happens to be in a valley between two Gaussians, the term in equation 4 will be negative. It then looks like the Kullback-Leibler information can be decreased by increasing $\tilde{s}_i(t)$. This only holds for small $\tilde{s}_i(t)$, however. At some point after the distribution of $s_i(t)$ has become broader than the separation between the two Gaussians, the Kullback-Leibler information starts to increase as $\tilde{s}_i(t)$ increases.

In order to avoid the problem, only positive terms of equations 3 and 4 will be included. The final approximation is thus equation 2 plus the positive terms of equations 3 and 4.