# Using Mixtures of Gaussians to Compare Approaches to Signal Separation

Kaare Brandt Petersen
Technical University of Denmark

## 1 Introduction

In the signal separation technique called Independent Component Analysis (ICA), one refers to a problem as being "square" if the there is as many measurements as sources to separate and "overcomplete" or "under-determined" if there is more sources than measurements. While square ICA is thoroughly investigated through many different approaches with well understood differences and similarities, the overcomplete case is still posing a difficult problem. Below is a short overview of some of the more interesting or illustrative of the approaches.

In 1999, Hagai Attias presented in [1] a Maximum Likelihood approach assuming the generative model $\mathbf{x} = \mathbf{A}\mathbf{s} + \epsilon$. In this approach a model distribution is constructed and using Mixture of Gaussians as priors makes it possible to complete the relevant integrals and obtain a closed form expression for the distribution over $\mathbf{x}$. The model distribution is approximated to the data distribution through the Kullback Leibler divergence. This approach is extremely flexible while still having appealing analytical properties and the only real drawback is the bad scaling behavior: A sum over $K^D$ must be computed, which can be rather larger for e.g. image data.

In 2000, Lewicki and Sejnowksi presented in [5] a Maximum Likelihood approach assuming the generative model $\mathbf{x} = \mathbf{A}\mathbf{s} + \epsilon$. In this approach, the log likelihood of $\mathbf{A}$ is Taylor expanded to second order around the maximum a posteriori estimate of the sources, i.e. one approximate the likelihood with a gaussian, and the sources are in turn are estimated using the updated estimate of $\mathbf{A}$. In this approach we see the difficulty that most overcomplete techniques is trying to work around: The likelihood, or some other suitable cost-function, contains some integral involving the source prior and is therefore in general hard to solve. The approach of Lewicki and Sejnowski from 2000 substitutes the integral with a second order expansion and we shall see other possibilities in the following.

In 2001, Girolami presented in [2] a Maximum Likelihood approach assuming the generative model $\mathbf{x} = \mathbf{A}\mathbf{s} + \epsilon$. In this approach, Girloami assumes Laplacian priors and the integral associated with the log likelihood is approximated through a variational scheme: The laplacian priors are reformulated in dual space which provides a lower bound of the likelihood to be optimized. The drawback of this approach is that the trick only works for laplacian priors and that the algorithm optimizes a lower boudn in stead of the log likelihood it self.

In 2002, Hojen-Soerensen, Winther and Hansen presented in [4], the so-called "Mean Field Approach" assuming the generative model $\mathbf{x} = \mathbf{As} + \epsilon$. The parameters $\mathbf{A}$ and possibly the noise covariance $\mathbf{W}$ are estimated through maximum likelihood assuming knowledge of the mean values of the sources and vice versa. That is, the integral of the log likelihood, translates into mean values of the sources, which are approximated with estimates of the mean values obtained from Mean Field theory. The nice feature of this approach is that we substitute a very complicated integral with an easier non-linear equation and that one can do this for any prior. The problem, of course is that al though Mean Field estimates are fairly accurate they are still approximations.

Also in 2002, Shriki, Sampolinski and Lee presented in [7] an interesting variant of Infomax on the filtering model $\hat{\mathbf{s}} = \mathbf{Wx}$. In the setup $\mathbf{y} = g(\mathbf{Wx})$, Shriki et. al. obtains a relation between $p(\mathbf{x})$ and $p(\mathbf{y})$ by assuming a noisy relation $\mathbf{y} = g(\mathbf{Wx}) + \epsilon$ and letting the noise go to zero. The main problem is that the limit is not taken properly care of and that the certainty of the result therefore is doubtful.

And finally in 2003, Teh, Welling, Osindero and Hinton presented in [8] the Energy Based Model on the filtering model $\hat{\mathbf{s}} = \mathbf{Wx}$. Through a setup using inspiration from physics, a model distribution for $\mathbf{x}$ is constructed and adjusted to be as close to the data distribution as possible through a Kullback Leibler divergence. Again the authors are faced with an intractable integral and this time it is approximated with the so-called "n-step Learning" which is a Hybrid Monte Carlo technique using $n$ steps in the estimation of the integral. The remarkable claim of the paper is that very few steps such as $n = 1$, or $n = 3$ often will be sufficient to obtain overall convergence of the algorithm.

A common feature for most of the approaches is that they assume a noisy model and obtain a likelihood which involves a difficult integral which is then approximated in some way or another. The exception from this is the approach of Hagai Attias, but one can then argue that assuming priors to be mixtures of gaussians either is a restriction or an approximation.

## 1.1 This Paper

Thus, Independent Component Analysis (ICA) can be performed by a vast range of different methods. These can differ from each other by assumed properties such as noise or time-correlation, but also by the fundamental issue on whether they attempt to estimate the generative mixing matrix, denoted $\mathbf{A}$, or a filtering matrix denoted $\mathbf{W}$. In case of the same number of observations and sources, the square case, most if not all of these methods can be proven to be equivalent. But in the overcomplete case, where the number of observations are smaller than the number of sources, their differences becomes apparent and it is not easy to compare the results of generative and filtering approaches.

This paper makes an attempt to compare the result of two different methods in the overcomplete case: The Maximum Likelihood (ML), which estimates the generative $\mathbf{A}$, and the Energy Based Models (EBM) which seek to estimate a filtering matrix $\mathbf{W}$. This is done by assuming the priors to be centered mixtures of gaussians which makes it possible to compare the optimization schemes. This approach, with respect to ML, is closely related to the work of Hagai Attias in

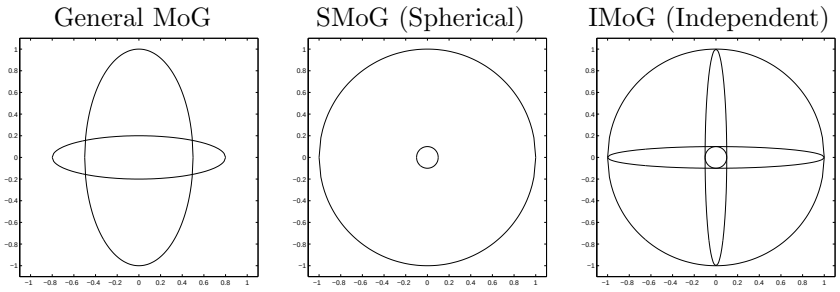


Figure 1: Contour plot of probability densities in 2D source space. and this is the only of the the three which in fact is independent in its variables, i.e. a Independent Mixture of Gaussians, IMoG.

[1], but where Attias is assuming completely general mixtures of gaussians and a noisy mixture model, the mixtures of gaussians in this paper are assumed centered for simplicity and, more importantly, the limit of zero noise is derived to be able to compare with the noiseless EBM.

The structure of the paper is as follows: In Sec 2, we introduce the reader to the mixture of gaussians to be used, and in Sec 3 we apply the mixtures of gaussians to the EBM and the ML. In Sec 4 we compare the results found and finally in Sec 5, we make a short summary.

## 2 Mixture of Gaussians

In this paper we make extensive use of the family of distributions known as mixtures of gaussians (MoG). To clear up some common misconceptions about MoG we introduce the general MoG and then discuss two important subsets of distributions. The density of a $D$-dimensional centered MoG is

$$p(\mathbf{s}) = \sum_{\kappa} \frac{\rho_\kappa}{\sqrt{|2\pi \mathbf{D}_\kappa|}} \exp\Big[ -\frac{1}{2} \mathbf{s}^T \mathbf{D}_\kappa^{-1} \mathbf{s} \Big] \tag{1}$$

where the weights $\rho_\kappa$ sum to one. The matrix $\mathbf{D}_\kappa$ can be any positive definite matrix, but in this context often assumed diagonal, $\mathbf{D}_\kappa = \mathrm{diag}(\sigma_{1\kappa}^2, \sigma_{2\kappa}^2, ..., \sigma_{D\kappa}^2)$. The marginal distribution for each $s_i$ is itself a mixture of gaussians, $s_i \sim \sum_\kappa \rho_\kappa \mathcal{N}(0, \sigma_{i\kappa}^2)$, but note about the joint distribution that even if the coordinates $s_i$ in each component are independent, i.e. if $\mathbf{D}_\kappa$ all are diagonal, the coordinates $s_i$ are *not* in general independent in the bigger joint distribution $p(\mathbf{s})$.

One subset of interest is the MoG constructed as a product of $D$ 1-dimensional mixtures of gaussians, and therefore independent by construction. Since the variables are independent we call this kind of MoG for Independent Mixtures of Gaussians (IMoG). Denoting the parameters of $i'th$ marginal by $\sum_\kappa \rho_{i\kappa} \mathcal{N}(0, \sigma_{i\kappa}^2)$, the density of the joint distribution is given by

$$p(\mathbf{s}) = \sum_{\{\mathbf{k}\}} \frac{\tilde{\rho}_{\mathbf{k}}}{\sqrt{|2\pi \mathbf{D}_{\mathbf{k}}|}} \exp\Big[ -\frac{1}{2} \mathbf{s}^T \mathbf{D}_{\mathbf{k}}^{-1} \mathbf{s} \Big] \tag{2}$$

The symbol $\{\mathbf{k}\}$ denotes all combinations of the vector $\mathbf{k}$ having length $D$, consisting of integers from 1 to $K$, $\mathbf{D_k} = \text{diag}(\sigma^2_{1k_1}, \sigma^2_{2k_2}, ..., \sigma^2_{Dk_D})$, and $\tilde{\rho}_\mathbf{k}$ is defined by $\tilde{\rho}_\mathbf{k} = \prod_{i=1}^{D} \rho_{ik_i}$.

Another subset of interest is the MoG, where the matrices are not only diagonal, but all some constant times the identity $\mathbf{D}_\kappa = \sigma^2_\kappa \mathbf{I}$. In this case the density is spherical symmetric and we call this kind of distributions for Spherical Mixtures of Gaussians (SMoG).

Obvious from Eq. 2, IMoG is itself a MoG, but the reverse is not true in general. As visualized on Figure 1, the density contour of a MoG with diagonal covariances is some weighted sum of axis-aligned ellipsoids. The length of axis of the ellipsoids corresponds to variances of gaussian components, i.e. the diagonal elements of the covariance matrices. This is also true for the product of 1-dimensional MoG, but in this case *all* combinations of the marginal variances are present as sets of axis of some ellipsoid. Thus knowing the density of a MoG with diagonal covariances, its variables are independent if and only if all combinations of axis lengths are present. Therefore, since this by definition is not the case for SMoG, SMoG and IMoG must be disjoint subsets of MoG.

# 3 Models and Derivation

We now use the MoG as source priors for two different approaches to ICA: Energy Based Models (EBM) and Maximum Likelihood (ML).

We consider a situation in which $D$ sources $\mathbf{s}_t$ are mixed into a set of $M$ measurements $\mathbf{x}_t$, expressed by the equation $\mathbf{x}_t = \mathbf{A}\mathbf{s}_t$. The signals are $N$ time steps long and can be arranged into the matrices $\mathbf{S}$ and $\mathbf{X}$, such that mixing of all $N$ vectors can be expressed in one equation $\mathbf{X} = \mathbf{AS}$. The sources are white and since EBM is restricted to square and overcomplete mixing, we assume $M \leq D$.

## 3.1 Energy Based Models

The EBM method, presented in [8], aim at demixing the measurements $\mathbf{X}$ by a filtering with $\mathbf{W}$ in the traditional way $\hat{\mathbf{S}} = \mathbf{WX}$. That this, in the overcomplete case, is not producing independent estimated sources, is discussed in more detail in Sec 4. The filtering coefficients is determined through a construction of a model distribution $p_\mathbf{W}(\mathbf{x})$ which is approximated to the data distribution

$$p_0(\mathbf{x}) = \frac{1}{N} \sum_{t=1}^{N} \delta(\mathbf{x} - \mathbf{x}_t) \tag{3}$$

through the Kullback-Leibler divergence. The model distribution is constructed in the following way: An energy is defined by $E(\mathbf{x}; \mathbf{W}) = -\ln p_s(\mathbf{Wx})$, which ensures the property that choosing $\mathbf{W}$ such that the resulting estimated sources are not too unlikely, is rewarded through low energy levels and unlikely source values penalized with high energy levels. Other definitions of energy could be made, but this is especially appealing due to its calculational properties. The

energy $E$ is used in a Gibbs distribution, i.e.

$$p_{\mathbf{W}}(\mathbf{x}) = \frac{e^{-E(\mathbf{x};\mathbf{W})}}{Z(\mathbf{W})} = \frac{p_s(\mathbf{W}\mathbf{x})}{\int p_s(\mathbf{W}\mathbf{x})d\mathbf{x}} \tag{4}$$

which is chosen as our model distribution. In [8], Teh et. al. make no assumption on the prior, which makes the normalization part more difficult. To deal with this they use so-called $n$-step learning, a variant of a hybrid monte carlo approach, do approximately estimate the normalization part of the optimization. In this paper we in stead choose the prior to be a MoG, which enables us to calculate the integral of Eq. 4 and obtain an closed form expression for the model distribution. The result is

$$p_{\mathbf{W}}(\mathbf{x}) = \sum_\kappa \gamma_\kappa \frac{\exp(-\frac{1}{2}\mathbf{x}^T\mathbf{W}^T\mathbf{D}_\kappa^{-1}\mathbf{W}\mathbf{x})}{\sqrt{|2\pi(\mathbf{W}^T\mathbf{D}_\kappa^{-1}\mathbf{W})^{-1}|}} \tag{5}$$

where $\gamma_\kappa$ in general is dependent on $\mathbf{W}$ and $\mathbf{D}_{\kappa'}$ for all $\kappa'$ through the following relation: Setting $\xi_\kappa = (|2\pi(\mathbf{W}^T\mathbf{D}_\kappa^{-1}\mathbf{W})^{-1}|/|2\pi\mathbf{D}_\kappa|)^{1/2}$, we can write $\gamma_\kappa = \rho_\kappa\xi_\kappa/\sum_{\kappa'}\rho_{\kappa'}\xi_{\kappa'}$. Note that in the square case $\xi_\kappa = 1$ and therefore $\gamma_\kappa = \rho_\kappa$ and in the case of SMoG priors we obtain $\xi_\kappa = (2\pi/\sigma_\kappa^2)^{(D-M)/2}$. The estimated optimal filtering matrix $\hat{\mathbf{W}}$ is determined by

$$\hat{\mathbf{W}} = \min_{\mathbf{W}}\Big[KL(p_{\mathbf{W}}||p_0)\Big]$$

in which the gradient of the KL-divergence can be calculated analytically when the prior is chosen to be mixtures of gaussians or some other analytically appealing distribution.

## 3.2 Maximum Likelihood

In the ML setup presented here we assume a generative model $\mathbf{X} = \mathbf{A}\mathbf{S} + \mathbf{\Gamma}$, where we, in order to be able to deal with the overcomplete case, have added white gaussian noise, $\mathbf{\Gamma}$. In the end we let the noise variance go to zero to obtain the noiseless result.

Assuming white gaussian noise and the prior on the sources $p_s(\mathbf{s})$ to be a MoG with covariance matrices $\mathbf{D}_\kappa$ and weights $\rho_\kappa$, we can complete the integration and write the distribution over $\mathbf{x}$ as

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{s})p(\mathbf{s})d\mathbf{s} = \sum_\kappa \rho_\kappa \frac{\sqrt{|2\pi\mathbf{\Phi}_\kappa^{-1}|}\exp(-\frac{1}{2}\mathbf{x}^T\mathbf{\Psi}_\kappa\mathbf{x})}{\sqrt{|2\pi\mathbf{\Sigma}|}\sqrt{|2\pi\mathbf{D}_\kappa|}}$$

where $\mathbf{\Sigma} = \sigma^2\mathbf{I}$ is the noise covariance matrix, $\mathbf{\Phi}_\kappa = \mathbf{A}^T\mathbf{\Sigma}^{-1}\mathbf{A} + \mathbf{D}_\kappa^{-1}$ and $\mathbf{\Psi}_\kappa = \mathbf{\Sigma}^{-1} - \mathbf{\Sigma}^{-1}\mathbf{A}\mathbf{\Phi}_\kappa^{-1}\mathbf{A}^T\mathbf{\Sigma}^{-1}$. We now want to consider the limit of $\sigma^2 \to 0$, but we need to do this with great care, since otherwise crucial details will vanish in the approximation. Using the Woodbury identity, singular value decomposition and some very good approximations (see the appendix for details), we obtain the following limits

$$\begin{aligned} \mathbf{\Psi}_\kappa &\to (\mathbf{A}\mathbf{D}_\kappa\mathbf{A}^T)^{-1} \\ \sqrt{|2\pi\mathbf{\Phi}_\kappa^{-1}|}/\sqrt{|2\pi\mathbf{\Sigma}|} &\to w_\kappa/\sqrt{|\mathbf{A}\mathbf{A}^T|} \end{aligned}$$

Here $w_\kappa$ is a constant with respect to $\sigma^2$, but depends on $\mathbf{D}_\kappa$ and on $\mathbf{A}$ through the unique orthogonal matrix $\mathbf{V}$. $\mathbf{V}$ which fulfills the eigen value equation $\mathbf{A}^T\mathbf{A}\mathbf{V} = \mathbf{V}\mathbf{\Lambda}$, such that the eigen values in $\mathbf{\Lambda}$ are decreasing in size down the diagonal. The expression for $w_\kappa$ is $w_\kappa = \prod_{i=M+1}^{D}(2\pi/(\mathbf{V}^T\mathbf{D}_\kappa^{-1}\mathbf{V})_{ii})^{1/2}$. With this limit taken care of, we can write the maximum likelihood expression for $p(\mathbf{x})$ as

$$p_{\mathbf{A}}(\mathbf{x}) = \sum_\kappa \rho_\kappa \frac{\exp(-\frac{1}{2}\mathbf{x}^T(\mathbf{A}\mathbf{D}_\kappa\mathbf{A}^T)^{-1}\mathbf{x})}{\sqrt{|2\pi\mathbf{A}\mathbf{D}_\kappa\mathbf{A}^T|}} \tag{6}$$

where the weights somewhat surprisingly turns out to be same as those of the prior (see the appendix for details). The estimated generative mixing matrix $\hat{\mathbf{A}}$ is the matrix maximizing the log likelihood

$$\hat{\mathbf{A}} = \max_{\mathbf{A}}\Big[\ln P(\mathbf{X}|\mathbf{A})\Big]$$

Note that we have not estimated the sources in this process, only the generative mixing matrix.

# 4 Comparing EBM and ML

Now we compare the EBM and ML approaches derived in the previous section and discuss the significance of the differences and similarities. In the square case, we obtain total equivalence of all expressions setting $\mathbf{W} = \mathbf{A}^{-1}$ and thus not surprisingly we can conclude, as Teh et .al. does in [8], that the two approaches are equivalent when the number of observations equal the number of sources. Therefore the discussion and comparison in this sections is almost entirely concerned with the overcomplete case.

In the overcomplete case it is not obvious how one should compare results on the generative $\mathbf{A}$ and the filtering matrix $\mathbf{W}$. The filtering approach does not retrieve the original sources, since for any matrix $\mathbf{W}$ we have $\mathbf{W}\mathbf{A}_g \neq \mathbf{I}$ because of the dimensionality: It is impossible to construct $D$ $M$-dimensional orthogonal matrices when $D > M$. And we cannot in general compare the filter matrix $\mathbf{W}$ with the pseudo-inverse of $\mathbf{A}$, since this is not the optimal solution in all cases [5]. But using MoG as prior, both the model distribution $p_{\mathbf{W}}(\mathbf{x})$ of the EBM and the loglikelihood $p_{\mathbf{A}}(\mathbf{x})$ of the ML becomes MoG's with parameters which must be estimated to fit a common data set $\mathbf{X}$. In fact we end up with two optimizations which look rather similar

$$0 = \frac{\partial}{\partial\mathbf{W}}\sum_{t=1}^{N}\ln p_{\mathbf{W}}(\mathbf{x}_t) \qquad 0 = \frac{\partial}{\partial\mathbf{A}}\sum_{t=1}^{N}\ln p_{\mathbf{A}}(\mathbf{x}_t)$$

The similarity is to some extend both genuine and deceptive: Both distributions are MoG, but the dependency of the weights and covariances on $\mathbf{W}$ and $\mathbf{A}$ are different. In this section we compare EBM and ML by comparing the covariances and weights of their MoG's.

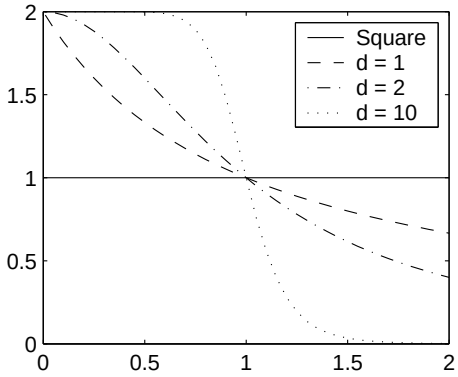


Figure 2: The Weights for EBM and ML. The plot demonstrates for SMoG priors that the differences in weights increase dramatically when the settings gets strongly overcomplete. (See the text for details on the plots)

## 4.1 Spherical MoG

We now consider the special case when the priors are SMoG, i.e. the variances are given as $\mathbf{D}_\kappa = \sigma_\kappa^2 \mathbf{I}$. In this case the sources are not assumed to be independent, which is interesting in its on right, but it also serves as a clear example of properties which holds for the more general cases as well. When we assume the priors to be SMoG, the model distribution simplifies significantly,

$$p_{\mathbf{W}}(\mathbf{x}) = \sum_\kappa \gamma_\kappa \frac{\exp(-\frac{1}{2}\mathbf{x}^T\mathbf{W}^T\mathbf{D}_\kappa^{-1}\mathbf{W}\mathbf{x})}{\sqrt{|2\pi(\mathbf{W}^T\mathbf{D}_\kappa^{-1}\mathbf{W})^{-1}|}}, \qquad \gamma_\kappa = \frac{\rho_\kappa (1/\sigma_\kappa^2)^d}{\sum_{\kappa'} \rho_{\kappa'} (1/\sigma_{\kappa'}^2)^d} \tag{7}$$

where $d = (D - M)/2$.

### The Covariances

The covariances of EBM and ML respectively can in fact become equal in the case of SMoG. The equation setting the covariances equal

$$(\mathbf{W}^T\mathbf{D}_\kappa^{-1}\mathbf{W})^{-1} = \mathbf{A}\mathbf{D}_\kappa\mathbf{A}^T \qquad \forall\kappa \tag{8}$$

translates into $\mathbf{W}^T\mathbf{W}\mathbf{A}\mathbf{A}^T = \mathbf{I}$, which is fulfilled for $\mathbf{W} = \mathbf{A}^+$, where $\mathbf{A}^+$ denotes the pseudo inverse (Moore-Penrose) of the matrix $\mathbf{A}$. When $\mathbf{A}$ has full rank, the pseudo-inverse is given by $\mathbf{A}^+ = \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}$. But the equation is also fulfilled for any matrix $\mathbf{W} = \mathbf{U}\mathbf{A}^+$, where $\mathbf{U}$ is orthogonal and thus, there is an entire family of matrices which would make the covariances of the EBM equal to the covariances of the ML for a given $\mathbf{A}$. Conversely for any $\mathbf{W}$ we can choose $\mathbf{A} = \mathbf{W}^+$ to obtain the same covariances and and in this sense the two approaches have equal flexibility with respect to adjusting the covariances to the data. This is evident in the $4 \times 2$ example in Fig 3 a) and b).

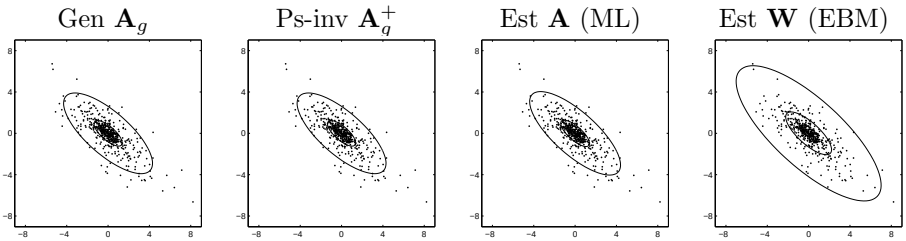


Figure 3: The covariances in case of SMoG priors for different settings. Plot a) Generative $\mathbf{A}$. Plot b) Pseudo-inverse of the generative $\mathbf{A}$. Plot c) Estimated $\mathbf{A}$. Plot d) Estimated $\mathbf{W}$.

**The Weights**

The weights can be compared by examining the ratio $\gamma_\kappa/\rho_\kappa$ and as we shall see, this ratio differs strongly from 1 in most cases. From the expression of $\gamma_\kappa$ in Eq. 7, we see that the constrain making the weights of ML and EBM equal is

$$\frac{(1/\sigma_\kappa^2)^d}{\sum_{\kappa'} \rho_{\kappa'}(1/\sigma_{\kappa'}^2)^d} \stackrel{?}{=} 1 \qquad \forall \kappa$$

which is clearly impossible when the weights $\sigma_\kappa^2$ must be different for different $\kappa$. Thus, in the SMoG case, the weights of EBM and ML cannot be equal and furthermore the ratio $\gamma_\kappa/\rho_\kappa$ becomes relatively large for those $\kappa$ where $\sigma_\kappa^2$ is very small and vice versa.

Fig 2 is a general illustration of this. Here we assume a SMoG prior with two components which has $\sigma_1^2 = 1$ and $\rho_1 = \rho_2 = 0.5$ and let $\sigma_2^2$ vary between 0 and 2. The resulting ratio $\gamma_2/\rho_2$ is shown in Fig 2: The $x$-axis is $\sigma_2^2$, the $y$-axis is $\gamma_2/\rho_2$ and the 4 curves are plottet for different degrees of overcompleteness $d$=0 (Square) and $d = 1, 2, 10$ (Overcomplete). Clearly only in the square case and for $\sigma_2^2$ values close to 1, is the ratio reasonably close to 1.

Another more specific example is shown in Fig 3 c) and d), which is the estimated covariances for ML and EBM in a $4 \times 2$ case. In this example the effect of the enhanced weight on smaller covariances is clear: When the weights of the smaller covariance is strong, the points far from origin is considered extreme, and the covariances are expanded accordingly. Thus, for this reason, EBM seem to favor larger covariances compared to ML.

# 5 Summary and Acknowledgements

Conclusively the use of Mixtures of Gaussians made it possible to compare Maximum Likelihood with Energy Based Models. The results show that in the overcomplete case with Spherical Mixtures of Gaussians as priors, the Energy Based Model is biased toward larger covariances compared to the Maximum Likelihood. One can show that this effect is also present for IMoG priors, though not at all as strong.

Finally I need to give due credit: This paper is closely related to the result of earlier work together with Jiucang Hao and Te-Won Lee from University of California San Diego (UCSD).

# A Details of the Calculations

When calculating the limit of $\mathbf{\Psi}_\kappa$, we first set in the definition $\mathbf{\Sigma} = \sigma^2\mathbf{I}$ and simplify the expression into

$$\mathbf{\Psi}_\kappa = \frac{1}{\sigma^2}\left\{\mathbf{I} - \mathbf{A}(\mathbf{A}^T\mathbf{A} + \sigma^2\mathbf{D}_\kappa^{-1})\mathbf{A}^T\right\}$$

Now defining $\mathbf{Q} = \mathbf{A}\mathbf{D}_\kappa\mathbf{A}^T/\sigma^2$ we can use Woodburys identity for inverse matrices and obtain

$$\mathbf{A}(\mathbf{A}^T\mathbf{A} + \sigma^2\mathbf{D}_\kappa^{-1})\mathbf{A}^T = \mathbf{Q} - \mathbf{Q}(\mathbf{I} + \mathbf{Q})^{-1}\mathbf{Q}$$

Since $\mathbf{Q}$ is symmetric and very large compared to $\mathbf{I}$, the right hand side can be approximated by $\mathbf{I} - \mathbf{Q}^{-1}$. To see this, write $\mathbf{Q}$ as $\mathbf{Q} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$, for an orthogonal $\mathbf{V}$ and diagonal $\mathbf{\Lambda}$ and remember the identity $x - x^2/(1+x) = 1 - 1/(1+x)$ to obtain $\mathbf{I} - \mathbf{V}\mathrm{diag}(1/(1 + \Lambda_{ii}/\sigma^2))\mathbf{V}^T \cong \mathbf{I} - \mathbf{V}\mathrm{diag}(1/(\Lambda_{ii}/\sigma^2))\mathbf{V}^T = \mathbf{I} - \mathbf{Q}^{-1}$. Inserting this into the equation containing $\mathbf{\Psi}_\kappa$, gives the desired result.

When calculating the limit of the fraction containing the determinant $|2\pi\mathbf{\Phi}_\kappa^{-1}|$, we use the fact that since $\mathbf{A}^T\mathbf{A}$ is symmetric there exists an orthogonal $\mathbf{V}$ such that $\mathbf{A}^T\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$. Since $|\mathbf{V}| = 1$ we get

$$|2\pi\mathbf{\Phi}_\kappa^{-1}|/|2\pi\mathbf{\Sigma}| = (2\pi\sigma^2)^{(D-M)}/|\mathbf{\Lambda} + \sigma^2\mathbf{V}^T\mathbf{D}^{-1}\mathbf{V}|$$

Since $\sigma^2$ are assumed arbitrary small, only the diagonal of the sum of matrices will contributed significantly to the determinant. And since further more $\mathbf{A}^T\mathbf{A}$ has rank $M$, the $M$ first elements of the diagonal matrix will dominate together with the remaining $M - D$ factors

$$|\mathbf{\Lambda} + \sigma^2\mathbf{V}^T\mathbf{D}^{-1}\mathbf{V}| \cong \prod_{i=1}^{M}\mathbf{\Lambda}_{ii}\prod_{j=M+1}^{D}(\sigma^2\mathbf{V}^T\mathbf{D}^{-1}\mathbf{V})_{jj}$$

inserting this into the fraction above gives the desired result.

The coefficients $\alpha_\kappa$ has the structure $\rho_\kappa w_\kappa\sqrt{|2\pi\mathbf{A}\mathbf{D}_\kappa\mathbf{A}^T|}/\sqrt{|\mathbf{A}\mathbf{A}^T|\cdot|2\pi\mathbf{D}_\kappa|}$. In the square case much of the difficulties of taking the noiseless limit disappears, $w_\kappa = 1$ and we easily obtain $\alpha_\kappa = \rho_\kappa$. In the case of SMoG, setting $\mathbf{D}_\kappa = \sigma_\kappa^2\mathbf{I}$, we obtain $w_\kappa = (2\pi\sigma_\kappa^2)^{(D-M)/2}$ and therefore $\alpha_\kappa = \rho_\kappa$. Supported by numerical results we conjecture that this is also the case for the general overcomplete case.

# References


[1] H. Attias, *Independent Factor Analysis*, Neural Computation 11, 803-851, 1999.

[2] M. Girolami, *A Variational Method for Learning Sparse and Overcomplete Representations*, Neural Computation, 13(11), pp 2517 - 2532, 2001.

[3] G. E. Hinton, *Training products of Experts by minimizing contrastive divergence*, Neural Computation, 14(8):1771-1800, 2002.

[4] P. Hojen-Sorensen, O. Winther, L. K. Hansen, *Mean-Field Approaches to Independent Component Analysis*, Neural Computation Volume 14, Issue 4, April 2002.

[5] M. S. Lewicki, T. J. Sejnowski, *Learning Overcomplete Representations*, Neural Computation, 12(2):337-65, 2000.

[6] R. M. Neal, *Probabilistic Inference Using Markov Chain Monte Carlo Methods*, Techinal Report CRG-TR-93-I, Department of Computer Science, ujniversity of Toronto, 1993.

[7] O. Shriki, H. Sampolinski, D.D. Lee, *An information maximization approach to overcomplete and recurrent representations* Neural Info. Proc. Sys. 13 (2001).

[8] Y. W. Teh, M. Welling, S. Osindero, G. Hinton *"Energy-Based Models for Sparse Overcomplete Representations"*, The Journal of Machine Learning Research - Special Issue on Independent Components Analysis, guest edited by Te-Won Lee, Jean-Franois Cardoso, Erkki Oja and Shun-ichi Amari. 2003.