Journal of Machine Learning Research 6 (2005) 695–709 Submitted 11/04; Revised 3/05; Published 4/05

Estimation of Non-Normalized Statistical Models by Score Matching

Aapo Hyvärinen AAPO.HYVARINEN@HELSINKI.FI
Helsinki Institute for Information Technology (BRU)
Department of Computer Science
FIN-00014 University of Helsinki, Finland

Editor: Peter Dayan

Abstract

One often wants to estimate statistical models where the probability density function is known only up to a multiplicative normalization constant. Typically, one then has to resort to Markov Chain Monte Carlo methods, or approximations of the normalization constant. Here, we propose that such models can be estimated by minimizing the expected squared distance between the gradient of the log-density given by the model and the gradient of the log-density of the observed data. While the estimation of the gradient of log-density function is, in principle, a very difficult non-parametric problem, we prove a surprising result that gives a simple formula for this objective function. The density function of the observed data does not appear in this formula, which simplifies to a sample average of a sum of some derivatives of the log-density given by the model. The validity of the method is demonstrated on multivariate Gaussian and independent component analysis models, and by estimating an overcomplete filter set for natural image data.

Keywords: statistical estimation, non-normalized densities, pseudo-likelihood, Markov chain Monte Carlo, contrastive divergence

1. Introduction

In many cases, probabilistic models in machine learning, statistics, or signal processing are given in the form of non-normalized probability densities. That is, the model contains an unknown normalization constant whose computation is too difficult for practical purposes.

Assume we observe a random vector $\mathbf{x} \in \mathbb{R}^n$ which has a probability density function (pdf) denoted by $p_{\mathbf{x}}(.)$. We have a parametrized density model $p(.;\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is an m-dimensional vector of parameters. We want to estimate the parameter $\boldsymbol{\theta}$ from $\mathbf{x}$, i.e. we want to approximate $p_{\mathbf{x}}(.)$ by $p(.;\hat{\boldsymbol{\theta}})$ for the estimated parameter value $\hat{\boldsymbol{\theta}}$. (We shall here consider the case of continuous-valued variables only.)

The problem we consider here is that we only are able to compute the pdf given by the model up to a multiplicative constant $Z(\boldsymbol{\theta})$:

$$p(\boldsymbol{\xi};\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} q(\boldsymbol{\xi};\boldsymbol{\theta}).$$

That is, we do know the functional form of q as an analytical expression (or any form that can be easily computed), but we do *not* know how to easily compute Z which is given by

©2005 Aapo Hyvärinen.

an integral that is often analytically intractable:

$$Z(\boldsymbol{\theta}) = \int_{\boldsymbol{\xi} \in \mathbb{R}^n} q(\boldsymbol{\xi}; \boldsymbol{\theta}) \, d\boldsymbol{\xi}.$$

In higher dimensions (in fact, for almost any $n > 2$), the numerical computation of this integral is practically impossible as well.

Usually, estimation of non-normalized models is approached by Markov Chain Monte Carlo (MCMC) methods, which are very slow, or by making some approximations, which may be quite poor (Mackay, 2003).

Non-normalized models are often encountered in continous-valued Markov random fields, which are widely used in image modelling, see e.g. (Bouman and Sauer, 1993; Li, 2001). In general, undirected graphical models cannot be normalized except in the Gaussian case. Other recent work in image modelling also includes non-normalized models (Hyvärinen and Hoyer, 2001; Teh et al., 2003). Presumably, the number of useful applications for non-normalized models is much larger than the present literature suggests. Non-normalized models have been avoided because their estimation has been considered too difficult; the advent of efficient estimation methods may significantly increase their utility.

In this paper, we propose a simple method for estimating such non-normalized models. This is based on minimizing the expected squared distance of the score function of $\mathbf{x}$ and the score function given by the model. (By score function, we mean here the gradient of log-density.) We show that this distance can be estimated by a very simple formula involving only sample averages of some derivatives of the logarithm of the pdf given by the model. Thus, the computations involved are essentially not more complicated than in the case where we know an analytical expression for the normalization constant. The proposed formula is exact and does not involve any approximations, which is why we are able to prove the local consistency of the resulting method. Minimization of the proposed objective function thus provides an estimation method that is computationally simple yet statistically locally consistent.

2. Estimation by Score Matching

In the following, we use extensively the gradient of the log-density with respect to the data vector. For simplicity, we call this the score function, although according the conventional definition, it is actually the score function with respect to a hypothetical location parameter (Schervish, 1995). For the model density, we denote the score function by $\boldsymbol{\psi}(\boldsymbol{\xi}; \boldsymbol{\theta})$:

$$\boldsymbol{\psi}(\boldsymbol{\xi}; \boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial \log p(\boldsymbol{\xi};\boldsymbol{\theta})}{\partial \xi_1} \\ \vdots \\ \frac{\partial \log p(\boldsymbol{\xi};\boldsymbol{\theta})}{\partial \xi_n} \end{pmatrix} = \begin{pmatrix} \psi_1(\boldsymbol{\xi}; \boldsymbol{\theta}) \\ \vdots \\ \psi_n(\boldsymbol{\xi}; \boldsymbol{\theta}) \end{pmatrix} = \nabla_{\boldsymbol{\xi}} \log p(\boldsymbol{\xi}; \boldsymbol{\theta}).$$

The point in using the score function is that it does not depend on $Z(\boldsymbol{\theta})$. In fact we obviously have

$$\boldsymbol{\psi}(\boldsymbol{\xi}; \boldsymbol{\theta}) = \nabla_{\boldsymbol{\xi}} \log q(\boldsymbol{\xi}; \boldsymbol{\theta}). \tag{1}$$

Likewise, we denote by $\boldsymbol{\psi}_{\mathbf{x}}(.) = \nabla_{\boldsymbol{\xi}} \log p_{\mathbf{x}}(.)$ the score function of the distribution of observed data $\mathbf{x}$. This could in principle be estimated by computing the gradient of the logarithm of a non-parametric estimate of the pdf—but we will see below that no such computation is necessary. Note that score functions are mappings from $\mathbb{R}^n$ to $\mathbb{R}^n$.

We now propose that the model is estimated by minimizing the expected squared distance between the model score function $\boldsymbol{\psi}(.;\boldsymbol{\theta})$ and the data score function $\boldsymbol{\psi}_{\mathbf{x}}(.)$. We define this squared distance as

$$J(\boldsymbol{\theta}) = \frac{1}{2}\int_{\boldsymbol{\xi}\in\mathbb{R}^n} p_{\mathbf{x}}(\boldsymbol{\xi})\|\boldsymbol{\psi}(\boldsymbol{\xi};\boldsymbol{\theta}) - \boldsymbol{\psi}_{\mathbf{x}}(\boldsymbol{\xi})\|^2 d\boldsymbol{\xi}. \tag{2}$$

Thus, our *score matching* estimator of $\boldsymbol{\theta}$ is given by

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} J(\boldsymbol{\theta}).$$

The motivation for this estimator is that the score function can be directly computed from q as in (1), and we do not need to compute Z. However, this may still seem to be a very difficult way of estimating $\boldsymbol{\theta}$, since we might have to compute an estimator of the data score function $\boldsymbol{\psi}_{\mathbf{x}}$ from the observed sample, which is basically a non-parametric estimation problem. However, no such non-parametric estimation is needed. This is because we can use a simple trick of partial integration to compute the objective function very easily, as shown by the following theorem:

Theorem 1 *Assume that the model score function $\boldsymbol{\psi}(\boldsymbol{\xi};\boldsymbol{\theta})$ is differentiable, as well as some weak regularity conditions.*[1]

Then, the objective function J in (2) can be expressed as

$$J(\boldsymbol{\theta}) = \int_{\boldsymbol{\xi}\in\mathbb{R}^n} p_{\mathbf{x}}(\boldsymbol{\xi}) \sum_{i=1}^{n} \left[\partial_i \psi_i(\boldsymbol{\xi};\boldsymbol{\theta}) + \frac{1}{2}\psi_i(\boldsymbol{\xi};\boldsymbol{\theta})^2\right] d\boldsymbol{\xi} + \textit{const.} \tag{3}$$

where the constant does not depend on $\boldsymbol{\theta}$,

$$\psi_i(\boldsymbol{\xi};\boldsymbol{\theta}) = \frac{\partial \log q(\boldsymbol{\xi};\boldsymbol{\theta})}{\partial \xi_i}$$

is the i-th element of the model score function, and

$$\partial_i \psi_i(\boldsymbol{\xi};\boldsymbol{\theta}) = \frac{\partial \psi_i(\boldsymbol{\xi};\boldsymbol{\theta})}{\partial \xi_i} = \frac{\partial^2 \log q(\boldsymbol{\xi};\boldsymbol{\theta})}{\partial \xi_i^2}$$

is the partial derivative of the i-th element of the model score function with respect to the i-th variable.

The proof, given in the Appendix, is based a simple trick of partial integration that has previously been used in the theory of independent component analysis for modelling the densities of the independent components (Pham and Garrat, 1997).

We have thus proven the remarkable fact that the squared distance of the model score function from the data score function can be computed as a simple expectation of certain

[1] Namely: the data pdf $p_{\mathbf{x}}(\boldsymbol{\xi})$ is differentiable, the expectations $E_{\mathbf{x}}\{\|\boldsymbol{\psi}(\mathbf{x};\boldsymbol{\theta})\|^2\}$ and $E_{\mathbf{x}}\{\|\boldsymbol{\psi}_{\mathbf{x}}(\mathbf{x})\|^2\}$ are finite for any $\boldsymbol{\theta}$, and $p_{\mathbf{x}}(\boldsymbol{\xi})\boldsymbol{\psi}(\boldsymbol{\xi};\boldsymbol{\theta})$ goes to zero for any $\boldsymbol{\theta}$ when $\|\boldsymbol{\xi}\| \to \infty$.

functions of the non-normalized model pdf. If we have an analytical expression for the non-normalized density function q, these functions are readily obtained by derivation using (1) and taking further derivatives.

In practice, we have T observations of the random vector $\mathbf{x}$, denoted by $\mathbf{x}(1), \ldots, \mathbf{x}(T)$. The sample version of J is obviously obtained from (3) as

$$\tilde{J}(\boldsymbol{\theta}) = \frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{n} \left[\partial_i \psi_i(\mathbf{x}(t); \boldsymbol{\theta}) + \frac{1}{2} \psi_i(\mathbf{x}(t); \boldsymbol{\theta})^2 \right] + \text{const.} \tag{4}$$

which is asymptotically equivalent to J due to the law of large numbers. We propose to estimate the model by minimization of $\tilde{J}$ in the case of a real, finite sample.

One may wonder whether it is enough to minimize J to estimate the model, or whether the distance of the score functions can be zero for different parameter values. Obviously, if the model is degenerate in the sense that two different values of $\boldsymbol{\theta}$ give the same pdf, we cannot estimate $\boldsymbol{\theta}$. If we assume that the model is not degenerate, and that $q > 0$ always, we have local consistency as shown by the following theorem and the corollary:

Theorem 2 *Assume the pdf of $\mathbf{x}$ follows the model: $p_{\mathbf{x}}(.) = p(.; \boldsymbol{\theta}^*)$ for some $\boldsymbol{\theta}^*$. Assume further that no other parameter value gives a pdf that is equal*[2] *to $p(.; \boldsymbol{\theta}^*)$, and that $q(\boldsymbol{\xi}; \boldsymbol{\theta}) > 0$ for all $\boldsymbol{\xi}, \boldsymbol{\theta}$. Then*

$$J(\boldsymbol{\theta}) = 0 \Leftrightarrow \boldsymbol{\theta} = \boldsymbol{\theta}^*.$$

For a proof, see the Appendix.

Corollary 3 *Under the assumptions of the preceding Theorems, the score matching estimator obtained by minimization of $\tilde{J}$ is consistent, i.e. it converges in probability towards the true value of $\boldsymbol{\theta}$ when sample size approaches infinity, assuming that the optimization algorithm is able to find the global minimum.*

The corollary is proven by applying the law of large numbers. As sample size approaches infinity, $\tilde{J}$ converges to J (in probability). Thus, the estimator converges to a point where J is globally minimized. By Theorem 2, the global minimum is unique and found at the true parameter value (obviously, J cannot be negative).

This result of consistency assumes that the global minimum of $\tilde{J}$ is found by the optimization algorithm used in the estimation. In practice, this may not be true, in particular because there may be several local minima. Then, the consistency is of local nature, i.e., the estimator is consistent if the optimization iteration is started sufficiently close to the true value. Note that consistency implies asymptotic unbiasedness.

3. Examples

Here, we provide three simulations to illustrate how score matching works, as well as to confirm its consistency and applicability to real data.

[2] In this theorem and its proof, equalities of pdf's are to be taken in the sense of equal almost everywhere with respect to the Lebesgue measure.

3.1 Multivariate Gaussian Density

As a very simple illustrative example, we consider estimation of the parameters of the multivariate Gaussian density.

3.1.1 Estimation

The probability density function is given by

$$p(\mathbf{x};\mathbf{M},\boldsymbol{\mu}) = \frac{1}{Z(\mathbf{M},\boldsymbol{\mu})}\exp(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{M}(\mathbf{x}-\boldsymbol{\mu})),$$

where $\mathbf{M}$ is a symmetric positive-definite matrix (the inverse of the covariance matrix). Of course, the expression for Z is well-known in this case, but this serves as an illustration of the method. As long as there is no chance of confusion, we use $\mathbf{x}$ here as the general n-dimensional vector. Thus, here we have

$$q(\mathbf{x}) = \exp(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{M}(\mathbf{x}-\boldsymbol{\mu})), \tag{5}$$

and we obtain

$$\boldsymbol{\psi}(\mathbf{x};\mathbf{M},\boldsymbol{\mu}) = -\mathbf{M}(\mathbf{x}-\boldsymbol{\mu}),$$

and

$$\partial_i\boldsymbol{\psi}(\mathbf{x};\mathbf{M},\boldsymbol{\mu}) = -m_{ii}.$$

Thus, we obtain

$$\tilde{J}(\mathbf{M},\boldsymbol{\mu}) = \frac{1}{T}\sum_{t=1}^{T}[\sum_i -m_{ii} + \frac{1}{2}(\mathbf{x}(t)-\boldsymbol{\mu})^T\mathbf{M}\mathbf{M}(\mathbf{x}(t)-\boldsymbol{\mu})]. \tag{6}$$

To minimize this with respect to $\boldsymbol{\mu}$, it is enough to compute the gradient

$$\nabla_{\boldsymbol{\mu}}\tilde{J} = \mathbf{M}\mathbf{M}\boldsymbol{\mu} - \mathbf{M}\mathbf{M}\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}(t),$$

which is obviously zero if and only if $\boldsymbol{\mu}$ is the sample average $\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}(t)$. This is truly a minimum because the matrix $\mathbf{MM}$ that defines the quadratic form is positive-definite.

Next, we compute the gradient with respect to $\mathbf{M}$, which gives

$$\nabla_{\mathbf{M}}\tilde{J} = -\mathbf{I} + \mathbf{M}\frac{1}{2T}\sum_{t=1}^{T}(\mathbf{x}(t)-\boldsymbol{\mu})(\mathbf{x}(t)-\boldsymbol{\mu})^T + \frac{1}{2T}[\sum_{t=1}^{T}(\mathbf{x}(t)-\boldsymbol{\mu})(\mathbf{x}(t)-\boldsymbol{\mu})^T]\mathbf{M},$$

which is zero if and only if $\mathbf{M}$ is the inverse of the sample covariance matrix $\frac{1}{T}\sum_{t=1}^{T}(\mathbf{x}(t)-\boldsymbol{\mu})(\mathbf{x}(t)-\boldsymbol{\mu})^T$, which thus gives the score matching estimate.

Interestingly, we see that score matching gives exactly the same estimator as maximum likelihood estimation. In fact, the estimators are identical for any sample (and not just asymptotically). The maximum likelihood estimator is known to be consistent, so the score matching estimator is consistent as well.

3.1.2 Intuitive Interpretation

This example also gives some intuitive insight into the principle of score matching. Let us consider what happened if we just maximized the non-normalized log-likelihood, i.e., log of q in (5). It is maximized when the scale parameters in $\mathbf{M}$ are zero, i.e., the model variances are infinite and the pdf is completely flat. This is because then the model assigns the same probability to all possible values of $\mathbf{x}(t)$, which is equal to 1. In fact, the same applies to the second term in (6), which thus seems to be closely connected to maximization of the non-normalized log-likelihood.

Therefore, the first term in (3) and (6), involving second derivatives of the logarithm of q, seems to act as a kind of a normalization term. Here it is equal to $-\sum_i m_{ii}$. To minimize this, the m_{ii} should be made as large (and positive) as possible. Thus, this term has the opposite effect to the second term. Since the first term is linear and the second term polynomial in $\mathbf{M}$, the minimum of the sum is different from zero.

A similar interpretation applies to the general non-Gaussian case. The second term in (3), expectation of the norm of score function, is closely related to maximization of non-normalized likelihood: if the norm of this gradient is zero, then in fact the data point is in a local extremum of the non-normalized log-likelihood. The first term then measures what kind of an extremum this is. If it is a minimum, the first term is positive and the value of J is increased. To minimize J, the first term should be negative, in which case the extremum is a maximum. In fact, the extremum should be as steep a maximum (as opposed to a flat maximum) as possible to minimize J. This counteracts, again, the tendency to assign the same probability to all data points that is often inherent in the maximization of the non-normalized likelihood.

3.2 Estimation of Basic Independent Component Analysis Model

Next, we show the validity of score matching in estimating the following model

$$\log p(\mathbf{x}) = \sum_{k=1}^{n} G(\mathbf{w}_k^T \mathbf{x}) + Z(\mathbf{w}_1, \ldots, \mathbf{w}_n), \tag{7}$$

which is the basic form of the independent component analysis (ICA) model. Again, the normalization constant is well-known and equal to $-\log|\det \mathbf{W}|$ where the matrix $\mathbf{W}$ has the vectors $\mathbf{w}_i$ as rows, but this serves as an illustration of our method.

The nice thing about this model is that we can easily generate data that follows this model. In fact, if latent variables $s_i, i = 1 \ldots, n$ are independently distributed and have the pdf given by $\exp(G(s_i))$, the linear transformation

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{8}$$

with $\mathbf{A} = \mathbf{W}^{-1}$ follows the pdf's given in (7), see e.g. (Hyvärinen et al., 2001). Thus, we will be estimating the generative model in (8) using the non-normalized likelihood in (7).

Here, we choose the distribution of the components s_i to be so-called logistic with

$$G(s) = -2 \log \cosh(\frac{\pi}{2\sqrt{3}} s) - \log 4.$$

This distribution is normalized to unit variance as typical in the theory of ICA. The score function of the model in (7 is given by

$$\boldsymbol{\psi}(\mathbf{x};\mathbf{W}) = \sum_{k=1}^{n} \mathbf{w}_k g(\mathbf{w}_k^T \mathbf{x}), \tag{9}$$

where the scalar nonlinear function g is given by

$$g(s) = -\frac{\pi}{3}\tanh(\frac{\pi}{2\sqrt{3}}s).$$

The relevant derivatives of the score function are given by:

$$\partial_i \psi_i(x) = \sum_{k=1}^{n} w_{ki}^2 g'(\mathbf{w}_k^T \mathbf{x}),$$

and the sample version of the objective function $\tilde{J}$ is given by

$$\begin{aligned}\tilde{J} &= \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{n}\left[\sum_{k=1}^{n} w_{ki}^2 g'(\mathbf{w}_k^T\mathbf{x}(t)) + \frac{1}{2}\sum_{j=1}^{n} w_{ji} g(\mathbf{w}_j^T\mathbf{x}(t))\sum_{k=1}^{n} w_{ki} g(\mathbf{w}_k^T\mathbf{x}(t))\right]\\ &= \sum_{k=1}^{n}\|\mathbf{w}_k\|^2\frac{1}{T}\sum_{t=1}^{T} g'(\mathbf{w}_k^T\mathbf{x}(t)) + \frac{1}{2}\sum_{j,k=1}^{n}\mathbf{w}_j^T\mathbf{w}_k\frac{1}{T}\sum_{t=1}^{T} g(\mathbf{w}_k^T\mathbf{x}(t))g(\mathbf{w}_j^T\mathbf{x}(t)).\end{aligned} \tag{10}$$

We performed simulations to validate the consistency of score matching estimation, and to compare its efficiency with respect to maximum likelihood estimation. We generated data following the model as described above, where the dimension was chosen to be $n = 4$. Score matching estimation consisted of minimizing $\tilde{J}$ in (10) by a simple gradient descent; likelihood was maximized using a natural gradient method (Amari et al., 1996; Hyvärinen et al., 2001), using the true value of Z. We repeated the estimation for several different sample sizes: 500, 1000, 2000, 4000, 8000, and 16000. For each sample size, the estimation was repeated 11 times using different random initial points in the optimization, and different random data sets. For each estimate, a measure of asymptotic variance was computed as follows. The matrix $\hat{\mathbf{W}}\mathbf{A}$, where $\hat{\mathbf{W}}$ is the estimate was normalized row-by-row so that the largest value on each row had an absolute value of 1. Then, the sum of squares of all the elements was computed, and 4 (i.e. the sum of the squares of the four elements equal to one) was subtracted. This gives a measure of the squared error of the estimate (we cannot simply compare $\hat{\mathbf{W}}\mathbf{A}$ with identity because the order of the components is not well-defined). For each sample size and estimator type (score matching vs. maximum likelihood) we then computed the median error.

Figure 1 shows the results. The error of score matching seems to go to zero, which validates the theoretical consistency result of Theorem 2. Score matching gives slightly larger errors than maximum likelihood, which is to be expected because of the efficiency results of maximum likelihood estimation (Pham and Garrat, 1997).

In the preceding simulation, we knew exactly the proper function g to be used in the score function. To investigate the robustness of the method to misspecification of the score

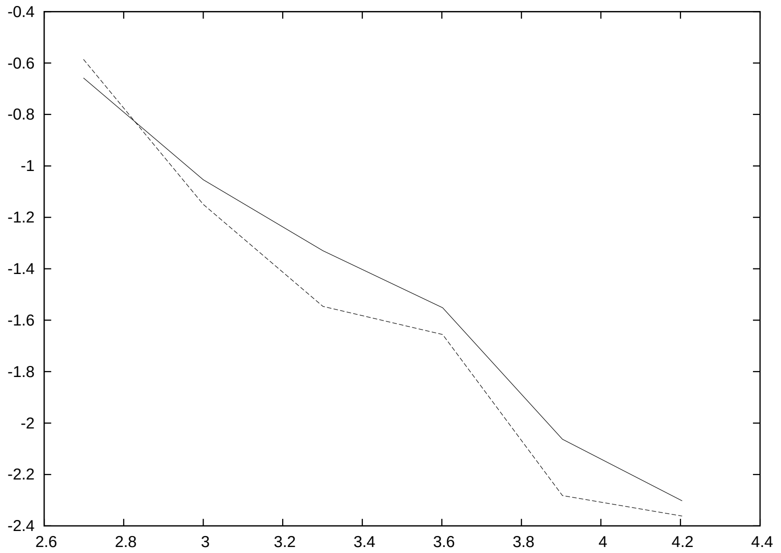

Figure 1: The estimation errors of score matching (solid line) compared with errors of maximum likelihood estimation (dashed line) for the basic ICA model. Horizontal axis: $\log_{10}$ of sample size. Vertical axis: $\log_{10}$ of estimation error.

function (a well-known problem in ICA estimation), we ran the same estimation methods, score matching and maximum likelihood, for data that was generated by a slightly different distribution. Specifically, we generated the data so that the independent components s_i had Laplacian distributions of unit variance (Hyvärinen et al., 2001). We then estimated the model using exactly the same g as before, which was not theoretically correct. The estimation errors are shown in Figure 2. We see that score matching still seems consistent. Interestingly, it now performs slightly better than maximum likelihood estimation (which would more properly be called quasi-maximum likelihood estimation due to the misspecification (Pham and Garrat, 1997)).

3.3 Estimation of an Overcomplete Model for Image Data

Finally, we show image analysis results using an overcomplete version of the ICA model. The likelihood is defined almost as in (7), but the number of components m is larger than the dimension of the data n, see e.g. (Teh et al., 2003), and we introduce some extra parameters. The likelihood is given by

$$\log p(\mathbf{x}) = \sum_{k=1}^{m} \alpha_k G(\mathbf{w}_k^T \mathbf{x}) + Z(\mathbf{w}_1, \dots, \mathbf{w}_n, \alpha_1, \dots, \alpha_n), \tag{11}$$

where the vectors $\mathbf{w}_k = (w_{k1}, \dots, w_{kn})$ are constrained to unit norm (unlike in the preceding example), and the α_k are scaling parameters. We introduce here the extra parameters α_k to account for different distributions for different projections. Constraining $\alpha_k = 1$ and $m = n$

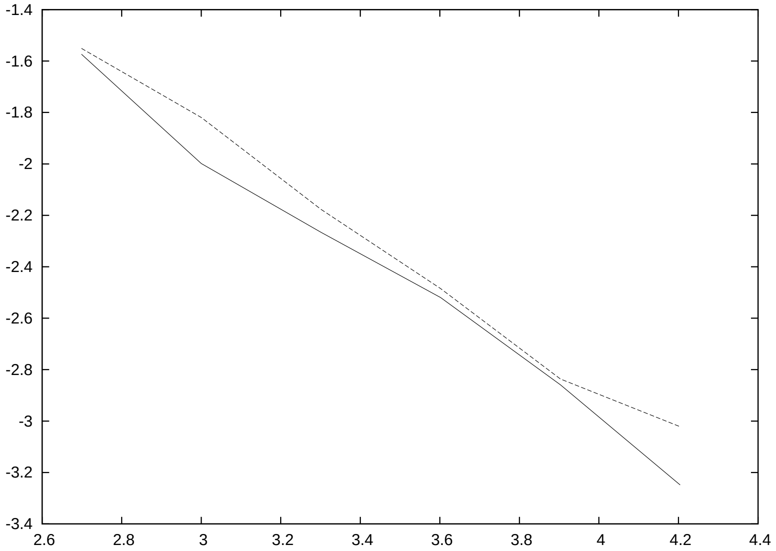

Figure 2: The estimation errors of score matching compared with errors of maximum likelihood estimation for the basic ICA model. This time, the pdf of the independent components was slightly misspecified. Legend as in Fig. 1.

and allowing the $\mathbf{w}_k$ to have any norm, this becomes the basic ICA model of the preceding subsection.

The model is related to ICA with overcomplete bases (Hyvärinen et al., 2001; Hyvärinen and Inki, 2002; Olshausen and Field, 1997), i.e. the case where there are more independent components and basis vectors than observed variables. In contrast to most ICA models, the overcompleteness is expressed as overcompleteness of *filters* $\mathbf{w}_k$ which seems to make the problem a bit simpler because no latent variables need to be inferred. However, the normalization constant Z is not known when G is non-quadratic, i.e. when the model is non-Gaussian, which is why previous research had to resort to MCMC methods (Teh et al., 2003) or some approximations (Hyvärinen and Inki, 2002).

We have the score function

$$\boldsymbol{\psi}(\mathbf{x}; \mathbf{W}, \alpha_1, \dots, \alpha_m) = \sum_{k=1}^{m} \alpha_k w_k g(\mathbf{w}_k^T \mathbf{x}),$$

where g is the first derivative of G. Going through similar developments as in the case of the basic ICA model, the sample version of the objective function $\tilde{J}$ can be shown to equal

$$\tilde{J} = \sum_{k=1}^{m} \alpha_k \frac{1}{T} \sum_{t=1}^{T} g'(\mathbf{w}_k^T \mathbf{x}(t)) + \frac{1}{2} \sum_{j,k=1}^{m} \alpha_j \alpha_k \mathbf{w}_j^T \mathbf{w}_k \frac{1}{T} \sum_{t=1}^{T} g(\mathbf{w}_k^T \mathbf{x}(t)) g(\mathbf{w}_j^T \mathbf{x}(t)). \tag{12}$$

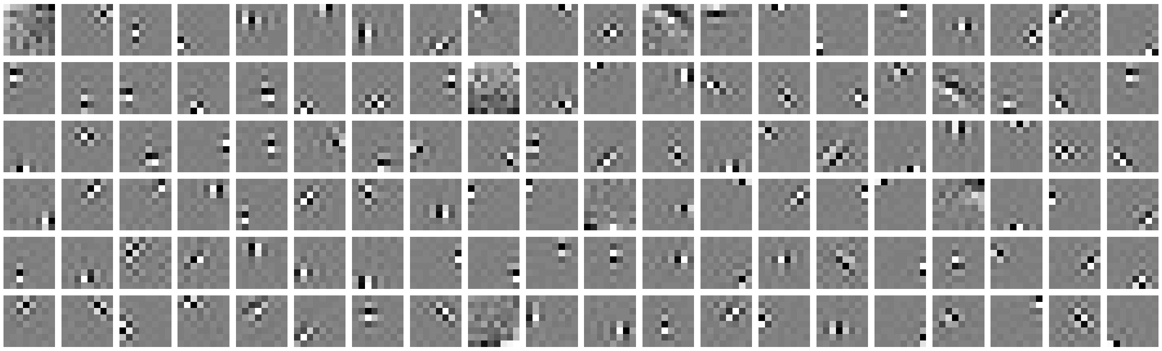

Figure 3: The overcomplete set of filters $\mathbf{w}_i$ estimated from natural image data. Note that no dimension reduction was performed, and we show filters instead of basis vectors, which is why the results are much less smooth and "beautiful" than some published ICA results (Hyvärinen et al., 2001).

We estimated the model for image patches of 8×8 pixels taken from natural images, see P.O. Hoyer's *imageica* package.[3] As preprocessing, the DC component (i.e. the mean gray-scale value) was removed from each image patch, reducing the effective dimensionality of the data to $n = 63$. The data was also whitened, i.e. the model was used in a linearly transformed space (the exact method of whitening has no significance). We set $m = 200$. We also took the tanh function as g, which corresponds to $G(u) = \log\cosh(u)$ (we did not bother to find the right scaling as in the basic ICA case). The objective function $\tilde{J}$ in (12) was optimized by gradient descent. The $\mathbf{w}_i$ were set to random initial values, and the α_i were all set to the initial value 1.5 that was found to be close to the optimal value in pilot experiments.

The obtained vectors $\mathbf{w}_i$ are shown in Figure 3. For the purposes of visualization, the vectors were converted back to the original space from the whitened space. The optimal α_i were in the range $0.5 \ldots 2$.

To show that the method correctly found different vectors and not duplicates of a smaller set of vectors, we computed the dot-products between the vectors, and for each $\mathbf{w}_i$, we selected the largest absolute value of dot-product $|\mathbf{w}_i^T \mathbf{w}_j|, j \neq i$. The dot-products were computed in the whitened space. The histogram of these maximal dot-products is shown in Figure 4. They are all much smaller than 1 (in absolute value), in fact all are smaller than 0.5. Since the vectors $\mathbf{w}_i$ were normalized to unit norm, this shows that no two $\mathbf{w}_i$ were close to equal, and we did find m different vectors.

4. Discussion

Here we discuss the connections of our method to two well-known methods before concluding the paper.

[3] The package can be downloaded at http://www.cs.helsinki.fi/patrik.hoyer/.

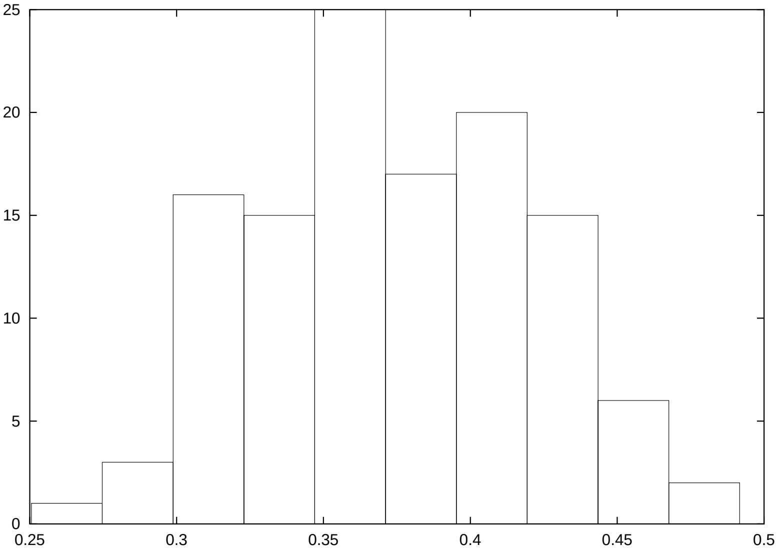

Figure 4: The distribution of maximal dot-products of a filter $\mathbf{w}_i$ with all other filters, computed in the whitened space.

4.1 Comparison with Pseudo-Likelihood Estimation

A related method for estimating non-normalized models is maximization of pseudo-likelihood (Besag, 1974). The idea is to maximize the product of marginal conditional likelihoods. The pdf is approximated by

$$\log p_{pseudo}(\mathbf{x}) = \sum_{i=1}^{n} p(x_i | x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n), \tag{13}$$

and the likelihood is computed using this approximation. The idea was originally developed in connection with Markov random fields, in which context it is quite natural because the conditional probabilities are often given as part of the model specification. The idea can still be used in the general case considered in this article. However, the conditional probabilities in (13) are not necessarily readily available and need to be computed. In particular, these conditional densities need to be normalized. The computational burden needed in the normalization is reduced from the original problem since we only need to numerically compute n one-dimensional integrals which is far more feasible than a single n-dimensional integral. However, compared to score matching, this is a computationally expensive method since score matching avoids the need for numerical integration altogether.

The question of consistency of pseudo-likelihood estimation seems to be unclear. Some consistency proofs were provided by Besag (1974, 1977), but these only apply to special cases such as Gaussian or binary random fields. Sufficiently general consistency results on pseudo-likelihood estimation seem to be lacking. This is another disadvantage with respect to score matching, which was shown above to be (locally) consistent.

4.2 Comparison with Contrastive Divergence

An interesting approximative MCMC method called contrastive divergence was recently proposed by Hinton (2002). The basic principle is to use an MCMC method for computing the derivative of the logarithm of the normalization factor Z, but the MCMC is allowed to run for only a single iteration (or a few iterations) before doing the gradient step.

The method is generally biased, even asymptotically (Carreira-Perpiñán and Hinton, 2005b), except in some special cases such as the multivariate Gaussian distribution (Carreira-Perpiñán and Hinton, 2005a). Score matching is thus preferable if a consistent estimator is wanted.

The computational efficiency of contrastive divergence is difficult to evaluate since it is not really a single method but a family of methods, depending on the MCMC method used. For the case of continuous-valued variables that we consider here, a Metropolis-type algorithm would probably be the method of choice, but there is a large number of different variants whose performances are likely to be quite different.

Nevertheless, contrastive divergence is a much more general method than score matching since it is applicable to intractable latent variable models. It can also handle binary/discrete variables—in fact, it is probably much easier to implement, using Gibbs sampling, for binary variables than for continous-valued variables. Extension of score matching to these two cases is an important problem for future research.

4.3 Conclusion

We have proposed a new method, score matching, to estimate statistical models in the case where the normalization constant is unknown. Although the estimation of the score function is computationally difficult, we showed that the distance of data and model score functions is very easy to compute. The main assumptions in the method are: 1) all the variables are continuous-valued and defined over $\mathbb{R}^n$, 2) the model pdf is smooth enough. Score matching provides a computationally simple yet locally consistent alternative to existing methods, such as MCMC and various approximative methods.

Acknowledgments

I am grateful to Patrik Hoyer, Jarmo Hurri, and Shohei Shimizu for comments on the manuscript, to Sam Roweis for interesting discussions, and to Miguel Carreira-Perpiñán and Geoffrey Hinton for providing access to unpublished results. The work was supported by the Academy of Finland, Academy Research Fellow position and project #48593.

Appendix A. Proof of Theorem 1

Definition (2) gives

$$J(\boldsymbol{\theta}) = \int p_{\mathbf{x}}(\boldsymbol{\xi}) \left[\frac{1}{2} \|\boldsymbol{\psi}_{\mathbf{x}}(\boldsymbol{\xi})\|^2 + \frac{1}{2} \|\boldsymbol{\psi}(\boldsymbol{\xi};\boldsymbol{\theta})\|^2 - \boldsymbol{\psi}_{\mathbf{x}}(\boldsymbol{\xi})^T \boldsymbol{\psi}(\boldsymbol{\xi};\boldsymbol{\theta}) \right] d\boldsymbol{\xi}. \tag{14}$$

(For simplicity, we omit the integration domain in here.) The first term in brackets does not depend on $\boldsymbol{\theta}$, and can be ignored. The integral of the second term is simply the integral of the sum of the second terms in brackets in (3). Thus, the difficult thing to prove is that integral of the third term in brackets in (14) equals the integral of the sum of the first terms in brackets in (3). This term equals

$$-\sum_i \int p_{\mathbf{x}}(\boldsymbol{\xi})\psi_{\mathbf{x},i}(\boldsymbol{\xi})\psi_i(\boldsymbol{\xi};\theta)d\boldsymbol{\xi},$$

where $\psi_{\mathbf{x},i}(\boldsymbol{\xi})$ denotes the i-th element of the vector $\boldsymbol{\psi}_{\mathbf{x}}(\boldsymbol{\xi})$. We can consider the integral for a single i separately, which equals

$$-\int p_{\mathbf{x}}(\boldsymbol{\xi})\frac{\partial \log p_{\mathbf{x}}(\boldsymbol{\xi})}{\partial \xi_i}\psi_i(\boldsymbol{\xi};\boldsymbol{\theta})d\boldsymbol{\xi} = -\int \frac{p_{\mathbf{x}}(\boldsymbol{\xi})}{p_{\mathbf{x}}(\boldsymbol{\xi})}\frac{\partial p_{\mathbf{x}}(\boldsymbol{\xi})}{\partial \xi_i}\psi_i(\boldsymbol{\xi};\boldsymbol{\theta})d\boldsymbol{\xi} = -\int \frac{\partial p_{\mathbf{x}}(\boldsymbol{\xi})}{\partial \xi_i}\psi_i(\boldsymbol{\xi};\boldsymbol{\theta})d\boldsymbol{\xi}.$$

The basic trick of partial integration needed the proof is simple: for any one-dimensional pdf p and any function f, we have

$$\int p(x)(\log p)'(x)f(x)dx = \int p(x)\frac{p'(x)}{p(x)}f(x)dx = \int p'(x)f(x)dx = -\int p(x)f'(x)dx$$

under some regularity assumptions that will be dealt with below.

To proceed with the proof, we need to use a multivariate version of such partial integration:

Lemma 4

$$\lim_{a\to\infty,b\to-\infty} f(a,\xi_2,\ldots,\xi_n)g(a,\xi_2,\ldots,\xi_n) - f(b,\xi_2,\ldots,\xi_n)g(b,\xi_2,\ldots,\xi_n)$$
$$= \int_{-\infty}^{\infty} f(\boldsymbol{\xi})\frac{\partial g(\boldsymbol{\xi})}{\partial \xi_1}d\xi_1 + \int_{-\infty}^{\infty} g(\boldsymbol{\xi})\frac{\partial f(\boldsymbol{\xi})}{\partial \xi_1}d\xi_1,$$

assuming that f and g are differentiable. The same applies for all indices of ξ_i, but for notational simplicity we only write the case $i = 1$ here.

Proof of lemma:

$$\frac{\partial f(\boldsymbol{\xi})g(\boldsymbol{\xi})}{\partial \xi_1} = f(\boldsymbol{\xi})\frac{\partial g(\boldsymbol{\xi})}{\partial \xi_1} + g(\boldsymbol{\xi})\frac{\partial f(\boldsymbol{\xi})}{\partial \xi_1}.$$

We can now consider this as a function of ξ_1 alone, all other variables being fixed. Then, integrating over $\xi_1 \in \mathbb{R}$, we have proven the lemma.

Now, we can apply this lemma on $p_{\mathbf{x}}$ and $\psi_1(\boldsymbol{\xi};\boldsymbol{\theta})$ which were both assumed to be differentiable in the theorem, and we obtain:

$$\begin{aligned}
-\int \frac{\partial p_{\mathbf{x}}(\boldsymbol{\xi})}{\partial \xi_1}\psi_1(\boldsymbol{\xi};\boldsymbol{\theta})d\boldsymbol{\xi} &= -\int \left[\int \frac{\partial p_{\mathbf{x}}(\boldsymbol{\xi})}{\partial \xi_1}\psi_1(\boldsymbol{\xi};\boldsymbol{\theta})d\xi_1\right] d(\xi_2,\ldots,\xi_n) \\
&= -\int \Bigg[\lim_{a\to\infty,b\to-\infty}[p_{\mathbf{x}}(a,\xi_2,\ldots,\xi_n)\psi_1(a,\xi_2,\ldots,\xi_n;\boldsymbol{\theta}) \\
&\qquad - p_{\mathbf{x}}(b,\xi_2,\ldots,\xi_n)\psi_1(b,\xi_2,\ldots,\xi_n;\boldsymbol{\theta})] \\
&\qquad\qquad - \int \frac{\partial \psi_1(\boldsymbol{\xi};\boldsymbol{\theta})}{\partial \xi_1}p_{\mathbf{x}}(\boldsymbol{\xi})d\xi_1\Bigg] d(\xi_2,\ldots,\xi_n).
\end{aligned}$$

For notational simplicity, we consider the case of $i = 1$ only, but this is true for any i.

The limit in the above expression is zero for any $\xi_2, \ldots, \xi_n, \boldsymbol{\theta}$ because we assumed that $p_{\mathbf{x}}(\boldsymbol{\xi})\boldsymbol{\psi}(\boldsymbol{\xi}; \boldsymbol{\theta})$ goes to zero at infinity. Thus, we have proven that

$$-\int \frac{\partial p_{\mathbf{x}}(\boldsymbol{\xi})}{\partial \xi_i} \psi_i(\boldsymbol{\xi}; \boldsymbol{\theta}) d\boldsymbol{\xi} = \int \frac{\partial \psi_i(\boldsymbol{\xi}; \boldsymbol{\theta})}{\partial \xi_i} p_{\mathbf{x}}(\boldsymbol{\xi}) d\boldsymbol{\xi},$$

that is, integral of the the third term in brackets in (14) equals the integral of the sum of the first terms in brackets in (3), and the proof of the theorem is complete.

Appendix B. Proof of Theorem 2

Assume $J(\boldsymbol{\theta}) = 0$. Then, the assumption $q > 0$ implies $p_{\mathbf{x}}(\boldsymbol{\xi}) > 0$ for all $\boldsymbol{\xi}$, which implies that $\boldsymbol{\psi}_{\mathbf{x}}(.)$ and $\boldsymbol{\psi}(.; \boldsymbol{\theta})$ are equal. This implies $\log p_{\mathbf{x}}(.) = \log p(.; \boldsymbol{\theta}) + c$ for some constant c. But c is necessarily 0 because both $p_{\mathbf{x}}$ and $p(.; \boldsymbol{\theta})$ are pdf's. Thus, $p_{\mathbf{x}} = p(.; \boldsymbol{\theta})$. By assumption, only $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ fulfills this equality, so necessarily $\boldsymbol{\theta} = \boldsymbol{\theta}^*$, and we have proven the implication from left to right. The converse is trivial.

References

S.-I. Amari, A. Cichocki, and H. H. Yang. A new learning algorithm for blind source separation. In *Advances in Neural Information Processing Systems 8*, pages 757–763. MIT Press, 1996.

J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B*, 36(2):192–236, 1974.

J. Besag. Efficiency of pseudolikelihood estimation for simple gaussian fields. *Biometrika*, 64(3):616–618, 1977.

C. Bouman and K. Sauer. A generalized gaussian image model for edge-preserving MAP estimation. *IEEE Transactions on Image Processing*, 2(3):296–310, 1993.

M. Á. Carreira-Perpiñán and G. E. Hinton. On contrastive divergence (CD) learning. Technical report, Dept of Computer Science, University of Toronto, 2005a. In preparation.

M. Á. Carreira-Perpiñán and G. E. Hinton. On contrastive divergence learning. In *Proceedings of the Workshop on Artificial Intelligence and Statistics (AISTATS2005)*, Barbados, 2005b.

G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

A. Hyvärinen and P. O. Hoyer. A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images. *Vision Research*, 41(18):2413–2423, 2001.

A. Hyvärinen and M. Inki. Estimating overcomplete independent component bases from image windows. *Journal of Mathematical Imaging and Vision*, 17:139–152, 2002.

A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis.* Wiley Interscience, 2001.

S. Z. Li. *Markov Random Field Modeling in Image Analysis.* Springer, 2nd edition, 2001.

D. J. C. Mackay. *Information Theory, Inference and Learning Algorithms.* Cambridge University Press, 2003.

B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37:3311–3325, 1997.

D.-T. Pham and P. Garrat. Blind separation of mixture of independent sources through a quasi-maximum likelihood approach. *IEEE Transactions on Signal Processing*, 45(7): 1712–1725, 1997.

M. Schervish. *Theory of Statistics.* Springer, 1995.

Y. W. Teh, M. Welling, S. Osindero, and G. E. Hinton. Energy-based models for sparse overcomplete representations. *Journal of Machine Learning Research*, 4:1235–1260, 2003.