Optimal approximation of signal priors

Aapo Hyvärinen [*]
Helsinki Institute for Information Technology (BRU)
Dept of Computer Science
University of Helsinki, Finland

27th October 2006

Abstract

In signal restoration by Bayesian inference, one typically uses a parametric model of the prior distribution of the signal. Here, we consider how the parameters of a prior model should be estimated from observations of uncorrupted signals. A lot of recent work has implicitly assumed that maximum likelihood estimation is the optimal estimation method. Our results imply that this is not the case. We first obtain an objective function that approximates the error occurred in signal restoration due to an imperfect prior model. Next, we show that in an important special case (small gaussian noise), the error is the same as the score matching objective function, which was previously proposed as an alternative for likelihood based on purely computational considerations. Our analysis thus shows that score matching combines computational simplicity with statistical optimality in signal restoration, providing a viable alternative to maximum likelihood methods. We also show how the method leads to a new intuitive and geometric interpretation of "structure" inherent in probability distributions.

1 Introduction

1.1 Empirical Bayes and signal restoration

An approach that has gained increasing acceptance in machine learning, computational neuroscience, and signal processing is based on hierarchical Bayesian modelling. The typical setting for modelling the observed multivariate continuous-valued data vector, denoted by $\mathbf{x}$, is as follows. The vector $\mathbf{x}$ follows a distribution with probability density function (pdf) $p(\mathbf{x}|\mathbf{s})$, where $\mathbf{s}$ is a vector of latent variables or parameters. The vector $\mathbf{s}$ in its turn follows a prior distribution $p(\mathbf{s}|\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is a vector of (hyper)parameters. Typically, $\mathbf{x}$ is a somehow *corrupted* or *incomplete* version of $\mathbf{s}$ which is the real quantity of interest, and $\boldsymbol{\theta}$ gives some kind of *features.* The joint probability is obtained by concatenating these probabilities:

$$p(\mathbf{x}, \mathbf{s}, \boldsymbol{\theta}) = p(\mathbf{x}|\mathbf{s})p(\mathbf{s}|\boldsymbol{\theta}) \tag{1}$$

where we assume a flat prior for $\boldsymbol{\theta}$.

The central idea is that in such methods, the hyperparameters or features $\boldsymbol{\theta}$ are not set subjectively, but estimated (learned) from the data. Methods in which the hyperparameters are estimated from the data $\mathbf{x}$ are usually called Empirical Bayes. In this paper, we consider a setting that is slightly different from conventional Empirical Bayes. We assume that a separate sample of $\mathbf{s}$, denoted by $\mathbf{s}(1), \ldots, \mathbf{s}(T)$ can be observed, and the hyperparameters $\boldsymbol{\theta}$ are estimated from such a

[*] Dept of Computer Science, P.O.Box 68, FIN-00014 University of Helsinki, Finland. Fax: +358-9-191 51120, email: aapo.hyvarinen@helsinki.fi

sample. The prior $p(\mathbf{s}|\boldsymbol{\theta})$ is then used for Bayesian inference of $\mathbf{s}$ when an $\mathbf{x}$ is observed for unknown $\mathbf{s}$. (In what follows, we shall simply call $p(\mathbf{s}|\boldsymbol{\theta})$ the "prior" and $\boldsymbol{\theta}$ the "parameter" vector, omitting the prefix "hyper".)

There are many applications in which such a formalism with observed $\mathbf{s}$ has been applied. The prime example is signal restoration, see e.g. (O'Ruanaidh and Fitzgerald, 1996; Chipman et al., 1997; Johnstone and Silverman, 2005). The vector $\mathbf{x}$ corresponds to a corrupted version of a signal, and $\mathbf{s}$ corresponds to the original uncorrupted signal. In many cases, we can observe a sample of the distribution of $p(\mathbf{s}|\boldsymbol{\theta})$ by measuring the signal under circumstances where the corrupting process is not present. For example, when denoising natural images it is not a problem to find practically noise-free natural images (Simoncelli and Adelson, 1996; Hyvärinen, 1999); the same applies for restoration of audio signals (Godsill and Rayner, 1995). A prior estimated from noise-free signals can then be used for denoising noisy signals.

Another application can be found in Bayesian perception, where the $\mathbf{s}$ correspond to some perceptual quantities of a scene (speed and direction of motion, depth etc.) that are sometimes difficult to instantly infer from the data $\mathbf{x}$ that is measured by the retina (Knill and Richards, 1996). However, if such scenes are observed for a longer period of time, and information from different perceptual cues are combined, the perceptual system can often obtain virtually exact observations of those latent quantities, and these can be used, in the long run, to learn the model parameters. The prior with these parameters can then enhance the performance of the system in more difficult situations where few cues are available and/or the observation period is very short.

1.2 Point estimates vs. full Bayesian treatment

The goal in such inference is typically to obtain a point estimate of $\mathbf{s}$. This is because in practical applications, the posterior must typically be output as a point estimate (e.g. a denoised image). The typical, and computationally most feasible, point estimate to summarize the posterior of $\mathbf{s}$ is the maximum a posteriori (MAP) estimator (see below).

If computational resources were not an issue, one could use the theoretically sound treatment based on integrating out the parameters, considering their full posterior distributions. That is, the full posterior $p(\boldsymbol{\theta}|\mathbf{s}(1),\ldots,\mathbf{s}(T))$, given the separate sample of $\mathbf{s}$, is used to compute the posterior of $\mathbf{s}$ as in

$$p(\mathbf{s}|\mathbf{x},\mathbf{s}(1),\ldots,\mathbf{s}(T)) = \int p(\mathbf{x}|\mathbf{s})p(\mathbf{s}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{s}(1),\ldots,\mathbf{s}(T))\,d\boldsymbol{\theta}\,/p(\mathbf{x}) \tag{2}$$

where the normalizing constant equals

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{s})p(\mathbf{s}|\boldsymbol{\theta})\,d\mathbf{s}\,d\boldsymbol{\theta} \tag{3}$$

The problem is that the computation of (2) requires multidimensional integration which is computationally most demanding. In order to reduce the computational load by avoiding multidimensional integration, many methods use a point estimate for $\boldsymbol{\theta}$. In the context of signal restoration, this means fixing the signal features and other parameters to a single value, which is obviously a widespread approach.

Thus, we consider here a computationally simplified setting where a point estimate $\hat{\boldsymbol{\theta}}$ of parameters is first obtained, and it is used in MAP estimation of $\mathbf{s}$. MAP estimation simply means finding the value that maximizes the posterior density of $\mathbf{s}$, given an estimate $\hat{\boldsymbol{\theta}}$:

$$\hat{\mathbf{s}}_{MAP}(\hat{\boldsymbol{\theta}},\mathbf{x}) = \arg\max_{\mathbf{s}} p(\mathbf{x}|\mathbf{s})p(\mathbf{s}|\hat{\boldsymbol{\theta}}) = \arg\max_{\mathbf{s}} \log p(\mathbf{x}|\mathbf{s}) + \log p(\mathbf{s}|\hat{\boldsymbol{\theta}}) \tag{4}$$

where the notation with $\hat{\boldsymbol{\theta}}$ and $\mathbf{x}$ in parentheses emphasizes that the estimate is a function of both the observed data $\mathbf{x}$ and the (previously) obtained parameter estimate $\boldsymbol{\theta}$. Such a framework is often used with very high-dimensional data where computational considerations are of central importance.

1.3 Optimal approximation of prior

The question we attempt to answer in this paper is how the parameters in $\boldsymbol{\theta}$ should be estimated from a sample of uncorrupted signals $\mathbf{s}(1), \ldots, \mathbf{s}(T)$ in this context.

Most work on Bayesian inference in signal restoration and computational neuroscience seems to implicitly assume that maximum likelihood estimation (MLE) is the optimal way of estimating the parameters. However, this does not follow from the classic optimality criteria of MLE. The main justification for MLE is that it is, under certain assumptions, asymptotically Fisher-efficient, i.e. gives asymptotically the most exact estimates for parameters, in terms of squared error (Schervish, 1995). In our case, this would mean that the error in the estimate of $\boldsymbol{\theta}$ is a small as possible.

However, what we want to minimize here is rather the error in the MAP estimate of $\mathbf{s}$, and not the error in $\boldsymbol{\theta}$. It is possible that some estimation methods give a large error in $\boldsymbol{\theta}$, but this error does not induce a large error in $\mathbf{s}$. As a common example of a related situation consider multicollinearity in prediction by linear regression: if the predicting variables are highly correlated, their individual regression coefficients have large estimation errors; yet, the prediction might be quite good. So, if we are not interested in the values of the parameters themselves, but only the quality of the Bayesian inference that they provide, estimation errors in $\boldsymbol{\theta}$ may be irrelevant, and there seems to be no reason to consider MLE of $\boldsymbol{\theta}$ optimal.

Furthermore, the prior model $p(\mathbf{s}|\boldsymbol{\theta})$ might only be a rough *approximation* of the true prior distribution of $\mathbf{s}$; the real prior might not belong to the family $p(\mathbf{s}|\boldsymbol{\theta})$. In such a case, any considerations of squared error in $\boldsymbol{\theta}$ may be of little use. Then, estimation of $\boldsymbol{\theta}$ should be based on a direct measure of how good the ensuing MAP estimation of $\mathbf{s}$ is.

Information theory provides another justification for MLE in terms of optimal compression, see e.g. (Cover and Thomas, 1991). However, such considerations seem to be irrelevant if the goal is Bayesian (MAP) inference of $\mathbf{s}$.

In this paper, we analyze the performance of the MAP estimator of $\mathbf{s}$. This is a function of the parameter value $\hat{\boldsymbol{\theta}}$ used in the prior, which are assumed to be estimated from a sample $\mathbf{s}(1), \ldots, \mathbf{s}(T)$. We derive a first-order approximation of the error, and show that it consists of two parts. Only one of those parts depends on the $\hat{\boldsymbol{\theta}}$. Optimal estimation of parameters should thus be based on minimization of the objective function given by that part. Such an objective function is quite different from likelihood. Interestingly, a special case of the objective function leads to the score matching distance previously proposed in (Hyvärinen, 2005) based on a completely different motivation. Furthermore, we give a geometric interpretation of the resulting estimation process and show how this is related to a measure of "structure" of probability distributions.

2 Optimality criterion for estimation

2.1 Hierachical data model

We shall first rigorously define the whole process of data generation and parameter estimation in a hierarchical model where a separate sample of uncorrupted signals can be observed.

1. Estimation of parameters: A sample $\mathbf{s}(1), \ldots, \mathbf{s}(T)$ is generated from a prior distribution $p_0(\mathbf{s})$. From this sample, we compute an estimate $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$, using a method to be specified.

2. Generation of $\mathbf{s}$ underlying for observed data: A single vector $\mathbf{s}_0$ is generated from the prior distribution $p_0(\mathbf{s})$.

3. Generation of observed data: A data vector $\mathbf{x}$ is generated from the data distribution $p(\mathbf{x}|\mathbf{s}_0)$.

4. MAP inference: Using $\hat{\boldsymbol{\theta}}$ and $\mathbf{x}$, an estimate $\hat{\mathbf{s}}$ for $\mathbf{s}_0$ is obtained by MAP estimation as in (4).

In step 4, the data generating process $p(\mathbf{x}|\mathbf{s})$ is assumed known; its estimation would be a completely different problem. The prior distribution p_0 is approximated by a parameterized family of pdf's, $p(.|\boldsymbol{\theta})$. We do *not* assume that p_0 belongs to the family $p(.|\boldsymbol{\theta})$.

The goal is now to minimize the error $\|\boldsymbol{\Delta}\mathbf{s}\| = \|\hat{\mathbf{s}} - \mathbf{s}_0\|$ that is due to the error in the approximation of the prior $p_0(\mathbf{s})$ by $p(\mathbf{s}|\hat{\boldsymbol{\theta}})$. Even with a perfect estimate for the prior, there will, of course, be an estimation error in $\hat{\mathbf{s}}$ due the randomness in the process of sampling the data from $p(\mathbf{x}|\mathbf{s}_0)$, which corresponds to the process corrupting the signal. However, we will see below that it is possible to separate these two kinds of errors.

We emphasize that it is the error in $\hat{\mathbf{s}}$ and not in $\hat{\boldsymbol{\theta}}$ that we fundamentally want to minimize. Actually, the error in $\hat{\boldsymbol{\theta}}$ is not even a properly defined quantity because the prior $p_0(\mathbf{s})$ need not belong to the family $p(\mathbf{s}|\boldsymbol{\theta})$ used in its approximation. Thus, we shall ultimately define the optimal method of parameter estimation, or prior approximation, as the one that minimizes the error in $\hat{\mathbf{s}}$.

2.2 Analysis of estimation error

First, we need some notation. Denote the derivatives of the log-pdf of $\mathbf{s}$ given $\boldsymbol{\theta}$ by

$$\boldsymbol{\psi}(\mathbf{s}|\boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial \log p(\mathbf{s}|\boldsymbol{\theta})}{\partial s_1} \\ \vdots \\ \frac{\partial \log p(\mathbf{s}|\boldsymbol{\theta})}{\partial s_n} \end{pmatrix} = \begin{pmatrix} \psi_1(\mathbf{s}|\boldsymbol{\theta}) \\ \vdots \\ \psi_n(\mathbf{s}|\boldsymbol{\theta}) \end{pmatrix} = \nabla_{\mathbf{s}} \log p(\mathbf{s}|\boldsymbol{\theta})$$

and the corresponding Hessian matrix by

$$H(\mathbf{s}|\boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial \log p(\mathbf{s}|\boldsymbol{\theta})}{\partial s_1 s_1} \cdots \frac{\partial \log p(\mathbf{s}|\boldsymbol{\theta})}{\partial s_1 s_n} \\ \vdots \\ \frac{\partial \log p(\mathbf{s}|\boldsymbol{\theta})}{\partial s_n s_1} \cdots \frac{\partial \log p(\mathbf{s}|\boldsymbol{\theta})}{\partial s_n s_n} \end{pmatrix} = \nabla_{\mathbf{s}} \boldsymbol{\psi}(\mathbf{s}|\boldsymbol{\theta})^T$$

Similary, denote by $\boldsymbol{\psi}(\mathbf{x}|\mathbf{s})$ and $H(\mathbf{x}|\mathbf{s})$ the gradient and the Hessian matrix of $\log p(\mathbf{x}|\mathbf{s})$, where the differentiation is still done with respect to $\mathbf{s}$, and denote by $\boldsymbol{\psi}_0(\mathbf{s})$ and $H_0(\mathbf{s})$ the corresponding gradient and Hessian of $\log p_0(\mathbf{s})$. In the following, we use the shorter notation $\hat{\mathbf{s}} = \hat{\mathbf{s}}_{MAP}(\hat{\boldsymbol{\theta}}, \mathbf{x})$.

Our main result is given in the following theorem, proven in Appendix A:

Theorem 1 *Assume that all the the log-pdf's in (4) are differentiable. Assume further that the estimation error* $\boldsymbol{\Delta}\mathbf{s} = \hat{\mathbf{s}} - \mathbf{s}_0$ *is small. Then the first-order approximation of the error is*

$$\|\boldsymbol{\Delta}\mathbf{s}\|^2 = \|\mathcal{E}_1 + \mathcal{E}_2\|^2 + o(\|\mathbf{M}^{-1}\boldsymbol{\Delta}\mathbf{s}\|^2) \tag{5}$$

where

$$\mathcal{E}_1 = \mathbf{M}^{-1}\left[\boldsymbol{\psi}_0(\mathbf{s}_0) - \boldsymbol{\psi}(\mathbf{s}_0|\hat{\boldsymbol{\theta}})\right] \tag{6}$$

$$\mathcal{E}_2 = \mathbf{M}^{-1}\left[\boldsymbol{\psi}_0(\mathbf{s}_0) + \boldsymbol{\psi}(\mathbf{x}|\mathbf{s}_0)\right] \tag{7}$$

with

$$\mathbf{M} = H_0(\mathbf{s}_0) + H(\mathbf{x}|\mathbf{s}_0) \tag{8}$$

Now, the matrix $\mathbf{M}$ and the error vector in $\mathcal{E}_2$ are functions of $\mathbf{s}_0$ and $\mathbf{x}$ only, i.e. the data generating parts (steps 3 and 4) above. Thus, they do not depend on our estimate for $\boldsymbol{\theta}$. In contrast, $\boldsymbol{\psi}_0(\mathbf{s}_0) - \boldsymbol{\psi}(\mathbf{s}_0|\hat{\boldsymbol{\theta}})$ in $\mathcal{E}_1$ *does* depend on $\hat{\boldsymbol{\theta}}$ which is a function of the sample $\mathbf{s}(1), \ldots, \mathbf{s}(T)$ (step 2 above).

If the errors $\mathcal{E}_1$ and $\mathcal{E}_2$ were orthogonal, we could decompose the expected error as

$$E\{\|\boldsymbol{\Delta}\mathbf{s}\|^2\} = E\{\|\mathcal{E}_1\|^2\} + E\{\|\mathcal{E}_2\|^2\} + o(\|\boldsymbol{\Delta}\mathbf{s}\|^2) \tag{9}$$

and we would see a clear decomposition of the error in two parts (definition of the expectation will be specified later):

- The first part, $E\{\|\mathcal{E}_1\|^2\}$, is the error in the estimate $\hat{\mathbf{s}}$ due to an error in our approximation $p(.|\boldsymbol{\theta})$ of the prior p_0. In fact, if the approximation of the prior is exact, $\boldsymbol{\psi}_0(\mathbf{s}_0) = \boldsymbol{\psi}(\mathbf{s}_0|\hat{\boldsymbol{\theta}})$ for any $\mathbf{s}_0$, and this term is zero.
- The second part, $E\{\|\mathcal{E}_2\|^2\}$, does not depend on the sample $\mathbf{s}(1), \ldots, \mathbf{s}(T)$ or $\hat{\boldsymbol{\theta}}$ at all. It is related to the error that the MAP estimator has even when the prior p_0 is known perfectly. This can be seen from the fact that if $\mathbf{s}_0$ were equal to the MAP estimator, $\mathcal{E}_2$ would be zero because the gradients cancel each other as in the definition of the MAP estimator.

While the two errors do not seem to be orthogonal in general, we do have an orthogonality result in an important special case, which is infinitesimal gaussian noise. This shall be treated in Section 4 and Theorem 3. Thus, we do have some justification for considering the two errors independently from each other: $\mathcal{E}_2$ would be only dependent on the model and not on the estimator $\hat{\boldsymbol{\theta}}$, in which case computation of $\hat{\boldsymbol{\theta}}$ should be based on $\mathcal{E}_1$ alone.

3 Proposal of optimal estimator

3.1 Direct minimization of approximate restoration error

Based on Theorem 1, we propose to minimize $\|\mathcal{E}_1\|^2$ in order to minimize the estimation (restoration) error in $\mathbf{s}$. Such an estimator should be optimal in the sense of minimizing squared error, at least if the two errors in the Theorem are orthogonal enough.

One further problem is that $\|\mathcal{E}_1\|^2$ depends also on $p_0(\mathbf{s})$ via $\boldsymbol{\psi}_0$ and H_0 whose estimation may be very difficult. For reasons that will become apparent later, the occurrence of $\boldsymbol{\psi}_0$ is actually not a problem. Regarding H_0, we use a first-order approximation, replacing it by its estimate $H(\mathbf{s}|\hat{\boldsymbol{\theta}})$.

Thus, taking the expected value of the error $\|\mathcal{E}_1\|^2$ over all $\mathbf{s}$ with respect to p_0, we arrive at the following objective function:

$$\mathcal{J}(\boldsymbol{\theta}) = \frac{1}{2} \int p_0(\mathbf{s}) \| \left[H(\mathbf{s}|\boldsymbol{\theta}) + H(\mathbf{x}|\mathbf{s}) \right]^{-1} \left[\boldsymbol{\psi}_0(\mathbf{s}) - \boldsymbol{\psi}(\mathbf{s}|\boldsymbol{\theta}) \right] \|^2 d\mathbf{s} \tag{10}$$

Since we have a sample of $\mathbf{s}$, the practical estimation will use a sample version, which equals

$$\tilde{\mathcal{J}}(\boldsymbol{\theta}) = \frac{1}{2} \sum_{t=1}^{T} \| \left[H(\mathbf{s}(t)|\boldsymbol{\theta}) + H(\mathbf{x}|\mathbf{s}(t)) \right]^{-1} \left[\boldsymbol{\psi}_0(\mathbf{s}(t)) - \boldsymbol{\psi}(\mathbf{s}(t)|\boldsymbol{\theta}) \right] \|^2 \tag{11}$$

So, we conclude that optimal estimation of the parameters is based, at least approximatively, on minimization or $\tilde{\mathcal{J}}$ with respect to $\boldsymbol{\theta}$

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

Basically, the objective function is a weighted squared error between the gradient of the log-density $\boldsymbol{\psi}_0$ of the sample $\mathbf{s}(t)$ and the gradient of the log-density given by the model, $\boldsymbol{\psi}(.|\hat{\boldsymbol{\theta}})$. This is actually rather natural because the definition of the MAP estimator (4) implies that the sum of the gradients of the log-densities $p(\mathbf{x}|\mathbf{s})$ and $p(\mathbf{s}|\hat{\boldsymbol{\theta}})$ must be zero; only the latter gradient depends on the parameter estimate $\hat{\boldsymbol{\theta}}$. So, to minimize the error in the MAP estimator, one should find an $\boldsymbol{\theta}$ that gives an accurate model of that gradient.

3.2 Simple computation of objective function

It may seem that the objective function $\tilde{\mathcal{J}}$ is computationally intractable because it uses $\boldsymbol{\psi}_0(\mathbf{s}(t))$ which depends on the unknown prior p_0. However, it turns out that the objective function is very closely related to the "score matching" objective function proposed in (Hyvärinen, 2005), see also (Pham and Garrat, 1997; Taleb and Jutten, 1999). Here, we present a generalization of the result in (Hyvärinen, 2005) that allows simple computation of $\tilde{\mathcal{J}}$. This is given by the following Theorem:

Theorem 2 *Denote the i, j-th element of the square* $\mathbf{MM}$ *of the pre-multiplying matrix* $[H(\mathbf{s}|\boldsymbol{\theta}) + H(\mathbf{x}|\mathbf{s})]^{-1}$ *in (10) by* $G_{ij}(\mathbf{s})$. *Assume some regularity conditions on the Hessians.*[1] *Then, the objective function in (10) can be expressed as*

$$\mathcal{J}(\boldsymbol{\theta}) = \int p_0(\mathbf{s}) \left\{ \sum_{ij} \partial_i \psi_i(\mathbf{s}|\boldsymbol{\theta}) G_{ij}(\mathbf{s}) + \psi_i(\mathbf{s}|\boldsymbol{\theta}) \partial_i G_{ij}(\mathbf{s}) + \frac{1}{2} G_{ij}(\mathbf{s}) \psi_i(\mathbf{s}|\boldsymbol{\theta}) \psi_j(\mathbf{s}|\boldsymbol{\theta}) \right\} d\mathbf{s} + \mathit{const.} \quad (12)$$

where ∂_i denotes differentiation with respect to the i-th element, and the constant term does not depend on $\boldsymbol{\theta}$. Moreover, this holds for any arbitrary functions G_{ij} fulfilling the regularity constraints.

The Theorem is proven in Appendix B, see also (Dawid and Lauritzen, 2005) for a related result.

Obviously, the sample version of this expression for the objective function is obtained as

$$\tilde{\mathcal{J}}(\boldsymbol{\theta}) = \sum_{t=1}^{T} \sum_{ij} \partial_i \psi_i(\mathbf{s}(t)|\boldsymbol{\theta}) G_{ij}(\mathbf{s}(t)) + \psi_i(\mathbf{s}(t)|\boldsymbol{\theta}) \partial_i G_{ij}(\mathbf{s}(t)) + \frac{1}{2} G_{ij}(\mathbf{s}(t)) \psi_i(\mathbf{s}(t)|\boldsymbol{\theta}) \psi_j(\mathbf{s}(t)|\boldsymbol{\theta}) \quad (13)$$

where we have omitted the irrelevant constant. Here, we see the remarkable fact that this sample version is easy to compute: it only contains sample averages of some functions which are all part of the model specification and can be simply computed, provided that the model is defined using functions $\log p(.|\boldsymbol{\theta})$ whose derivatives can be given in closed form or otherwise simply computed.

3.3 Relationship to score matching

In fact, in (Hyvärinen, 2005) a special case of our estimation method was proposed based on purely computational considerations. The problem considered in that paper was what to do if the normalization constant of the pdf is not known. In other words, the prior pdf is defined using a function q in a form that is simple to compute, but q does not integrate to unity. Thus, the pdf is given by

$$p(\mathbf{s}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} q(\mathbf{s}; \boldsymbol{\theta})$$

where we do *not* know how to easily compute Z which is given by an integral that is often analytically intractable:

$$Z(\boldsymbol{\theta}) = \int q(\mathbf{s}; \boldsymbol{\theta})\, d\mathbf{s}$$

Now, the derivatives of the log-density ("score function") with respect to the s_i do not depend on Z at all, so the problem of computing the normalization constant disappears when we consider only the score functions. It is natural to try to estimate the model by looking at the Euclidean distance between the score function of the data and the score function given by the model. This leads to a special case of the present objective function, where the matrix $\mathbf{M}$ is replaced by identity. In (Hyvärinen, 2005), it was further proven that such an estimator is (locally) consistent.

Thus, we see that our proposed estimator combines statistical optimality, in the sense discussed in this paper, with computational simplicity, in the sense that the prior model $p(\mathbf{s}|\boldsymbol{\theta})$ does not need to integrate to unity, as was originally shown in (Hyvärinen, 2005) for a special case. In the next section, we will see that this special case emerges when we take a particular form for $p(\mathbf{x}|\mathbf{s})$.

[1] The regularity conditions are: G_{ij} is differentiable and $p_0(\mathbf{s})G_{ij}(\mathbf{s})\psi_i(\mathbf{s})$ vanishes when $\|\mathbf{s}\| \to \infty$ for all i, j, and the integrals given in (40) are finite.

4 Case of gaussian infinitesimally small noise

A very interesting special case is obtained when $\mathbf{x}$ is equal to $\mathbf{s}$ plus infinitesimally small gaussian i.i.d. (white) noise:

$$\log p(\mathbf{x}|\mathbf{s}) = -\frac{1}{2\sigma^2}\|\mathbf{x}-\mathbf{s}\|^2 - \frac{n}{2}\log 2\pi\sigma^2 \tag{14}$$

where n is the dimension of both $\mathbf{x}$ and $\mathbf{s}$, and we consider the limit of

$$\sigma^2 \rightarrow 0 \tag{15}$$

Such additive gaussian noise is an important practical case in signal processing and computational neuroscience. It can also be considered a theoretical archetype of signal corruption. In this case the matrix $\mathbf{M}$ is of the form

$$\mathbf{M} = H_0(\mathbf{s}_0) - \frac{1}{\sigma^2}\mathbf{I} \tag{16}$$

Taking the limit of $\sigma^2 \rightarrow 0$, we see that $\mathbf{M}$ approches the identity matrix multiplied by $-1/\sigma^2$. Our objective function is thus simplified to the Euclidean distance of the score functions, if we ignore the scaling by $1/\sigma^2$. This simplifies the computations very much, and gives the original score matching distance proposed in (Hyvärinen, 2005) and discussed in the preceding section. The sample version of the objective function, in the present notation, is then given by

$$\tilde{\mathcal{J}} = \sum_{t=1}^{T}\sum_{i} \partial_i \psi_i(\mathbf{s}(t)|\boldsymbol{\theta}) + \frac{1}{2}\psi_i(\mathbf{s}(t)|\boldsymbol{\theta})^2 \tag{17}$$

In the case of infinitesimal gaussian noise, we also have exact orthogonality of the two errors $\mathcal{E}_1$ and $\mathcal{E}_2$ in Theorem 1. In Appendix C we prove the following:

Theorem 3 *Assume that $p(\mathbf{x}|\mathbf{s})$ is as in (14–15). Then, we have for any $\mathbf{s}_0$:*

$$E_{\mathbf{x}|\mathbf{s}_0}\{\langle \mathcal{E}_1, \mathcal{E}_2\rangle\} = 0 \textit{ for all } \mathbf{s}_0 \tag{18}$$

That is, the two errors in Theorem 1 are orthogonal, and (9) holds when the expectations are taken over $\mathbf{x}$ given any $\mathbf{s}_0$.

Thus, this theorem gives some justification for considering the $\mathcal{E}_1$ and $\mathcal{E}_2$ separately, and estimating parameters by simply minimizing $\mathcal{E}_1$.

5 Interpretations and projection and structure

In this section, we will propose two intuitive interpretations of the estimation performed by score matching. The interpretations is based on two ideas:

- Score matching estimator is obtained by minimizing a Euclidean distance, which leads to an interpretation as *projection*.
- The amount of noise that can be removed from data is dependent on the amount of *structure* inherent in the data vector. Such structure is often associated with information-theoretical quantities such as (neg)entropy, but our analysis provides an alternative measure of structure.

The word "structure" is used loosely in what follows, intuitively it means a lack of complete randomness in the data distribution. This is similar to the intuitive principle of information theory, in which the structure present in the data distribution allows it to be represented more compactly, i.e. compressed. Here, we show how the proportion of gaussian noise that can be removed from noisy observations leads to a similar measure of structure.

5.1 Definition of geometry

We begin by defining basic geometrical concepts based on the score functions. Consider the space S of probability density functions which are sufficiently smooth in the sense that the assumptions given in the theorems above are fulfilled. Assume that $p_{\mathbf{s}}$ in S is fixed once and for all. Given any two pdf's p_1 and p_2 in S, we define their dot-product as

$$\langle p_1, p_2 \rangle_{\mathbf{s}} = \int p_{\mathbf{s}}(\boldsymbol{\xi}) \left[\sum_{i=1}^{n} \psi_{1,i}(\boldsymbol{\xi}) \psi_{2,i}(\boldsymbol{\xi}) \right] d\boldsymbol{\xi} \tag{19}$$

where $\psi_{1,i}$ denotes the i-th element in the score function of p_1, and likewise for $\psi_{2,i}$. (For a bit more mathematical rigour, we use $\boldsymbol{\xi}$ as the integrating variable instead of $\mathbf{s}$.) The norm of a pdf is then given by

$$\|p_1\|_{\mathbf{s}}^2 = \langle p_1, p_1 \rangle_{\mathbf{s}} = \int p_{\mathbf{s}}(\boldsymbol{\xi}) \left[\sum_{i=1}^{n} \psi_{1,i}(\boldsymbol{\xi})^2 \right] d\boldsymbol{\xi} = \int p_{\mathbf{s}}(\boldsymbol{\xi}) \|\boldsymbol{\psi}_1(\boldsymbol{\xi})\|^2 d\boldsymbol{\xi} \tag{20}$$

where the notation $\|.\|$, without a subscript, in the right-most integral denotes the ordinary Euclidean norm.

The norm we have just defined is closely related to Fisher information. The multidimensional Fisher information matrix is defined here as

$$I_F(\mathbf{s}) = E\{\boldsymbol{\psi}(\mathbf{s})\boldsymbol{\psi}(\mathbf{s})^T\}. \tag{21}$$

Strictly speaking, this is the Fisher information matrix w.r.t. a hypothetical location parameter. Obviously, we have

$$\|p_{\mathbf{s}}\|_{\mathbf{s}}^2 = \mathrm{tr}(I_F(\mathbf{s})) \tag{22}$$

Using the norm, we can also naturally define the distance:

$$\|p_1 - p_2\|_{\mathbf{s}}^2 = \int p_{\mathbf{s}}(\boldsymbol{\xi}) \left[\sum_{i=1}^{n} (\psi_{1,i}(\boldsymbol{\xi}) - \psi_{2,i}(\boldsymbol{\xi}))^2 \right] d\boldsymbol{\xi} \tag{23}$$

Now we proceed to show how these geometric concepts can be interpreted as measures of the structure of a prior distribution in Bayesian inference.

5.2 Denoising capacity using perfect model

First of all, the norm $\|.\|_{\mathbf{s}}$ defined in (20) is closely related to denoising capacity. In previous work, we proved the following:

Theorem 4 *Assume that $p(\mathbf{x}|\mathbf{s})$ is a gaussian distribution with mean $\mathbf{s}$ and covariance $\sigma^2\mathbf{I}$. The quadratic error of the MAP estimator $\hat{\mathbf{s}}$, when the distribution $p_{\mathbf{s}}$ is exactly known,is given by*

$$tr(E\{(\mathbf{s} - \hat{\mathbf{s}})(\mathbf{s} - \hat{\mathbf{s}})^T\}) = n\sigma^2 - \sigma^4 \|p_{\mathbf{s}}\|_{\mathbf{s}}^2 + \textit{ terms of higher order in } \sigma^2 \tag{24}$$

where σ^2 is the noise level.

This is a simple corollary of Theorem 2 in (Hyvärinen, 1999).

Thus, we can interpret $\|p_{\mathbf{s}}\|_{\mathbf{s}}^2$ as the *amount of structure* that is present in the data vector $\mathbf{s}$. It determines the amount of noise reduction that we can achieve by MAP estimation when we have a perfect model of the distribution of $\mathbf{s}$. (The dominant term $n\sigma^2$ does not depend on the distribution of the data so it is irrelevant as a measure of structure.) The case of an imperfect model will be considered in the next section. Now we show some examples of different distributions and the amounts of structure they contain.

Example 1 *A flat distibution*

$$p_f(\boldsymbol{\xi}) = c \text{ for all } \boldsymbol{\xi} \in \mathbb{R}^n \tag{25}$$

has no information that could be used in denoising. In fact, it corresponds to a score function that is identically zero, so the norm $\|p_f\|_\mathbf{s}$ *is zero.*

Example 2 *The gaussian distribution has minimum structure in the sense of* $\|.\|_\mathbf{s}$ *for a fixed covariance structure (Cover and Thomas, 1991). This holds for both our Fisher-information based measure and the more widely used Shannon entropy.*

Example 3 *Take any* $\mathbf{s}$ *with smooth pdf. Consider the variable rescaled variable* $\sigma\mathbf{s}$*. When* $\sigma \to 0$*, the* $\|p_\mathbf{s}\|_\mathbf{s}$ *goes to infinity. The structure becomes infinitely "strong" in the sense that we then know that* $\mathbf{s}$ *does not take any other values than zero. Conversely, if* $\sigma \to \infty$*,* $\|p_\mathbf{s}\|_\mathbf{s}$ *goes to zero, because the limit is the flat prior. On the other hand, translating the distribution as* $\mathbf{s} + \boldsymbol{\nu}$ *for a constant* $\boldsymbol{\nu}$ *does not change* $\|.\|_\mathbf{s}$*.*

5.3 Denoising capacity using imperfect model

In practice, we do not have a perfect model of $p_\mathbf{s}$, which was the case treated in previous sections. Denote by $\hat{p}$ our approximation of $p_\mathbf{s}$. Simple combination of the proofs of Theorems 1 and 4 gives the following general result

Theorem 5 *Assume that* $p(\mathbf{x}|\mathbf{s})$ *is as in Theorem 4. Assume we use* $\hat{p}$ *as the approximation of the prior* $p(\mathbf{s}|\hat{\boldsymbol{\theta}})$ *in the MAP estimator defined in (4). The denoising error can then be decomposed as*

$$tr(E\{(\mathbf{s}-\hat{\mathbf{s}})(\mathbf{s}-\hat{\mathbf{s}})^T\}) = n\sigma^2 - \sigma^4\|p_\mathbf{s}\|_\mathbf{s}^2 + \sigma^4\|\hat{p} - p_\mathbf{s}\|_\mathbf{s}^2 + \text{ terms of higher order in } \sigma^2 \tag{26}$$

We see that the error is increased by a factor proportional to the distance $\|\hat{p} - p_\mathbf{s}\|_\mathbf{s}^2$. Thus, it is this distance between $\hat{p}$ and $p_\mathbf{s}$ that gives the reduction of denoising capacity due to imperfect model. This enables us to interpret the distance as the amount of structure of $\mathbf{s}$ which is *not modelled* by $\hat{p}$. Thus, the metric we have defined is the *metric of optimal estimation* if the purpose is to construct a prior model of the data to be used in Bayesian inference such as denoising.

5.4 Orthogonal decomposition in exponential families

A particularly illustrative decomposition can be obtained for exponential families. Assume our model comes from an exponential family, i.e.

$$\log p(\mathbf{s}|\boldsymbol{\theta}) = \sum_{i=1}^{k} \theta_i g_i(\mathbf{s}) + \log Z(\boldsymbol{\theta}) \tag{27}$$

where the parameter vector $\boldsymbol{\theta}$ can take all values in $\mathbb{R}^k$, and Z is a normalizing constant that makes the integral equal to unity. The score functions are simply obtained:

$$\boldsymbol{\psi}(\mathbf{s}|\boldsymbol{\theta}) = \sum_{i=1}^{k} \theta_i \nabla g_i(\mathbf{s}) \tag{28}$$

which shows that the space of score functions in the model family is a linear subspace. This implies that estimation by minimization of $\|p(.|\boldsymbol{\theta}) - p_\mathbf{s}\|_\mathbf{s}$ is an orthogonal projection. In an orthogonal projection, the residual is orthogonal to the result of the projection. Denote the estimator minimizing $\|.\|_\mathbf{s}$ by $\hat{p}$. Then this orthogonality means

$$\langle \hat{p} - p_\mathbf{s}, \hat{p} \rangle_\mathbf{s} = 0 \tag{29}$$

which also implies the following Pythagorean decomposition

$$\|p_{\mathbf{s}}\|^2_{\mathbf{s}} = \|\hat{p} - p_{\mathbf{s}}\|^2_{\mathbf{s}} + \|\hat{p}\|^2_{\mathbf{s}} \tag{30}$$

This decomposition has a very interesting interpretation. We have by Theorem 5 and (30)

$$\mathrm{tr}(E\{(\mathbf{s}-\hat{\mathbf{s}})(\mathbf{s}-\hat{\mathbf{s}})^T\}) = \sigma^2 n + \sigma^4[\|\hat{p} - p_{\mathbf{s}}\|^2_{\mathbf{s}} - \|p_{\mathbf{s}}\|^2_{\mathbf{s}}] + o(\sigma^4) = \sigma^2 n - \sigma^4 \|\hat{p}\|^2_{\mathbf{s}} + o(\sigma^4) \tag{31}$$

So, we see that the terms in (30) can be intuitively interpreted so that the decomposition reads

$$\begin{array}{ccccc} \|p_{\mathbf{s}}\|^2_{\mathbf{s}} & = & \|\hat{p} - p_{\mathbf{s}}\|^2_{\mathbf{s}} & + & \|\hat{p}\|^2_{\mathbf{s}} \\ \text{Structure in data} & = & \text{Structure not modelled} & + & \text{Structure modelled} \end{array} \tag{32}$$

The interpretation of the first two terms here has already been discussed. The third term in (32) measures, according to (31), the denoising capacity when $\hat{p}$ is used as a model of the data. This is why, in general, we call it the *amount of structure modelled*. However, this decomposition is strictly true only in the case of the exponential family.

6 Simulations

We performed some simulations to investigate validate the approximations made in deriving our main Theorem (Theorem 1) and our method. In our simulations, the one-dimensional quantity s was corrupted by additive white gaussian noise. Four different distribution $p_o(s)$ were used:

1. The so-called "logistic" distribution of zero mean and unit variance, whose pdf is given by

$$\log p_0(s) = -2 \log \cosh(\frac{\pi}{2\sqrt{3}} s) - \log 4$$

2. A Gamma distribution with (4,1) degrees of freedom
3. A Chi-square distribution with 4 degrees of freedom
4. An asymmetric Gaussian density that was obtained by first taking a standardized Gaussian variable, and then multiplying the positive values by two.

All the four distributions were further standardized to zero mean and unit variance. All these four distibutions were modelled (approximated) by a logistic distribution with a location parameter θ:

$$\log p(s|\theta) = -2 \log \cosh(\frac{\pi}{2\sqrt{3}}(s-\theta)) - \log 4$$

which is very good for some of the four p_0 (perfect in the first case) and not very good in others.

A sample of 2,000 data points was obtained from each of the four prior distributions. The parameter θ was estimated using score matching, as well as maximum likelihood for comparison. Another sample of 10,000 data points was generated and Gaussian noise of different variances was added to it, which gave the corrupted data x. The MAP estimator $\hat{s}_{MAP}$ for s was then computed, for the two estimates of θ given by score matching estimation (SME) and maximum likelihood estimation (MLE) and for each of the 10,000 observed x's. The errors in the denoising inference were computed as $|\hat{s} - s|$ for the two estimators.[2]

[2] The results in the simulation use the absolute error, whereas the theorem considered squared error. This discrepancy is due to the fact that when we did simulations with squared error, no differences in the estimation performances could be made significant with standard tests (see next paragraph in main text for the testing procedure). It seems that the squared errors depend too strongly on a few outliers, and thus the sampling errors are too large to show significant differences between SME and MLE in a reasonable computing time.

The procedure was repeated four times with different noise levels.

Table 1 shows the obtained results. First, we see that the estimates obtained for θ are quite different for the two estimation methods, except in the case of the logistic distribution because it is symmetric around the mean and both methods are consistent. Note that because the logistic distribution is only an approximation for the other three pdf's, no "correct" value for θ is available for them, so the values of $\hat{\theta}$ cannot be compared with any ground truth.

The errors in s are what we essentially want to compare. The table shows that these are very close to each other, and the errors for SME are usually slightly smaller. This is to be understood from the decomposition given in Theorem 1 which shows that the error is a sum of two terms, which are approximately orthogonal. The second term $\mathcal{E}_2$ in (5), which does not depend on the estimation method for θ, is quite large. So, the differences in the errors between SME and MLE constitute only a small fraction of the total error.

The difference between the errors for SME and MLE is so small that one might doubt its statistical significance (it could be due to the limited sample used in the simulations). So, we performed a one-sided Wilcoxon signed-rank test. The test used the null hypothesis that the error in SME is larger than the error in MLE with a probability larger than 50%. All the p-values are smaller than 0.01 for other distributions than the logistic one.

For the logistic one, the p-values are neither very small nor very large, showing that MLE had neither better nor worse performance than SME: the differences in the errors are so small that even with the 10,000 samples no significant error can be seen. This is the case where p_0 belongs to the model family $p(.|\theta)$. This implies a reservation with respect to the applicability of our theorem: if the p_0 belongs to the model family $p(.|\theta)$, the errors approach zero in the limit of a large sample, and the approximation made in the theorem does not seem useful. We conjecture this is because the error $\mathcal{E}_1$ is then smaller than the error in the approximation due to lack of exact orthogonality of the errors. Thus, our theorem is interesting only when we are *approximating* the prior density p_0, and the approximation does not converge to the right values even for infinite samples.

The main result is, however, that for all the distributions where the family $p(\mathbf{s}|\boldsymbol{\theta})$ only gives an approximation to the true distribution of $\mathbf{s}$, the difference of average errors corresponding to SME and average errors corresponding to MLE are negative and the p-values are all smaller than 0.01. This confirms the utility of the approximation given in Theorem 1: using $\hat{\boldsymbol{\theta}}$ given by SME leads to smaller errors in the estimation of $\mathbf{s}$. However, the simulations also show that the advantage of SME is mainly of theoretical interest since the improvement in the estimates is quite small.

7 Conclusion

We have considered the estimation problem encountered in Bayesian perception and signal processing: the estimation of a prior model of a signal, based on a sample of such signals. If the objective is to have a prior that is optimal in Bayesian inference, the optimal estimation method is not maximum likelihood — at least not in the limit of very weak signal corruption which we analyzed. Rather, it turns out to be a generalization of the "score matching" estimator originally proposed purely on computational grounds in (Hyvärinen, 2005). Thus, we see that score matching has also some statistical optimality properties in signal restoration, although our simulations showed that the advantage with respect to maximum likelihood may be negligible in practice and only of theoretical interest. Moreover, it leads to a new geometric interpretation of statistical estimation, as well as a new approach to the measurement of how much "interesting structure" there is in a probability distribution.

	gamma	chi-square	asymm gauss	logistic
SM: value of $\hat{\theta}$	-0.249	-0.317	-0.299	-0.028
ML: value of $\hat{\theta}$	-0.088	-0.117	-0.081	-0.012
noise variance = 0.05				
SM: error in s	0.1765	0.175781	0.17279	0.173143
ML: error in s	0.176613	0.176128	0.173118	0.173136
mean of diffs	-0.00011331	-0.000346634	-0.000327916	7.44455e-06
p-value	0.00939945	9.78615e-08	3.84953e-06	0.831596
noise variance = 0.1				
SM: error in s	0.239926	0.233253	0.244252	0.238763
ML: error in s	0.24027	0.234266	0.244951	0.238748
mean of diffs	-0.000343373	-0.0010133	-0.000698879	1.45104e-05
p-value	0.000323944	3.99902e-13	1.6945e-08	0.658968
noise variance = 0.2				
SM: error in s	0.322587	0.319896	0.325484	0.323635
ML: error in s	0.322932	0.321494	0.326497	0.323639
mean of diffs	-0.000344532	-0.00159834	-0.001013	-4.09519e-06
p-value	0.000702169	2.28817e-13	1.45232e-09	0.409028
noise variance = 0.5				
SM: error in s	0.456851	0.444344	0.455303	0.454624
ML: error in s	0.457837	0.446792	0.455844	0.454601
mean of diffs	-0.000985527	-0.00244719	-0.00054089	2.29635e-05
p-value	1.20853e-06	2.77556e-16	1.4589e-09	0.524602

Table 1: Results for the simulations on denoising. For each of the four distributions, the estimates of θ are first given. Next, for each of the four noise levels, the errors in estimation of s using $\hat{\theta}$ from score matching (SM) or from maximum likelihood (ML) are given. For the errors in s, the mean of the error in SM minus error in ML is also given ("mean of diffs"). The p-values are from a one-sided Wilcoxon signed-rank test of the null hypothesis that the errors for SM are larger than the errors for ML with probability larger than 50%, i.e. the null hypothesis that SM leads to larger errors than ML.

A Proof of Theorem 1

Due to differentiablity of the functions, the gradient is zero at the point of MAP estimate. We obtain by definition of MAP:

$$\boldsymbol{\psi}(\hat{\mathbf{s}}|\hat{\boldsymbol{\theta}}) + \boldsymbol{\psi}(\mathbf{x}|\hat{\mathbf{s}}) = \mathbf{0} \tag{33}$$

Trivially, this can be manipulated to give

$$\boldsymbol{\psi}_0(\hat{\mathbf{s}}) + [\boldsymbol{\psi}(\hat{\mathbf{s}}|\hat{\boldsymbol{\theta}}) - \boldsymbol{\psi}_0(\hat{\mathbf{s}})] + \boldsymbol{\psi}(\mathbf{x}|\hat{\mathbf{s}}) = \mathbf{0} \tag{34}$$

We make a first-order Taylor expansion with respect to $\hat{\mathbf{s}}$ for the first and last terms on the left-hand side of (34) to yield

$$\begin{aligned} \boldsymbol{\psi}_0(\mathbf{s}_0 + \boldsymbol{\Delta}\mathbf{s}) + [\boldsymbol{\psi}(\hat{\mathbf{s}}|\hat{\boldsymbol{\theta}}) - \boldsymbol{\psi}_0(\hat{\mathbf{s}})] + \boldsymbol{\psi}(\mathbf{x}|\mathbf{s}_0 + \boldsymbol{\Delta}\mathbf{s}) \\ = \boldsymbol{\psi}_0(\mathbf{s}_0) + H_0(\mathbf{s}_0)\boldsymbol{\Delta}\mathbf{s} + [\boldsymbol{\psi}(\hat{\mathbf{s}}|\hat{\boldsymbol{\theta}}) - \boldsymbol{\psi}_0(\hat{\mathbf{s}})] \\ + \boldsymbol{\psi}(\mathbf{x}|\mathbf{s}_0) + H(\mathbf{x}|\mathbf{s}_0)\boldsymbol{\Delta}\mathbf{s} + o(\|\boldsymbol{\Delta}\mathbf{s}\|) = \mathbf{0} \end{aligned} \tag{35}$$

which gives, after reordering terms:

$$[H_0(\mathbf{s}_0) + H(\mathbf{x}|\mathbf{s}_0)]\boldsymbol{\Delta}\mathbf{s} = [\boldsymbol{\psi}_0(\hat{\mathbf{s}}) - \boldsymbol{\psi}(\hat{\mathbf{s}}|\hat{\boldsymbol{\theta}})] - [\boldsymbol{\psi}_0(\mathbf{s}_0) + \boldsymbol{\psi}(\mathbf{x}|\mathbf{s}_0)] + o(\|\boldsymbol{\Delta}\mathbf{s}\|) \tag{36}$$

Now, make a first-order approximation for the first term in brackets on the right-hand side:

$$\boldsymbol{\psi}_0(\hat{\mathbf{s}}) - \boldsymbol{\psi}(\hat{\mathbf{s}}|\hat{\boldsymbol{\theta}}) = \boldsymbol{\psi}_0(\mathbf{s}_0) - \boldsymbol{\psi}(\mathbf{s}_0|\hat{\boldsymbol{\theta}}) + o(\|\boldsymbol{\Delta}\mathbf{s}\|) \tag{37}$$

Thus, we can solve the estimation error by multiplying both sides of (36) by $\mathbf{M}^{-1}$. Taking the norm of both side then yields

$$\|\boldsymbol{\Delta}\mathbf{s}\|^2 = \|\mathcal{E}_1 + \mathcal{E}_2\|^2 + o(\|\mathbf{M}^{-1}\boldsymbol{\Delta}\mathbf{s}\|^2) \tag{38}$$

with $\mathcal{E}_1$ and $\mathcal{E}_2$ as given by the theorem. This holds for a given estimate $\hat{\boldsymbol{\theta}}$ and a given data sample $\mathbf{x}$, which then define the estimate $\hat{\mathbf{s}}$.

B Proof of Theorem 2

From (10), we obtain simply

$$\mathcal{J} = \frac{1}{2}\int p_0(\mathbf{s}) \sum_{ij} G_{ij}(\mathbf{s})[\psi_{0,i}(\mathbf{s}) - \psi_i(\mathbf{s}|\boldsymbol{\theta})][\psi_{0,j}(\mathbf{s}) - \psi_j(\mathbf{s}|\boldsymbol{\theta})]d\mathbf{s} \tag{39}$$

where $\psi_{0,i}$ denotes the i-th element of $\boldsymbol{\psi}_0$, i.e. the derivative of $\log p_0$ with respect to s_i. We will prove the theorem in the general case, for any functions G_{ij} that fulfill the regularity constraints. The proof is a simple variant of the partial integration trick used in basic score matching (Hyvärinen, 2005) based on earlier work by (Pham and Garrat, 1997; Taleb and Jutten, 1999). Simple manipulations give

$$\mathcal{J} \;=\; -\int p_0(\mathbf{s}) \sum_{ij} G_{ij}(\mathbf{s})\psi_{0,i}(\mathbf{s})\psi_j(\mathbf{s}|\boldsymbol{\theta})d\mathbf{s} \;+\; \frac{1}{2}\int p_0(\mathbf{s}) \sum_{ij} G_{ij}\psi_i(\mathbf{s}|\boldsymbol{\theta})\psi_j(\mathbf{s}|\boldsymbol{\theta})d\mathbf{s} \;+\; \text{const.} \tag{40}$$

where the constant only depends on p_0 and not on $\boldsymbol{\theta}$. The latter term on the right-hand side of (40) is clearly equal to the last term of $\mathcal{J}$ given in the theorem. What really needs to be proven is that the first term on the righ-hand side of (40) equals the sum of the first two terms of $\mathcal{J}$ in the theorem. Now, we use partial integration as follows:

$$\begin{aligned}\int p_0(\mathbf{s})G_{ij}(\mathbf{s})\psi_{0,i}(\mathbf{s})\psi_j(\mathbf{s}|\boldsymbol{\theta})d\mathbf{s} &= \int p_0(\mathbf{s})\frac{\partial_i p_0(\mathbf{s})}{p_0(\mathbf{s})}\psi_j(\mathbf{s}|\boldsymbol{\theta})G_{ij}(\mathbf{s})d\mathbf{s} \\ &= \int \partial_i p_0(\mathbf{s})\psi_i(\mathbf{s}|\boldsymbol{\theta})G_{ij}(\mathbf{s})d\mathbf{s} \\ = p_0(\mathbf{s})\psi_i(\mathbf{s}|\boldsymbol{\theta})G_{ij}(\mathbf{s})\big|_{s_i=\infty} - p_0(\mathbf{s})\psi_i(\mathbf{s}|\boldsymbol{\theta})G_{ij}(\mathbf{s})\big|_{s_i=-\infty} &- \int p_0(\mathbf{s})\partial_i(\psi_i(\mathbf{s}|\boldsymbol{\theta})G_{ij}(\mathbf{s}))d\mathbf{s} \\ &= -\int p_0(\mathbf{s})[(\partial_i G_{ij}(\mathbf{s}))\psi_i(\mathbf{s}|\boldsymbol{\theta}) + G_{ij}(\mathbf{s})\partial_i\psi_i(\mathbf{s}|\boldsymbol{\theta})]d\mathbf{s} \quad (41)\end{aligned}$$

where the disappearance of the two terms evaluated at infinity is due to the regularity assumptions of the theorem. (A more rigorous justification for this partial intergation element-by-element is given in Lemma 4 of (Hyvärinen, 2005)). In (40), we have a sum of such terms over i and j. When we take the sum, we obtain the first two terms in curly brackets in (12). Thus we have shown the theorem.

C Proof of Theorem 3

Actually, the theorem holds even for gaussian noise that is not i.i.d. We shall prove the theorem in this general case where

$$\boldsymbol{\psi}(\mathbf{x}|\mathbf{s}_0) = -\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \mathbf{s}_0) \quad (42)$$

which implies $H(\mathbf{x}|\mathbf{s}_0) = -\boldsymbol{\Sigma}^{-1}$. We assume that $\boldsymbol{\Sigma}^{-1}$ grows infinitely large with respect to some matrix norm, which is a generalization of $\sigma^2 \to 0$. We have

$$\begin{aligned}\langle\mathcal{E}_1, \mathcal{E}_2\rangle = \left[\boldsymbol{\psi}_0(\mathbf{s}_0) - \boldsymbol{\psi}(\mathbf{s}_0|\hat{\boldsymbol{\theta}})\right]^T &[H_0(\mathbf{s}_0) + H(\mathbf{x}|\mathbf{s}_0)]^{-2}\,[\boldsymbol{\psi}_0(\mathbf{s}_0) + \boldsymbol{\psi}(\mathbf{x}|\mathbf{s}_0)] \\ &\longrightarrow \left[\boldsymbol{\psi}_0(\mathbf{s}_0) - \boldsymbol{\psi}(\mathbf{s}_0|\hat{\boldsymbol{\theta}})\right]^T (\boldsymbol{\Sigma}^{-1})^{-2}\left[-\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \mathbf{s}_0)\right] \quad (43)\end{aligned}$$

because the terms with $\boldsymbol{\Sigma}^{-1}$, i.e. $H(\mathbf{x}|\mathbf{s}_0)$ and $\boldsymbol{\psi}(\mathbf{x}|\mathbf{s}_0)$ grow to be infinitely large with respect to the other terms. Now, we can take the expectation with respect to $\mathbf{x}$ to obtain

$$E_{\mathbf{x}}\{\langle\mathcal{E}_1, \mathcal{E}_2\rangle\} \longrightarrow \left[\boldsymbol{\psi}_0(\mathbf{s}_0) - \boldsymbol{\psi}(\mathbf{s}_0|\hat{\boldsymbol{\theta}})\right]^T \boldsymbol{\Sigma}^2\left[-\boldsymbol{\Sigma}^{-1}(\mathbf{s}_0 - \mathbf{s}_0)\right] = \mathbf{0} \quad (44)$$

because $E\{\mathbf{x}\} = \mathbf{s}$ and no other term except for $\mathbf{x}$ depends on $\mathbf{x}$, i.e. the sampling of the data. Thus we have proven the orthogonality.

References

Chipman, H. A., Kolczyk, E. D., and McCulloch, R. E. (1997). Adaptive Bayesian wavelet shrinkage. *J. of the American Statistical Association*, 92:1413–1421.

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory.* Wiley.

Dawid, A. P. and Lauritzen, S. L. (2005). The geometry of decision theory. In *Proc. 2nd Int. Symposium on Information Geometry and its Applications*, Tokyo, Japan.

Godsill, S. and Rayner, P. (1995). A Bayesian approach to restoration of degraded audio signals. *IEEE Trans. on Speech and Audio Processing*, 3:267–278.

Hyvärinen, A. (1999). Sparse code shrinkage: Denoising of nongaussian data by maximum likelihood estimation. *Neural Computation*, 11(7):1739–1768.

Hyvärinen, A. (2005). Estimation of non-normalized statistical models using score matching. *J. of Machine Learning Research*, 6:695–709.

Johnstone, I. and Silverman, B. (2005). Empirical Bayes selection of wavelet thresholds. *Annals of Statistics*, 33(4):1700–1752.

Knill, D. C. and Richards, W., editors (1996). *Perception as Bayesian Inference*. Cambridge University Press.

O'Ruanaidh, J. J. K. and Fitzgerald, W. J. (1996). *Numerical Bayesian Methods Applied to Signal Processing*. Springer.

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

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

Simoncelli, E. P. and Adelson, E. H. (1996). Noise removal via Bayesian wavelet coring. In *Proc. Third IEEE Int. Conf. on Image Processing*, pages 379–382, Lausanne, Switzerland.

Taleb, A. and Jutten, C. (1999). Source separation in post-nonlinear mixtures. *IEEE Trans. on Signal Processing*, 47(10):2807–2820.