Optimal approximation of signal priors

Aapo Hyvärinen *
Helsinki Institute for Information Technology
Dept of Computer Science, University of Helsinki
Finland

Revised submission

7th June 2007

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

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

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 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]

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, which is actually the target of the analysis in this paper, any considerations of squared error in $\boldsymbol{\theta}$ may be of little use and even ill-defined. In fact, the error in this case may not have anything to do with Fisher-efficiency, because even in the limit of an infinite sample, when the variance of the estimator goes to zero, the prior model will not be equal to the distribution of the data. 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 Hierarchical 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] Since the analysis developed below uses the mean-squared error, it might be suggested that the minimum mean-squared error (MMSE) estimator should be used instead. The main justification for our choice of the MAP estimation is that it is often much simpler computationally, and therefore much more widely used.

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.

2.2 Goals and limitations of the analysis

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 $\hat{\mathbf{s}}$.

An important choice we make in this analysis is that we completely neglect finite-sample effects, in other words, we assume that we have an infinite sample of $\mathbf{s}$. Thus, there is an error in the approximation of p_0 by our model $p(.|\boldsymbol{\theta})$ simply because p_0 does not belong to the model family, and *not* because of random fluctuations in the estimator $\hat{\boldsymbol{\theta}}$. This approach is quite realistic in the case of neural computation and signal processing, where the number of observations can often be made arbitrarily large (e.g. by just sampling more image patches) but the distributions are extremely complex and any model is only a rough approximation. This is, in fact, why we prefer to call this problem "approximation" of signal priors instead of estimation.

A limitation that was already pointed out in the introduction is that we assume we can access an uncorrupted sample of the original signals $\mathbf{s}$. This may be easy in some cases, but impossible in others. Many Empirical Bayes methods actually estimate parameters from corrupted signals, so our analysis is not applicable to them. Some examples on this are methods based on Stein's unbiased risk estimation, see Section 4.3.

2.3 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 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 using a perfect prior model, $\mathcal{E}_2$ would be zero (because according to the definition of the MAP estimator, the sum of these gradients has to be zero).

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

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

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

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

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.

[2] 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 (51) are finite.

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 functions") 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 $p(\mathbf{x}|\mathbf{s})$.

4 Case of gaussian infinitesimally small noise

4.1 Simplification of the estimator

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

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

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 $\sigma^2 \to 0$, we see that $\mathbf{M}$ approaches 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

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

4.2 Exact orthogonality of errors

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 $\mathbf{s}_0$:*

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

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 $\mathcal{E}_1$.

4.3 Relation to Stein's unbiased risk estimator

In the case of infinitesimal gaussian noise, we also see an interesting connection to Stein's unbiased risk estimator (SURE).[3]

Let us consider the following problem: Assume that the random variable x follows a normal distribution with unit variance and unknown mean μ. We consider estimators of μ of the following form

$$\hat{\mu} = x + g(x) \tag{19}$$

for some function g to be chosen. This can be interpreted in terms of denoising by considering μ to be the original signal and x a noisy observation. Stein (1981) showed that in this case, an unbiased estimator of the risk (i.e. expected squared error) of the estimator is obtained as

$$E_x\{(x + g(x) - \mu)^2\} = 1 + E_x\{g^2(x) + 2g'(x)\} \tag{20}$$

To see the connection with our framework, assume that the estimator is obtained by MAP estimation using a prior $p(.|\boldsymbol{\theta})$ for μ, with parameters $\boldsymbol{\theta}$. Further, assume that the noise is infinitesimal with respect to the signal; since (20) assumes that noise variance is unity, this means that we assume that the variance of μ is very large. Then, a first-order approximation of the MAP estimator gives (Hyvärinen, 1999)

$$\hat{\mu} = x - (\log p)'(x|\boldsymbol{\theta}) \tag{21}$$

In other words, MAP estimation leads to

$$g_{\boldsymbol{\theta}}(x) = -(\log p)'(x|\boldsymbol{\theta}) = -\psi(x|\boldsymbol{\theta}) \tag{22}$$

where $\psi(.|\boldsymbol{\theta})$ is the derivative of the logarithm of the prior pdf of μ, which depends on the parameters $\boldsymbol{\theta}$. Now, Stein's risk estimator gives, when averaged over the distribution of μ:

$$E_\mu E_x\{(x + g_{\boldsymbol{\theta}}(x) - \mu)^2\} = 1 + E_\mu E_x\{\psi^2(x|\boldsymbol{\theta}) + 2\psi'(x|\boldsymbol{\theta})\} \tag{23}$$

Since the variance of x is infinitesimal compared to the variance of μ, the expectation with respect to x can be ignored, and we can simply take $x = \mu$. Thus, minimization of this risk is equivalent to minimizing

$$E_\mu\{\frac{1}{2}\psi^2(\mu|\boldsymbol{\theta}) + \psi'(\mu|\boldsymbol{\theta})\} \tag{24}$$

which is nothing else than the original score matching objective proposed in (Hyvärinen, 2005). For notational simplicity, we considered here the one-dimensional case, but the result holds in n dimensions because we simply need to take the sum of the errors (risks) in different dimensions.

[3] I'm grateful to an anonymous referee for pointing out this connection.

Thus, SURE provides another way of deriving score matching estimation as the optimal prior in the special case of infinitesimal gaussian noise.

SURE can be estimated from noisy samples; it assumes noise variance is known but this can usually be estimated. It was applied for wavelet shrinkage by Donoho and Johnstone (1995) . The connection between score matching and SURE was pointed out by (Raphan and Simoncelli, 2007) in a rather different framework.

5 Interpretations as 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

$$\langle p_1, p_2 \rangle_{\mathbf{s}} = \int p_{\mathbf{s}}(\boldsymbol{\xi}) \left[\sum_{i=1}^{n} \psi_{1,i}(\boldsymbol{\xi}) \psi_{2,i}(\boldsymbol{\xi}) \right] d\boldsymbol{\xi} \tag{25}$$

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

$$\|p_1\|_{\mathbf{s}}^2 = \langle p_1, p_1 \rangle_{\mathbf{s}} = \int p_{\mathbf{s}}(\boldsymbol{\xi}) \left[\sum_{i=1}^{n} \psi_{1,i}(\boldsymbol{\xi})^2 \right] d\boldsymbol{\xi} = \int p_{\mathbf{s}}(\boldsymbol{\xi}) \|\boldsymbol{\psi}_1(\boldsymbol{\xi})\|^2 d\boldsymbol{\xi} \tag{26}$$

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

$$I_F(\mathbf{s}) = E\{\boldsymbol{\psi}(\mathbf{s}) \boldsymbol{\psi}(\mathbf{s})^T\}. \tag{27}$$

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

$$\|p_{\mathbf{s}}\|_{\mathbf{s}}^2 = \mathrm{tr}(I_F(\mathbf{s})) \tag{28}$$

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

$$\text{dist}_{\mathbf{s}}^2(p_1, p_2) = \int p_{\mathbf{s}}(\boldsymbol{\xi}) \left[\sum_{i=1}^{n} (\psi_{1,i}(\boldsymbol{\xi}) - \psi_{2,i}(\boldsymbol{\xi}))^2 \right] d\boldsymbol{\xi} = \int p_{\mathbf{s}}(\boldsymbol{\xi}) \|\boldsymbol{\psi}_1(\boldsymbol{\xi}) - \boldsymbol{\psi}_2(\boldsymbol{\xi})\|^2 d\boldsymbol{\xi} \tag{29}$$

Basically, we are defining something similar to a Hilbertian structure in the space of score functions ψ. 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 (26) 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*

$$tr(E\{(\mathbf{s} - \hat{\mathbf{s}})(\mathbf{s} - \hat{\mathbf{s}})^T\}) = n\sigma^2 - \sigma^4 \|p_{\mathbf{s}}\|_{\mathbf{s}}^2 + \text{ terms of higher order in } \sigma^2 \tag{30}$$

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 distribution*

$$p_f(\boldsymbol{\xi}) = c \text{ for all } \boldsymbol{\xi} \in \mathbb{R}^n \tag{31}$$

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 $p_{\mathbf{s}}$. 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*

$$tr(E\{(\mathbf{s} - \hat{\mathbf{s}})(\mathbf{s} - \hat{\mathbf{s}})^T\}) = n\sigma^2 - \sigma^4 \|p_{\mathbf{s}}\|_{\mathbf{s}}^2 + \sigma^4 dist_{\mathbf{s}}^2(\hat{p}, p_{\mathbf{s}}) + \text{ terms of higher order in } \sigma^2 \tag{32}$$

We see that the error is increased proportionally to the distance $\text{dist}^2_{\mathbf{s}}(\hat{p}, p_{\mathbf{s}})$. Thus, it is this distance between $\hat{p}$ and $p_{\mathbf{s}}$ that gives the reduction of denoising capacity due to an imperfect model. This enables us to interpret this 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.

$$\log p(\mathbf{s}|\boldsymbol{\theta}) = \sum_{i=1}^{k} \theta_i g_i(\mathbf{s}) + \log Z(\boldsymbol{\theta}) \tag{33}$$

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:

$$\boldsymbol{\psi}(\mathbf{s}|\boldsymbol{\theta}) = \sum_{i=1}^{k} \theta_i \nabla g_i(\mathbf{s}) \tag{34}$$

which shows that the space of score functions in the model family is a linear subspace. This implies that estimation by minimization of $\text{dist}^2_{\mathbf{s}}(p(.|\boldsymbol{\theta}), p_{\mathbf{s}})$ 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

$$\langle \hat{p} - p_{\mathbf{s}}, \hat{p} \rangle_{\mathbf{s}} = 0 \tag{35}$$

and it also implies the following Pythagorean decomposition

$$\|p_{\mathbf{s}}\|^2_{\mathbf{s}} = \text{dist}^2_{\mathbf{s}}(\hat{p}, p_{\mathbf{s}}) + \|\hat{p}\|_{\mathbf{s}} \tag{36}$$

This decomposition has a very interesting interpretation. We have by Theorem 5 and (36)

$$\text{tr}(E\{(\mathbf{s} - \hat{\mathbf{s}})(\mathbf{s} - \hat{\mathbf{s}})^T\}) = \sigma^2 n + \sigma^4[\text{dist}^2_{\mathbf{s}}(\hat{p}, p_{\mathbf{s}}) - \|p_{\mathbf{s}}\|_{\mathbf{s}}] + o(\sigma^4) = \sigma^2 n - \sigma^4 \|\hat{p}\|^2_{\mathbf{s}} + o(\sigma^4) \tag{37}$$

So, we see that the terms in (36) can be intuitively interpreted so that the decomposition reads

$$\begin{array}{rcccc} \|p_{\mathbf{s}}\|^2_{\mathbf{s}} & = & \text{dist}^2_{\mathbf{s}}(\hat{p}, p_{\mathbf{s}}) & + & \|\hat{p}\|^2_{\mathbf{s}} \\ \text{Structure in data} & = & \text{Structure not modelled} & + & \text{Structure modelled} \end{array} \tag{38}$$

The interpretation of the first two terms here has already been discussed. The third term in (38) measures, according to (37), 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 the validity of the approximations made in deriving our main Theorem (Theorem 1) and our estimation method. In our simulations, the one-dimensional quantity s was corrupted by additive gaussian noise. Four different distribution $p_0(s)$ were used:

1. A gaussian mixture model with different probabilities for the two kernels:

$$p_0(s) = \frac{3}{4}\varphi(s) + \frac{1}{4}\varphi(s-5) \tag{39}$$

   where φ is the standardized gaussian pdf.

2. A second gaussian mixture model which has a strong peak due to small variance of one of the kernels:

$$p_0(s) = \frac{1}{2\sigma 1}\varphi(s/\sigma_1) + \frac{1}{2}\varphi(s-5) \tag{40}$$

   where φ is the standardized gaussian pdf. The width of the first kernel was set to $\sigma_1 = 0.2$.

3. A Chi-square distribution with 4 degrees of freedom.

4. The Laplacian (double-exponential) distribution of zero mean and unit variance, whose pdf is given by

$$p_0(s) = \frac{1}{\sqrt{2}}\exp(\sqrt{2}|s|) \tag{41}$$

All the four distributions were further standardized to zero mean and unit variance. All these distributions were modelled (approximated) by a smoothened version of the Laplacian distribution with a location parameter θ to be estimated:

$$\log p(s|\theta) = -\frac{\sqrt{2}}{\gamma}\log\cosh(\gamma(s-\theta)) - Z(\gamma) \tag{42}$$

which is a very good model for the Laplacian data, but not good in others. The parameter γ controls the smoothness of the pdf: for $\gamma = \infty$, this distribution becomes the Laplacian distribution of unit variance. We used the value $\gamma = 10$.

A sample of 10,000 data points was obtained from each of the four prior distributions. Thus, any finite-sample effects were reduced to a minimum, and the effects investigated were almost exclusively due to the fact that p_0 does not belong to the model distribution family. The parameter θ was estimated using score matching, as well as maximum likelihood for comparison.

Another sample of 20,000 data points was generated from each distribution, 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 20,000 observed x's. The squared errors in the denoising inference were computed as $(\hat{s}-s)^2$ for the two estimators. The procedure was repeated four times with different noise levels.

In some cases, 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 t-test on the differences. (It was checked that the distributions are gaussian enough for the t-test to be valid.) The test used the null hypothesis that the mean error in SME is larger than the mean error in MLE. If the p-value is very small (< 0.01), SME is significantly better than MLE, and if it is very large (> 0.99), MLE is significantly better than SME.

To get a better idea of the scale of the errors, we compared the obtained denoising errors with the one obtained using a perfect prior, i.e. the true distribution s in the inference. We quantified the denoising performance using a performance index which was the amount of noise removed (reduction in squared error), based on either MLE or SME, as a proportion of the amount of noise removed using the perfect prior, expressed in percentages:

$$\text{Performance index} = 100 \times \frac{\text{Noise variance} - \text{Squared error using prior model } p(s|\hat{\theta})}{\text{Noise variance} - \text{Squared error using perfect prior model}} \tag{43}$$

This performance index can also take negative values, which means that the denoising estimator is so bad that it actually increases the noise in the data.

Table 1 shows the obtained results. First, we see that the estimates obtained for θ are quite different for the two estimation methods. The only exception is the case of the Laplacian distribution, because it is symmetric around the mean and both methods are consistent, eventually converging to the real location value of 0. Note that because the (smoothed) Laplacian distribution $p(s|\theta)$ is only a very rough 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 squared errors in $\hat{s}$ are what we essentially want to compare. For all the distributions except the Laplacian, SME gives a smaller squared error for the smaller noise levels, and the difference is statistically significant. The improvement is typically 5%-30% of the maximum noise reduction, as measured by the performance index. In some cases, MLE gives such a bad approximation that the noise is actually increased, which leads to a negative value for the performance index. For the largest noise level, MLE gives a smaller error in some cases. This is because the theory developed in this paper considers the limit of infinitesimal noise, and the first-order approximations used in the analysis break down at large noise levels.

For the Laplacian distribution, the p-values were some cases close to significance, and in others far from it: The difference in the errors is so small that even with the 20,000 samples hardly any significant error can be seen. Actually, the p-values here are irrelevant from the viewpoint our analysis, because in this case, the differences are completely due to finite-sample effects, since both methods are consistent, and converge on the value $\theta = 0$. This is the case where p_0 belongs (up to a small approximation error due to using (42) with finite γ) to the model family $p(.|\theta)$. A reservation with respect to the applicability of our theorem is thus necessary: if p_0 belongs to the model family $p(.|\theta)$, any improvement obtained by SME is negligible. This is presumably because the error $\mathcal{E}_1$ then approaches zero in the limit of a large sample, whereas $\mathcal{E}_2$ does not. So, the contribution of $\mathcal{E}_1$ becomes negligible and its minimization has no real effect on the denoising error. Thus, our method is interesting only when we are *approximating* the prior density p_0, and the approximation does not converge to the right data distribution even for an infinite sample.

The same results are shown in Table 2 for MMSE inference. That is, the parameter estimation was done as above, but the denoising was performed by taking the mean of the posterior distribution of $p(s|x)$. The results are qualitatively quite similar to those of Table 1, although slightly less favourable to SME, especially for the largest noise level. Note that there is no contradiction in MMSE estimation being less favourable to SME, even though SME was shown to approximately minimize mean squared error. This is because SME finds parameter values which minimize mean squared error for MAP inference, which is different from minimizing mean squared error for MMSE inference.

Thus, the simulations show that when 1) the data distribution is not very well approximated by the model, and 2) noise level is small, the average errors corresponding to SME are significantly smaller than average errors corresponding to MLE. The difference is significant statistically, and possibly also in practice for some distributions. 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 $\mathbf{s}$. Nevertheless, it could be argued that the advantage of SME is mainly of theoretical interest since the improvement is rather small and limited to the smallest noise levels. Future research is needed to see if the difference is important enough in some practical applications.

7 Conclusion

We 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. Our analysis is based on the assumption that we can observe a sample of uncorrupted signals to estimate the model.

Denoise by MAP	gauss mixt 1	gauss mixt 2	chi square	Laplacian
SM: value of $\hat{\theta}$	-0.447	-0.961	-0.505	-0.027
ML: value of $\hat{\theta}$	-0.335	-0.385	-0.225	0.002
noise variance = 0.05				
SM: squared error in $\hat{s}$	0.0451832	0.0437251	0.0475274	0.0458882
ML: squared error in $\hat{s}$	0.0458262	0.0539467	0.0481536	0.0458603
PP: error in s	0.0409871	0.0232552	0.0445455	0.0458615
p-value of difference	0	0	1.41838e-08	0.983707
SM: performance index	53.4441	23.4623	45.3304	99.3535
ML: performance index	46.3097	-14.757	33.8515	100.027
noise variance = 0.1				
SM: squared error in $\hat{s}$	0.0872924	0.0862113	0.0905881	0.0884577
ML: squared error in $\hat{s}$	0.0889856	0.109962	0.0928176	0.0884553
PP: error in s	0.0765966	0.0491676	0.0840644	0.0884539
p-value of difference	0	0	2.22045e-16	0.534107
SM: performance index	54.298	27.1258	59.0617	99.967
ML: performance index	47.0631	-19.5983	45.071	99.9879
noise variance = 0.2				
SM: squared error in $\hat{s}$	0.157888	0.175383	0.168148	0.163455
ML: squared error in $\hat{s}$	0.162068	0.214278	0.171554	0.163322
PP: error in s	0.141524	0.129202	0.153341	0.163327
p-value of difference	0	0	2.00357e-09	0.989493
SM: performance index	72.0148	34.7715	68.2662	99.6507
ML: performance index	64.8668	-20.1669	60.9658	100.014
noise variance = 0.5				
SM: squared error in $\hat{s}$	0.359553	0.495427	0.353038	0.336505
ML: squared error in $\hat{s}$	0.363577	0.48297	0.344779	0.33623
PP: error in s	0.331788	0.45787	0.307364	0.336233
p-value of difference	0	0.999999	1	0.984166
SM: performance index	83.4939	10.8552	76.2902	99.8336
ML: performance index	81.1015	40.4231	80.5778	100.002

Table 1: Results for the simulations on denoising using MAP estimation. 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 comparison, the error for MAP denoising using the perfect prior (PP) is shown. The p-value is for the null hypothesis that the mean error for SME is larger than the mean error for MLE; at the same time, it tests the opposite hypothesis, so that one minus the p-value is the p-value for the null hypothesis that the mean SME error is smaller. The performance index shows the denoising obtained as a percentage of the denoising using perfect prior.

Denoise by MMSE	gauss mixt 1	gauss mixt 2	chi square	Laplacian
SM: value of $\hat{\theta}$	-0.447	-0.961	-0.505	-0.027
ML: value of $\hat{\theta}$	-0.335	-0.385	-0.225	0.002
noise variance = 0.05				
SM: squared error in $\hat{s}$	0.0452207	0.045682	0.0470846	0.0457491
ML: squared error in $\hat{s}$	0.0457162	0.052363	0.0475836	0.0457327
PP: error in s	0.0406318	0.0227558	0.044132	0.0457332
p-value of difference	0	0	5.25571e-10	0.968615
SM: performance index	51.0166	15.8492	49.684	99.6284
ML: performance index	45.7274	-8.67333	41.1795	100.012
noise variance = 0.1				
SM: squared error in $\hat{s}$	0.0879704	0.0915913	0.0890099	0.0878702
ML: squared error in $\hat{s}$	0.0891757	0.103574	0.090115	0.0878727
PP: error in s	0.0742282	0.0434066	0.0829422	0.087871
p-value of difference	0	0	6.7579e-10	0.446729
SM: performance index	46.6775	14.8582	64.4289	100.007
ML: performance index	42.0005	-6.31553	57.9501	99.9861
noise variance = 0.2				
SM: squared error in $\hat{s}$	0.160096	0.185088	0.162824	0.159106
ML: squared error in $\hat{s}$	0.16198	0.192796	0.162506	0.159044
PP: error in s	0.128385	0.0957017	0.149551	0.159045
p-value of difference	0	0	0.805858	0.954085
SM: performance index	55.7204	14.2974	73.689	99.8501
ML: performance index	53.0903	6.90742	74.3201	100.002
noise variance = 0.5				
SM: squared error in $\hat{s}$	0.346242	0.455679	0.328342	0.313556
ML: squared error in $\hat{s}$	0.343421	0.394174	0.313815	0.313384
PP: error in s	0.278469	0.26963	0.287739	0.313387
p-value of difference	1	1	1	0.979591
SM: performance index	69.4072	19.239	80.8715	99.9093
ML: performance index	70.6803	45.9376	87.7152	100.002

Table 2: Same as Table 1 but using MMSE estimator instead of MAP. The parameter estimates are identical to Table 1 and repeated for convenience only.

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, in addition to its original motivation, which was computational simplicity. Our simulations confirmed that signal restoration based on score matching estimation has smaller errors when compared to maximum likelihood estimation, although the difference may be small in practice and mainly of theoretical interest.

Moreover, the analysis 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, based on the capacity of denoising using that structure.

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:

$$\boldsymbol{\psi}(\hat{\mathbf{s}}|\hat{\boldsymbol{\theta}}) + \boldsymbol{\psi}(\mathbf{x}|\hat{\mathbf{s}}) = \mathbf{0} \tag{44}$$

Trivially, this can be manipulated to give

$$\boldsymbol{\psi}_0(\hat{\mathbf{s}}) + [\boldsymbol{\psi}(\hat{\mathbf{s}}|\hat{\boldsymbol{\theta}}) - \boldsymbol{\psi}_0(\hat{\mathbf{s}})] + \boldsymbol{\psi}(\mathbf{x}|\hat{\mathbf{s}}) = \mathbf{0} \tag{45}$$

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 (45) to yield

$$\begin{aligned}\boldsymbol{\psi}_0(\mathbf{s}_0 + \boldsymbol{\Delta}\mathbf{s}) + [\boldsymbol{\psi}(\hat{\mathbf{s}}|\hat{\boldsymbol{\theta}}) - \boldsymbol{\psi}_0(\hat{\mathbf{s}})] + \boldsymbol{\psi}(\mathbf{x}|\mathbf{s}_0 + \boldsymbol{\Delta}\mathbf{s}) \\ = \boldsymbol{\psi}_0(\mathbf{s}_0) + H_0(\mathbf{s}_0)\boldsymbol{\Delta}\mathbf{s} + [\boldsymbol{\psi}(\hat{\mathbf{s}}|\hat{\boldsymbol{\theta}}) - \boldsymbol{\psi}_0(\hat{\mathbf{s}})] \\ + \boldsymbol{\psi}(\mathbf{x}|\mathbf{s}_0) + H(\mathbf{x}|\mathbf{s}_0)\boldsymbol{\Delta}\mathbf{s} + o(\|\boldsymbol{\Delta}\mathbf{s}\|) = \mathbf{0}\end{aligned} \tag{46}$$

which gives, after reordering terms:

$$[H_0(\mathbf{s}_0) + H(\mathbf{x}|\mathbf{s}_0)]\boldsymbol{\Delta}\mathbf{s} = [\boldsymbol{\psi}_0(\hat{\mathbf{s}}) - \boldsymbol{\psi}(\hat{\mathbf{s}}|\hat{\boldsymbol{\theta}})] - [\boldsymbol{\psi}_0(\mathbf{s}_0) + \boldsymbol{\psi}(\mathbf{x}|\mathbf{s}_0)] + o(\|\boldsymbol{\Delta}\mathbf{s}\|) \tag{47}$$

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

$$\boldsymbol{\psi}_0(\hat{\mathbf{s}}) - \boldsymbol{\psi}(\hat{\mathbf{s}}|\hat{\boldsymbol{\theta}}) = \boldsymbol{\psi}_0(\mathbf{s}_0) - \boldsymbol{\psi}(\mathbf{s}_0|\hat{\boldsymbol{\theta}}) + o(\|\boldsymbol{\Delta}\mathbf{s}\|) \tag{48}$$

Thus, we can solve the estimation error by multiplying both sides of (47) by $\mathbf{M}^{-1}$. Taking the norm of both side then yields

$$\|\boldsymbol{\Delta}\mathbf{s}\|^2 = \|\mathcal{E}_1 + \mathcal{E}_2\|^2 + o(\|\mathbf{M}^{-1}\boldsymbol{\Delta}\mathbf{s}\|^2) \tag{49}$$

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

$$\mathcal{J} = \frac{1}{2}\int p_0(\mathbf{s}) \sum_{ij} G_{ij}(\mathbf{s})[\psi_{0,i}(\mathbf{s}) - \psi_i(\mathbf{s}|\boldsymbol{\theta})][\psi_{0,j}(\mathbf{s}) - \psi_j(\mathbf{s}|\boldsymbol{\theta})]d\mathbf{s} \tag{50}$$

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

$$\mathcal{J} \;=\; -\int p_0(\mathbf{s}) \sum_{ij} G_{ij}(\mathbf{s}) \psi_{0,i}(\mathbf{s}) \psi_j(\mathbf{s}|\boldsymbol{\theta}) d\mathbf{s} \;+\; \frac{1}{2} \int p_0(\mathbf{s}) \sum_{ij} G_{ij} \psi_i(\mathbf{s}|\boldsymbol{\theta}) \psi_j(\mathbf{s}|\boldsymbol{\theta}) d\mathbf{s} \;+\; \text{const.} \quad (51)$$

where the constant only depends on p_0 and not on $\boldsymbol{\theta}$. The latter term on the right-hand side of (51) 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 right-hand side of (51) equals the sum of the first two terms of $\mathcal{J}$ in the theorem. Now, we use partial integration as follows:

$$\begin{aligned}
\int p_0(\mathbf{s}) G_{ij}(\mathbf{s}) \psi_{0,i}(\mathbf{s}) \psi_j(\mathbf{s}|\boldsymbol{\theta}) d\mathbf{s} &= \int p_0(\mathbf{s}) \frac{\partial_i p_0(\mathbf{s})}{p_0(\mathbf{s})} \psi_j(\mathbf{s}|\boldsymbol{\theta}) G_{ij}(\mathbf{s}) d\mathbf{s} \\
&= \int \partial_i p_0(\mathbf{s}) \psi_i(\mathbf{s}|\boldsymbol{\theta}) G_{ij}(\mathbf{s}) d\mathbf{s} \\
= p_0(\mathbf{s}) \psi_i(\mathbf{s}|\boldsymbol{\theta}) G_{ij}(\mathbf{s})|_{s_i=\infty} - p_0(\mathbf{s}) \psi_i(\mathbf{s}|\boldsymbol{\theta}) G_{ij}(\mathbf{s})|_{s_i=-\infty} &- \int p_0(\mathbf{s}) \partial_i (\psi_i(\mathbf{s}|\boldsymbol{\theta}) G_{ij}(\mathbf{s})) d\mathbf{s} \\
&= -\int p_0(\mathbf{s}) [(\partial_i G_{ij}(\mathbf{s})) \psi_i(\mathbf{s}|\boldsymbol{\theta}) + G_{ij}(\mathbf{s}) \partial_i \psi_i(\mathbf{s}|\boldsymbol{\theta})] d\mathbf{s} \quad (52)
\end{aligned}$$

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 integration element-by-element is given in Lemma 4 of (Hyvärinen, 2005)). In (51), 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

$$\boldsymbol{\psi}(\mathbf{x}|\mathbf{s}_0) = -\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \mathbf{s}_0) \quad (53)$$

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

$$\begin{aligned}
\langle \mathcal{E}_1, \mathcal{E}_2 \rangle = \left[\boldsymbol{\psi}_0(\mathbf{s}_0) - \boldsymbol{\psi}(\mathbf{s}_0|\hat{\boldsymbol{\theta}})\right]^T &\left[H_0(\mathbf{s}_0) + H(\mathbf{x}|\mathbf{s}_0)\right]^{-2} \left[\boldsymbol{\psi}_0(\mathbf{s}_0) + \boldsymbol{\psi}(\mathbf{x}|\mathbf{s}_0)\right] \\
&\longrightarrow \left[\boldsymbol{\psi}_0(\mathbf{s}_0) - \boldsymbol{\psi}(\mathbf{s}_0|\hat{\boldsymbol{\theta}})\right]^T (\boldsymbol{\Sigma}^{-1})^{-2} \left[-\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \mathbf{s}_0)\right] \quad (54)
\end{aligned}$$

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 $\mathbf{x}$, given $\mathbf{s}_0$, to obtain

$$E_{\mathbf{x}|\mathbf{s}_0}\{\langle \mathcal{E}_1, \mathcal{E}_2 \rangle\} \longrightarrow \left[\boldsymbol{\psi}_0(\mathbf{s}_0) - \boldsymbol{\psi}(\mathbf{s}_0|\hat{\boldsymbol{\theta}})\right]^T \boldsymbol{\Sigma}^2 \left[-\boldsymbol{\Sigma}^{-1}(\mathbf{s}_0 - \mathbf{s}_0)\right] = \mathbf{0} \quad (55)$$

because $E\{\mathbf{x}|\mathbf{s}_0\} = \mathbf{s}_0$ and no other term except for $\mathbf{x}$ in the second brackets depends on $\mathbf{x}$, i.e. the sampling of the observed 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.

Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. of the American Statistical Association*, 90:1200–1224.

Godsill, S. J. and Rayner, P. J. W. (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. M. and Silverman, B. W. (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.

Raphan, M. and Simoncelli, E. P. (2007). Learning to be Bayesian without supervision. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press.

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.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9:1135–1151.

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