ESTIMATION THEORY AND INFORMATION GEOMETRY BASED ON DENOISING

Aapo Hyvärinen

Dept of Computer Science, Dept of Mathematics & Statistics, and HIIT
University of Helsinki, Finland.

ABSTRACT

We consider a new estimation method ("score matching") for parametric statistical models. It is based on optimal denoising using empirical Bayes. The ensuing method has the additional advantage that it does not require the model probability densities to be properly normalized, unlike maximum likelihood. In fact, it does not even require the model densities to be integrable, so one can use improper model densities. Furthermore, the method leads to a new geometric intepretation of estimation.

1. INTRODUCTION

1.1. Signal denoising by Empirical Bayes

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 (e.g. an *image*), 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 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".)

The starting point of our analysis is to consider 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.

1.2. Denoising with infinitesimal gaussian noise

To simplify the analysis, we make here the following assumptions:

1. *A point estimate of $\boldsymbol{\theta}$ is used.* This is because computational considerations usually make integration over the parameter space (e.g. feature space in the case of images) too expensive.

2. *A point estimate of the $\mathbf{s}$ is used.* This is because in practical applications, the posterior must typically be output as a point estimate (e.g. a denoised image).

3. *The point estimate is the maximum a posteriori (MAP) estimate.* This is the typical, and computationally most feasible, point estimate to summarize the posterior of $\mathbf{s}$. MAP estimation simply means finding the value that maximizes the posterior density of $\mathbf{s}$, given an estimate $\hat{\boldsymbol{\theta}}$:

$$\begin{aligned}\hat{\mathbf{s}}_{MAP}(\hat{\boldsymbol{\theta}}, \mathbf{x}) &= \arg\max_{\mathbf{s}} p(\mathbf{x}|\mathbf{s})p(\mathbf{s}|\hat{\boldsymbol{\theta}}) \\ &= \arg\max_{\mathbf{s}} \log p(\mathbf{x}|\mathbf{s}) + \log p(\mathbf{s}|\hat{\boldsymbol{\theta}}) \end{aligned} \tag{2}$$

   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.

4. *The corrupting process is additive gaussian noise with infinitesimal variance.* Additive gaussian noise is the archetypal corrupting process, and an infinitesimal variance allows first-order approximations which are the core of the analysis given here.

Thus, $\mathbf{x}$ is the sum of an n-dimensional nongaussian random vector $\mathbf{s}$ and the noise vector $\mathbf{n}$:

$$\mathbf{x} = \mathbf{s} + \mathbf{n}. \tag{3}$$

where the noise $\mathbf{n}$ is gaussian and of covariance $\sigma^2\mathbf{I}$, where σ^2 is infinitesimal. The maximum a posteriori (MAP) estimator for $\mathbf{s}$ is

$$\hat{\mathbf{s}} = \arg\min_{\mathbf{u}} \frac{1}{2\sigma^2}\|\mathbf{x} - \mathbf{u}\|^2 + \log p_{\mathbf{s}}(\mathbf{u}) \tag{4}$$

which is the basic method of Bayesian denoising. We obtain (e.g. [1])

$$\hat{\mathbf{s}} = \mathbf{g}(\mathbf{x}) \tag{5}$$

where the function $\mathbf{g}$ is defined by

$$\mathbf{g}^{-1}(\mathbf{u}) = \mathbf{u} - \sigma^2 \nabla \log p_{\mathbf{s}}(\mathbf{u}). \tag{6}$$

1.3. Why maximum likelihood may not be optimal

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. 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 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. [2]. However, such considerations seem to be irrelevant if the goal is Bayesian (MAP) inference of $\mathbf{s}$.

1.4. Our approach

In this paper, we summarize the theory developed in [3, 4, 5] and present it in a simplified, unified setting. We show how to obtain the optimal estimator of $\boldsymbol{\theta}$ in terms of denoising, i.e. inference of $\mathbf{s}$ (Sections 2 and 3). The ensuing solution turns out to have an important computational advantage as well: it enables the consistent estimation of non-normalized models without computation of the normalization constant (Section 4). Since the estimator is based on minimization of a Euclidean distance, we propose an intuitive intepretation of the estimator in terms of a geometrical projection (Section 5). The quantities involved can also be interpreted in terms of "structure" inherent in the data, as has hitherto been done in the case of Shannon entropy. Finally, we point out an interesting aspect of the theory, which is that the model densities do not need to be integrable at all, i.e. they need not be proper densities (Section 6).

2. ANALYSIS OF ESTIMATION ERROR

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. 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 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 $p(\mathbf{x}|\mathbf{s}_0)$. In our simplified setting, this means we add gaussian noise of infinitesimal variance (see [5] for a more general approach).

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 (2).

In step 4, the data generating process $p(\mathbf{x}|\mathbf{s})$ is assumed to be exactly 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.

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

$$\boldsymbol{\psi}(\mathbf{s}|\boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial \log p(\mathbf{s}|\boldsymbol{\theta})}{\partial s_1} \\ \vdots \\ \frac{\partial \log p(\mathbf{s}|\boldsymbol{\theta})}{\partial s_n} \end{pmatrix} = \begin{pmatrix} \psi_1(\mathbf{s}|\boldsymbol{\theta}) \\ \vdots \\ \psi_n(\mathbf{s}|\boldsymbol{\theta}) \end{pmatrix} = \nabla_{\mathbf{s}} \log p(\mathbf{s}|\boldsymbol{\theta}) \tag{7}$$

and the corresponding Hessian matrix by

$$H(\mathbf{s}|\boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial \log p(\mathbf{s}|\boldsymbol{\theta})}{\partial s_1 s_1} & \cdots & \frac{\partial \log p(\mathbf{s}|\boldsymbol{\theta})}{\partial s_1 s_n} \\ & \vdots & \\ \frac{\partial \log p(\mathbf{s}|\boldsymbol{\theta})}{\partial s_n s_1} & \cdots & \frac{\partial \log p(\mathbf{s}|\boldsymbol{\theta})}{\partial s_n s_n} \end{pmatrix} = \nabla_{\mathbf{s}} \boldsymbol{\psi}(\mathbf{s}|\boldsymbol{\theta})^T \quad (8)$$

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 $\hat{\mathbf{s}} = \hat{\mathbf{s}}_{MAP}(\hat{\boldsymbol{\theta}}, \mathbf{x})$.

2.2. Decomposition of error

Our first result is given in the following theorem, obtained by combining Theorems 1 and 2 in [5]:

Theorem 1 *Assume that all the log-pdf's in (2) 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*

$$E\{\|\boldsymbol{\Delta}\mathbf{s}\|^2\} = \sigma^4 E\{\|\mathcal{E}_1\|^2\} + \sigma^4 E\{\|\mathcal{E}_2\|^2\} + o(\sigma^2 \|\boldsymbol{\Delta}\mathbf{s}\|^2) \quad (9)$$

where

$$\mathcal{E}_1 = \boldsymbol{\psi}_0(\mathbf{s}_0) - \boldsymbol{\psi}(\mathbf{s}_0|\hat{\boldsymbol{\theta}}) \quad (10)$$

$$\mathcal{E}_2 = \boldsymbol{\psi}_0(\mathbf{s}_0) + \boldsymbol{\psi}(\mathbf{x}|\mathbf{s}_0) \quad (11)$$

and the expectation is taken over the distribution p_0 for the $\mathbf{s}_0$.

Now, the error vector in $\mathcal{E}_2$ is a function of $\mathbf{s}_0$ and $\mathbf{x}$ only, i.e. the data generating parts (steps 3 and 4) above. Thus, it does 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). Thus, we see a clear decomposition of the error in two parts

- 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).

3. PROPOSAL OF OPTIMAL ESTIMATOR

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.

Thus, taking the expected value of the error $\|\mathcal{E}_1\|^2$ over all $\mathbf{s}$ with respect to p_0, and introducing the factor $1/2$ for notational simplicity, we arrive at the following objective function:

$$\mathcal{J}(\boldsymbol{\theta}) = \frac{1}{2} \int p_0(\mathbf{s}) \|\boldsymbol{\psi}_0(\mathbf{s}) - \boldsymbol{\psi}(\mathbf{s}|\boldsymbol{\theta})\|^2 d\mathbf{s} \quad (12)$$

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 (2) 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.

It may seem that the objective function $\tilde{\mathcal{J}}$ is computationally intractable because it uses $\boldsymbol{\psi}_0(\mathbf{s})$ 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 [3], see also [6]. Thus, we can use the following equivalent form:

Theorem 2 *Under some regularity constraints [3], the objective function in (12) can be expressed as*

$$\mathcal{J}(\boldsymbol{\theta}) = \int p_0(\mathbf{s}) \left[\sum_i \partial_i \psi_i(\mathbf{s}|\boldsymbol{\theta}) + \frac{1}{2} \psi_i(\mathbf{s}|\boldsymbol{\theta})^2 \right] d\mathbf{s} + \textit{const.} \quad (13)$$

where ∂_i denotes differentiation with respect to s_i, and the constant term does not depend on $\boldsymbol{\theta}$.

The proof, reproduced in the Appendix, is based on a simple trick of partial integration.

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

$$\tilde{\mathcal{J}}(\boldsymbol{\theta}) = \sum_{t=1}^{T} \sum_i \partial_i \psi_i(\mathbf{s}(t)|\boldsymbol{\theta}) + \frac{1}{2} \psi_i(\mathbf{s}(t))^2 \quad (14)$$

where we have omitted the irrelevant constant.

The sample version in (14) 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.

4. COMPUTATIONAL ADVANTAGE OF PROPOSED ESTIMATOR

In fact, the objective function in (14) was originally proposed in [3] because it solves an unrelated computational

problem. 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

$$p(\mathbf{s}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} q(\mathbf{s}|\boldsymbol{\theta}) \tag{15}$$

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

$$Z(\boldsymbol{\theta}) = \int q(\mathbf{s}|\boldsymbol{\theta})\, d\mathbf{s} \tag{16}$$

Now, the important point is that the derivatives of the log-density with respect to the s_i ($\boldsymbol{\psi}$, "score functions") do not depend on Z at all, so the problem of computing the normalization constant disappears when we consider only the score functions. This is because

$$\begin{aligned} \nabla_{\mathbf{s}} \log p(\mathbf{s}|\boldsymbol{\theta}) &= \nabla_{\mathbf{s}} \log q(\mathbf{s}|\boldsymbol{\theta}) - \nabla_{\mathbf{s}} \log Z(\boldsymbol{\theta}) \\ &= \nabla_{\mathbf{s}} \log q(\mathbf{s}|\boldsymbol{\theta}) \end{aligned} \tag{17}$$

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 as in (12). This leads to the present objective function, independently of any considerations of statistical optimality .

Thus, the computational advantage of the objective function in (14) is that it does not contain Z, or any other any integrals or other expressions which would be difficult to compute. This is in stark contrast to the maximum likelihood estimator, which would require a numerical evaluation of the integral in (16). In [3], it was further proven that such an estimator is (locally) consistent.

We can thus conclude that our proposed estimator combines statistical optimality, in the sense of denoising, with computational simplicity, in the sense that the prior model $p(\mathbf{s}|\boldsymbol{\theta})$ does not need to integrate to unity.

5. AN ALTERNATIVE APPROACH TO INFORMATION GEOMETRY

In this section, we will propose two intuitive interpretations of the estimation performed by the proposed estimator, i.e. score matching. The interpretations are 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 needed 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

$$\langle p_1, p_2 \rangle_{\mathbf{s}} = \int p_{\mathbf{s}}(\boldsymbol{\xi}) \left[\sum_{i=1}^{n} \psi_{1,i}(\boldsymbol{\xi}) \psi_{2,i}(\boldsymbol{\xi}) \right] d\boldsymbol{\xi} \tag{18}$$

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

$$\begin{aligned} \|p_1\|_{\mathbf{s}}^2 = \langle p_1, p_1 \rangle_{\mathbf{s}} &= \int p_{\mathbf{s}}(\boldsymbol{\xi}) \left[\sum_{i=1}^{n} \psi_{1,i}(\boldsymbol{\xi})^2 \right] d\boldsymbol{\xi} \\ &= \int p_{\mathbf{s}}(\boldsymbol{\xi}) \|\boldsymbol{\psi}_1(\boldsymbol{\xi})\|^2 d\boldsymbol{\xi} \end{aligned} \tag{19}$$

where the notation $\|.\|$, without a subscript, in the rightmost 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

$$I_F(\mathbf{s}) = E\{\boldsymbol{\psi}(\mathbf{s}) \boldsymbol{\psi}(\mathbf{s})^T\}. \tag{20}$$

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

$$\|p_{\mathbf{s}}\|_{\mathbf{s}}^2 = \mathrm{tr}(I_F(\mathbf{s})) \tag{21}$$

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

$$\begin{aligned} &\mathrm{dist}_{\mathbf{s}}^2(p_1, p_2) \\ &\quad = \int p_{\mathbf{s}}(\boldsymbol{\xi}) \left[\sum_{i=1}^{n} (\psi_{1,i}(\boldsymbol{\xi}) - \psi_{2,i}(\boldsymbol{\xi}))^2 \right] d\boldsymbol{\xi} \\ &\quad = \int p_{\mathbf{s}}(\boldsymbol{\xi}) \|\boldsymbol{\psi}_1(\boldsymbol{\xi}) - \boldsymbol{\psi}_2(\boldsymbol{\xi})\|^2 d\boldsymbol{\xi} \end{aligned} \tag{22}$$

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 (19) is closely related to denoising capacity. In previous work, we proved the following:

Theorem 3 *Assume that $p(\mathbf{x}|\mathbf{s})$ is a gaussian distribution with mean $\mathbf{s}$ and covariance $\sigma^2\mathbf{I}$. The squared error of the MAP estimator $\hat{\mathbf{s}}$, when the distribution $p_{\mathbf{s}}$ is exactly known,is given by*

$$tr(E\{(\mathbf{s}-\hat{\mathbf{s}})(\mathbf{s}-\hat{\mathbf{s}})^T\}) = n\sigma^2 - \sigma^4\|p_{\mathbf{s}}\|_{\mathbf{s}}^2 + \text{ terms of higher order in } \sigma^2 \quad (23)$$

This is a simple corollary of Theorem 2 in [1].

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*

$$p_f(\boldsymbol{\xi}) = c \text{ for all } \boldsymbol{\xi} \in \mathbb{R}^n \quad (24)$$

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 [2]. 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 $\mathbf{s} + \boldsymbol{\mu}$ for a constant $\boldsymbol{\mu}$ 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 $p_{\mathbf{s}}$. Combinining the proofs of Theorems 1 and 3 gives the following general result

Theorem 4 *Assume that $p(\mathbf{x}|\mathbf{s})$ is as in Theorem 3. Assume we use $\hat{p}$ as the approximation of the prior $p(\mathbf{s}|\hat{\boldsymbol{\theta}})$ in the MAP estimator defined in (2). The denoising error can then be decomposed as*

$$tr(E\{(\mathbf{s}-\hat{\mathbf{s}})(\mathbf{s}-\hat{\mathbf{s}})^T\}) = n\sigma^2 - \sigma^4\|p_{\mathbf{s}}\|_{\mathbf{s}}^2 + \sigma^4 dist_{\mathbf{s}}^2(\hat{p}, p_{\mathbf{s}}) + \text{ terms of higher order in } \sigma^2 \quad (25)$$

We see that the error is increased proportionally to the distance $\text{dist}_{\mathbf{s}}^2(\hat{p}, p_{\mathbf{s}})$. 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. For exponential families, the decomposition is orthogonal as will be shown next.

5.4. Exponential families

Now we show how the theory is simplified for exponential families.

5.4.1. Orthogonal decomposition

First we show how a particularly illustrative geometric decomposition can be obtained in the case of exponential families.

Assume our model comes from an exponential family defined as:

$$\log p(\mathbf{s}|\boldsymbol{\theta}) = \sum_{i=1}^{n} \theta_i F_i(\mathbf{s}) + \log Z(\boldsymbol{\theta}) \quad (26)$$

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

$$\boldsymbol{\psi}(\mathbf{s}|\boldsymbol{\theta}) = \sum_{i=1}^{n} \theta_i \nabla F_i(\mathbf{s}) \quad (27)$$

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

$$\langle \hat{p} - p_{\mathbf{s}}, \hat{p} \rangle_{\mathbf{s}} = 0 \quad (28)$$

and it also implies the following Pythagorean decomposition

$$\|p_{\mathbf{s}}\|_{\mathbf{s}}^2 = \text{dist}_{\mathbf{s}}^2(\hat{p}, p_{\mathbf{s}}) + \|\hat{p}\|_{\mathbf{s}}^2 \quad (29)$$

This decomposition has a very interesting interpretation. Above, we interpreted the term on the left-hand side as the amount of structure in the data, and the first term on the right-hand side as the amount of structure not modelled. On the other hand, we have by Theorem 4 and (29)

$$\text{tr}(E\{(\mathbf{s}-\hat{\mathbf{s}})(\mathbf{s}-\hat{\mathbf{s}})^T\}) = \sigma^2 n - \sigma^4\|\hat{p}\|_{\mathbf{s}}^2 + o(\sigma^4) \quad (30)$$

which suggests an intepretation of $\|\hat{p}\|_{\mathbf{s}}^2$ as the structure (succesfully) modelled, because it is the reduction in denoising error when using the model. So, we see that the

terms in (29) can be intuitively interpreted so that the decomposition reads

$$\begin{aligned}\text{Structure in data} \\ = \text{ Structure not modelled} \\ + \text{ Structure modelled}\end{aligned} \quad (31)$$

each term in this "equation" correspondinf to one term in Eq. (29).

5.4.2. Closed-form solution

Another interesting property of the exponential family in the context of score matching was shown in [4]: the estimator can be obtained in closed form. Again, we assume that the parameter space is $\mathbb{R}^m$, i.e. $\boldsymbol{\theta}$ can take all possible real values.

Let us denote the matrix of partial derivatives of F, i.e. its Jacobian, by $\mathbf{K}(\xi)$, with elements defined as:

$$K_{ki}(\boldsymbol{\xi}) = \frac{\partial F_k}{\partial \xi_i}, \quad (32)$$

and the required matrix of second derivatives by

$$H_{ki}(\boldsymbol{\xi}) = \frac{\partial^2 F_k}{\partial \xi_i^2}. \quad (33)$$

Now, we have

$$\psi_i(\boldsymbol{\xi}; \boldsymbol{\theta}) = \sum_{k=1}^{m} \theta_k K_{ki}(\boldsymbol{\xi}), \quad (34)$$

and the objective function $\tilde{\mathcal{J}}$ in (14) becomes

$$\begin{aligned}\tilde{\mathcal{J}}(\boldsymbol{\theta}) &= \frac{1}{T}\sum_{t=1}^{T}\sum_{i} \\ &\left[\frac{1}{2}(\sum_{k=1}^{m}\theta_k K_{ki}(\mathbf{s}(t)))^2 + \sum_{k=1}^{m}\theta_k H_{ki}(\mathbf{s}(t))\right] \\ &= \frac{1}{2}\boldsymbol{\theta}^T(\frac{1}{T}\sum_{t=1}^{T}\mathbf{K}(\mathbf{s}(t))\mathbf{K}(\mathbf{s}(t))^T)\boldsymbol{\theta} \\ &\quad + \boldsymbol{\theta}^T(\frac{1}{T}\sum_{t=1}^{T}\sum_{i} H_{ki}(\mathbf{s}(t))).\end{aligned} \quad (35)$$

This is a simple quadratic form of $\boldsymbol{\theta}$. Thus, the minimizing $\boldsymbol{\theta}$ can be easily solved by computing the gradient and setting it to zero. This gives $\hat{\boldsymbol{\theta}}$ is closed form as

$$\hat{\boldsymbol{\theta}} = -\left[\hat{E}\{\mathbf{K}(\mathbf{s})\mathbf{K}(\mathbf{s})^T\}\right]^{-1}(\sum_{i}\hat{E}\{\mathbf{h}_i(\mathbf{s})\}), \quad (36)$$

where $\hat{E}$ denotes the sample average (i.e. expectation over the sample distribution), and the vector $\mathbf{h}_i(x)$ is the i-th column of the matrix $\mathbf{H}$ defined in (33).

6. NON-INTEGRABLE DENSITIES

To conclude this review, we point out an interesting property of the this theory which has hitherto escaped our attention. The model densities are not only allowed to be non-normalized (which is merely a computational simplification), but in fact, there is no need to assume that they are integrable at all. Consider, for example, the following model density on the real line:

$$p(x; \mu, \sigma) = [1 + \exp(-\frac{x-\mu}{\sigma})]^{-1} \quad (37)$$

with $\sigma > 0$. For any values of μ and σ, we have

$$\lim_{x\to\infty} x = 1 \text{ and } \int p(x; \mu, \sigma)dx = \infty \quad (38)$$

However, there is no reason why we could not use this model in the estimation theory based on score matching. We can estimate the model parameters by minimization of (14) without any modification. The geometric interpretation presented in Section 5 is completely valid, and Fisher information is well-defined.

This observation greatly relaxes the constraints on the functional forms that one can use in specifying the models. In practice, the constraint that the densities must go to zero at infinity often forces the modeller to change the algebraically, conceptually, or computationally simplest functional forms to satisfy this constraint. However, that may not be necessary using score matching theory.

In our example in (37), it is intuitively clear what the "meaning" of the density and the parameters is: almost all the data should have greater values than μ, and σ expresses how strict this constraint is. Using classic theory, the modeller would have to figure out some simple way of making the density go to zero as $x \to \infty$, which would complicate this model considerably.

Nevertheless, not any function can be used as density even in score matching. First of all, the density function must be positive, and their log-derivatives must not grow too fast, in order to make all the integrals involved in the objective function finite. A sufficient condition is that the log-derivatives of the model density are all bounded.

7. CONCLUSION

We started by considering 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. The corruption process was assumed to be additive gaussian noise with infinitesimal variance.

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 the "score matching" estimator originally proposed purely on computational grounds in [3].

The score matching estimator does not require the probability density to be normalized to unit integral. In contrast to maximum likelihood estimation, score matching thus avoids computational problems related to numerical integration of the probability density. In fact, the model densities are not required to be integrable at all. Furthermore, for exponential families, the estimator can be obtained in closed form.

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.

Moreover, the analysis leads to a new geometric interpretation of statistical estimation as projection, 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. For exponential families, the projection is orthogonal.

An alternative viewpoint on this estimation framework is provided by [7]. See also [8] for connections to other methods, and [9] for a practical application of the framework.

A. PROOF OF THEOREM 2

The proof is reproduced from [3]. Definition (12) gives

$$\mathcal{J}(\boldsymbol{\theta}) = \int p_0(\boldsymbol{\xi}) \left[\frac{1}{2} \|\boldsymbol{\psi}_0(\boldsymbol{\xi})\|^2 + \frac{1}{2} \|\boldsymbol{\psi}(\boldsymbol{\xi};\boldsymbol{\theta})\|^2 - \boldsymbol{\psi}_0(\boldsymbol{\xi})^T \boldsymbol{\psi}(\boldsymbol{\xi};\boldsymbol{\theta}) \right] d\boldsymbol{\xi} \quad (39)$$

The first term in brackets does not depend on $\boldsymbol{\theta}$, and can be ignored. The integral of the second term is simply integral of the sum of the second terms in brackets in (13). Thus, the difficult thing to prove is that integral of the third term in brackets in (39) equals the integral of the sum of the first terms in brackets in (13). This term equals

$$-\sum_i \int p_0(\boldsymbol{\xi}) \psi_{0,i}(\boldsymbol{\xi}) \psi_i(\boldsymbol{\xi};\theta) d\boldsymbol{\xi}$$

where $\psi_{0,i}(\boldsymbol{\xi})$ denotes the i-th element of the vector $\boldsymbol{\psi}_0(\boldsymbol{\xi})$. We can consider the integral for a single i separately, which equals

$$\begin{aligned} &- \int p_0(\boldsymbol{\xi}) \frac{\partial \log p_0(\boldsymbol{\xi})}{\partial \xi_i} \psi_i(\boldsymbol{\xi};\boldsymbol{\theta}) d\boldsymbol{\xi} \\ &\quad = - \int \frac{p_0(\boldsymbol{\xi})}{p_0(\boldsymbol{\xi})} \frac{\partial p_0(\boldsymbol{\xi})}{\partial \xi_i} \psi_i(\boldsymbol{\xi};\boldsymbol{\theta}) d\boldsymbol{\xi} \\ &\quad = - \int \frac{\partial p_0(\boldsymbol{\xi})}{\partial \xi_i} \psi_i(\boldsymbol{\xi};\boldsymbol{\theta}) d\boldsymbol{\xi} \end{aligned}$$

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

$$\begin{aligned} &\int p(x) (\log p)'(x) f(x) dx \\ &\quad = \int p(x) \frac{p'(x)}{p(x)} f(x) dx \\ &\quad = \int p'(x) f(x) dx \\ &\quad = - \int p(x) f'(x) dx \end{aligned}$$

under some regularity assumptions that will be dealt with below.

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

Lemma 1

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

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

Proof of lemma:

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

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

Now, we can apply this lemma on p_0 and $\psi_1(\boldsymbol{\xi};\boldsymbol{\theta})$ which were both assumed to be differentiable in the theorem, and we obtain:

$$\begin{aligned} &- \int \frac{\partial p_0(\boldsymbol{\xi})}{\partial \xi_1} \psi_1(\boldsymbol{\xi};\boldsymbol{\theta}) d\boldsymbol{\xi} \\ &\quad = - \int \left[\int \frac{\partial p_0(\boldsymbol{\xi})}{\partial \xi_1} \psi_1(\boldsymbol{\xi};\boldsymbol{\theta}) d\xi_1 \right] d(\xi_2, \ldots, \xi_n) \\ &\quad = - \int \Big[\lim_{a \to \infty, b \to -\infty} \\ &\quad [p_0(a, \xi_2, \ldots, \xi_n) \psi_1(a, \xi_2, \ldots, \xi_n; \boldsymbol{\theta}) \\ &\quad - p_0(b, \xi_2, \ldots, \xi_n) \psi_1(b, \xi_2, \ldots, \xi_n; \boldsymbol{\theta})] \\ &\quad - \int \frac{\partial \psi_1(\boldsymbol{\xi};\boldsymbol{\theta})}{\partial \xi_1} p_0(\boldsymbol{\xi}) d\xi_1 \Big] d(\xi_2, \ldots, \xi_n) \end{aligned}$$

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

The limit in the above expression is zero for any $\xi_2, \ldots, \xi_n, \boldsymbol{\theta}$ because we assumed that $p_0(\boldsymbol{\xi}) \boldsymbol{\psi}(\boldsymbol{\xi};\boldsymbol{\theta})$ goes to zero at infinity. Thus, we have proven that

$$- \int \frac{\partial p_0(\boldsymbol{\xi})}{\partial \xi_i} \psi_i(\boldsymbol{\xi};\boldsymbol{\theta}) d\boldsymbol{\xi} = \int \frac{\partial \psi_i(\boldsymbol{\xi};\boldsymbol{\theta})}{\partial \xi_i} p_0(\boldsymbol{\xi}) d\boldsymbol{\xi}$$

that is, integral of the the third term in brackets in (39) equals the integral of the sum of the first terms in brackets in (13), and the proof of the theorem is complete.

B. REFERENCES

[1] A. Hyvärinen, "Sparse code shrinkage: Denoising of nongaussian data by maximum likelihood estimation," *Neural Computation*, vol. 11, no. 7, pp. 1739–1768, 1999.

[2] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, 2nd edition, 2006.

[3] A. Hyvärinen, "Estimation of non-normalized statistical models using score matching," *J. of Machine Learning Research*, vol. 6, pp. 695–709, 2005.

[4] A. Hyvärinen, "Some extensions of score matching," *Computational Statistics & Data Analysis*, vol. 51, pp. 2499–2512, 2007.

[5] A. Hyvärinen, "Optimal approximation of signal priors," *Neural Computation*, In press.

[6] D.-T. Pham and P. Garrat, "Blind separation of mixture of independent sources through a quasi-maximum likelihood approach," *IEEE Trans. on Signal Processing*, vol. 45, no. 7, pp. 1712–1725, 1997.

[7] A. P. Dawid and S. L. Lauritzen, "The geometry of decision theory," in *Proc. 2nd Int. Symposium on Information Geometry and its Applications*, Tokyo, Japan, 2005.

[8] A. Hyvärinen, "Connections between score matching, contrastive divergence, and pseudolikelihood for continuous-valued variables," *IEEE Transactions on Neural Networks*, vol. 18, pp. 1529–1531, 2007.

[9] U. Köster and A. Hyvärinen, "A two-layer ICA-like model estimated by score matching," in *Proc. Int. Conf. on Artificial Neural Networks (ICANN2007)*, Porto, Portugal, 2007, pp. 798–807.