Information Geometry, Bayesian Inference, Ideal Estimates and Error Decomposition

Huaiyu Zhu *

Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501.

Email: zhuh@santafe.edu

Richard Rohwer

HNC Software, Inc., 5930 Cornerstone Court West, San Diego, CA, 92121

Email: rjr@hnc.com

April 24, 1998

Running title: Bayesian information geometric inference

Abstract

In statistics it is necessary to study the relation among many probability distributions. Information geometry elucidates the geometric structure on the space of all distributions. When combined with Bayesian decision theory, it leads to the new concept of "ideal estimates". They uniquely exist in the space of finite measures, and are generally sufficient statistic. The optimal estimate on any model is given by projecting the ideal estimate onto that model. An error decomposition theorem splits the error of an estimate into the sum of statistical error and approximation error. They can be expanded to yield higher order asymptotics. Furthermore, the ideal estimates under certain uniform priors, invariantly defined in information geometry, corresponds to various optimal non-Bayesian estimates, such as the MLE.

Keywords: Bayesian inference, ideal estimate, information geometry, error decomposition, nonparametric estimation

1 Introduction

Say we have a sample z taken from an unknown true distribution p. It is required to find an estimated distribution q which is as close to p as possible. This is a typical estimation problem, and there exist many criteria for evaluating the goodness of a given solution.

Two problems present themselves immediately. The first is that the true distribution p is unknown, so it should be clarified what is meant by the closeness between q and an unknown object. The second is that the closeness between p and q generally cannot be measured by their distance in an arbitrary parameter space, as the latter is not invariant with respect to the change of variables. Aitchison (1975) considered this problem in the Bayesian framework, using the Kullback-Leibler deviation [1] as a cost function. The KL-deviation is invariant, solving the second problem. The posterior mean deviation between p and q is used in place of the distance between the unknown p and q, solving the first problem.

In this paper we demonstrate that a generalized version of this approach describes most known criteria for statistical estimation in a unified way. The main purpose of this paper is not to develop the theory to its fullest extent, but to demonstrate how it is possible to "have the best of both worlds" of Bayesian and asymptotic theories: coherence and covariance.

Let $\mathcal{P}$ be the space of all probability distributions on the sample space Z, and $\mathcal{Q} \subseteq \mathcal{P}$ be the model in which the estimate q is sought. Assume for the moment that all our knowledge about the true distribution p, excluding the sample z, can be represented as a prior distribution $P(p)$ on $\mathcal{P}$. Other assumptions can be treated later. It is not required that $P(p)$ is supported on $\mathcal{Q}$. For otherwise the question of robustness would not exist. The Bayes theorem gives the posterior $P(p|z)$, which is also a distribution on $\mathcal{P}$.

The goal of estimation is to find $q \in \mathcal{Q}$ that contains the most amount of information in the

[1] We use the term "deviation" in place of the more conventional "divergence", following the reasons given in (Čencov, 1982, §8, Note (1)).

sample (and the prior). It will be shown that a family of information deviations D_δ, $\delta \in [0,1]$, work extremely well as cost functions. The corresponding optimal estimate will be called the δ-estimate.

An essential technical step in this approach is to first seek the optimal estimate q not in the model space $\mathcal{Q}$ or even the space $\mathcal{P}$, but in the space $\widetilde{\mathcal{P}}$ of all the finite measures (normalizable but not necessarily normalized distributions). The reward for doing so will become clear in §5, but here are some highlights:

- The δ-optimal estimate $\widehat{p} \in \widetilde{\mathcal{P}}$ always uniquely exists. It is given explicitly as a posterior average. It can be called a δ-"ideal estimate" for the following reasons.
- The δ-optimal estimate $\widehat{q} \in \mathcal{Q}$ for any model $\mathcal{Q}$ is given by the δ-projection of $\widehat{p}$ onto $\mathcal{Q}$. In particular, the projection onto the space of probabilities $\mathcal{P}$ is a simple normalization.
- There exists an "error decomposition" theorem: The error of any estimate q is the sum of uncertainty (error of $\widehat{p}$) and approximation error (deviation of q from $\widehat{p}$.)
- The δ-ideal estimate $\widehat{p}$ is a sufficient statistic of the posterior under most of the usual assumptions. For example, it is so for exponential family models.
- The projection onto $\mathcal{Q}$ is not guaranteed to be unique, unless $\mathcal{Q}$ is $(1-\delta)$-flat, but an enclosing flat model can keep all the necessary ancillary information. This is the case for curved exponential families.
- Many "nice" results resembling those properties of Gaussian measures on linear spaces now apply to any statistical problems without regularity conditions.

The magic number δ above is simply an index of the affine structures of statistical manifolds. Its special instances have been well known in statistics, from exponential families (0-flat or flat in log-likelihood) to mixture families (1-flat or flat in likelihood). The exponential families admit sufficient statistics simply because they are 0-flat; The MLE has best second order efficiency

(smallest deficiency or information loss) simply because it is 1-straight; The 1-deviation is the KL-deviation; The $1/2$-deviation corresponds to the Hellinger distance (Amari, 1985). The Jeffreys' prior is $(1/2)$-uniform, which is uniform in the parameterization stabilizing asymptotic variances; The "normal likelihood" parameterization is $2/3$-uniform; The "asymptotically non-skewed" parameterization is $1/3$-uniform (Hougaard, 1982; Kass, 1984). Note that our $\delta \in [0,1]$ follows (Kass, 1984; Hougaard, 1982), and corresponds to $(1-\alpha)/2$ in (Amari, 1982, 1985), $1/u$ in (Čencov, 1982), α in (LeCam, 1970; Rényi, 1961; Hartigan, 1965; Ferguson, 1973), t in (Chernoff, 1952), k in (Hartigan, 1967).

Statistics has always been a science of methodology as well as mathematics. Information geometry helps to reduce some choices of *principles* into choices of *indices*. Hopefully this will help to clarify the mathematics behind the methodology.

In §2 we use an example, perhaps the simplest possible, to illustrate the main results and techniques required. Sections 3 and 4 reviews necessary background of information geometry. Although the geometric language may be unfamiliar, most of the concrete concepts are well-known in statistics. Our main results are in §5. The generality and limitations of this approach is discussed in §6.

2 A Simple Example and Some Observations

To make things more concrete, suppose we want to estimate the parameter of the binomial model $P(\boldsymbol{n}|\boldsymbol{p}) = C(n_1,n_2)p_1^{n_1}p_2^{n^2}$, where $C(n_1,n_2) = (n_1+n_2)!/n_1!n_2!$. Assuming a Beta prior $P(\boldsymbol{p}) = B(\boldsymbol{p}|\boldsymbol{a}) = p_1^{a_1-1}p_2^{a_2-1}/B(a_1,a_2)$, the posterior is also Beta, $P(\boldsymbol{p}|\boldsymbol{n}) = B(\boldsymbol{p}|\boldsymbol{b})$ with $\boldsymbol{b} = \boldsymbol{a}+\boldsymbol{n}$. Using the Kullback-Leibler deviation

$$K(\boldsymbol{p},\boldsymbol{q}) = \sum_i p_i \log \frac{p_i}{q_i}, \tag{2.1}$$

two posterior mean losses of an estimate $\boldsymbol{q}$ can be constructed

$$E_1(\boldsymbol{q}|\boldsymbol{n}) = \int_p P(\boldsymbol{p}|\boldsymbol{n})K(\boldsymbol{p},\boldsymbol{q}), \qquad E_0(\boldsymbol{q}|\boldsymbol{n}) = \int_p P(\boldsymbol{p}|\boldsymbol{n})K(\boldsymbol{q},\boldsymbol{p}), \tag{2.2}$$

and their minimization leads to the optimal estimates

$$\operatorname*{Min}_{\boldsymbol{q}} E_1(\boldsymbol{q}|\boldsymbol{n}) \implies q_i = \int_p P(\boldsymbol{p}|\boldsymbol{n})p_i = b_i/(b_1+b_2), \tag{2.3}$$

$$\operatorname*{Min}_{\boldsymbol{q}} E_0(\boldsymbol{q}|\boldsymbol{n}) \implies \log q_i = \int_p P(\boldsymbol{p}|\boldsymbol{n}) \log p_i + \text{constant} = \Psi(b_i) + \text{constant}, \tag{2.4}$$

where Ψ is the the digamma function, the logarithmic derivative of the Γ function.

As will be seen shortly, $K(\boldsymbol{p},\boldsymbol{q}) = D_1(\boldsymbol{p},\boldsymbol{q}) = D_0(\boldsymbol{q},\boldsymbol{p})$ are special cases of the family of information deviations D_δ, while $l_1(\boldsymbol{p}) = \boldsymbol{p}$, $l_0(\boldsymbol{p}) = \log \boldsymbol{p}$ are special instances of a family of coordinates l_δ, $\delta \in [0,1]$. The examples (2.3)–(2.4) are special cases of a general result:

$$\operatorname*{Min}_{\boldsymbol{q}} E_\delta(\boldsymbol{q}|\boldsymbol{n}) := \int_p P(\boldsymbol{p}|\boldsymbol{n})D_\delta(\boldsymbol{p},\boldsymbol{q}) \implies q_i \propto \widehat{p}_i := l_\delta^{-1}\left(\int_p P(\boldsymbol{p}|\boldsymbol{n})l_\delta(p_i)\right). \tag{2.5}$$

More details about this example can be found in (Zhu and Rohwer, 1995b).

One thing that is not satisfactory about this result is that $\boldsymbol{q} \propto \widehat{\boldsymbol{p}}$ instead of $\boldsymbol{q} = \widehat{\boldsymbol{p}}$. This is because $\boldsymbol{q} \in \mathcal{P}$, the space of probabilities, while in general $\widehat{\boldsymbol{p}} \in \widetilde{\mathcal{P}}$, the space of finite measures, and $\widehat{\boldsymbol{p}} \notin \mathcal{P}$ unless $\delta = 1$. We call $\widehat{\boldsymbol{p}}$ the ideal estimate, and we would like to draw statisticians' attention to many of its nice properties. Statistical theories in the space $\widetilde{\mathcal{P}}$ (as a flat enveloping space) are simpler and more intuitive, in the same way Euclidean geometry compares with spherical geometry. Results thus obtained can be easily translated back to $\mathcal{P}$ since the projection of $\widehat{\boldsymbol{p}}$ onto $\mathcal{P}$ is a simple normalization, rendering (2.5) a special case. Nonetheless, this does require the extension of information deviation and other concepts to $\widetilde{\mathcal{P}}$, as will be given in the next section. In the section that follows several pertinent concepts from information geometry will be reviewed.

The 1-optimal estimate has several interesting and well-known special cases: The Bayes-Laplace prior ($\boldsymbol{a} = [1,1]$) (Fisher, 1936) gives "Laplace's rule of succession" $q_i = (n_i+1)/(n+2)$. The Fisher-Jeffreys prior ($\boldsymbol{a} = [1/2,1/2]$) (Fisher, 1922, p. 325) (Jeffreys, 1961, p. 63, p. 179)

gives $q_i = (n_i + 1/2)/(n + 1)$. The Haldane prior ($\boldsymbol{a} = [0, 0]$) (Jeffreys, 1961, p. 123) gives the maximum likelihood estimate (MLE) $q_i = n_i/n$. See also (Good, 1965) for some historical literature. The interpretations of these priors will be clear in §5.4.

3 Certain Functions of Finite Measures

3.1 Spaces constructed from measures and homogeneous function

We generally follow Halmos and Savage (1949) for notation and terminology concerning measures, except the following caveat. While it is not always crucial in statistics to distinguish between measures and functions, although Bayes and Fisher never failed in their writings to draw attention to it, such distinctions are important here. We use notations like $p = f\,dx$ where p and dx are measures while f is a function. For example, let dx be the Lebesgue measure on $\mathbb{R}$, and

$$f(x) := (2\pi)^{-1/2} e^{-x^2/2}, \qquad F(x) := \int_{-\infty}^{x} f(t)dt. \tag{3.1}$$

Then

$$p = f\,dx = dF \tag{3.2}$$

is the standard Gaussian measure on $\mathbb{R}$,

$$f = p/dx = dF/dx \tag{3.3}$$

is its density function (relative to Lebesgue measure), and

$$F(x) = \int_{[-\infty,x]} dF = p([-\infty, x]) \tag{3.4}$$

is the corresponding probability distribution function. The notation $\int p = 1$ is well-defined here while the more standard $\int dp$ is not. In the latter notation p behaves more like F, but because F is only defined when X is a region in an Euclidean space, it will not be considered in this paper.

Now consider a measurable space $[Z, \mathcal{F}]$, where Z is a sample space and $\mathcal{F}$ is a σ-algebra of measurable sets. Reference to $\mathcal{F}$ or $(Z, \mathcal{F})$ are omitted when there is no risk of confusion. Denote by $\mathcal{M}_+$, $\widetilde{\mathcal{P}}$, and $\mathcal{P}$ the space of measures, finite measures and probability measures on $[Z, \mathcal{F}]$, respectively. In other words, $\widetilde{\mathcal{P}} = \left\{p \in \mathcal{M}_+ : \int p < \infty\right\}$, $\mathcal{P} = \left\{p \in \widetilde{\mathcal{P}} : \int p = 1\right\}$.

A function $F : \mathbb{R}_+^m \to \mathbb{R}_+ \cup 0$ is called homogeneous if

$$\forall a_1, \dots a_m, c \in \mathbb{R}_+ : \; F(ca_1, \dots, ca_m) = cF(a_1, \dots, a_m). \tag{3.5}$$

It can be naturally extended to a homogeneous function for finite measures, $F : \widetilde{\mathcal{P}}^m \to \widetilde{\mathcal{P}}$, by

$$F(p_1, \dots, p_m) := rF\left(\frac{p_1}{r}, \dots, \frac{p_m}{r}\right), \qquad \forall p_i, r \in \widetilde{\mathcal{P}}, \; r \equiv p_1 + \dots + p_m. \tag{3.6}$$

where p/r is the Radon-Nikodým derivative. The expression $F(p_1, \dots, p_m)$ is independent of the choice of $r \equiv p_1 + \dots + p_m$.

Most statistical theories deal with a dominated family of measures, with much regularity assumptions. They can be avoided by using the concept of fractional power of measures as introduced by Neveu (1965, IV.1.4, p. 112–113). Let $p \in \widetilde{\mathcal{P}}$, $\delta \in (0, 1]$,

$$f \in L_{1/\delta}(p) := \left\{f : \int p|f|^{1/\delta} < \infty\right\}. \tag{3.7}$$

Define an equivalence relation among the couples $[f, p]$ by $[f, pg] = [fg^\delta, p]$. Then the equivalence class of $[f, p]$ may be unambiguously denoted $p^\delta f$. The space of δth power of finite measures,

$$L_{1/\delta} := \left\{p^\delta f : \; p \in \widetilde{\mathcal{P}}, f \in L_{1/\delta}(p)\right\}, \tag{3.8}$$

is a Banach space with obvious definition of addition, multiplication and norm. In particular, L_2 is a Hilbert space. For any $p \in \widetilde{\mathcal{P}}$, the space $L_{1/\delta}(p) \subset L_{1/\delta}$ isometrically. The mapping $p \to p^\delta$ maps $\widetilde{\mathcal{P}}$ onto $\widetilde{\mathcal{P}}^\delta \subset L_{1/\delta}$.

It should be pointed out that the geometries useful for statistics when $\delta \in \{0, 1\}$ is not described by the Lebesgue spaces L_1 and L_∞, but instead by certain Orlicz spaces. They are Banach spaces with the same linear structure but different norms.

3.2 Information deviation

Information deviations are the appropriate substitutes for squared distances in the space of probabilities. Let $\delta \in [0,1]$. For finite measures $p, q \in \widetilde{\mathcal{P}}$, the δ-deviation D_δ is defined as

$$F_\delta(p,q) := \begin{cases} \dfrac{\delta p + (1-\delta)q - p^\delta q^{1-\delta}}{\delta(1-\delta)} & \delta \in (0,1), \\ \lim\limits_{\delta\to 0} F_\delta(p,q) = p - q + q\log(q/p), & \delta = 0, \\ \lim\limits_{\delta\to 1} F_\delta(p,q) = q - p + p\log(p/q), & \delta = 1. \end{cases} \tag{3.9}$$

$$D_\delta(p,q) := \int F_\delta(p,q). \tag{3.10}$$

The motivation of this definition was given in (Zhu and Rohwer, 1995c), based on considerations of the dual affine geometry (Amari, 1985, Chap. 3). Here it is better taken as just an arbitrary definition which is only justified by its usefulness. The following properties shows its resemblance to squared distance.

Proposition 3.1 (Algebraic properties)

$$F_\delta(p,q) = F_{1-\delta}(q,p), \qquad D_\delta(p,q) = F_{1-\delta}(q,p). \tag{3.11}$$

$$F_\delta(ap,aq) = aF_\delta(p,q), \qquad D_\delta(ap,aq) = aD_\delta(p,q), \qquad \forall a > 0. \tag{3.12}$$

$$F_\delta(p,q) \geq 0, \qquad D_\delta(p,q) \geq 0. \tag{3.13}$$

$$F_\delta(p,q) = 0 \iff D_\delta(p,q) = 0 \iff p = q. \tag{3.14}$$

In the above, (3.11)–(3.13) follows the definitions trivially. (3.14) follows from the fact that the integration of a measure vanishes if and only if the measure itself vanishes, and that $F_\delta(p,q)$ is the difference between the arithmetic and geometric averages of p and q which vanishes if and only if $p = q$ (Hardy et al., 1952; Halmos, 1950).

Proposition 3.2 (Special examples) *For finite measures $p, q \in \widetilde{\mathcal{P}}$,*

$$F_1(p,q) = q - p + p\log\frac{p}{q}. \qquad D_1(p,q) = \int \left(q - p + p\log\frac{p}{q}\right). \tag{3.15}$$

$$F_{1/2}(p,q) = 2(\sqrt{p} - \sqrt{q})^2. \qquad D_{1/2}(p,q) = 2\int (\sqrt{p} - \sqrt{q})^2. \tag{3.16}$$

For probability measures $p, q \in \mathcal{P}$:

$$D_\delta(p,q) = \frac{1}{\delta(1-\delta)}\left(1 - \int p^\delta q^{1-\delta}\right), \tag{3.17}$$

$$D_1(p,q) = D_0(q,p) = \int p\log\frac{p}{q}. \tag{3.18}$$

In the above, (3.15) is verified by l'Hopital rule. The rest are obvious.

The quantity $K(p,q) := D_1(p,q)$ is an extension of the Kullback-Leibler deviation (cross entropy) (Kullback and Leibler, 1951). The quantity $D_{1/2}(p,q)$ is the (square of) the Hellinger distance.

The δ-deviation also defines a topology on $\widetilde{\mathcal{P}}$. For all $\delta \in (0,1)$, they are equivalent to each other, and to that of $\mathcal{P}^\delta \subseteq L_{1/\delta}$. We shall take this topology as default unless stated otherwise, as it corresponds to most of the convergence concepts in asymptotic theories. Given the topology on $\widetilde{\mathcal{P}}$, the σ-algebra of Borel sets are well defined (Neveu, 1965; Kolmogorov, 1956; Halmos, 1950), so that $\widetilde{\mathcal{P}}$ itself is a measurable space, on which the prior and posterior are well-defined. On the other hand, it was shown by Csiszár (1967b) that the D_0- and D_1-topologies are much stronger and more difficult to deal with.

4 Information geometry

This section reviews some known results and essential concepts in information geometry, as will be needed in our main theory (§5). Many special cases of these concepts are well-known and form the backbone of theoretical statistics.

4.1 Metric and affine connection

The familiar concepts of consistency, efficiency (first and second order), sufficiency, exponential family, maximum likelihood, non-informative priors, and so on, can be expressed in terms of metric and affine connections (Amari, 1985). These geometrical concepts might be easier to understand for parametric models. Consider an m-dimensional model $\mathcal{Q} = \{p(\theta) : \theta = [\theta^1, \ldots, \theta^m] \in \mathbb{R}^m\}$. The metric is expressed in Fisher information matrix g_{ij}, while the δ-affine connection is expressed in the Riemann-Christoffel symbol $\overset{\delta}{\Gamma}_{ijk}$, defined by

$$g_{ij} = \int p \partial_i l \partial_j l, \qquad T_{ijk} = \int p \partial_i l \partial_j l \partial_k l, \tag{4.1}$$

$$\Gamma_{ijk} = \int p \partial_i \partial_j l \partial_k l, \qquad \overset{\delta}{\Gamma}_{ijk} = \Gamma_{ijk} + \delta T_{ijk}, \tag{4.2}$$

where $l := \log p$, and $\partial_i := \partial / \partial \theta^i$.

The metric was introduced in (Fisher, 1922, 1925; Rao, 1945; Cramér, 1946) and characterizes the distance between two infinitesimally nearby measures. Under regularity conditions guaranteeing asymptotic normality, the asymptotic squared distance between the MLE and the true distribution is m/n when measured in the metric, where m is the dimension of the model and n is the sample size. The efficiency of an estimate is simply the ratio between the metric of the whole sample and that of the estimate. See (Amari, 1985, Ch. 7) for details. See also (Kass, 1980, 1989; Murray and Rice, 1993) for a more intuitive exposition.

For small sample asymptotics, or for the second order efficiency (Rao, 1962) (also called information loss (Fisher, 1925)), the concept of δ-affine connections ($\delta \in [0, 1]$) is also needed. An affine connection defines an affine structure (the meaning of "parallel translation") on the tangent space. It therefore also implies the concepts of uniformness (length invariant along translation) and straightness (direction invariant along translation). The latter together with the metric also defines curvature.

For example, the 0-affine structure defines the log-likelihood to be flat and uniform, while the 1-affine structure defines the likelihood to be flat and uniform. The $1/2$-connection is the

metric connection. Under regularity conditions, the asymptotic variance of an estimate on a model is given as $A^2/2 + B^2 + C^2/2$, where A is the 1-curvature of the parameterization (parameter-effect curvature, naming curvature, Bhattacharyya curvature), B is the 0-curvature of the model (it is zero for exponential families), and C is the 1-curvature of the estimator (it is zero for MLE) (Efron, 1975; Dawid, 1975; Reed, 1975; Amari, 1982, 1985, 1987). This provides a quantitative expression of information loss, and explains the sufficiency of exponential families and MLEs. See also (Kass, 1987) for an introduction.

The δ- and $(1-\delta)$-connections are dually affine to each other, with respect to the metric. This uniquely defines the δ-deviation through dual-affine geometry (Amari, 1985, §3.5). Conversely, The metric and the δ-affine connections can be recovered from $D_\delta(p,q)$ through differentiation (Eguchi, 1983),

$$g_{ij} = -\underset{p}{\partial_i}\underset{q}{\partial_j} D_\delta(p,q)\big|_{q=p}, \qquad \overset{\delta}{\Gamma}_{ijk} = -\underset{p}{\partial_i}\underset{p}{\partial_j}\underset{q}{\partial_k} D_\delta(p,q)\big|_{q=p}, \tag{4.3}$$

where $\underset{p}{\partial_i} := \partial/\partial_i$ with θ being the parameter for p. Similar definition holds for $\underset{q}{\partial_i}$.

The following expansion can also be verified

$$D_\delta(p+dp,p) \approx \sum_{ij} \frac{1}{2} g_{ij} d\theta^i d\theta^j + \frac{1}{6}\sum_{ijk} \left(\Gamma_{ijk} + \Gamma_{jki} + \Gamma_{kij} + (\delta+1)T_{ijk}\right) d\theta^i d\theta^j d\theta^k. \tag{4.4}$$

It is important to note that the above are only examples in finite dimensional parametric models. The concepts of the metric and the δ-affine connection themselves do not require any regularity conditions. The metric is simply induced by the metric of the Hilbert space L_2. See also (Koshevnik and Levit, 1976). The δ-affine connection is simply given by the linear structure of the Banach spaces $L_{1/\delta}$. As defined in §3.1, these spaces do not depend on any dominating measure.

4.2 Other concepts relating to δ-affine structure

The following geometric concepts were studied by Amari (1985, Chap 2, 3) for parametric models. Here we give a simplified version, which is possible because we are only interested in $\widetilde{\mathcal{P}}$. The δ-

coordinate $l_\delta(p) \in L_{1/\delta}$ is defined as

$$l_\delta(p) := \frac{1}{\delta} p^\delta, \qquad \delta \in (0, 1], \qquad l_0(p) := \log p. \tag{4.5}$$

A curve is called a δ-geodesic if and only if its δ-coordinates form a straight line in the Banach space $L_{1/\delta}$. A manifold is called δ-flat if its δ-coordinates form a flat submanifold in $L_{1/\delta}$. Note that this concept concerns the external curvature instead of intrinsic curvature. A manifold $\mathcal{Q} \subseteq \widetilde{\mathcal{P}}$ is called δ-convex if all the δ-geodesics connecting two points on $\mathcal{Q}$ are contained in $\mathcal{Q}$. In other words, $\mathcal{Q}$ is δ-convex if its δ-coordinate is a convex set in $L_{1/\delta}$.

Two curves c_1 and c_2 are orthogonal to each other if their $1/2$-coordinates are orthogonal at the point of intersection. This is equivalent to the δ-coordinate of c_1 being orthogonal to the $(1-\delta)$-coordinate of c_2. A curve is orthogonal to a manifold if it is orthogonal to all the smooth curves within that manifold passing through the point of intersection. A point $q \in \mathcal{Q}$ is called a δ-projection of a point $p \in \widetilde{\mathcal{P}}$ onto $\mathcal{Q}$ if the δ-geodesic connecting p and q is orthogonal to $\mathcal{Q}$.

As mentioned in the introduction, studies on invariance priors (Jeffreyes, 1946; Hartigan, 1964) leads to the family of uniform prior which are uniform in the δ-affine structure. The corresponding differential equation in (Hartigan, 1965) contains perhaps the first appearance of the expression for $\overset{\delta}{\Gamma}_{ijk}$ (although not called as such). Furthermore, studies on uniform parameterization reveal five different concepts of uniformity (Hougaard, 1982), corresponding to uniformity in the δ-affine structure with $\delta \in \{0, 1/3, 1/2, 2/3, 1\}$ (Kass, 1984).

4.3 Pythagorean Theorems

Most of the nice aspects statistical inference of the Gaussian measures on linear spaces stem from the Euclidean (Hilbert) structure of the spaces. The measure spaces are not linear, but they still possess one extremely useful property which underlies our theory, namely the relation between minimization and orthogonal projection.

Theorem 4.1 (Generalized cosine) *Let $\delta \in [0,1]$, $p, r, q \in \widetilde{\mathcal{P}}$. Then*

$$D_\delta(p,r) + D_\delta(r,q) = D_\delta(p,q) + \int \left(l_\delta(p) - l_\delta(r)\right)\left(l_{1-\delta}(q) - l_{1-\delta}(r)\right). \tag{4.6}$$

Proof: For $\delta \in (0,1)$, this is a consequence of the following straight forward calculation

$$\begin{aligned}&\left(\delta p + (1-\delta)r - p^\delta r^{1-\delta}\right) + \left(\delta r + (1-\delta)q - r^\delta q^{1-\delta}\right) - \left(\delta p + (1-\delta)q - p^\delta q^{1-\delta}\right)\\ =&r - p^\delta r^{1-\delta} - r^\delta q^{1-\delta} + p^\delta q^{1-\delta} = \left(r^\delta - p^\delta\right)\left(r^{1-\delta} - q^{1-\delta}\right).\end{aligned}$$

For $\delta \in \{0,1\}$ the conclusion follows by taking limit of δ, or using (3.15). □

This theorem is a generalization of the classical "cosine theorem" for a Hilbert space

$$\|a-c\|^2 + \|b-c\|^2 = \|a-b\|^2 + 2(a-c)\cdot(b-c). \tag{4.7}$$

The generalization may be understood in two senses. On the one hand, taking $\delta = 1/2$, the space $\widetilde{\mathcal{P}}^{1/2}$ is a subspace of the Hilbert space L_2. On the other hand, for two Gaussian measures p_1, p_2 with the same covariance kernels V on a reproducing kernel Hilbert space with inner product H, the deviation $D_1(p_1,p_2) = D_0(p_1,p_2) = \|\mu_1 - \mu_2\|_H^2$. This reduces all the Hilbert space function approximation, smoothing, filtering and regression problems into the problem of Bayesian estimation of Gaussians. The regularizer or smoothing kernel are given by prior information kernel (matrix).

Corollary 4.2 (Generalized Pythagoras) *If the δ-geodesic connecting p and r is orthogonal to the $(1-\delta)$-geodesic connecting r and q, then*

$$D_\delta(p,r) + D_\delta(r,q) = D_\delta(p,q). \tag{4.8}$$

These results rely essentially on the geometric structures. Special cases of the Pythagorean theorem were proved by Čencov (1968) on $\mathcal{P}$ with $\delta \in \{0,1\}$, and by Amari (1985) on $\mathcal{P}$ with $\delta \in [0,1]$.

Theorem 4.3 (Projection) *Let $\mathcal{Q} \subseteq \widetilde{\mathcal{P}}$ be a submanifold, then a local minimum $q \in \mathcal{Q}$ of $D_\delta(p, q)$ is a δ-projection of p onto $\mathcal{Q}$. If $\mathcal{Q}$ is $(1-\delta)$-flat then the minimum is unique.*

The proof is similar to classical proof of unique projection to a linear subspace under quadratic distance. See, e.g., (Amari, 1985, Th. 3.8) or (Zhu and Rohwer, 1995c, Th. 3.2).

5 Optimal and Ideal Estimates

5.1 General framework

We now return to the problem posed at the beginning of this paper. Consider a sample space Z, the space $\widetilde{\mathcal{P}}$ of finite measures on Z, and the space $\mathcal{P}$ of probability measures on Z (Cf. Figure 1).

A computational model (or simply a model) is a subspace $\mathcal{Q} \subseteq \widetilde{\mathcal{P}}$. An estimator $\tau : Z \to \mathcal{Q}$ maps a sample $z \in Z$, taken from the true distribution $p \in \mathcal{P}$, to an estimate $q = \tau(z) \in \mathcal{Q}$. Conventionally, the model $\mathcal{Q}$ is often parameterized such that each measure $q \in \mathcal{Q}$ is represented by some parameter $\theta \in \Theta$, the parameter space. Usually it is also assumed that $p \in \mathcal{Q} \subseteq \mathcal{P}$. Such assumptions are not necessary here.

In the Bayesian framework, one assumes there is a prior $P(p)$ of the true distribution $p \in \mathcal{P}$ (or, more restrictively, $p \in \mathcal{Q}$). The posterior $P(p|z)$ is given by the Bayes theorem. In the following $\langle\cdot\rangle_z$ denotes the average over the posterior $P(p|z)$. The posterior δ-average of p is defined by $l_\delta(a_\delta(p)) = \langle l_\delta(p)\rangle_z$, or explicitly,

$$a_\delta(p) := \begin{cases} \left\langle p^\delta \right\rangle_z^{1/\delta}, & \delta \in (0, 1], \\ \exp \langle \log p \rangle_z, & \delta = 0. \end{cases} \tag{5.1}$$

It is also called Hölder's δ-mean weighted by $P(p|z)$ (Hardy et al., 1952; Aczél and Daróczy, 1975).

Using the δ-deviation $D_\delta(p, q)$ between distributions p and q acting as a loss function, the

Bayes risk of the estimator τ and estimate q are defined as

$$E_\delta(\tau) := \int_{p,z} P(p,z) D_\delta(p, \tau(z)), \qquad E_\delta(q|z) := \int_p P(p|z) D_\delta(p,q). \tag{5.2}$$

The δ-(optimal) estimate on $\mathcal{Q}$, $\tau_{\delta,\mathcal{Q}}(z)$, is defined by minimizing $E_\delta(q|z)$ over $q \in \mathcal{Q}$. If $\mathcal{Q} = \widetilde{\mathcal{P}}$ it is called the δ-ideal estimate and denoted $\tau_\delta(z)$. The δ-optimal estimator and δ-ideal estimator are defined correspondingly. The following well-known result, which follows from Bayes' theorem and Fubini's theorem, is often called the coherence of Bayesian methods: Suppose $E_\delta(\tau) < \infty$, then an estimator τ is optimal if and only if $\tau(z)$ is an optimal estimate for almost every sample $z \in Z$.

This framework is schematically illustrated in Figure 1. It also applies to sample size larger than one by taking Z and $\mathcal{P}$ to be the appropriate product spaces; When necessary a sample of size n will be denoted $z^n = [z_1, \ldots, z_n]$.

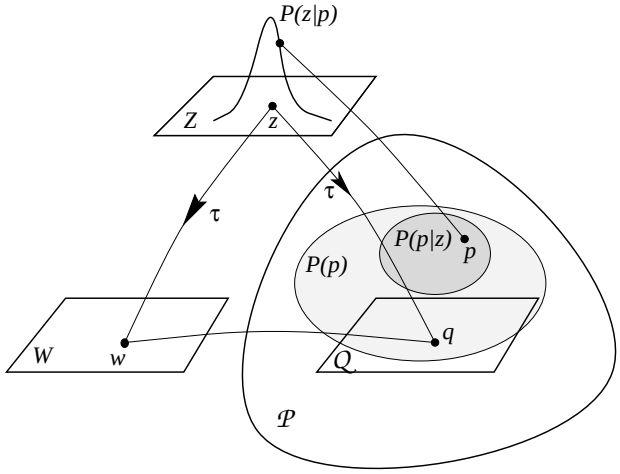

Figure 1: Spaces, distributions and mappings associated with estimation

5.2 Ideal estimates and error decomposition

The main theorem of this paper is the following "error decomposition theorem".

Theorem 5.1 *Let $\delta \in [0,1]$, $z \in Z$, $\widehat{p} := a_\delta(p)$ and $q \in \widetilde{\mathcal{P}}$. Then*

$$\langle D_\delta(p,q)\rangle_z = \langle D_\delta(p,\widehat{p})\rangle_z + D_\delta(\widehat{p},q), \tag{5.3}$$

$$\langle D_\delta(p,\widehat{p})\rangle_z = \begin{cases} \dfrac{1}{1-\delta}\displaystyle\int \left(\langle p\rangle_z - \widehat{p}\right), & \delta \in [0,1) \\ \displaystyle\int \left(\langle p\log p\rangle_z - \widehat{p}\log\widehat{p}\right), & \delta = 1. \end{cases} \tag{5.4}$$

$$\tau_\delta(z) = \widehat{p}. \tag{5.5}$$

Proof: The proof involves some straightforward calculations. First consider $\delta \in (0,1)$. From $\widehat{p}^\delta = \left\langle p^\delta \right\rangle_z$, it follows that

$$\begin{aligned} \delta(1-\delta)\left\langle D_\delta(p,\widehat{p})\right\rangle_z &= \int \left(\delta\left\langle p\right\rangle_z + (1-\delta)\widehat{p} - \left\langle p^\delta\right\rangle_z \widehat{p}^{1-\delta}\right) \\ &= \int \left(\delta\left\langle p\right\rangle_z + (1-\delta)\widehat{p} - \widehat{p}^\delta\widehat{p}^{1-\delta}\right) \\ &= \delta\int \left(\langle p\rangle_z - \widehat{p}\right). \end{aligned}$$

Therefore,

$$\begin{aligned} \delta(1-\delta)\left\langle D_\delta(p,q)\right\rangle_z &= \int \left(\delta\left\langle p\right\rangle_z + (1-\delta)q - \left\langle p^\delta\right\rangle_z q^{1-\delta}\right) \\ &= \int \left(\delta\left\langle p\right\rangle_z - \delta\widehat{p} + \delta\widehat{p} + (1-\delta)q - \widehat{p}^\delta q^{1-\delta}\right) \\ &= \delta(1-\delta)\left\langle D_\delta(p,\widehat{p})\right\rangle_z + \delta(1-\delta)D_\delta(\widehat{p},q). \end{aligned}$$

Now consider $\delta = 1$. From $\widehat{p} = \langle p\rangle_z$, we obtain

$$\begin{aligned} \langle D_1(p,\widehat{p})\rangle_z &= \int \left(\widehat{p} - \langle p\rangle_z + \langle p\log p\rangle_z - \langle p\rangle_z \log\widehat{p}\right) \\ &= \int \left(\langle p\log p\rangle_z - \langle p\rangle_z \log\widehat{p}\right) \\ &= \langle D_1(p,q)\rangle_z - D_1(\widehat{p},q). \end{aligned}$$

Similarly, for $\delta = 0$, we have $\log \widehat{p} = \langle \log p \rangle_z$, and

$$\begin{aligned}
\langle D_0(p, \widehat{p}) \rangle_z &= \int (\langle p \rangle_z - \widehat{p} + \widehat{p} \log \widehat{p} - \widehat{p} \langle \log p \rangle_z) \\
&= \int (\langle p \rangle_z - \widehat{p}) \\
&= \langle D_0(p, q) \rangle_z - D_0(\widehat{p}, q).
\end{aligned}$$

The proof is then completed by assembling these results together. □

In other words, the δ-ideal estimate uniquely exists, and is given by the posterior δ-mean. The error of any other estimate is larger and can be decomposed into "uncertainty" and "approximation error". This generalizes the well-known "mean squared error equals variance plus bias squared" formula for linear Gaussian models. Compare (5.1, 5.3–5.5) with

$$\bar{x} = \langle x \rangle , \tag{5.6}$$

$$\left\langle \|x - a\|^2 \right\rangle = \left\langle \|x - \bar{x}\|^2 \right\rangle + \|\bar{x} - a\|^2, \tag{5.7}$$

$$\left\langle \|x - \bar{x}\|^2 \right\rangle = \left\langle \|x\|^2 \right\rangle - \|\bar{x}\|^2, \tag{5.8}$$

$$\operatorname*{Min}_a \left\langle \|x - a\|^2 \right\rangle \implies a = \bar{x}. \tag{5.9}$$

See Figure 2 for a schematic illustration.

It can be shown that with natural conjugate priors, $\widehat{p}$ is a sufficient statistic for the commonly used statistical models, such as exponential families and uniform distributions. See (Aitchison, 1975) for Gamma models and (Zhu and Rohwer, 1995a) for Gaussian models. This is likely to be true under very general assumptions, even for statistical models without finite dimensional sufficient statistics, although in that case the ideal estimate would not be contained in any finite dimensional models even if the prior is supported on a finite dimensional model $\mathcal{Q}$. If this conjecture is true, it would be a realization of model-free reduction of data without losing information as envisioned by Fisher (1936).

For $\delta \neq 1$ the exact error decomposition (5.3) in a closed form is a new result without comparable classical counterparts, because it only holds in $\widetilde{\mathcal{P}}$ but not in $\mathcal{P}$, as $\mathcal{P}$ is not δ-convex for any

$\delta \in [0,1)$, while $\widetilde{\mathcal{P}}$ is δ-convex for any $\delta \in [0,1]$ (Amari, 1985, Ch 3). This was our motivation for extending the δ-deviation to finite measures. Considering the uniqueness of δ-geometry and the important role played by the Pythagorean theorems, it is unlikely that other formulation of statistics would support the concept of an ideal estimate (Hartigan, 1967, See also). Our proof is also new and simpler than previous ones for special cases. It is gratifying to see that such a fundamental theorem can be proved in a purely algebraic manner.

The error decomposition formula may be expanded to obtain asymptotic results, with the help of formulas like (4.4). The extensive literature on asymptotics is mostly concerned with second or third order expansions. Methods for asymptotic expansion up to arbitrary order in a curved exponential families and corresponding asymptotic theory for estimation and test were developed in (Amari, 1985, Ch. 4,5,6). See also (Amari, 1982, 1987). From third-order upwards the results depend on δ. In a broader sense, the method of least squares by Legendre, Laplace and Gauss and the χ^2 method by Pearson may be considered as the forerunners of second order approximation.

Corollary 5.2 *The* 1*-ideal estimate is the posterior marginal distribution.*

$$\widehat{p}(z'|z) = \int_p P(z', p, z), \tag{5.10}$$

where z' *is a sample independent of* z *conditional on* p*.*

This appears to be first recognized by Clarke and Barron (1990, §III.A).

5.3 Optimal estimates on a model

In practice the ideal estimate may be intractable to compute. It is often of interest to know the optimal estimate within a (usually finite dimensional) model $\mathcal{Q}$. In general there are both local and global optima on a model, so $\tau_{\delta,\mathcal{Q}}$ is usually a set, allowing local minima in the definition. As usual, assume $P(p)$ to be a prior over $\widetilde{\mathcal{P}}$, $z \in Z$, $\delta \in [0,1]$, and $\widehat{p} = \tau_\delta(z)$. The following propositions characterize the optimal estimates on various models. They follow trivially from

either the definitions or the error decomposition theorem.

Proposition 5.3 *Let* $\mathcal{Q} \subseteq \widetilde{\mathcal{P}}$. *Then* $\tau_{\delta,\mathcal{Q}}(z)$ *is obtained by* $\mathrm{Min}_{q\in\mathcal{Q}}\, D_\delta(\widehat{p}, q)$.

Its similarity with classical results on linear spaces is unmistakable (See Figure 2). This result and the sufficiency of the ideal estimate suggests that statistical inferences can be separated into three levels.

- Estimation: If the prior information is represented as a prior distribution, and the task is to summarize the most amount of information, then the ideal estimate $\widehat{p}$ is the unique and sufficient solution.

- Approximation: If, furthermore, the solution has to be represented by parameters in a model $\mathcal{Q}$, then it is given by the projection $\widehat{q}$ of the ideal estimate $\widehat{p}$ into $\mathcal{Q}$. Whether $\widehat{q}$ is unique depends on the model. The information loss is determined by the deviation of $\widehat{q}$ from $\widehat{p}$.

- Decision: For decision problems with an arbitrary cost (utility) function, the optimal decision would be a function of the ideal estimate $\widehat{p}$, since it is sufficient.

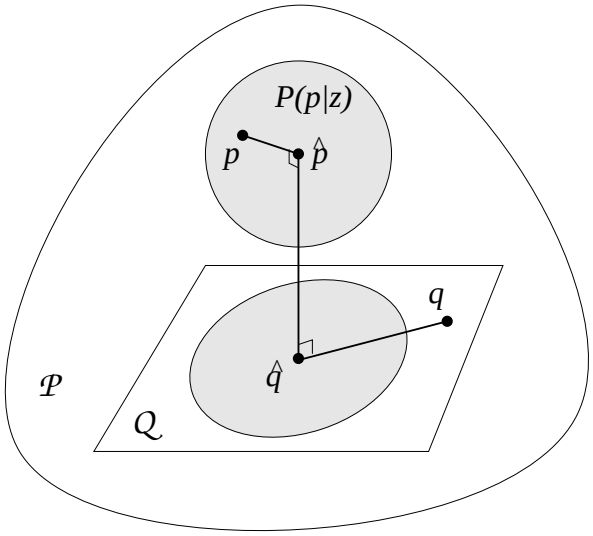

Figure 2: Decomposition of errors

Proposition 5.4 *Let $\mathcal{Q}_1 \subseteq \mathcal{Q}_2 \subseteq \widetilde{\mathcal{P}}$, $\widehat{q}_i = \tau_{\delta,\mathcal{Q}_i}(z)$. Then $E_\delta(\widehat{q}_1|z) \geq E_\delta(\widehat{q}_2|z)$.*

This is to say that a larger computational model $\mathcal{Q}$ is always better than a smaller one if computational cost is not taken into account. On the other hand, a prior with a larger support usually contains less information, and would result in less precise inference. Traditionally, the model used as the support of the prior is usually not distinguished from the model used to represent the estimates. These two conflicting requirements is often used to decide an "optimal model size", which can be quite misleading sometimes. There is no intrinsic statistical reason why this should be so, and we strongly propose the separation of statistical models (support of the prior) and the computational models. In fact, because the computational model $\mathcal{Q}$ is necessarily finite dimensional, the proper treatment of statistical problems in practice almost always requires the separation of prior from the model. This issue was discussed in detail in (Zhu and Rohwer, 1996) with many practical examples.

This issue can be viewed from another perspective. Traditionally a statistical problem is usually approached first by assuming a finite dimensional parametric model $\mathcal{Q}$, which may be recanted later: If the data is overwhelming to show the inadequacy of the model, alternative models are usually considered, and the issue of robustness arises. A more natural and general treatment would be to simply assumes that the prior is spilled over the whole space $\mathcal{P}$, while being more concentrated around the model $\mathcal{Q}$. The robustness of estimation on a model is then characterized by comparing $\langle E_\delta(\widehat{q}|z)\rangle = \langle \min_{q\in\mathcal{Q}} \langle D_\delta(p,q)\rangle_z \rangle$ with $\langle E_\delta(\widehat{p}|z)\rangle$.

Proposition 5.5 *Let $\mathcal{Q} \subseteq \widetilde{\mathcal{P}}$ be a $(1-\delta)$-convex subset of $\widetilde{\mathcal{P}}$. Then $\tau_{\delta,\mathcal{Q}}(z) = \{\widehat{q}\}$, with*

$$D_\delta(\widehat{p}, q) = D_\delta(\widehat{p}, \widehat{q}) + D_\delta(\widehat{q}, q), \qquad \forall q \in \mathcal{Q}. \tag{5.11}$$

The 0-projection onto a 1-flat manifold was considered by Csiszár (1975) while the 1-projection onto a 0-flat manifold was considered by Čencov (1982). Both proved uniqueness. It was shown by Amari (1985) that the δ-projection onto a $(1-\delta)$-convex manifold is unique, which motivates

our consideration of $\widetilde{\mathcal{P}}$ instead of $\mathcal{P}$, since $\widetilde{\mathcal{P}}$ is δ-flat for any δ (Amari, 1985, Ch. 3). Csiszár and Tusnady (1984) used both 0- and 1-projections to characterize the very useful EM algorithm. Projections according to various f-deviations are also studied in (Eguchi, 1983; Vajda, 1984).

In these previous studies the role of the "ideal estimate" was assumed by either the empirical distribution or the MLE. Since the true distribution is unknown, the choice of the "target" poses a logical dilemma in itself. It is usually chosen according to the asymptotic properties, and its appropriateness for small samples is often difficult to elucidate. The concept of ideal estimate removes this circular argument and supplies information geometry with a nonparametric object to approximate. It justifies using the empirical distribution or the MLE for approximation under D_1, and provides other more suitable targets to approximate under other D_δ, as the next example shows.

Proposition 5.6 *Let* $c \in \mathbb{R}_+$, $\mathcal{Q}_c := \left\{ q \in \widetilde{\mathcal{P}} : \int q = c \right\}$. *Then* $\tau_{\delta,\mathcal{Q}_c}(z) = \{\widehat{q}\}$, $\widehat{q} := c\widehat{p}/\int \widehat{p}$. *In particular,* $\tau_{\delta,\mathcal{P}}(z) = \left\{ \widehat{p}/\int \widehat{p} \right\}$.

This means it is trivial to translate many results obtained for $\widetilde{\mathcal{P}}$ into results for $\mathcal{P}$.

5.4 Finite sample space example and uniform priors

The multinomial model provides a good example to illustrate the properties of the ideal and optimal estimates without regularity assumptions. These were first given by Zhu and Rohwer (1995b). For other examples see also (Aitchison, 1975) for Gamma models, and (Zhu and Rohwer, 1995a) for Gaussian models.

Let $\boldsymbol{n} \in \mathbb{N}^m$, $\boldsymbol{a} \in \mathbb{R}_+^m$, and $\boldsymbol{p} \in \mathcal{P} := \left\{ \boldsymbol{p} \in \mathbb{R}_+^m : \sum \boldsymbol{p} = 1 \right\}$. Denote $n := \sum \boldsymbol{n}$, $a := \sum \boldsymbol{a}$. Consider the multinomial family of distributions

$$M(\boldsymbol{n}|\boldsymbol{p}) = C(\boldsymbol{n})\boldsymbol{p}^{\boldsymbol{n}} := \frac{n!}{\prod_i n_i!} \prod p_i^{n_i}. \tag{5.12}$$

with a Dirichlet distribution prior

$$D(\boldsymbol{p}|\boldsymbol{a}) = \frac{\boldsymbol{p}^{\boldsymbol{a}-1}}{B(\boldsymbol{a})} := \frac{\Gamma(a)}{\prod_i \Gamma(a_i)} \prod_i p_i^{a_i-1}. \tag{5.13}$$

The posterior is also a Dirichlet distribution $D(\boldsymbol{p}|\boldsymbol{b})$, with $\boldsymbol{b} = \boldsymbol{a} + \boldsymbol{n}$. Denote $b := \sum \boldsymbol{b}$.

The δ-ideal estimate $\widehat{\boldsymbol{p}} \in \widetilde{\mathcal{P}} = \mathbb{R}_+^m$ is given by

$$\widehat{p}_i^{\delta} = \frac{(b_i)_\delta}{(b)_\delta} = \frac{\Gamma(b_i+\delta)/\Gamma(b_i)}{\Gamma(b+\delta)/\Gamma(b)}, \tag{5.14}$$

where $(x)_y := \Gamma(x+y)/\Gamma(x)$. In particular,

$$\begin{cases} \widehat{p}_i = b_i/b, & \delta = 1, \\ \widehat{p}_i = \exp(\Psi(b_i) - \Psi(b)), & \delta = 0. \end{cases} \tag{5.15}$$

It is obvious that $\widehat{p}$ is a sufficient statistic. The δ-optimal estimate $\widehat{\boldsymbol{q}} = \widehat{\boldsymbol{p}}/\sum \widehat{\boldsymbol{p}} \in \mathcal{P}$ is given by

$$\widehat{q}_i = \frac{\widehat{p}_i}{\sum_i \widehat{p}_i} = \frac{(b_i)_\delta^{1/\delta}}{\sum_i (b_i)_\delta^{1/\delta}}. \tag{5.16}$$

For this finite sample space example, the δ-uniform prior turns out to be the Dirichlet prior $D(p|\delta\mathbf{1})$ (Zhu and Rohwer, 1995c), $\delta \in [0,1]$. For $\delta = 0$ it is uniform in the log-likelihood, i.e. in the natural parameters of exponential families. For $\delta = 1$ it is uniform in the likelihood, i.e. in the mixture parameters. For $\delta = 1/2$ it is Jeffreys prior, uniform in the Fisher information metric. With the 0-uniform prior, the 1-ideal estimate is the empirical distribution, and its 1-projection onto any model is the MLE on that model. The projection is unique if the model is 0-flat (exponential model).

One plausible generalization of δ-uniform priors to infinite sample space is the Dirichlet process prior (Ferguson, 1973), but some technical details need to be worked out. See also (Sibisi and Skilling, 1997) for other related priors on measure space.

There is generally no unique way to define δ-uniform priors on a model $\mathcal{Q}$ which is not δ-flat, unless $\delta = 1/2$. The "projection" of a δ-uniform prior from $\widetilde{\mathcal{P}}$ to $\mathcal{Q}$ depends implicitly on a metric (Amari, private communication), which can be arbitrary for $\delta \neq 1/2$. This appears to be related to

the “marginalisation paradoxes” (Dawid et al., 1973). See also (Stone and Dawid, 1972; Akaike, 1980). If we take the view point that non-Bayesian theories corresponds to limit of Bayesians with invariant priors, then there may be multiple non-Bayesian theories for the same statistical problem.

A word of caution here: The term “invariance” has two distinct meanings when used in statistical context. The meaning adopted here is based on Markov morphisms on the whole space of probability measures, which always leads to one of the δ-uniform priors if the uniqueness of the δ-affine structures is true. This explicitly excludes any structure in the sample space. This is in line with the usage in (Morse and Sacksteder, 1966; Čencov, 1982; Amari, 1985).

Another meaning of invariance is based on group structures on the sample space, corresponding to classical invariant priors (Box and Tiao, 1973; Berger, 1985; Bernardo and Smith, 1994). Note that uniformness according to one group generally does not consistently lead to uniformness according to all its subgroups, giving rise to the marginalization paradoxes (Dawid et al., 1973).

6 Generality and limitations

Estimation in the framework presented has two desirable properties:

Coherence Optimality of estimator is evaluated on the estimates it produces for all samples.

Covariance Optimality is independent of the naming of either the samples or the data generators.

The coherence is due to the Bayesian framework which evaluates estimators by the average performance of the estimates it produces, while the covariance is due to geometric language, which sees parameters as coordinates and deal with concepts which change properly under change of coordinates. It is therefore interesting to see what kind of limitations are imposed by this approach.

6.1 Generality of the framework

The Bayesian decision theory framework does not impose a serious restriction, because the complete class theorem of Wald states that any admissible decision is a (limit of) Bayes decision (Ferguson, 1967). Information geometry provides Bayesian methods with an invariant way to proceed with statistical inference once the posterior is obtained. Expressed in a coherent and covariant form, non-Bayesian methods can be considered as Bayesian methods with an (improper) prior which is invariant to re-parameterization. However, additional regularity conditions may be required for such limiting cases as Fubini's theorem is no longer applicable, so technically some non-Bayesian results may not be special instances of the general results.

Many statistical problems are not stated in covariant forms. For example, a well-known Bayesian method, "maximum posterior method", actually seeks the maximum of the posterior density function relative to an implicitly assumed dominating measure. This is not well-defined because changing the dominating measure can move the posterior maximum to any given point. It is therefore always advantageous to explicitly specify such assumptions. As Fisher observed, if the dominating measure is taken to be the prior, then the method reduces to the MLE which is both non-Bayesian and invariant. There is no loss of generality by requiring any special coordinates to be explicitly specified.

Furthermore, the nonparametric treatment frees information geometry from the practical confinement of finite dimensional models, parametric models, exponential family models, models dominated by one measure, and asymptotic expansions. Since the estimates live in an infinite dimensional space, this framework is more like a non-parametric theory in its flexibility and generality, while specializabe to parameterized problems with additional assumptions.

For estimations where prior assumptions can be summarized into the form of a prior distribution, this framework is more manageable than classical approaches, with its minimal requirement for regularity conditions. However, it would not be preferable in practice if for concrete problems

its conclusions were weaker than special theories developed for the particular problems. At least for one important class of practical problems, Gaussian measures on Hilbert spaces as mentioned in §4.3, the new framework is capable to deliver identical results as classical theories.

6.2 Uniqueness of the geometric concepts

On finite sample spaces the metric and the δ-connections are unique up to a constant factor, when constrained by the invariance of Markov morphisms (including change of variables both in the sample space and the parameter space) (Čencov, 1982, §11,12). An elementary proof of uniqueness of metric was given by Campbell (1985). This was also generally believed to be true for infinite sample spaces, possibly under mild regularity conditions (Amari, 1985, §3.8).

One the other hand, it is known that any f-deviation as defined by Csiszár (1967a) is invariant. Many interesting results for f-deviation in general and δ-deviation were obtained by Vajda (1989), some of them were also generalized to $\widetilde{\mathcal{P}}$. It is unknown if it is possible to obtain similar results with other f-deviations, but results relating to ideal estimates are highly unlikely. For first and second order efficiencies there is no need to consider other deviations because they all agree to some D_δ up to third order expansion (Amari, 1985, §3.8). On the other hand, the exact error decomposition relies heavily on the dual affine structure and the Pythagorean theorem, which is unlikely to exist under other formulations. Furthermore, any function $D(p, q)$ deserves to be called "*information* deviation" only if it measures the information contents of p relative to q and nothing else. This is so for any D_δ in the following technical sense: The δ-deviation cannot be increased by any Markov morphism; It remains unchanged if and only the morphism is sufficient. This is highly unlikely to hold for any other $D(p, q)$.

In addition it is interesting to note that Hartigan (1967) had proved that, in a parametric family under certain regularity conditions, the only invariant (both in the sample space and in the parameter space) statistical inference must be in the form of p^δ.

6.3 The role of the ideal estimate

In the sense of keeping sufficient statistics, ideal estimates play the same role as empirical distributions and are usually as awkward to work with, but since they live in the same space as other estimates with a rich geometrical structure, they provide new insight when analyzing the inevitable loss of information by ordinary estimates. They are easier starting points for asymptotic theories.

The ideas of an estimate extracting all the information in the sample and of comparing estimates by the amounts of information they extract date back to (Fisher, 1922, 1925). One exact result was known to Fisher: If sufficient estimates exist then the maximum likelihood estimate (1-straight estimate) is sufficient. It also became known that under some regularity conditions the finite dimensional models admitting sufficient estimates are exactly the exponential families (0-flat models). Other than these two, all the other results are asymptotic. LeCam (1953) discussed some difficulties involved. See Amari (1985) for asymptotic theory for parametric models. We can now see this as simply due to the interplay between $\widehat{p} \in \mathcal{P} \iff \delta = 1$ and the requirement $q \in \mathcal{P}$.

Statisticians are not accustomed to deal with unnormalized measures. Since the 1-ideal estimate does reside in $\mathcal{P}$, one might wonder whether it is really necessary to consider other δ. Following are several circumstances in which the full generality might be useful. First, a theory intended to include all "good" methods should not reject some methods which satisfy all the desiderata, especially if the reason for rejection is only a normalizing constant (Cf. Prop. 5.6). Secondly, consideration for all δ reveals the reason for the optimality of MLE: It is the optimal solution of the only method which works entirely within $\mathcal{P}$, as tradition demands. Thirdly, for $\delta \neq 1$, the ideal estimate holds one additional ancillary information, the sample size, which Fisher called the natural ancillary. Finally, many properties for $\delta \in \{0, 1\}$ are much more difficult to analyze, and taking limit from $\delta \in (0, 1)$ might be the easiest way around.

Acknowledgement

We are grateful and thank S. Amari for many enlightening discussions. This manuscript has existed in various forms for several years. We thank everyone who made comments, some anonymously, especially S. Eguchi, C. Williams, I. Nabney, D. Lowe, C. Bishop. This work was partially performed in NCRG, Aston, supported by EPSRC GR/J17814, partially at Santa Fe Institute, supported by TXN, Inc, and partially at Newton Institute, Cambridge.

References

Aczél, J. and Daróczy, Z. (1975). *On Measures of Information and Their Characteristics*. Academic Press, New York.

Aitchison, J. (1975). Goodness of prediction fit. *Biometrika*, 62(3):547–554.

Akaike, H. (1980). The interpretation of improper prior distributions as limits of data dependent proper prior distributions. *J. R. Statist. Soc., B*, 42(1):46–52.

Amari, S. (1982). Differential geometry of curved exponential families—curvature and information loss. *Ann. Statist.*, 10(2):357–385.

Amari, S. (1985). *Differential-Geometrical Methods in Statistics*, volume 28 of *Springer Lecture Notes in Statistics*. Springer-Verlag, New York.

Amari, S. (1987). Differential geometrical theory of statistics. In Amari et al. (1987), chapter 2, pages 19–94.

Amari, S., Barndorff-Nielsen, O. E., Kass, R. E., Lauritzen, S. L., and Rao, C. R., editors (1987). *Differential Geometry in Statistical Inference*, volume 10 of *IMS Lecture Notes Monograph*. Inst. Math. Statist., Hayward, CA.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian theory*. Wiley series in probability and mathematical statistics. Wiley, Chichester.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. J. Wiley, New York.

Campbell, L. L. (1985). An extended Čencov characterization of a Riemannian metric. Technical Report #1985-3, Dept. Math. Stat., Queen's Univ.

Čencov, N. N. (1968). Nonsymmetrical distance between probability distributions, entropy and the theorem of Pythagoras. *Mathematical Notes*, 4:686–691. (Matematicheskie Zametki, 4:3, 323-332).

Čencov, N. N. (1982). *Optimal Decision Rules and Optimal Inference*. Amer. Math. Soc., Rhode Island. Translation from Russian, 1972, Nauka, Moscow.

Chernoff, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Statist.*, 23:493–507.

Clarke, B. S. and Barron, A. R. (1990). Information-theoretic asymptotics of Bayes methods. *IEEE Trans. Info. Th.*, 36(3):453–471.

Cramér, H. (1946). A contribution to the theory of statistical estimation. *Skandinavisk Aktuarietidskrift*, 29:85–94.

Csiszár, I. (1967a). Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, 2:299–318.

Csiszár, I. (1967b). On topological properties of f-divergences. *Studia Sci. Math. Hungar.*, 2:329–339.

Csiszár, I. (1975). I-divergence geometry of probability distributions and minimization problems. *Ann. Prob.*, 3(1):146–158.

Csiszár, I. and Tusnady, G. (1984). Information geometry and alternating minimization procedures. *Statistics & Decisions, Supplement Issue No 1*, pages 205–237.

Dawid, A. P. (1975). Discussion of Efron's paper. *Ann. Statist.*, 3:1231–1234.

Dawid, A. P., Stone, M., and Zidek, J. V. (1973). Marginalization paradoxes in Bayesian and structural inference (with discussion). *J. R. Statist. Soc., B*, 35:189–233.

Efron, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency) (with discussion). *Ann. Statist.*, 3:1189–1242.

Eguchi, S. (1983). Second order efficiency of minimum contrast estimators in a curved exponential family. *Ann. Statist.*, 11:793–803.

Ferguson, T. S. (1967). *Mathematical Statistics : A Decision Theoretic Approach*. Academic Press, New York.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, 1(2):209–230.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Phil. Trans. R. Soc., A*, 222:309–368.

Fisher, R. A. (1925). Theory of statistical estimation. *Proc. Camb. Phi. Soc.*, 22:700–725.

Fisher, R. A. (1936). Uncertain inference. *Proc. Amer. Acad. Arts Sci.*, 71(4):245–258.

Good, I. J. (1965). *The Estimation of Probabilities : An Essay on Modern Bayesian Methods*. MIT Press Research Monograph, 30. MIT, Cambridge, MA.

Halmos, P. R. (1950). *Measure Theory*. Van Nostrand, New York.

Halmos, P. R. and Savage, L. J. (1949). Application of the Radon-Nikodym theorem to the theory of sufficient statistics. *Ann. Math. Statist.*, pages 225–241.

Hardy, G. H., Littlewood, J. E., and Polya, G. (1952). *Inequalities*. Cambridge Univ., 2 edition.

Hartigan, J. A. (1964). Invariant prior distributions. *Ann. Math. Statist.*, 35:836–845.

Hartigan, J. A. (1965). The asymptotically unbiased prior distribution. *Ann. Math. Statist.*, 36:1137–1152.

Hartigan, J. A. (1967). The likelihood and invariance principles. *J. R. Statist. Soc., B*, 29:533–539.

Hougaard, P. (1982). Parameterization of non-linear models. *J. R. Statist. Soc., B*, 44:244–252.

Jeffreyes, H. (1946). An invariant form for the prior probability in etimation problems. *Proc. R. Soc. Lond., A*, 196:453–461.

Jeffreys, H. (1961). *Theory of Probability*. Clarendon Press, Oxford. (First ed. 1939).

Kass, R. E. (1980). *The Riemannian structure of model spaces*. PhD thesis, U. Chicago.

Kass, R. E. (1984). Canonical parameterization and zero parameter effects curvature. *J. R. Statist. Soc., B*, 46:86–92.

Kass, R. E. (1987). Introduction. In Amari et al. (1987), chapter 1, pages 1–17.

Kass, R. E. (1989). The geometry of asymptotic inference (with discussion). *Statist. Sci.*, 4(3):188–234.

Kolmogorov, A. N. (1956). *Foundations of the Theory of Probability*. Chelsea., New York. Translation of *Grundbegriffe der Wahrscheinlichkeitsrechnung, 1933*.

Koshevnik, Y. A. and Levit, B. Y. (1976). On a non-parametric analogue of the information matrix. *Th. Prob. Appl.*, 21:738–753.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Statist.*, 22:79–86.

LeCam, L. (1953). On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates. *U. Calif. Public. Statist.*, 1:277–330.

LeCam, L. (1970). On the assumptions used to prove asymptotic normality of maximum likelihood estimates. *Ann. Math. Statist.*, 41(3):802–828.

Morse, N. and Sacksteder, R. (1966). Statistical isomorphism. *Ann. Math. Statist.*, 37:203–214.

Murray, M. K. and Rice, J. W. (1993). *Differential Geometry and Statistics*. Monographs on Statistics and Applied Probability. Chapman and Hall, London.

Neveu, J. (1965). *Mathematical Foundations of the Calculus of Probability*. Holden-Day, San Francisco. Translated from French, 1964, Masson.

Rao, C. R. (1945). Information and accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.*, 37:81–91.

Rao, C. R. (1962). Efficient estimates and optimum inference procedures in large samples (with discussion). *J. R. Statist. Soc., B*, 24:46–72.

Reed, J. (1975). Discussion on Professor Efron's paper. *Ann. Statist.*, 3(6):1234–1238.

Rényi, A. (1961). On measures of entropy and information. In *Proc. 4th Berkeley Symp. on Math. Statist. Prob.*, volume 1, pages 547–561. Univ. California.

Sibisi, S. and Skilling, J. (1997). Prior distributions on measure space. *J. R. Statist. Soc., B*, 59(1):217–235.

Stone, M. and Dawid, A. P. (1972). Un-Bayesian implications of improper Bayes inference in routine statistical problems. *Biometrika*, 59:369–373.

Vajda, I. (1984). Minimum divergence principle in statistical estimation. *Statistics & Decisions, Supplement Issue No 1*, pages 239–261.

Vajda, I. (1989). *Theory of Statistical Inference and Information*. Kluwer Academic, Dordrecht. Translation of: Teoria informacie a statistickeho rozhodovania.

Zhu, H. and Rohwer, R. (1995a). Bayesian invariant measurements of generalisation for continuous distributions. Technical Report NCRG/4352, Aston University. `ftp://cs.aston.ac.uk/neural/zhuh/continuous.ps.Z`.

Zhu, H. and Rohwer, R. (1995b). Bayesian invariant measurements of generalisation for discrete distributions. Technical Report NCRG/4351, Aston University. `ftp://cs.aston.ac.uk/neural/zhuh/discrete.ps.Z`.

Zhu, H. and Rohwer, R. (1995c). Information geometric measurements of generalisation. Technical Report NCRG/4350, Aston University. `ftp://cs.aston.ac.uk/neural/zhuh/generalisation.ps.Z`.

Zhu, H. and Rohwer, R. (1996). Bayesian regression filters and the issue of priors. *Neural Comp. Appl.*, 4(3):130–142.