SORT 30 (2) July-December 2006, 125-170

What does intrinsic mean in statistical estimation?*

Gloria García[1] and Josep M. Oller[2]

[1] *Pompeu Fabra University, Barcelona, Spain* [2] *University of Barcelona, Barcelona, Spain.*

Abstract

In this paper we review different meanings of the word *intrinsic* in statistical estimation, focusing our attention on the use of this word in the analysis of the properties of an estimator. We review the intrinsic versions of the bias and the mean square error and results analogous to the Cramér-Rao inequality and Rao-Blackwell theorem. Different results related to the Bernoulli and normal distributions are also considered.

MSC: 62F10, 62B10, 62A99.

Keywords: Intrinsic bias, mean square Rao distance, information metric.

1 Introduction

Statistical estimation is concerned with specifying, in a certain framework, a plausible probabilistic mechanism which explains observed data. The inherent nature of this problem is inductive, although the process of estimation itself is derived through mathematical deductive reasoning.

In parametric statistical estimation the probability is assumed to belong to a class indexed by some parameter. Thus the inductive inferences are usually in the form of point or region estimates of the probabilistic mechanism which has generated some specific data. As these estimates are provided through the estimation of the parameter, a label of the probability, different estimators may lead to different methods of induction.

*This research is partially sponsored by CGYCIT, PB96-1004-C02-01 and 1997SGR-00183 (Generalitat de Catalunya), Spain.

Address for correspondence: J. M. Oller, Departament d'Estadística, Universitat de Barcelona, Diagonal 645, 08028-Barcelona, Spain. **e-mail:** joller@ub.edu

Received: April 2006

Under this approach an estimator should not depend on the specified parametrization of the model: this property is known as the *functional invariance* of an estimator. At this point, the notion of intrinsic estimation is raised for the first time: an estimator is *intrinsic* if it satisfies this functional invariance property, and in this way is a real probability measure estimator. On the other hand, the bias and the mean square error (MSE) are the most commonly accepted measures of the performance of an estimator. Nevertheless these concepts are clearly dependent on the model parametrization and thus unbiasedness and uniformly minimum variance estimation are *non-intrinsic*.

It is also convenient to examine the goodness of an estimator through *intrinsic* conceptual tools: this is the object of the *intrinsic analysis of statistical estimation* introduced by Oller & Corcuera (1995) (see also Oller (1993b) and Oller (1993a)). These papers consider an intrinsic measure for the bias and the square error taking into account that a parametric statistical model with suitable regularity conditions has a natural Riemannian structure given by the information metric. In this setting, the square error loss is replaced by the square of the corresponding Riemannian distance, known as the *information distance* or the *Rao distance*, and the bias is redefined through a convenient vector field based on the geometrical properties of the model. It must be pointed out that there exist other possible intrinsic losses but the square of the Rao distance is the most natural intrinsic version of the square error.

In a recent paper of Bernardo & Juárez (2003), the author introduces the concept of intrinsic estimation by considering the estimator which minimizes the Bayesian risk, taking as a loss function a symmetrized version of Kullback-Leibler divergence (Bernardo & Rueda (2002)) and considering a reference prior based on an information-theoretic approach (Bernardo (1979) and Berger & Bernardo (1992)) which is independent of the model parametrization and in some cases coincides with the Jeffreys uniform prior distribution. In the latter case the prior, usually improper, is proportional to the Riemannian volume corresponding to the information metric (Jeffreys (1946)). This estimator is intrinsic as it does not depend on the parametrization of the model.

Moreover, observe that both the loss function and the reference prior are derived just from the model and this gives rise to another notion of intrinsic: an estimation procedure is said to be *intrinsic* if it is formalized only in terms of the model. Observe that in the framework of information geometry, a concept is *intrinsic* as far as it has a well-defined geometrical meaning.

In the present paper we review the basic results of the above-mentioned intrinsic analysis of the statistical estimation. We also examine, for some concrete examples, the intrinsic estimator obtained by minimizing the Bayesian risk using as an intrinsic loss the square of the Rao distance and as a reference prior the Jeffrey's uniform prior. In each case the corresponding estimator is compared with the one obtained by Bernardo & Juárez (2003).

2 The intrinsic analysis

As we pointed out before, the bias and mean square error are not intrinsic concepts. The aim of the *intrinsic analysis* of the *statistical estimation*, is to provide intrinsic tools for the analysis of intrinsic estimators, developing in this way a theory analogous to the classical one, based on some natural geometrical structures of the statistical models. In particular, intrinsic versions of the Cramér–Rao lower bound and the Rao–Blackwell theorem have been established.

We first introduce some notation. Let $(\mathcal{X}, \mathcal{A}, \mu)$ be a measure space and Θ be a connected open set of $\mathbb{R}^n$. Consider a map $f : \mathcal{X} \times \Theta \longrightarrow \mathbb{R}$ such that $f(x,\theta) \geq 0$ and $f(x,\theta)\mu(dx)$ defines a probability measure on $(\mathcal{X}, \mathcal{A})$ to be denoted as P_θ. In the present paper a *parametric statistical model* is defined as the triple $\{(\mathcal{X}, \mathcal{A}, \mu)\,;\, \Theta\,;\, f\}$. We will refer to μ as the *reference measure of the model* and to Θ as the *parameter space*.

In a general framework Θ can be any manifold modelled in a convenient space as $\mathbb{R}^n$, $\mathbb{C}^n$, or any Banach or Hilbert space. So even though the following results can be written with more generality, for the sake of simplicity we consider the above-mentioned form for the parameter space Θ. In that case, it is customary to use the same symbol (θ) to denote points and coordinates.

Assume that the parametric statistical model is identifiable, i.e. there exists a one-to-one map between parameters θ and probabilities P_θ; assume also that f satisfies the regularity conditions to guarantee that the Fisher information matrix exists and is a strictly positive definite matrix. In that case Θ has a natural Riemannian manifold structure induced by its information metric and the parametric statistical model is said to be *regular*. For further details, see Atkinson & Mitchel (1981), Burbea (1986), Burbea & Rao (1982) and Rao (1945), among others.

As we are assuming that the model is identifiable, an *estimator $\mathcal{U}$ of the true probability measure* based on a k-size random sample, $k \in \mathbb{N}$, may be defined as a measurable map from $\mathcal{X}^k$ to the manifold Θ, which induces a probability measure on Θ known as the *image measure* and denoted as ν_k. Observe that we are viewing Θ as a manifold, not as an open set of $\mathbb{R}^n$.

To define the bias in an intrinsic way, we need the notion of mean or expected value for a random object valued on the manifold Θ. One way to achieve this purpose is through an affine connection on the manifold. Note that Θ is equipped with Levi–Civita connection, corresponding to the Riemannian structure supplied by the information metric.

Next we review the exponential map definition. Fix θ in Θ and let $T_\theta\Theta$ be the tangent space at θ. Given $\xi \in T_\theta\Theta$, consider a geodesic curve $\gamma_\xi : [0,1] \to \Theta$, starting at θ and satisfying $\frac{d\gamma_\xi}{dt}\big|_{t=0} = \xi$. Such a curve exists as far as ξ belongs to an open star-shaped neighbourhood of $0 \in T_\theta\Theta$. In that case, the exponential map is defined as $\exp_\theta(\xi) = \gamma_\xi(1)$. Hereafter, we restrict our attention to the Riemannian case, denoting by $\|\cdot\|_\theta$ the

norm at $T_\theta\Theta$ and by ρ the Riemannian distance. We define

$$\Im_\theta = \{\xi \in T_\theta\Theta : \|\xi\|_\theta = 1\} \subset T_\theta\Theta$$

and for each $\xi \in \Im_\theta$ we define

$$c_\theta(\xi) = \sup\{t > 0 : \rho\,(\theta, \gamma_\xi(t)) = t\}\,.$$

If we set

$$\mathfrak{D}_\theta = \{t\xi \in T_\theta\Theta : 0 \le t < c_\theta(\xi)\ ;\ \xi \in \Im_\theta\} \quad \text{and} \quad D_\theta = \exp_\theta(\mathfrak{D}_\theta)\,,$$

it is well known that $\exp_\theta$ maps $\mathfrak{D}_\theta$ diffeomorphically onto D_θ. Moreover, if the manifold is complete the boundary of $\mathfrak{D}_\theta$ is mapped by the exponential map onto the boundary of D_θ, called the *cut locus of* θ *in* Θ. For further details see Chavel (1993).

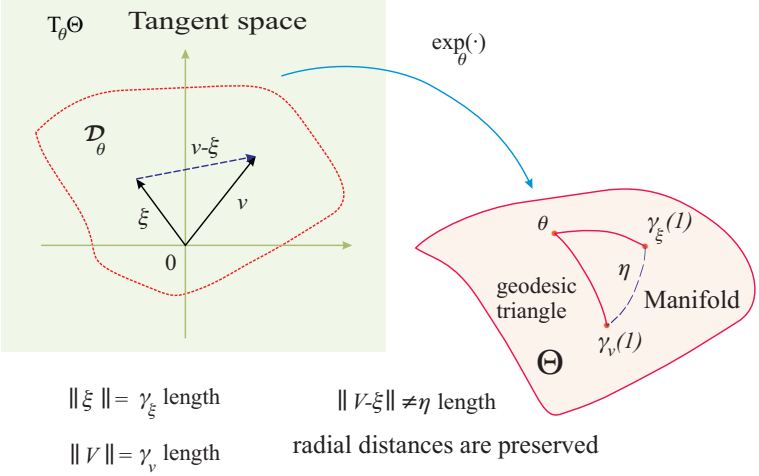

Figure 1: *The exponential map*

For the sake of simplicity, we shall assume that $\nu_k(\Theta \setminus D_\theta) = 0$, whatever true probability measure in the statistical model is considered. In this case, the inverse of the exponential map, $\exp_\theta^{-1}$, is defined ν_k–almost everywhere. For additional details see Chavel (1993), Hicks (1965) or Spivak (1979).

For a fixed sample size k, we define the *estimator vector field* A as

$$A_\theta(x) = \exp_\theta^{-1}(\mathcal{U}(x))\,, \quad \theta \in \Theta\,.$$

which is a C^∞ random vector field (first order contravariant tensor field) induced on the manifold through the inverse of the exponential map.

For a point $\theta \in \Theta$ we denote by E_θ the expectation computed with respect to the probability distribution corresponding to θ. We say that θ is a *mean value* of $\mathcal{U}$ if and

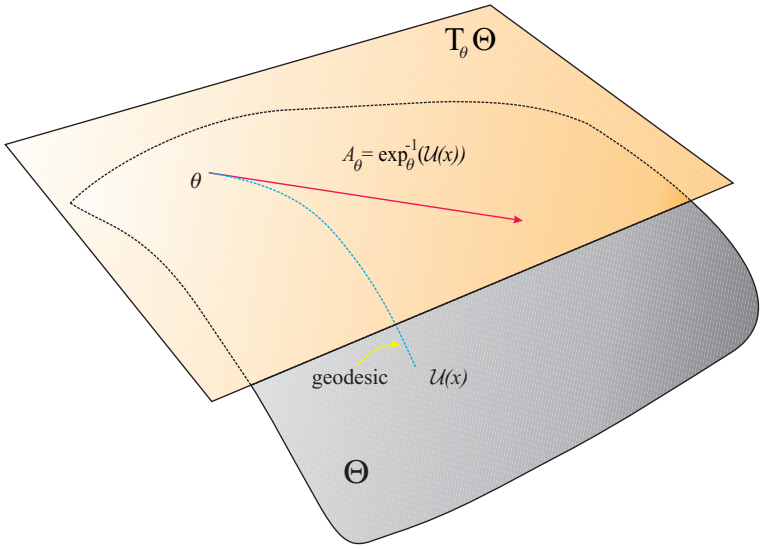

Figure 2: *Estimator vector field*

only if $E_\theta(A_\theta) = 0$. It must be pointed out that if a *Riemannian centre of mass* exists, it satisfies the above condition (see Karcher (1977) and Oller & Corcuera (1995)).

We say that an estimator $\mathcal{U}$ is *intrinsically unbiased* if and only if its mean value is the true parameter. A tensorial measure of the bias is the *bias vector field B*, defined as

$$B_\theta = E_\theta(A_\theta)\,, \quad \theta \in \Theta\,.$$

An *invariant bias measure* is given by the scalar field $\|B\|^2$ defined as

$$\|B_\theta\|^2_\theta\,, \quad \theta \in \Theta\,.$$

Notice that if $\|B\|^2 = 0$, the estimator is intrinsically unbiased.

The estimator vector field A also induces an intrinsic measure analogous to the mean square error. The *Riemannian risk of* $\mathcal{U}$, is the scalar field defined as

$$E_\theta\left(\|A_\theta\|^2_\theta\right) = E_\theta\left(\rho^2(\mathcal{U},\theta)\right)\,, \quad \theta \in \Theta\,.$$

since $\|A(x)\|^2_\theta = \rho^2(\mathcal{U}(x),\theta)$. Notice that in the Euclidean setting the Riemannian risk coincides with the mean square error using an appropriate coordinate system.

Finally note that if a mean value exists and is unique, it is natural to regard the expected value of the square of the Riemannian distance, also known as the *Rao distance*, between the estimated points and their mean value as an intrinsic version of the variance of the estimator.

To finish this section, it is convenient to note the importance of the selection of a loss function in a statistical problem. Let us consider the estimation of the probability of success $\theta \in (0,1)$ in a binary experiment where we perform independent trials until the first success. The corresponding density of the number of is given by

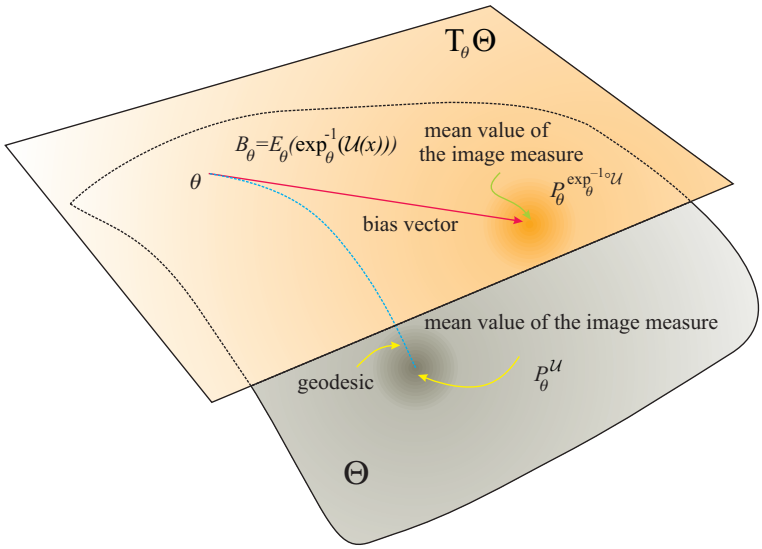

Figure 3: *Bias vector field*

$$f(k;\theta) = (1-\theta)^k\,\theta\,; \quad k = 0, 1, \ldots$$

If we restrict our attention to the class of unbiased estimators, a (classical) unbiased estimator U of θ, must satisfy

$$\sum_{k=0}^{\infty} U(k)\,(1-\theta)^k\,\theta = \theta, \quad \forall\theta \in (0,1),$$

where it follows that $\sum_{k=0}^{\infty} U(k)\,(1-\theta)^k$ is constant for all $\theta \in (0,1)$. So $U(0) = 1$ and $U(k) = 0$ for $k \geq 1$. In other words: when the first trial is a success, U assigns θ equal to 1; otherwise θ is taken to be 0.

Observe that, strictly speaking, there is no (classical) unbiased estimator for θ since U takes values in the boundary of the parameter space $(0,1)$. But we can still use the estimator U in a wider setting, extending both the sample space and the parameter space. We can then compare U with the maximum likelihood estimator, $V(k) = 1/(k+1)$ for $k \geq 0$, in terms of the mean square error. After some straightforward calculations, we obtain

$$\begin{aligned} E_\theta\left((U-\theta)^2\right) &= \theta - \theta^2 \\ E_\theta\left((V-\theta)^2\right) &= \theta^2 + \left(\theta\,Li_2(1-\theta) + 2\theta^2\ln(\theta)\right)/(1-\theta) \end{aligned}$$

where Li_2 is the dilogarithm function. Further details on this function can be found in Abramovitz (1970), page 1004. The next figure represents both mean square error of U and V.

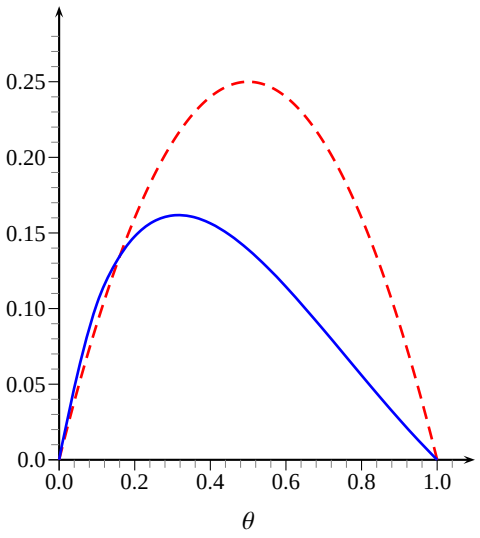

Figure 4: *MSE of U (dashed line) and V (solid line).*

It follows that there exist points in the parameter space for which the estimator U is preferable to V since U scores less risk; precisely for $\theta \in (0, 0.1606)$ where the upper extreme has been evaluated numerically. This admissibility contradicts the common sense that refuses U: this estimator assigns θ to be 0 even when the success occurs in a finite number of trials. This points out the fact that the MSE criterion is not enough to distinguish properly between estimators.

Instead of using the MSE we may compute the Riemannian risk for U and V. In the geometric model, the Rao distance ρ is given by

$$\rho(\theta_1, \theta_2) = 2 \left| \operatorname{arg\,tanh}\left(\sqrt{1-\theta_1}\right) - \operatorname{arg\,tanh}\left(\sqrt{1-\theta_2}\right) \right|, \ \theta_1, \theta_2 \in (0,1)$$

which tends to $+\infty$ when θ_1 or θ_2 tend to 0. So $E_\theta\left(\rho^2(U,\theta)\right) = +\infty$ meanwhile $E_\theta\left(\rho^2(V,\theta)\right) < +\infty$. The comparison in terms of Riemannian risk discards the estimator U in favour of the maximum likelihood estimator V, as is reasonable to expect.

Furthermore we can observe that the estimator U, which is classically unbiased, has infinite norm of the bias vector. So U is not even intrinsically unbiased, in contrast to V which has finite bias vector norm.

3 Intrinsic version of classical results

In this section we outline a relationship between the unbiasedness and the Riemannian risk obtaining an intrinsic version of the Cramér–Rao lower bound. These results are obtained through the comparison theorems of Riemannian geometry, see Chavel (1993)

and Oller & Corcuera (1995). Other authors have also worked in this direction, such as Hendricks (1991), where random objects on an arbitrary manifold are considered, obtaining a version for the Cramér–Rao inequality in the case of unbiased estimators. Recent developments on this subject can be found in Smith (2005).

Hereafter we consider the framework described in the previous section. Let $\mathcal{U}$ be an estimator corresponding to the regular model $\{(\chi, \mathcal{A}, \mu)\,;\ \Theta\,;\ f\}$, where the parameter space Θ is a n–dimensional real manifold and assume that for all $\theta \in \Theta$, $\nu_k(\Theta \setminus D_\theta) = 0$.

Theorem 3.1. *[Intrinsic Cramér–Rao lower bound] Let us assume that* $E\left(\rho^2(\mathcal{U}, \theta)\right)$ *exists and the covariant derivative of* $E(A)$ *exists and can be obtained by differentiating under the integral sign. Then,*

1. *We have*
$$E\left(\rho^2(\mathcal{U}, \theta)\right) \geq \frac{\left(\operatorname{div}(B) - E(\operatorname{div}(A))\right)^2}{kn} + \|B\|^2\ ,$$
where $\operatorname{div}(\cdot)$ *stands for the divergence operator.*
2. *If all the sectional Riemannian curvatures K are bounded from above by a non-positive constant* $\mathcal{K}$ *and* $\operatorname{div}(B) \geq -n$*, then*
$$E\left(\rho^2(\mathcal{U}, \theta)\right) \geq \frac{\left(\operatorname{div}(B) + 1 + (n-1)\sqrt{-\mathcal{K}}\,\|B\| \coth\left(\sqrt{-\mathcal{K}}\,\|B\|\right)\right)^2}{kn} + \|B\|^2\ .$$
3. *If all sectional Riemannian curvatures K are bounded from above by a positive constant* $\mathcal{K}$ *and* $d(\Theta) < \pi/2\sqrt{\mathcal{K}}$*, where* $d(\Theta)$ *is the diameter of the manifold, and* $\operatorname{div}(B) \geq -1$*, then*
$$E\left(\rho^2(\mathcal{U}, \theta)\right) \geq \frac{\left(\operatorname{div}(B) + 1 + (n-1)\sqrt{\mathcal{K}}\,d(\Theta)\cot\left(\sqrt{\mathcal{K}}\,d(\Theta)\right)\right)^2}{kn} + \|B\|^2\ .$$

In particular, for intrinsically unbiased estimators, we have:

4. *If all sectional Riemannian curvatures are non-positive, then*
$$E\left(\rho^2(\mathcal{U}, \theta)\right) \geq \frac{n}{k}$$
5. *If all sectional curvatures are less or equal than a positive constant* $\mathcal{K}$ *and* $d(\Theta) < \pi/2\sqrt{\mathcal{K}}$*, then*
$$E\left(\rho^2(\mathcal{U}, \theta)\right) \geq \frac{1}{kn}$$

The last result shows up the effect of the Riemannian sectional curvature on the precision which can be attained by an estimator.

Observe also that any one–dimensional manifold corresponding to one–parameter family of probability distributions is always Euclidean and $\operatorname{div}(B) = -1$; thus part 2 of

Theorem (3.1) applies. There are also some well known families of probability distributions which satisfy the assumptions of this last theorem, such as the multinomial, see Atkinson & Mitchel (1981), the negative multinomial distribution, see Oller & Cuadras (1985), or the extreme value distributions, see Oller (1987), among many others.

It is easy to check that in the n-variate normal case with known covariance matrix Σ, where the Rao distance is the Mahalanobis distance, the sample mean based on a sample of size k is an estimator that attains the intrinsic Cramér–Rao lower bound, since

$$\begin{aligned} E\left(\rho^2(\overline{X},\mu)\right) &= E\left((\overline{X}-\mu)^{\mathsf{T}}\Sigma^{-1}(\overline{X}-\mu)\right) = \\ &= E\left(\mathrm{tr}(\Sigma^{-1}(\overline{X}-\mu)(\overline{X}-\mu)^{\mathsf{T}})\right) = \\ &= \mathrm{tr}\left(\Sigma^{-1}E\left((\overline{X}-\mu)(\overline{X}-\mu)^{\mathsf{T}}\right)\right) = \mathrm{tr}(\frac{1}{k}\,I) = \frac{n}{k} \end{aligned}$$

where v^{T} is the transpose of a vector v.

Next we consider a tensorial version of the Cramér-Rao inequality. First we define the *dispersion tensor* corresponding to an estimator $\mathcal{U}$ as:

$$S_\theta = E_\theta(A_\theta \otimes A_\theta) \qquad \forall\theta \in \Theta$$

Theorem 3.2. *The dispersion tensor S satisfies*

$$S \geq \frac{1}{k}\mathrm{Tr}^{\,2,4}[\mathrm{G}^{\,2,2}[(\nabla B - E(\nabla A)) \otimes (\nabla B - E(\nabla A))]] + B \otimes B$$

where $\mathrm{Tr}^{\,i,j}$ *and* $\mathrm{G}^{\,i,j}$ *are, respectively, the contraction and raising operators on index i, j and ∇ is the covariant derivative. Here the inequality denotes that the difference between the right and the left hand side is non-negative definite.*

Now we study how we can decrease the mean square Rao distance of a given estimator. Classically this is achieved by taking the conditional mean value with respect to a sufficient statistic; we shall follow a similar procedure here. But now our random objects are valued on a manifold: we need to define the conditional mean value concept in this case and then obtain an intrinsic version of the Rao–Blackwell theorem.

Let $(\mathcal{X}, \mathcal{A}, P)$ be a probability space. Let M be a n–dimensional, complete and connected Riemannian manifold. Then M is a complete separable metric space (a Polish space) and we will have a regular version of the conditional probability of any M–valued random object f with respect to any σ-algebra $\mathcal{D} \subset \mathcal{A}$ on $\mathcal{X}$. In the case where the mean square of the Riemannian distance ρ of f exists, we can define

$$E(\rho^2(f,m)|\mathcal{D})(x) = \int_M \rho^2(t,m) P_{f|\mathcal{D}}(x,dt)\,,$$

where $x \in \mathcal{X}$, B is a Borelian set in M and $P_{f|\mathcal{D}}(x,B)$ is a regular conditional probability of f given $\mathcal{D}$.

If for each $x \in \mathcal{X}$ there exists a unique mean value $p \in M$ corresponding to the conditional probability $P_{f|\mathcal{D}}(x, B)$, i.e. a point $p \in M$ such that

$$\int_M \exp_p^{-1}(t) P_{f|\mathcal{D}}(x, dt) = 0_p,$$

we have a map from $\mathcal{X}$ to M that assigns, to each x, the mean value corresponding to $P_{f|\mathcal{D}}(x, B)$.

Therefore, if f is a random object on M and $\mathcal{D} \subset \mathcal{A}$ a σ–algebra on $\mathcal{X}$, we can define the conditional mean value of f with respect $\mathcal{D}$, denoted by $\mathfrak{M}(f|\mathcal{D})$, as a $\mathcal{D}$-measurable map, Z, such that

$$E(\exp_Z^{-1}(f(\cdot))|\mathcal{D}) = 0_Z$$

provided it exists. A sufficient condition to assure that the mean value exists and is uniquely defined, is the existence of an open geodesically convex subset $N \subset M$ such that $P\{f \in N\} = 1$. Finally, it is necessary to mention that $\mathfrak{M}(\mathfrak{M}(f|\mathcal{D})) \neq \mathfrak{M}(f)$, see for instance Kendall (1990).

Let us apply these notions to statistical point estimation. Given the regular parametric statistical model $\{(\mathcal{X}, \mathcal{A}, \mu); \Theta; f\}$, we assume that Θ is complete or that there exist a metric space isometry with a subset of a complete and connected Riemannian manifold. We recall now that a real valued function h on a manifold, equipped with an affine connection, is said to be *convex* if for any geodesic γ, $h \circ \gamma$ is a convex function. Then we have the following result.

Theorem 3.3. (Intrinsic Rao–Blackwell) *Let $\mathcal{D}$ be a sufficient σ–algebra for the statistical model. Consider an estimator $\mathcal{U}$ such that $\mathfrak{M}(\mathcal{U}|\mathcal{D})$ is well defined.*

If θ is such that $\rho^2(\theta, \cdot)$ is convex then

$$E_\theta\left(\rho^2(\mathfrak{M}(\mathcal{U}|\mathcal{D}), \theta)\right) \leq E_\theta(\rho^2(\mathcal{U}, \theta)) \ .$$

The proof is based on Kendall (1990). Sufficient conditions for the hypothesis of the previous theorem are given in the following result

Theorem 3.4. *If the sectional curvatures of N are at most 0, or $\mathcal{K} > 0$ with $d(N) < \pi/2\sqrt{\mathcal{K}}$, where $d(N)$ is the diameter of N, then $\rho^2(\theta, \cdot)$ is convex $\forall \theta \in \Theta$.*

It is not necessarily true that the mean of the square of the Riemannian distance between the true and estimated densities decreases when conditioning on $\mathcal{D}$. For instance, if some of the curvatures are positive and we do not have further information about the diameter of the manifold, we cannot be sure about the convexity of the square of the Riemannian distance.

On the other hand, the efficiency of the estimators can be improved by conditioning with respect to a sufficient σ-algebra $\mathcal{D}$ obtaining $\mathfrak{M}(\mathcal{U}|\mathcal{D})$. But in general the bias is

not preserved, in contrast to the classical Rao-Blackwell theorem; in other words, even if $\mathcal{U}$ were intrinsically unbiased, $\mathfrak{M}(\mathcal{U}|\mathcal{D})$ would not be in general intrinsically unbiased since,

$$\mathfrak{M}(\mathfrak{M}(\mathcal{U}|\mathcal{D})) \neq \mathfrak{M}(\mathcal{U}).$$

However the norm of the bias tensor of $\mathfrak{M}(\mathcal{U}|\mathcal{D})$ is bounded: if we let $B^{\mathfrak{M}(\mathcal{U}|\mathcal{D})}$ be the bias tensor, by the Jensen inequality,

$$||B_\theta^{\mathfrak{M}(\mathcal{U}|\mathcal{D})}||_\theta^2 \leq E_\theta\left(\rho^2(\mathfrak{M}(\mathcal{U}|\mathcal{D}), \theta)\right) \leq E_\theta(\rho^2(\mathcal{U}, \theta)).$$

4 Examples

This section is devoted to examine the goodness of some estimators for several models. Different principles apply in order to select a convenient estimator; here we consider the estimator that minimizes the Riemannian risk for a prior distribution proportional to the Riemannian volume. This approach is related to the ideas developed by Bernardo & Juárez (2003), where the authors consider as a loss function a symmetrized version of the Kullback-Leibler divergence instead of the square of the Rao distance and use a reference prior which, in some cases, coincides with the Riemannian volume. Once that estimator is obtained, we examine its intrinsic performance: we compute the corresponding Riemannian risk and its bias vector, precisely the square norm of the intrinsic bias. We also compare this estimator with the maximum likelihood estimator.

4.1 Bernoulli

Let $X_1, \ldots, X_k$ be a random sample of size k from a Bernoulli distribution with parameter θ, that is with probability density $f(x\,;\theta) = \theta^x(1-\theta)^{1-x}$, for $x \in \{0, 1\}$. In that case, the parameter space is $\Theta = (0, 1)$ and the metric tensor is given by

$$g(\theta) = \frac{1}{\theta(1-\theta)}$$

We assume the prior distribution π for θ be the Jeffreys prior, that is

$$\pi(\theta) \propto \frac{1}{\sqrt{\theta(1-\theta)}}$$

The corresponding joint density of θ and $(X_1, \ldots, X_k)$ is then proportional to

$$\frac{1}{\sqrt{\theta(1-\theta)}}\ \theta^{\sum_{i=1}^k X_i}\ (1-\theta)^{k-\sum_{i=1}^k X_i} = \theta^{\sum_{i=1}^k X_i - \frac{1}{2}}\ (1-\theta)^{k-\sum_{i=1}^k X_i - \frac{1}{2}}$$

which depends on the sample through the sufficient statistic $T = \sum_{i=1}^k X_i$. When $(X_1, \ldots, X_k) = (x_1, \ldots, x_k)$ put $T = t$. since,

$$\int_0^1 \theta^{\,t-\frac{1}{2}}(1-\theta)^{\,k-t-\frac{1}{2}}\,d\theta = \mathrm{Beta}\Big(t+\frac{1}{2},\ k-t+\frac{1}{2}\Big)$$

the posterior distribution $\pi(\cdot \mid t)$ based on the Jeffreys prior is as follows

$$\pi(\theta \mid t) = \frac{1}{\mathrm{Beta}\left(t+\frac{1}{2},\ k-t+\frac{1}{2}\right)}\,\theta^{t-\frac{1}{2}}\ (1-\theta)^{k-t-\frac{1}{2}}$$

where Beta is the Euler beta function.

The Bayes estimator related to the loss function given by the square of the Rao distance ρ^2 is

$$\theta^b(s) = \arg\min_{\theta^e\in(0,1)} \int_0^1 \rho^2(\theta^e,\theta)\ \ \pi(\theta \mid t)\,d\theta$$

Since an intrinsic estimation procedure is invariant under reparametrization, we perform the change of coordinates defined through the equation

$$1 = \left(\frac{d\theta}{d\xi}\right)^2 \frac{1}{\xi(1-\xi)}$$

in order to obtain a metric tensor equal to 1: the Riemannian distance expressed via this coordinate system, known as *Cartesian coordinate system*, will coincide with the Euclidean distance between the new coordinates. If we solve this differential equation, with the initial conditions equal to $\xi(0) = 0$, we obtain $\xi = 2\arcsin(\sqrt{\theta})$ and $\xi = -2\arcsin(\sqrt{\theta})$; we only consider the first of these two solutions. After some straightforward computations we obtain

$$\rho\,(\theta_1,\theta_2) = 2\arccos\Big(\sqrt{\theta_1\ \theta_1} + \sqrt{(1-\theta_1)\ (1-\theta_2)}\ \Big) = |\xi_1-\xi_2| \tag{1}$$

for $\xi_1 = 2\arcsin(\sqrt{\theta_1})$ and $\xi_2 = 2\arcsin(\sqrt{\theta_2})$ and $\theta_1,\theta_2 \in \Theta$.

In the Cartesian setting, the Bayes estimator $\xi^b(s)$ is equal to the expected value of ξ with respect to the posterior distribution

$$\pi(\xi \mid t) = \frac{1}{\mathrm{Beta}\left(t+\frac{1}{2},\ k-t+\frac{1}{2}\right)}\left(\sin^2\left(\frac{\xi}{2}\right)\right)^t\ \left(1-\sin^2\left(\frac{\xi}{2}\right)\right)^{k-t}$$

Once we apply the change of coordinates $\theta = \sin^2\left(\frac{\xi}{2}\right)$, the estimator $\xi^b(s)$ is

$$\xi^b(t) = \frac{1}{\mathrm{Beta}\left(t+\frac{1}{2},\ k-t+\frac{1}{2}\right)}\int_0^1 2\ \arcsin\Big(\sqrt{\theta}\ \Big)\ \theta^{\,t-\frac{1}{2}}\,(1-\theta)^{\,k-t-\frac{1}{2}}\,d\theta$$

Expanding $\arcsin(\sqrt{\theta}\,)$ in power series of θ,

$$\arcsin(\sqrt{\theta}) = \frac{1}{\sqrt{\pi}} \sum_{j=0}^{\infty} \frac{\Gamma\left(j+\frac{1}{2}\right)}{j!\,(2j+1)}\,\theta^{\,j+\frac{1}{2}}$$

where Γ is the Euler gamma function. After some computations, we obtain

$$\xi^b(t) = 2\,\frac{\Gamma(k+1)\;\Gamma(t+1)}{\Gamma\left(k+\frac{3}{2}\right)\;\Gamma\left(t+\frac{1}{2}\right)}\;{}_3F_2\left(\frac{1}{2},\frac{1}{2},t+1;k+\frac{3}{2},\frac{3}{2};1\right) \tag{2}$$

where ${}_3F_2$ denotes a generalized hypergeometric function. Further details on the gamma, beta and hypergeometric functions can be found on Erdélyi et al. (1955). Finally the Bayes estimator $\theta^b(t)$ of θ is given by

$$\theta^b(t) = \sin^2\left(\frac{\Gamma(k+1)\;\Gamma(t+1)}{\Gamma\left(k+\frac{3}{2}\right)\;\Gamma\left(t+\frac{1}{2}\right)}\;{}_3F_2\left(\frac{1}{2},\frac{1}{2},t+1;k+\frac{3}{2},\frac{3}{2};1\right)\right)$$

It is straightforward to prove that

$$\theta^b(k-t) = 1-\theta^b(t)$$

and can be approximated by

$$\theta^a(t) = \frac{t}{k} + \left(\frac{1}{2}-\frac{t}{k}\right)\left(\frac{0.63}{k}-\frac{0.23}{k^2}\right)$$

with relative errors less that 3.5% for any result based on sample size $k \le 100$.

The behaviour of these estimators, for different values of k and for small t, is shown in the following table.

	$\theta^b(0)$	$\theta^b(1)$	$\theta^b(2)$	$\theta^a(0)$	$\theta^a(1)$	$\theta^a(2)$
$k=1$	0.20276	0.79724	-	0.20000	0.80000	-
$k=2$	0.12475	0.50000	0.87525	0.12875	0.50000	0.87125
$k=5$	0.05750	0.23055	0.40995	0.05840	0.23504	0.41168
$k=10$	0.03023	0.12109	0.21532	0.03035	0.12428	0.21821
$k=20$	0.01551	0.06207	0.11037	0.01546	0.06392	0.11237
$k=30$	0.01043	0.04173	0.07420	0.01037	0.04301	0.07566
$k=50$	0.00630	0.02521	0.04482	0.00625	0.02600	0.04575
$k=100$	0.00317	0.01267	0.02252	0.00314	0.01308	0.02301

Observe that these estimators do not estimate θ as zero when $t=0$, similarly to the estimator obtained by Bernardo & Juárez (2003), which is particularly useful when we are dealing with rare events and small sample sizes.

The Riemannian risk of this intrinsic estimator has been evaluated numerically and is represented in Figure . Note that the results are given in terms of the Cartesian coordinates ξ^b, in order to guarantee that the physical distance in the plots is proportional to the Rao distance. The Riemannian risk of θ^b is given by

$$E_\theta\left(\rho^2(\theta^b,\theta)\right) = E_\xi\left((\xi^b-\xi)^2\right) = \sum_{t=0}^{k}(\xi^b(t)-\xi)^2\binom{k}{t}\sin^{2t}\left(\frac{\xi}{2}\right)\cos^{2(k-t)}\left(\frac{\xi}{2}\right)$$

which can be numerically computed through expression (2). This can be compared with the numerical evaluation of the Riemannian risk of the maximum likelihood estimator $\theta^* = t/k$, given by

$$\begin{aligned} E_\theta\left(\rho^2(\theta^*,\theta)\right) &= E_\xi\left((\xi^*-\xi)^2\right) \\ &= \sum_{t=0}^{k}\left(2\ \arcsin\left(\sqrt{\frac{t}{k}}\right)-\xi\right)^2\binom{k}{t}\sin^{2t}\left(\frac{\xi}{2}\right)\cos^{2(k-t)}\left(\frac{\xi}{2}\right) \end{aligned}$$

as we can see in Figure 5.
We point out that the computation of the Riemannian risk for the maximum likelihood estimator requires the extension by continuity of the Rao distance given in (1) to the closure of the parameter space Θ as θ^* takes values on $[0, 1]$.

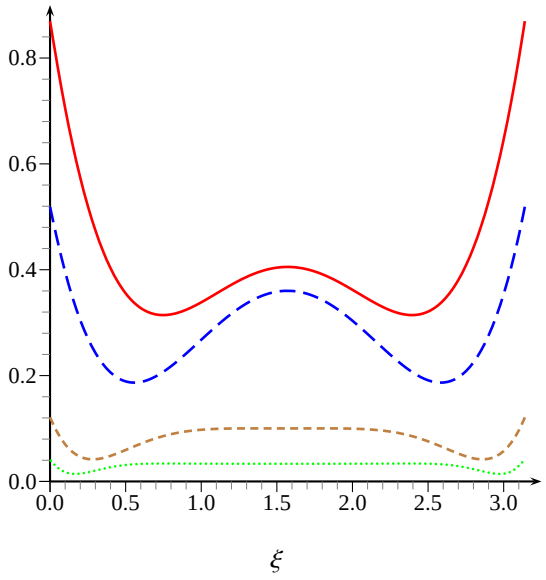

Figure 5: *Riemannian risk of ξ^b, for $k = 1$ (solid line), $k = 2$ (long dashed line), $k = 10$ (short dashed line) and $k = 30$ (dotted line).*

For a fixed sample size, observe that the Riemannian risk corresponding to ξ^b is lower than the Riemannian risk corresponding to ξ^* in a considerable portion of the parameter space, as it is clearly shown in Figure .

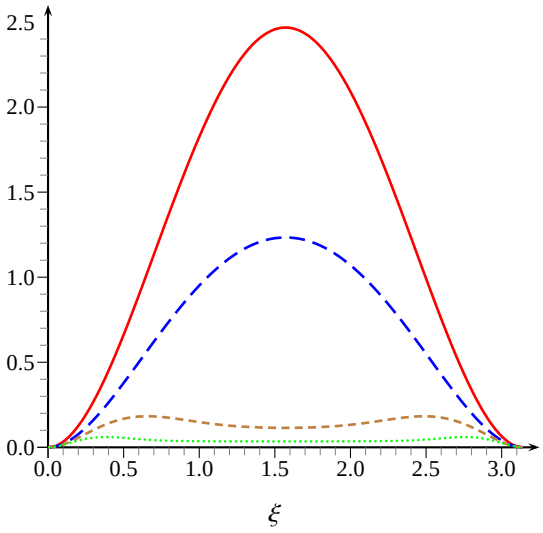

Figure 6: *Riemannian risk of ξ^*, for $k = 1$ (solid line), $k = 2$ (long dashed line), $k = 10$ (short dashed line) and $k = 30$ (dotted line).*

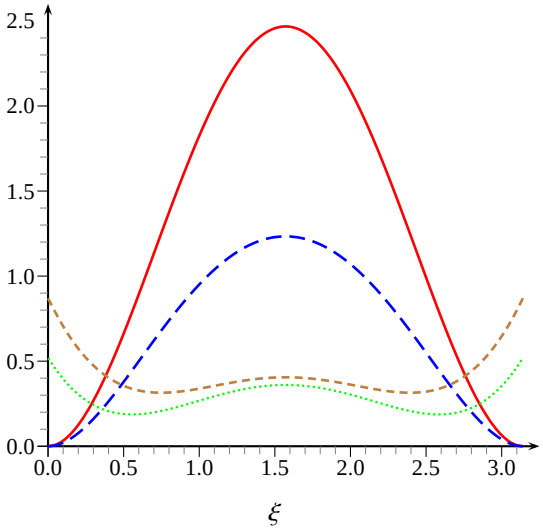

Figure 7: *Riemannian risk corresponding to θ^*, for $k = 1$ (solid line), $k = 2$ (long dashed line) and corresponding to θ^b, for $k = 1$ (short dashed line) and $k = 2$ (dotted line).*

Note that part of the Riemannian risk comes up through the bias of an estimator. Next the square of the norm of the bias vector B^b for θ^b and B^* for θ^* is evaluated numerically. Formally, in the Cartesian coordinate system ξ

$$B^b_\xi = E_\xi\left(\xi^b\right) - \xi = \sum_{t=0}^{k}(\xi^b(t) - \xi)\binom{k}{t}\sin^{2t}\left(\frac{\xi}{2}\right)\cos^{2(k-t)}\left(\frac{\xi}{2}\right)$$

$$B^*_\xi = E_\xi\left(\xi^*\right) - \xi = \sum_{t=0}^{k}\left(2\ \arcsin\left(\sqrt{\frac{t}{n}}\right) - \xi\right)\binom{k}{t}\sin^{2t}\left(\frac{\xi}{2}\right)\cos^{2(k-t)}\left(\frac{\xi}{2}\right)$$

The squared norm of the bias vector B^b and of B^* are represented in Figures and respectively.

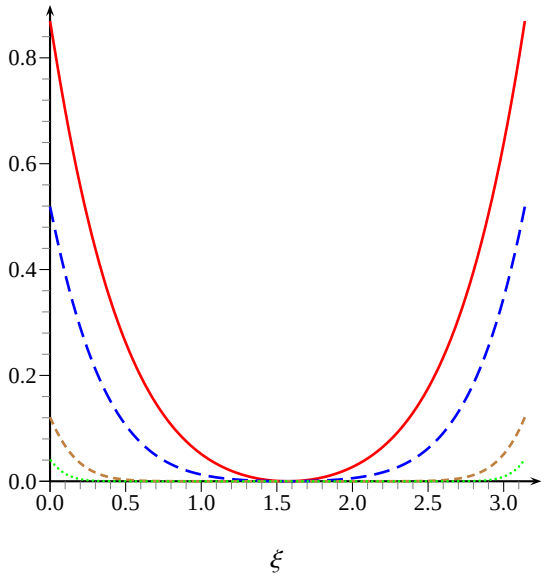

Figure 8: $\|B^b\|^2$ *for* $k = 1$ *(solid line),* $k = 2$ *(long dashed line),* $k = 10$ *(short dashed line) and* $k = 30$ *(dotted line).*

Now, when the sample size is fixed, the intrinsic bias corresponding to ξ^b is greater than the intrinsic bias corresponding to ξ^* in a wide range of values of the model parameter, that is the opposite behaviour showed up by the Riemannian risk.

4.2 Normal with mean value known

Let $X_1, \ldots, X_k$ be a random sample of size k from a normal distribution with known mean value μ_0 and standard deviation σ. Now the parameter space is $\Theta = (0, +\infty)$ and the metric tensor for the $N(\mu_0, \sigma)$ model is given by

$$g(\sigma) = \frac{2}{\sigma^2}$$

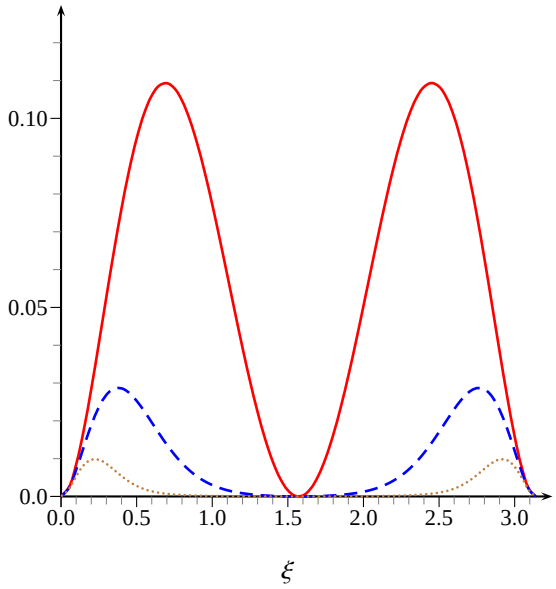

Figure 9: $\|B^*\|^2$ $k = 1, 2$ *(solid line) (the same curve),* $k = 10$ *(dashed line) and* $k = 30$ *(dotted line).*

We shall assume again the Jeffreys prior distribution for σ. Thus the joint density for σ and $(X_1, \ldots, X_k)$ is proportional to

$$\frac{1}{\sigma^{k+1}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{k}(X_i - \mu_0)^2\right)$$

depending on the sample through the sufficient statistic $S^2 = \frac{1}{k}\sum_{i=1}^{k}(X_i - \mu_0)^2$. When $(X_1, \ldots, X_k) = (x_1, \ldots, x_k)$ put $S^2 = s^2$. As

$$\int_0^\infty \frac{1}{\sigma^{k+1}} \exp\left(-\frac{k}{2\sigma^2}s^2\right) d\sigma = \frac{2^{\frac{k}{2}-1}}{(ks^2)^{\frac{k}{2}}}\,\Gamma\left(\frac{k}{2}\right)$$

the corresponding posterior distribution $\pi(\cdot \mid s^2)$ based on the Jeffreys prior satisfies

$$\pi(\sigma \mid s^2) = \frac{\left(ks^2\right)^{\frac{k}{2}}}{2^{\frac{k}{2}-1}\,\Gamma\left(\frac{k}{2}\right)} \frac{1}{\sigma^{k+1}} \exp\left(-\frac{k}{2\sigma^2}s^2\right)$$

Denote by ρ the Rao distance for the $N(\mu_0, \sigma)$. As we did in the previous example, instead of directly determining

$$\sigma^b(s) = \arg\min_{\sigma^e \in (0+\infty)} \int_0^{+\infty} \rho^2(\sigma^e, \sigma)\ \pi(\sigma \mid s^2)\, d\sigma$$

we perform a change of coordinates to obtain a Cartesian coordinate system. Then we compute the Bayes estimator for the new parameter's coordinate θ; as the estimator obtained in this way is intrinsic, we finish the argument recovering σ from θ. Formally, the

change of coordinates for which the metric tensor is constant and equal to 1 is obtained by solving the following differential equation:

$$1 = \left(\frac{d\sigma}{d\theta}\right) \frac{2}{\sigma^2}$$

with the initial conditions $\theta(1) = 0$. We obtain $\theta = \sqrt{2}\ \ln(\sigma)$ and $\theta = -\sqrt{2}\ \ln(\sigma)$; we only consider the first of these two solutions. We then obtain

$$\rho(\sigma_1, \sigma_2) = \sqrt{2}\ \left|\ln\left(\frac{\sigma_1}{\sigma_2}\right)\right| = |\theta_1 - \theta_2| \tag{3}$$

for $\theta_1 = \sqrt{2}\ln\sigma_1$ and $\theta_2 = \sqrt{2}\log\sigma_2$.

In the Cartesian setting, the Bayes estimator $\theta^b(s^2)$ for θ is the expected value of θ with respect to the corresponding posterior distribution, to be denoted as $\pi(\cdot \mid s^2)$. The integral can be solved, after performing the change of coordinates $\theta = \sqrt{2}\ \ln(\sigma)$ in terms of the gamma function Γ and the digamma function Ψ, that is the logarithmic derivative of Γ. Formally,

$$\begin{aligned}\theta^b(s^2) &= \frac{\left(ks^2\right)^{\frac{k}{2}}}{2^{\frac{n-3}{2}}\ \Gamma\left(\frac{k}{2}\right)} \int_0^{+\infty} \frac{\ln(\sigma)}{\sigma^{k+1}}\ \exp\left(-\frac{k}{2\sigma^2}s^2\right) d\sigma \\ &= \frac{\sqrt{2}}{2}\ \left(\ln\left(\frac{k}{2}\ s^2\right) - \Psi\left(\frac{k}{2}\right)\right)\end{aligned}$$

The Bayes estimator for σ is then

$$\sigma^b(s^2) = \sqrt{\frac{k}{2}}\ \exp\left(-\frac{1}{2}\Psi\left(\frac{k}{2}\right)\right)\ \sqrt{s^2}$$

Observe that this estimator, σ^b is a multiple of the maximum likelihood estimator $\sigma^* = \sqrt{s^2}$. We can evaluate the proportionality factor, for some values of n, obtaining the following table.

n	$\sqrt{\frac{k}{2}}\ \exp\left(-\frac{1}{2}\Psi\left(\frac{k}{2}\right)\right)$	n	$\sqrt{\frac{k}{2}}\ \exp\left(-\frac{1}{2}\Psi\left(\frac{k}{2}\right)\right)$
1	1.88736	10	1.05302
2	1.33457	20	1.02574
3	1.20260	30	1.01699
4	1.14474	40	1.01268
5	1.11245	50	1.01012
6	1.09189	100	1.00503
7	1.07767	250	1.00200
8	1.06725	500	1.00100
9	1.05929	1000	1.00050

The Riemannian risk of σ^b is given by

$$E_\sigma\left(\rho^2(\sigma^b,\sigma)\right) = E_\theta\left((\theta^b-\theta)^2\right) = \frac{1}{2}\Psi'(\frac{k}{2})$$

where Ψ' is the derivative of the digamma function. Observe that the Riemannian risk is constant on the parameter space.

Additionally we can compute the square of the norm of the bias vector corresponding to σ^b. In the Cartesian coordinate system θ and taking into account that $\frac{k\,S^2}{\sigma^2}$ is distributed as χ^2_k, we have

$$\begin{aligned} B^b_\theta &= E_\theta(\theta^b)-\theta = E_\sigma\left(\sqrt{2}\ \ln\left(\sqrt{\tfrac{k}{2}\,S^2}\right)-\Psi\left(\tfrac{k}{2}\right)\right)-\ \sqrt{2}\ln\sigma \\ &= \sqrt{2}\,E_\sigma\left(\ \ln\left(\sqrt{\frac{k\,S^2}{2\,\sigma^2}}\right)\right)-\Psi\left(\frac{k}{2}\right)=0 \end{aligned}$$

That is, the estimator σ^b is intrinsically unbiased. The bias vector corresponding to σ^* is given by

$$\begin{aligned} B^*_\theta &= E_\theta(\theta^*)-\theta = E_\sigma\left(\sqrt{2}\ \ln\left(\sqrt{S^2}\right)\right)-\ \sqrt{2}\ln\sigma \\ &= \sqrt{2}E_\sigma\left(\ \ln\left(\sqrt{\frac{S^2}{\sigma^2}}\right)\right)=\frac{1}{\sqrt{2}}\left(\Psi\left(\frac{k}{2}\right)\right)+\frac{1}{\sqrt{2}}\ln\left(\frac{k}{2}\right) \end{aligned}$$

which indicates that the estimator σ^* has a non-null intrinsic bias.

Furthermore, the Bayes estimator σ^b also satisfies the following interesting property related to the unbiasedness: it is the equivariant estimator under the action of the multiplicative group $\mathbb{R}_+$ that uniformly minimizes the Riemannian risk.

We can summarize the current statistical problem to the model corresponding to the sufficient statistic S^2 which follows a gamma distribution with parameters $\frac{n}{2\sigma^2}$ and $\frac{k}{2}$. This family is invariant under the action of the multiplicative group of $\mathbb{R}_+$ and it is straightforward to obtain that the equivariant estimators of σ which are function of $S=\sqrt{S^2}$ are of the form

$$T_\lambda(S)=\lambda S\ , \qquad \lambda\in(0,+\infty)$$

a family of estimators which contains $\sigma^b=\sqrt{\frac{k}{2}}\ \exp\left(-\frac{1}{2}\Psi\left(\frac{k}{2}\right)\right)S$, the Bayes estimator, and the maximum likelihood estimator $\sigma^*=S$. In order to obtain the equivariant estimator that minimizes the Riemannian risk, observe that the Rao distance (3) is an invariant loss function with respect to the induced group in the parameter space. This, together with the fact that the induced group acts transitively on the parameter space, makes the risk of any equivariant estimator to be constant on all the parameter space, among them the risk for σ^b and for S, as it was shown before. Therefore it is enough to minimize the risk at any point of the parameter space, for instance at $\sigma=1$. We want to

determine λ^* such that

$$\lambda^* = \arg \min_{\lambda \in (0,+\infty)} E_1\left(\rho^2(\lambda S, 1)\right)$$

It is easy to obtain

$$E_1\left(\rho^2(\lambda S,1)\right) = E_1\left(\left(\sqrt{2}\ \ln(\lambda S)\right)^2\right) = \frac{1}{2}\Psi'\left(\frac{k}{2}\right) + \frac{1}{2}\left(\Psi\left(\frac{k}{2}\right) + \ln\left(\frac{2\lambda^2}{k}\right)\right)^2$$

which attains an absolute minimum at

$$\lambda^* = \sqrt{\frac{k}{2}}\ \exp\left(-\frac{1}{2}\Psi\left(\frac{k}{2}\right)\right)$$

so that σ^b is the minimum Riemannian risk equivariant estimator.

Finally, observe that the results in Lehmann (1951) guarantee the unbiasedness of σ^b, as we obtained before, since the multiplicative group $\mathbb{R}_+$ is commutative, the induced group is transitive and σ^b is the equivariant estimator that uniformly minimizes the Riemannian risk.

4.3 Multivariate normal, Σ known

Let us consider now the case when the sample $X_1, \ldots, X_k$ comes from a n-variate normal distribution with mean value μ and known variance–covariance matrix Σ_0, positive definite. The joint density function can be expressed as

$$f(x_1, \ldots, x_k; \mu) = (2\pi)^{-\frac{nk}{2}}\ |\Sigma_0|^{-\frac{k}{2}}\ \mathrm{etr}\left(-\frac{k}{2}\Sigma_0^{-1}\left(s^2 + (\bar{x}-\mu)(\bar{x}-\mu)^{\mathsf{T}}\right)\right)$$

where $|A|$ denote the determinant of a matrix A, $\mathrm{etr}(A)$ is equal to the exponential mapping evaluated at the trace of the matrix A, $\bar{x}$ denotes $\frac{1}{k}\sum_{i=1}^k x_i$ and s^2 stands for $\frac{1}{k}\sum_{i=1}^k (x_i - \bar{x})(x_i - \bar{x})^{\mathsf{T}}$. In this case, the metric tensor G coincides with Σ_0^{-1}. Assuming the Jeffreys prior distribution for μ, the joint density for μ and $(X_1, \ldots, X_k) = (x_1, \ldots, x_k)$ is proportional to

$$|\Sigma_0|^{-\frac{1}{2}}\ \exp\left(-\frac{k}{2}\left((\bar{x}-\mu)^{\mathsf{T}}\Sigma_0^{-1}(\bar{x}-\mu)\right)\right)$$

Next we compute the corresponding posterior distribution. Since

$$\int_{\mathbb{R}^n} \exp\left(-\frac{k}{2}\left((\bar{x}-\mu)^{\mathsf{T}}\Sigma_0^{-1}(\bar{x}-\mu)\right)\right)\ |\Sigma_0|^{-\frac{1}{2}}\ d\mu = \left(\frac{2\pi}{k}\right)^{\frac{n}{2}}$$

the posterior distribution $\pi(\cdot \mid \bar{x})$ based on the Jeffreys prior is given by

$$\pi(\mu \mid \bar{x}) = (2\pi)^{-\frac{n}{2}}\ |\frac{1}{k}\Sigma_0|^{-\frac{1}{2}}\ \exp\left(-\frac{1}{2}(\mu - \bar{x})^{\mathsf{T}}\left(\frac{1}{k}\Sigma_0^{-1}\right)(\mu - \bar{x})\right)$$

Observe here that the parameter's coordinate μ is already Cartesian, the Riemannian distance expressed via this coordinate system coincides with the Euclidean distance between the coordinates. Therefore, the Bayes estimator for μ is precisely the sample mean $\overline{x}$.

$$\mu^b(\overline{x}) = \overline{x}$$

which coincides with the maximum likelihood estimator.

Arguments of invariance that are analogous to those in the previous example apply here, where $\mu^b = \overline{X}$ is the minimum Riemannian risk equivariant estimator under the action of the translation group $\mathbb{R}^n$. The induced group is again transitive so the risk is constant at any point of the parameter space; for simplicity we may consider $\mu = 0$. A direct computation shows that

$$E_0\left(\rho^2(\overline{X}, 0)\right) = E_0\left(\overline{X}^{\mathsf{T}}\, \Sigma_0^{-1}\, \overline{X}\right) = \frac{n}{k}$$

Following Lehmann, Lehmann (1951), and observing that the translation group is commutative, $\overline{X}$ is also unbiased, as can easily be verified.

References

Abramovitz, M. (1970). *Handbook of Mathematical Functions*. New York: Dover Publications Inc.

Atkinson, C. and Mitchel, A. (1981). Rao's distance measure. *Sankhyà*, 43, A, 345-365.

Berger, J. O. and Bernardo, J. M. (1992). Bayesian Statistics 4. (Bernardo, Bayarri, Berger, Dawid, Hackerman, Smith & West Eds.), *On the development of reference priors*, (pp. 35-60 (with discussion)). Oxford: Oxford University Press.

Bernardo, J. and Juárez, M. (2003). Bayesian Statistics 7. (Bernardo, Bayarri, Berger, Dawid, Hackerman, Smith & West Eds.), *Intrinsic Estimation*, (pp. 465-476). Berlin: Oxford University Press.

Bernardo, J. M. (1979). Reference posterior distributions for bayesian inference. *Journal of the Royal Statistical Society B*, 41, 113-147 (with discussion) Reprinted in Bayesian Inference 1 (G. C. Tiao and N. G. Polson, eds). Oxford: Edward Elgar, 229-263.

Bernardo, J. M. and Rueda, R. (2002). Bayesian hypothesis testing: A reference approach. *International Statistical Review*, 70, 351-372.

Burbea, J. (1986). Informative geometry of probability spaces. *Expositiones Mathematicae*, 4.

Burbea, J. and Rao, C. (1982). Entropy differential metric, distance and divergence measures in probability spaces: a unified approach. *Journal of Multivariate Analysis*, 12, 575-596.

Chavel, I. (1993). *Riemannian Geometry. A Modern Introduction.* Cambridge Tracts in Mathematics. Cambridge University Press.

Erdélyi, A., Magnus, W., Oberhettinger, F., and Tricomi, F. G. (1953-1955). *Higher Transcendental Functions*, volume 1,2,3. New York: McGraw–Hill.

Hendricks, H. (1991). A Cramér-Rao type lower bound for estimators with values in a manifold. *Journal of Multivariate Analysis*, 38, 245-261.

Hicks, N. (1965). *Notes on Differential Geometry*. New York: Van Nostrand Reinhold.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society*, 186 A, 453-461.

Karcher, H. (1977). Riemannian center of mass and mollifier smoothing. *Communications on Pure and Applied Mathematics*, 30, 509-541.

Kendall, W. (1990). Probability, convexity and harmonic maps with small image I: uniqueness and fine existence. *Proceedings of the London Mathematical Society*, 61, 371-406.

Lehmann, E. (1951). A general concept of unbiassedness. *Annals of Mathematical Statistics*, 22, 587-592.

Oller, J. (1987). Information metric for extreme value and logistic probability distributions. *Sankhyà*, 49 A, 17-23.

Oller, J. (1993a). Multivariate Analysis: Future Directions 2. (Cuadras and Rao Eds.), *On an Intrinsic analysis of statistical estimation* (pp. 421-437). Amsterdam: Elsevier science publishers B. V., North Holland.

Oller, J. (1993b). Stability Problems for Stochastic Models. (Kalasnikov and Zolotarev Eds.), *On an Intrinsic Bias Measure* (pp. 134-158). Berlin: Lect. Notes Math. 1546, Springer Verlag.

Oller, J. and Corcuera, J. (1995). Intrinsic analysis of statistical estimation. *Annals of Statistics*, 23, 1562-1581.

Oller, J. and Cuadras, C. (1985). Rao's distance for multinomial negative distributions. *Sankhyà*, 47 A, 75-83.

Rao, C. (1945). Information and accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, 37, 81-91.

Smith, S. (2005). Covariance, Subspace, and Intrinsic Cramér-Rao Bounds. *IEEE Transactions on Signal Processing*, 53, 1610-1630.

Spivak, M. (1979). *A Comprehensive Introduction to Differential Geometry*. Berkeley: Publish or Perish.

Discussion of "What does intrinsic mean in statistical estimation?" by Gloria García and Josep M. Oller

Jacob Burbea

University of Pittsburgh

burbea+@pitt.edu

In this review paper, García and Oller discuss and study the concept of intrinsicality in statistical estimation, where the attention is focused on the invariance properties of an estimator.

Statistical estimation is concerned with assigning a plausible probabilistic formalism that is supposed to explain the observed data. While the inference involved in such an estimation is inductive, the formulation and the derivation of the process of estimation are based on a mathematical deductive reasoning. In the analysis of this estimation, one likes to single out those estimates in which the associated estimator possesses a certain invariance property, known as the *functional invariance of an estimator*. Such an estimator essentially represents a well-defined probability measure, and as such it is termed as an *intrinsic estimator*. The present paper revolves around this concept within the framework of a parametric statistical estimation. In this context, the probability is assumed to belong to a family that is indexed by some parameter θ which ranges in a set Θ, known as the *parameter space of the statistical model*. In particular, the resulting inductive inferences are usually formulated in the form of a point or a region estimates which ensue from the estimation of the parameter θ. In general, however, such an estimation depend on the particular parametrization of the model, and thus different estimators may lead to different methods of induction. In contrast, and by definition, intrinsic estimators do not depend on the specific parametrization, a feature that is significant as well as desirable.

In order to develop a suitable analysis, called *intrinsic analysis*, of such a statistical estimation, it is required to assess the performance of the intrinsic estimators in terms of intrinsic measures or tools. At this stage, it is worthwhile to point out that, for example, the mean square error and the bias, which are commonly accepted measures of a performance of an estimator, are clearly dependent on the model parametrization. In particular, minimum variance estimation and unbiasedness are non intrinsic concepts. To avoid such situations, the intrinsic analysis exploits the natural geometrical structures that the statistical models, under some regularity conditions, possess to construct quantities which have a well-defined geometrical meaning, and hence also intrinsic.

More explicitly, let $(\chi, \mathcal{A}, \mu)$ be a measure space and consider the map $f : \chi \times \Theta \to \mathbb{R}^+$ such that $f(\cdot\,|\,\theta)d\mu$ defines a probability measure on the measure space $(\xi, \mathcal{A})$, to be denoted as P_θ, for each $\theta \in \Theta$. In this setting, the family $(P_\theta)_{\theta\in\Theta}$ is referred to as the *statistical model under the parameter space* Θ with μ as its *reference measure*. In general, the parameter space Θ is any smooth manifold that is modeled on some Banach space. It is also assumed that the mapping $\theta \to P_\theta$ of Θ onto $(P_\theta)_{\theta\in\Theta}$ is, at least locally, injective. Moreover, since the ensuing analysis is essentially local in nature, it is sufficient, as well as customary, to use the same symbol θ to denote points in θ and their coordinates in the Hilberian tangent space of the point. Some additional regularity conditions, which are quite mild, have to be imposed on the density f to guarantee that the Fisher information matrix of the model exists and is positive-definite on Θ. In this case, Θ admits a natural Riemannian manifold structure in which the Fisher information matrix serves as a metric tensor of a differential metric, known as the *information metric* of the model, and we say that the parametric statistical model $(P_\theta)_{\theta\in\Theta}$ is *regular*. In particular, such a model may well be identified with the now Riemannian manifold Θ, and hence we have at our disposal numerous intrinsic geometrical quantities as the Riemannian information metric, its Riemannian volume elements, its indicatrice, its Levi-Civita connection, its geodesics, and its Riemannian and sectional curvatures. The geodesic distance associated with the information metric is called the *information distance*, or the *Rao distance*, of the statistical model, and is usually denoted by f. We refer to Burbea (Burbea, 1986) for additional details.

As noted, this information geometrical structure enables the intrinsic analysis of a statistical estimation to develop intrinsic quantities and concepts that are parallel to the non-intrinsic ones. For example, the square error loss is replaced by the square ρ^2 of the information distance ρ of the statistical model. There exist, of course, other possible intrinsic losses, some of which even admit a simple expression in terms of easily computed quantities of the model. In contrast, and as a disadvantage, the information distance ρ does not, in general, admit a closed form expression. However, as far as the information content of a state is concerned, the square of the information distance should be regarded as the canonical intrinsic version of the square error. In a similar fashion, the intrinsic version of the mean, or the expected value, of a random object valued on the manifold Θ, is defined in terms of an affine correction on Θ, and is said to be *canonical* when the affine connection is the Levi-Civita connection associated with the information metric on Θ. In turn, such an intrinsic version of the mean gives rise to the intrinsic version of the bias. These intrinsic concepts were first developed in Oller and Corcuera (Oller and Corcuera, 1995), where the governing analysis patterned along differential geometric lines exhibited in Karcher (Karcher, 1977). Moreover, under the assumption that Θ is a finite dimensional real manifold, Oller and Corcuera (Oller and Corcuera, 1995) were able to establish an intrinsic version of Cramér-Rao inequality. The method of proof is based on comparison theorems of Riemannian geometry (see Chavel (Chavel, 1993)). A similar result, but which a different proof, seems to appear earlier and it is due

to Hendricks (Hendricks, 1991). Recent developments on the subject matter may be founding in Smith (Smith, 2005). The obtained intrinsic Cramér-Rao inequality also a tensorial version which, on following a method of proof due to Kendall (Kendall, 1990), leads to an intrinsic version of Rao-Blackwell theorem. A more detailed account of these and related results are exposed in the present discussed paper of García and Oller.

A somewhat different approach to intrinsic estimation is obtained by considering affine connections on the manifold Θ that differ from the Levi-Civita connection associated with the information metric. In this vein, García and Oller cite a recent work of Bernardo and Juárez (Bernardo and Juárez, M., 2003) in which the concept of intrinsic estimation is based on singling out the estimator that minimizes the Bayesian risk, where the symmetrized Kullback-Leibler divergence serves as an intrinsic loss, and where the so-called *information prior* serves as a reference prior. This information prior, which is derived from information theoretical arguments, is independent of the model parametrization, and in some cases coincides with the Jeffreys uniform prior distribution, in which the, usually improper, prior is proportional to the Riemannian volume element of the information metric on Θ. As such, the obtained estimator is indeed intrinsic, for it does not depend on the model parametrization.

To illustrate and elucidate matters, García and Oller conclude their paper by exploring, for some concrete cases, the intrinsic estimator obtained by minimizing the Bayesian risk, where the square of the information distance serves as an intrinsic loss, and where the Jeffreys prior serves as a reference prior. In each case, the obtained estimator is compared with the one obtained in Bernardo and Juárez (Bernardo and Juárez, M., 2003).

In conclusion, the review of García and Oller is presented in a lucid and clear manner. It provides a virtually self contained, and quite profound, account on intrinsic estimation. As such, the review should be regarded as a solid contribution to the subject matter.

References

Abramovitz, M. (1970). *Handbook of Mathematical Functions*. New York: Dover Publications Inc.

Atkinson, C. and Mitchel, A. (1981). Rao's distance measure. *Sankhyà*, 43, A, 345-365.

Berger, J. O. and Bernardo, J. M. (1992). Bayesian Statistics 4. (Bernardo, Bayarri, Berger, Dawid, Hackerman, Smith & West Eds.), *On the development of reference priors*, (pp. 35-60 (with discussion)). Oxford: Oxford University Press.

Bernardo, J. M. (1979). Reference posterior distributions for bayesian inference. *Journal of the Royal Statistical Society B*, 41, 113-147 (with discussion) Reprinted in Bayesian Inference 1 (G. C. Tiao and N. G. Polson, eds). Oxford: Edward Elgar, 229-263.

Bernardo, J. and Juárez, M. (2003). Bayesian Statistics 7. (Bernardo, Bayarri, Berger, Dawid, Hackerman, Smith & West Eds.), *Intrinsic Estimation*, (pp. 465-476). Berlin: Oxford University Press.

Bernardo, J. M. and Rueda, R. (2002). Bayesian hypothesis testing: A reference approach. *International Statistical Review*, 70, 351-372.

Burbea, J. (1986). Informative geometry of probability spaces. *Expositiones Mathematicae*, 4.

Burbea, J. and Rao, C. (1982). Entropy differential metric, distance and divergence measures in probability spaces: a unified approach. *Journal of Multivariate Analysis*, 12, 575-596.

Chavel, I. (1993). *Riemannian Geometry. A Modern Introduction.* Cambridge Tracts in Mathematics. Cambridge University Press.

Erdélyi, A., Magnus, W., Oberhettinger, F., and Tricomi, F. G. (1953-1955). *Higher Transcendental Functions*, volume 1,2,3. New York: McGraw–Hill.

Hendricks, H. (1991). A Cramér-Rao type lower bound for estimators with values in a manifold. *Journal of Multivariate Analysis*, 38, 245-261.

Hicks, N. (1965). *Notes on Differential Geometry*. New York: Van Nostrand Reinhold.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society*, 186 A, 453-461.

Karcher, H. (1977). Riemannian center of mass and mollifier smoothing. *Communications on Pure and Applied Mathematics*, 30, 509-541.

Kendall, W. (1990). Probability, convexity and harmonic maps with small image I: uniqueness and fine existence. *Proceedings of the London Mathematical Society*, 61, 371-406.

Lehmann, E. (1951). A general concept of unbiassedness. *Annals of Mathematical Statistics*, 22, 587-592.

Oller, J. (1987). Information metric for extreme value and logistic probability distributions. *Sankhyà*, 49 A, 17-23.

Oller, J. (1993a). Multivariate Analysis: Future Directions 2. (Cuadras and Rao Eds.), *On an Intrinsic analysis of statistical estimation* (pp. 421-437). Amsterdam: Elsevier science publishers B. V., North Holland.

Oller, J. (1993b). Stability Problems for Stochastic Models. (Kalasnikov and Zolotarev Eds.), *On an Intrinsic Bias Measure* (pp. 134-158). Berlin: Lect. Notes Math. 1546, Springer Verlag.

Oller, J. and Corcuera, J. (1995). Intrinsic analysis of statistical estimation. *Annals of Statistics*, 23, 1562-1581.

Oller, J. and Cuadras, C. (1985). Rao's distance for multinomial negative distributions. *Sankhyà*, 47 A, 75-83.

Rao, C. (1945). Information and accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, 37, 81-91.

Smith, S. (2005). Covariance, Subspace, and Intrinsic Cramér-Rao Bounds. *IEEE Transactions on Signal Processing*, 53, 1610-1630.

Spivak, M. (1979). *A Comprehensive Introduction to Differential Geometry*. Berkeley: Publish or Perish.

Joan del Castillo

Universitat Autónoma de Barcelona, Spain

castillo@mat.uab.es

The authors are to be congratulated for summarizing material from their remarkable mathematical work over recent years in a paper that is readable to a mathematical statistician.

The paper helps us to have a clearer understanding of certain everyday concepts such as bias, mean-square error or the parametrization invariant estimator. It is important to bear in mind that bias and mean-square errors are parametrization-dependent, most particularly if we are interested in estimating probability density functions rather than parameters.

The examples given in the paper are also remarkabe. The first shows that an unbiased estimator with less mean square error than the maximum-likelihood (ML) estimator, in certain points within the parameter space, is discarded from a differential geometric point of view. Whatever the circumstances, however, the ML estimator is once again a reasonable estimator, even in a non-reasonable situation. The examples also show that the Riemannian metric introduced is a complex tool for practical applications. The Kullback-Leibler distance is often a simpler way of performing a similar analysis, as Bernardo and Juárez (2003) carried out.

In Example 4.2, corresponding to the normal distribution with a known mean value, a Bayes estimator for the standard deviation was considered from a Jeffreys prior. The estimator was found by using a Riemannian distance, solving a differential equation and some integrals. Finally, a multiple of the sample standard deviation (the ML estimator) was found. The new estimator is the equivariant estimator that uniformly minimizes the Riemannian risk. Will we now change our view that the ML estimator bias-corrected is the best estimator we can have, on the basis of this mathematical result?

From my point of view, the main question is the following: if a model is not absolutely correct, then does the Riemannian metric have any practical sense? Moreover, in applied statistical work, who really believes that a statistical model is absolutely accurate?

Wilfrid S. Kendall

w.s.kendall@warwick.ac.uk

It is a pleasure to congratulate the authors on this paper, particularly as it is gratifying to see previous work on Riemannian mean-values being put to such good statistical use! Here are three comments stimulated by the reading of this most interesting paper:

1. Another interesting statistical use of Riemannian mean-values (of a more data-analytic nature) is to be found in the work of HuiLing Le and collaborators (for example, Kume and Le 2003; Le 2004). Here one is interested in computing summary shapes to represent a whole population of shapes (often derived from biological data); one is driven to use Riemannian mean-values because there is no natural Cartesian coordinate system for shapes.
2. The original motivation for my (1990) work was entirely probabilistic, and it was natural to continue investigations using barycentres based on non-metric connections (Kendall (1991, 1992)), with close links to convexity. Non-metric connections also arise in statistical asymptotics; have the authors considered whether there is anything to be gained from using these in their context?
3. The elegant discussion in Section 3 includes the important reminder that conditional centres of mass in the Riemannian context do not obey the classic commutativity of nested conditional expectations. For workers in stochastic differential geometry this leads to the consideration of Γ-martingales, which prove of great importance in probabilistic approaches to harmonic maps. It would be interesting if the work of the current paper could be extended in this way, perhaps to take account of time-series analysis (where it would be natural to consider a whole sequence of conditional Riemannian centres of mass). The work of Darling (2002) and Srivastava and Klassen (2004) may be relevant here.

References

Darling, R. W. R. (2002). Intrinsic location parameter of a diffusion process. *Electronic Journal of Probability*, 7, 23 pp. (electronic).

Kendall, W. S. (1990). Probability, convexity, and harmonic maps with small image. I. Uniqueness and fine existence. *Proceedings of the London Mathematical Society (3)*, 61, 371-406.

Kendall, W. S. (1991). Convex geometry and nonconfluent Γ-martingales. I. Tightness and strict convexity. In *Stochastic analysis (Durham, 1990)*, *The London Mathematical Society. Lecture Note Ser.*, 167, 163-178. Cambridge: Cambridge Univ. Press.

Kendall, W. S. (1992). Convex geometry and nonconfluent Γ-martingales. II. Wellposedness and G-martingale convergence. *Stochastics Stochastics Reports*, 38, 135-147.

Kume, A. and H. Le (2003). On Fréchet means in simplex shape spaces. *Advances in Applied Probability*, 35, 885-897.

Le, H. (2004). Estimation of Riemannian barycentres. *LMS Journal of Computation and Mathematics*, 7, 193-200 (electronic).

Srivastava, A. and E. Klassen (2004). Bayesian and geometric subspace tracking. *Advances in Applied Probability*, 36, 43-56.

Steven Thomas Smith[1]

MIT Lincoln Laboratory, Lexington, MA 02420

stsmith@ll.mit.edu

The intrinsic nature of estimation theory is fundamentally important, a fact that was realized very early in the field, as evidenced by Rao's first and seminal paper on the subject in 1945 (Rao, 1945). Indeed, intrinsic hypothesis testing, in the hands of none other than R. A. Fisher played a central role establishing one of the greatest scientific theories of the twentieth century: Wegener's theory of continental drift (Fisher, 1953). (Diaconis recounts the somewhat delightful way in which Fisher was introduced to this problem (Diaconis, 1988)). Yet in spite of its fundamental importance, intrinsic analysis in statistics, specifically in estimation theory, has in fact received relatively little attention, notwithstanding important contributions from Bradley Efron (Efron, 1975), Shun-ichi Amari (Amari, 1985), Josep Oller and colleagues (Oller, 1991), Harrie Hendriks (Hendriks, 1991), and several others. I attribute this limited overall familiarity with intrinsic estimation to three factors: (1) linear estimation theory, though in itself is implicitly intrinsic, is directly applicable to the vast majority of linear, or linearizable, problems encountered in statistics, physics, and engineering, obviating any direct appeal to the underlying coordinate invariance; (2) consequently, the number of problems demanding an intrinsic approach is limited, though in some fields, such as signal processing, nonlinear spaces abound (spheres, orthogonal and unitary matrices, Grassmann manifolds, Stiefel manifolds, and positive-definite matrices); (3) intrinsic estimation theory is really nonlinear estimation theory, which is hard, necessitating as it does facility with differential and Riemannian geometry, Lie groups, and homogeneous spaces—even Efron acknowledges this, admitting being "frustrated by the intricacies of the higher order differential geometry" [Efron, 1975, p. 1241]. García and Oller's review of intrinsic estimation is a commendable contribution to addressing this vital pedagogical matter, as well as providing many important insights and results on the application of intrinsic estimation. Their explanation of the significance of statistical invariance provides an excellent introduction to this hard subject.

1. This work was sponsored by DARPA under Air Force contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the author and are not necessarily endorsed by the United States Government.

Intrinsic estimation concerns itself with multidimensional quantities that are invariant to the arbitrary choice of coordinates used to describe these dimensions—when estimating points on the globe, what should it matter if we choose "Winkel's Tripel" projection, or the old Mercator projection? Just like the arbitrary choice of using the units of feet or meters, the numerical answers we obtain necessarily depend upon the choice of coordinates, but the answers are invariant to them because they transform according to a change of variables formula involving the Jacobian matrix.

Yet the arbitrary choice of metric—feet versus meters—is crucial in that it specifies the performance measure in which all results will be expressed numerically. There are two roads one may take in the intrinsic analysis of estimation problems: the purely intrinsic approach, in which the arbitrary metric itself is chosen based upon an invariance criterion, or the general case in which the statistician, physicist, or engineer chooses the metric they wish to use, invariant or not, and demands that answers be given in units expressed using this specific metric. García and Oller take the purely intrinsic path. They say, "the square of the Rao distance is the most natural intrinsic version of the square error," and proceed to compute answers using this Rao (or Fisher information) metric throughout their analysis. One may well point out that this Fisher information metric is itself based upon a statistical model chosen using various assumptions, approximations, or, in the best instances, the physical properties of the estimation problem itself. Thus the Fisher information metric is natural insofar as the measurements adhere to the statistical model used for them, indicating a degree of arbitrariness or uncertainty even in this "most natural" choice for the metric. In addition to the choice of metric, the choice of score function with which to evaluate the estimation performance is also important. This Fisher score yields intrinsic Cramér-Rao bounds, and other choices of score functions yield intrinsic versions of the Weiss-Weinstein, Bhattacharyya, Barankin and Bobrovsky-Zakai bounds (Smith, Scharf and McWhorter, 2006).

Moreover, there may be legitimately competing notions of natural invariance. The invariance that arises from the Fisher metric is one kind, as recognized. But in the cases where the parameter space is a Lie group $\boldsymbol{G}$ or a (reductive) homogeneous space $\boldsymbol{G}/\boldsymbol{H}$ ($\boldsymbol{H} \subset \boldsymbol{G}$ a Lie subgroup), such as found in the examples of unitary matrices, spheres, etc. cited above, invariance to transformations by the Lie group $\boldsymbol{G}$ is typically of principal physical importance to the problem. In such cases, we may wish to analyze the square error using the unique invariant Riemannian metric on this space, e.g., the square error on the sphere would be great circle distances, not distances measured using the Fisher information metric. In most cases, this natural, intrinsic metric (w.r.t. $\boldsymbol{G}$-group invariance) is quite different from the natural, intrinsic (w.r.t. the statistical model) Fisher metric. I am aware of only one nontrivial example where these coincide: the natural $\boldsymbol{Gl}(n,\mathbb{C})$-invariant Riemannian metric for the space of positive-definite matrices $\boldsymbol{Gl}(n,\mathbb{C})/\boldsymbol{U}(n)$ is the very same as the Fisher information metric for Gaussian covariance

matrix estimation [Smith, 2005, p. 1620]:

$$g_{\mathbf{R}}(\mathbf{A}, \mathbf{A}) = \mathrm{tr}(\mathbf{A}\mathbf{R}^{-1})^2 \tag{1}$$

(ignoring an arbitrary scale factor). In this example, the square error between two covariance matrices is given by, using Matlab notation and decibel units,

$$d(\mathbf{R}_1, \mathbf{R}_2) = \mathrm{norm}(10 * \log 10(\mathrm{eig}(\mathbf{R}_1, \mathbf{R}_2))) \tag{2}$$

(i.e., the logarithm of the 2-norm of a vector of generalized eigenvalues between the covariance matrices), precisely akin to the distance between variances in Equation (3) from García and Oller's review.

Furthermore, and perhaps most importantly, the statistician or engineer may well respond "So what?" to this metric's invariance, whether it be $\boldsymbol{G}$-invariance or invariance arising from the statistical model. For example, Hendriks (Hendriks, 1991) considers the embedded metric for a parameter manifold embedded in a higher dimensional Euclidean space; the extrinsic Euclidean metric possesses very nice invariance properties, but these are typically lost within arbitrary constrained submanifolds. The choice of metric is, in fact, arbitrary, and there may be many good practical reasons to express one's answers using some other metric instead of the most natural, invariant one. Or not—the appropriateness of any proposed metric must be assessed in the context of the specific problem at hand. For these reasons, an analysis of intrinsic estimation that allows for arbitrary distance metrics is of some interest (Smith, 2005), (Smith, Scharf and McWhorter, 2006), as well the special and important special case of the Fisher information metric.

Another critical factor affecting the results obtained is the weapon one chooses from the differential geometric arsenal. García and Oller present results obtained from the powerful viewpoint of comparison theory (Cheeger and Ebin, 1975), a global analysis that uses bounds on a manifold's sectional curvature to compare its global structure to various model spaces. The estimation bounds derived using these methods possess two noteworthy properties. First, Oller and Corcuera's expressions (Oller and Corcuera, 1995) are remarkably simple! It is worthwhile paraphrasing these bounds here. Let $\boldsymbol{\theta}$ be an unknown n-dimensional parameter, $\hat{\boldsymbol{\theta}}(\mathbf{z})$ an estimator that depends upon the data $\mathbf{z}$, $\boldsymbol{A}(\mathbf{z}|\boldsymbol{\theta}) = \exp_{\boldsymbol{\theta}}^{-1} \hat{\boldsymbol{\theta}}$ the (random) vector field representing the difference between the estimator $\hat{\boldsymbol{\theta}}$ and the truth $\boldsymbol{\theta}$, $\boldsymbol{b}(\boldsymbol{\theta}) = E[\boldsymbol{A}]$ the bias vector field, "$\exp_{\boldsymbol{\theta}}$" the Riemannian exponential map w.r.t. the Fisher metric, $\bar{K} \stackrel{\mathrm{def}}{=} \max_{\boldsymbol{\theta},H} K_{\boldsymbol{\theta}}(H)$ the maximum sectional curvature of the parameter manifold over all two-dimensional subspaces H, and $D = \sup_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2} d(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ be the manifold's diameter. Then, ignoring the sample size k which will always appear in the denominator for independent samples, Theorem 3.1 implies that the mean square error (MSE) about the bias—which is the random part of

the bound (add $\|\boldsymbol{b}\|^2$ for the MSE about $\boldsymbol{\theta}$)—is

$$E[\|\boldsymbol{A}-\boldsymbol{b}\|^2] \geq \begin{cases} n^{-1}(\operatorname{div} E[\boldsymbol{A}] - E[\operatorname{div} \boldsymbol{A}])^2, \\ \qquad\qquad \text{(general case);} \\ n^{-1}\Big(\operatorname{div}\boldsymbol{b} + 1 + (n-1)|\bar{K}|^{\frac{1}{2}}\|\boldsymbol{b}\|\coth(|\bar{K}|^{\frac{1}{2}}\|\boldsymbol{b}\|)\Big)^2, \\ \qquad\qquad \bar{K} \leq 0,\ \operatorname{div}\boldsymbol{b} \geq -n; \\ n^{-1}(\operatorname{div}\boldsymbol{b} + 1 + (n-1)\bar{K}^{\frac{1}{2}}D\cot(\bar{K}^{\frac{1}{2}}D))^2, \\ \qquad\qquad \bar{K} \geq 0,\ D < \pi/(2\bar{K}^{\frac{1}{2}}),\ \operatorname{div}\boldsymbol{b} \geq -1; \\ n, \qquad \bar{K} \leq 0; \\ n^{-1}, \qquad \bar{K} \geq 0,\ D < \pi/(2\bar{K}^{\frac{1}{2}}), \end{cases} \tag{3}$$

where

$$\operatorname{div}\boldsymbol{b}(\boldsymbol{\theta}) = \frac{1}{|g|^{\frac{1}{2}}}\sum_i \frac{\partial |g|^{\frac{1}{2}}\boldsymbol{b}^i(\boldsymbol{\theta})}{\partial\theta^i} \tag{4}$$

is the divergence of the vector field $\boldsymbol{b}(\boldsymbol{\theta})$ w.r.t. the Fisher metric $\mathbf{G}(\boldsymbol{\theta})$ and a particular choice of coordinates $(\theta^1, \theta^2, \ldots, \theta^n)$, and $|g|^{\frac{1}{2}} = |\det \mathbf{G}(\boldsymbol{\theta})|^{\frac{1}{2}}$ is the natural Riemannian volume form, also w.r.t. the Fisher metric.

Compare these relatively simple expressions to the ones I obtain for an arbitrary metric [Smith, 2005, Theorem 2]. In these bounds, the covariance $\mathbf{C}$ of $\boldsymbol{A}-\boldsymbol{b}$ about the bias is given by

$$\mathbf{C} \geq \mathbf{M}_{\boldsymbol{b}}\mathbf{G}^{-1}\mathbf{M}_{\boldsymbol{b}}^{\mathrm{T}} - \tfrac{1}{3}(\boldsymbol{R}_{\mathrm{m}}(\mathbf{M}_{\boldsymbol{b}}\mathbf{G}^{-1}\mathbf{M}_{\boldsymbol{b}}^{\mathrm{T}})\mathbf{G}^{-1}\mathbf{M}_{\boldsymbol{b}}^{\mathrm{T}} + \mathbf{M}_{\boldsymbol{b}}\mathbf{G}^{-1}\boldsymbol{R}_{\mathrm{m}}(\mathbf{M}_{\boldsymbol{b}}\mathbf{G}^{-1}\mathbf{M}_{\boldsymbol{b}}^{\mathrm{T}})^{\mathrm{T}}) \tag{5}$$

(ignoring negligible higher order terms), where

$$\mathbf{M}_{\boldsymbol{b}} = \mathbf{I} - \frac{1}{3}\|\boldsymbol{b}\|^2\boldsymbol{K}(\boldsymbol{b}) + \nabla\boldsymbol{b}, \tag{6}$$

$(\mathbf{G})_{ij} = g(\partial/\partial\theta^i, \partial/\partial\theta^j)$ is the Fisher information matrix, $\mathbf{I}$ is the identity matrix, $(\nabla\boldsymbol{b})^i{}_j = (\partial\boldsymbol{b}^i/\partial\theta^j) + \sum_k \Gamma^i_{jk}\boldsymbol{b}^k$ is the covariant differential of $\boldsymbol{b}(\boldsymbol{\theta})$, Γ^i_{jk} are the Christoffel symbols, and the matrices $\boldsymbol{K}(\boldsymbol{b})$ and $\boldsymbol{R}_{\mathrm{m}}(\mathbf{C})$ representing sectional and Riemannian curvature terms are defined by

$$(\boldsymbol{K}(\boldsymbol{b}))_{ij} = \begin{cases} \sin^2\alpha_i \cdot K(\boldsymbol{b}\wedge\boldsymbol{E}_i) + O(\|\boldsymbol{b}\|^3), & \text{if } i = j; \\ [\sin^2\alpha'_{ij}\cdot K(\boldsymbol{b}\wedge(\boldsymbol{E}_i+\boldsymbol{E}_j)) & \\ \quad -\sin^2\alpha''_{ij}\cdot K(\boldsymbol{b}\wedge(\boldsymbol{E}_i-\boldsymbol{E}_j))] & \text{if } i \neq j, \\ \quad + O(\|\boldsymbol{b}\|^3), & \end{cases} \tag{7}$$

α_i, α'_{ij}, and α''_{ij} are the angles between the tangent vector $\boldsymbol{b}$ and the orthonormal tangent basis vectors $\boldsymbol{E}_i = \partial/\partial\theta^i$, $\boldsymbol{E}_i + \boldsymbol{E}_j$, and $\boldsymbol{E}_i - \boldsymbol{E}_j$, respectively, and $\boldsymbol{R}_{\mathrm{m}}(\mathbf{C})$ is the mean Riemannian curvature defined by the equality

$$\langle \boldsymbol{R}_{\mathrm{m}}(\mathbf{C})\boldsymbol{\Omega}, \boldsymbol{\Omega}\rangle = E\big[\langle \boldsymbol{R}(\boldsymbol{X} - \boldsymbol{b}, \boldsymbol{\Omega})\boldsymbol{\Omega}, \boldsymbol{X} - \boldsymbol{b}\rangle\big], \tag{8}$$

where $\boldsymbol{R}(\boldsymbol{X}, \boldsymbol{Y})\boldsymbol{Z}$ is the Riemannian curvature tensor. Ignoring curvature, which is reasonable for small errors and biases, as well as the intrinsically local nature of the Cramér-Rao bound itself, the intrinsic Cramér-Rao bound simplifies to the expression

$$\mathbf{C} \geq (\mathbf{I} + \nabla\boldsymbol{b})\mathbf{G}^{-1}(\mathbf{I} + \nabla\boldsymbol{b})^{\mathrm{T}}. \tag{9}$$

The trace of Equations (5) and (9) (plus the square length $\|\boldsymbol{b}\|^2$) provides the intrinsic Cramér-Rao bound on the mean square error between the estimator $\hat{\boldsymbol{\theta}}$ and the true parameter $\boldsymbol{\theta}$.

Let's take a breath and step back to compare how these bounds relate to one another, beginning with the simplest, one-dimensional Euclidean case as the basis for our comparison. The biased Cramér-Rao bound for this case is provided in an exercise from Van Trees' excellent reference [Van Trees, 1968, p. 146f]: the variance of any estimator $\hat{\theta}$ of θ with bias $b(\theta) = E[\hat{\theta}] - \theta$ is bounded from below by

$$\mathrm{Var}(\hat{\theta} - \theta - b) \geq \frac{(1 + \partial b/\partial\theta)^2}{E\big[(\partial \log f(\mathbf{z}|\theta)/\partial\theta)^2\big]}. \tag{10}$$

The denominator is, of course, the Fisher information. The numerator, $(1 + \partial b/\partial\theta)^2$, is seen in all Equations (3)–(9) above. The term $\mathrm{div}\, E[\boldsymbol{A}] - E[\mathrm{div}\, \boldsymbol{A}]$ appearing in the numerator of Equation (3) is, in this context, precisely

$$\mathrm{div}\, E[\boldsymbol{A}] - E[\mathrm{div}\, \boldsymbol{A}] = \partial b/\partial\theta - E[(\partial/\partial\theta)(\hat{\theta} - \theta)] \tag{11}$$

$$= 1 + \partial b/\partial\theta, \tag{12}$$

i.e., the one-dimensional biased CRB, as promised. Likewise, the expression $\mathbf{I} + \nabla\boldsymbol{b}$ from Equation (9) reduces to this simplest biased CRB form as well. In the general Euclidean case with a Gaussian statistical model $\log f(\mathbf{z}|\boldsymbol{\mu}) = -(\mathbf{z} - \boldsymbol{\mu})^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{z} - \boldsymbol{\mu})$ + constants, the mean square error of any unbiased estimator of the mean $E[\mathbf{z}] = \boldsymbol{\mu}$ (with known covariance) is bounded by

$$E\big[(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})^{\mathrm{T}}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})\big] \geq \mathrm{tr}\, \mathbf{R}. \tag{13}$$

Note that this is the mean square error measured using the canonical Euclidean metric $\|\boldsymbol{\mu}\|_2^2 = \sum_i \mu_i^2$, i.e., the 2-norm. As discussed above, García and Oller use the intrinsic

Fisher metric to measure the mean square error, which results in the simpler expression

$$E[(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})^{\mathrm{T}}\mathbf{R}^{-1}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})] \geq \operatorname{tr}\mathbf{I} = n, \tag{14}$$

which is seen in part 4 of Equation (3), as well as in García and Oller's paper.

It may be argued that the relatively simple expressions for the bounds of Equation (3) actually conceal the true complexity found in these equations because these bounds depend on the sectional curvature and diameter of the underlying Riemannian manifold, which, especially for the case of the Fisher information metric, is typically quite involved, even for the simplest examples of Gaussian densities.

The second noteworthy property of the bounds of Equation (3) is their global, versus local, properties. It must be acknowledged that parts of these bounds, based as they are upon the global analysis of comparison theory (specifically, Bishop's comparison theorems I and II (Oller and Corcuera, 1995)), are not truly local bounds. As the square error becomes small (i.e., as either the sample size or signal-to-noise-ratio becomes large) relative to the inverse of the local curvature, the manifold appears to be locally flat (as the earth appears flat to us), and the curvature terms for a truly local Cramér-Rao bound should become increasingly negligible and approach the standard Euclidean case. This local effect is not captured by the global comparison theory, and explains why the curvature terms remain present for errors of all sizes in Equation (3). It also explains why a small change of curvature, i.e., a small change in the assumed statistical model, will result in very different numerical bounds, depending upon which of the five cases from Equation (3) applies, i.e., these bounds are not "tight"—they fall strictly below the asymptotic error of the (asymptotically efficient) maximum likelihood estimator. To see this, consider the 2-dimensional unit disk. It is flat; therefore, the fourth part of the bound applies, i.e., the lower bound on the square error is $2/k$. Now deform the unit disk a little so that it has a small amount of positive curvature, i.e., so that it is a small subset of a much larger 2-sphere. Now the fifth part of Equation (3) applies, and no matter how tiny the positive curvature, the lower bound on the square error is $1/(2k)$. This is a lower bound on the error which is discontinuous as a function of curvature. A tighter estimation bound for a warped unit disk, indeed for the case of general Riemannian manifolds, is possible. Nevertheless, the relative simplicity of these bounds lends themselves to rapid analysis, as well as insight and understanding, of estimation problems whose performance metric is Fisher information.

The tight bounds of Equations (5)–(9) can be achieved using Riemann's original local analysis of curvature (Spivak, 1999), which results in a Taylor series expansion for the Cramér-Rao bound with Riemannian curvature terms appearing in the second-order part. These are truly local Cramér-Rao bounds, in that as the square error becomes small relative to the inverse of the local curvature, the curvature terms become negligible, and the bounds approach the classical, Euclidean Cramér-Rao bound. Typical Euclidean

Cramér-Rao bounds take the form (Van Trees, 1968)

$$\mathrm{Var}(\hat{\theta} - \theta - b) \geq \frac{\mathrm{beamwidth}^2}{\mathrm{SNR}}, \tag{15}$$

where the "beamwidth" is a constant that depends upon the physical parameters of the measurement system, such as aperture and wavelength, and "SNR" denotes the signal-to-noise-ratio of average signal power divided by average noise power, typically the power in the deterministic part of the measurement divided by the power in its random part. The intrinsic Cramér-Rao bounds of (5)–(9) take the form (loosely speaking via dimensional analysis and an approximation of $\boldsymbol{A} = \exp_{\boldsymbol{\theta}}^{-1} \hat{\boldsymbol{\theta}} \approx \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}$ using local coordinates)

$$\mathrm{Cov}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} - \boldsymbol{b}) \gtrsim \frac{\mathrm{beamwidth}^2}{\mathrm{SNR}} \left(1 - \frac{\mathrm{beamwidth}^2 \cdot \mathrm{curvature}}{\mathrm{SNR}}\right) + O(\mathrm{SNR}^{-3}). \tag{16}$$

Note that the curvature term in Equation (16), the second term in a Laurent expansion about infinite SNR, approaches zero faster than the bound itself; therefore, this expression approaches the classical result as the SNR grows large—the manifold becomes flatter and flatter and its curvature becomes negligent. The same is true of the sample size. Also note that, as the curvature is a local phenomenon, these assertions all depend precisely where on the parameter manifold the estimation is being performed, i.e., the results are local ones. In addition, the curvature term decreases the Cramér-Rao bound by some amount where the curvature is positive, as should be expected because geodesics tend to coalesce in these locations, thereby decreasing the square error, and this term increases the bound where there is negative curvature, which is also as expected because geodesics tend to diverge in these areas, thereby increasing the square error. Finally, even though this intrinsic Cramér-Rao bound is relatively involved compared to Equation (3) because it includes local sectional and Riemannian curvature terms, the formulae for these curvatures is relatively simple in the case of the natural invariant metric on reductive homogeneous spaces (Cheeger and Ebin, 1975), arguably a desirable metric for many applications, and may also reduce to simple bounds [Smith, 2005, pp. 1623–24].

These comments have focused on but a portion of the wide range of interesting subjects covered well in García and Oller's review. Another important feature of their article that warrants attention is the discussion at the end of section 4.2 about obtaining estimators that minimize Riemannian risk. It is worthwhile comparing these results to the problem of estimating an unknown covariance matrix, analyzed using intrinsic methods on the space of positive definite matrices $\boldsymbol{Gl}(n, \mathbb{C}) / \boldsymbol{U}(n)$ (Smith, 2005). As noted above, the Fisher information metric and the natural invariant Riemannian metric on this space coincide, hence García and Oller's risk minimizing estimator analysis may be applied directly to the covariance matrix estimation problem as well. Furthermore, the

development of Riemannian risk minimization for arbitrary metrics, not just the Fisher one, appears promising. This body of work points to an important, yet unanswered question in this field: Aside from intrinsic estimation bounds using various distance metrics, does the intrinsic approach yield practical and useful results useful to the community at large? Proven utility will drive greater advances in this exciting field.

References

Amari, S. (1985). Differential-Geometrical Methods in Statistics. *Lecture Notes in Statistics*, 28. Berlin: Springer-Verlag.

Amari, S. (1993). *Methods of Information Geometry*, Translations of Mathematical Monographs, **191** (transl. D. Harada). Providence, RI: American Mathematical Society, 2000. Originally published in Japanese as "Joho kika no hoho" by Iwanami Shoten, Publishers, Tokyo.

Cheeger, J. and Ebin, D. G. (1975). *Comparison Theorems in Riemannian Geometry*. Amsterdam: North-Holland Publishing Company.

Diaconis, P. (1988). Group Representations in Probability and Statistics. *IMS Lecture Notes-Monograph Series*, 11, ed. S. S. Gupta. Hayward, CA: Institute of Mathematical Statistics.

Efron, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency) (with discussion). *Annals of Statistics*, 3, 1189-1242.

Fisher, R. A. (1953). Dispersion on a sphere. *Proceedings of the Royal Society London A*, 217, 295-305.

Hendriks, H. (1991). A Cramér-Rao type lower bound for estimators with values on a manifold. *Journal of Multivariate Analysis*, 38, 245-261.

Oller, J. M. (1991). On an intrinsic bias measure. In *Stability Problems for Stochastic Models*, Lecture Notes in Mathematics 1546, eds. V. V. Kalashnikov and V. M. Zolotarev, Suzdal, Russia, 134–158. Berlin: Springer-Verlag.

Oller, J. M. and Corcuera, J. M. (1995). Intrinsic analysis of statistical estimation. *Annals of Statistics*, 23, 1562–1581.

Rao, C. R. (1945). Information and the accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, 37, 81–89, 1945. See also *Selected Papers of C. R. Rao*, ed. S. Das Gupta, New York: John Wiley, Inc. 1994.

Smith, S. T. (2005). Covariance, subspace, and intrinsic Cramér-Rao bounds. *IEEE Transactions Signal Processing*, 53, 1610–1630.

Smith, S. T., Scharf, L. and McWhorter, L. T. (2006).
Intrinsic quadratic performance bounds on manifolds. In *Proceedings of the 2006 IEEE Conference on Acoustics, Speech, and Signal Processing* (Toulouse, France).

Spivak, M. (1999). *A Comprehensive Introduction to Differential Geometry*, 3d ed., vol. 2. Houston, TX: Publish or Perish, Inc.

Van Trees, H. L. (1968). *Detection, Estimation, and Modulation Theory*, Part 1. New York: John Wiley, Inc.

Rejoinder

It is not frequent to have the opportunity to discuss some fundamental aspects of statistics with mathematicians and statisticians such as Jacob Burbea, Joan del Castillo, Wilfrid S. Kendall and Steven T. Smith, in particular those aspects considered in the present paper. We are really very fortunate in this respect and thus we want to take this opportunity to thank the former Editor of SORT, Carles Cuadras, for making it possible.

Some questions have arisen on the naturalness of the information metric and thus on the adequacy of the intrinsic bias and Riemannian risk in measuring the behaviour of an estimator, see the comments of Professors S.T. Smith and J. del Castillo. Therefore it might be useful to briefly reexamine some reasons to select such metric structure for the parameter space.

The distance concept is widely used in data analysis and applied research to study the dissimilarity between physical objects: it is considered a measure on the information available about their differences. The distance is to be used subsequently to set up the relationship among the objects studied, whether by standard statistical inference or descriptive methods.

In order to define a distance properly, from a methodological point of view, it seems reasonable to pay attention to the formal properties of the underlying observation process rather than taking into account the physical nature of the objects studied. The distance is a property neither of physical objects nor of an observer: it is a property of the observation process.

From these considerations, a necessary first step is to associate to every physical object an adequate mathematical object. Usually suitable mathematical objects for this purpose are probability measures that quantify the propensity to happen of the different events corresponding to the underlying observation process.

In a more general context we may assume that we have some additional information concerning the observation process which allows us to restrict the set of probability measures that represents our possible physical objects. These ideas lead us to consider parametric statistical models to describe the knowledge of all our possible universe of study.

The question is now: which is the most convenient distance between the probability measures corresponding to a statistical model? Observe that this question pressuposes that the right answer depends on the statistical model considered: if we can assume that all the possible probability measures corresponding to a statistical analysis are of a given type (at least approximately) this information should be relevant in order to quantify, in a right scale, the differences between two probability measures of the model. In other

words, we are are interested not only in a distance between two probability mesures but in equipping with a metric structure the set of all possible probability measures which can describe our data. In the parametric case this is equivalent to equipping the parameter space with a metric structure.

Furthermore, in order to answer these questions we have to take into account some logical reasonable restrictions that such a distance must satisfy. The first requeriment should be that any reasonable distance have to be not increasing under *general data transformations*: if we change the data, once the data has been already obtained, the corresponding distance should not increase since it does not add any information on the differences between the objects compared. Here, by general data transformation we mean either any modification of the original algebra or carrying out randomization of the data or simply data transformations.

Moreover, the distance should be invariant under *admissible transformations*, i.e. transformations which induce an *equivalent* statistical problem. These admissible transformations generalize the Fisher and Blackwell sufficiency. Several interesting approaches to this concept may be found in Strasser (1985), Heyer (1982) and Čencov (1982).

It is well known that f-divergences, Csiszar (1967), are global invariant dissimilarity measures that satisfy the above-mentioned property. We may use any of these indexes to quantify how different are two probability measures, although in general they are not proper distances. But if we are interested in defining a metric structure on the parameter space it is important to bear in mind that all this divergences induce the same, up to a multiplicative constant, intrinsic metric in the parameter space: the information metric, which is a Riemannian metric, see Burbea & Rao (1982). Therefore, the parameter space became not only a metric space but a metric length space. Finally notice that other global distance measures, like the Hellinger distance, are useful to confer a metric structure to the parameter space but not a metric length space structure which is indeed a desirable property.

Response to Jacob Burbea

It is a pleasure for the authors that a so well-known expert on the information metric structures has invest time and effort in such a detailed review on the intrinsic estimation problem.

First it is worthwhile to point out that Professor Burbea has observed the *disadvantage of the information distance in not having a closed form*, while other possible intrinsic losses do accept such a form and are thus more readable and attractive in practice. But somehow this is an a-posteriori problem which can be solved by using proper numerical evaluations or well-known tight approximations of the information distance.

But the above-mentioned *canonical* status of the measures with regard to the information distance is even more remarkable: the selection of other intrinsic losses in an estimation problem lead to *a somewhat different approach to intrinsic estimation, obtained by considering affine connections on the manifold that differ from the Levi-Civita connection*, that is considering other connections which are not compatible with the metric.

Finally, and once again, we would like to thank Professor Burbea for the commendable explanation to the insights of this subject.

Response to Joan del Castillo

The authors would like to thank Professor del Castillo for his careful reading and suggesting comments on the paper.

As Prof. del Castillo has pointed out, bias and mean square error are parametrization dependent measures and thus not invariant under admissible transformations of the parameters as well as of the random variables.

The authors have considered a purely intrinsic approach to the estimation problem by setting the loss function to be the square of the Rao distance. It is possible then to obtain the explicit form for the Bayes estimator in the examples considered.

The maximum likelihood estimator is often a convenient estimator but in some situations is an estimator that can be improved, in terms of performance or, even more, appears to be a non-reasonable estimator, see Le Cam (1990). This is the case of the example 4.1 where the authors would like to point out that no *direct considerations of differential geometry* are involved to discard the ML estimator: the reasons are purely of the statistical kind. The ML estimator misbehave in the sense that scores 0 when the sample statistic T does. This is not the case of the Bayes estimator θ^b obtained, as it is shown by the table of page 137.

The situation in Example 4.2 is slightly different from the above. As Professor del Castillo observes, the ML estimator bias-corrected is *the best estimator we can have* as long as the considered measures of performance are the bias and mean square error. The estimator σ^b obtained in Example 4.2 is the equivariant estimator that and uniformly minimizes the Riemannian Risk. Since the acting group, the multiplicative group $\mathbb{R}_+$, is commutative σ^b, is also intrinsically unbiased. One could interpret the obtained estimator as a ML Riemannian risk-corrected but we would then omit the very remarkable properties of σ^b of being equivariant and intrinsically unbiased. Anyhow the term *best estimator* makes sense insofar as the performance of an estimator is fixed.

On the other hand we may observe that the criticism concerning the adequacy of the information metric, because the model is not exactly true, apply to all the methods in parametric statistical inference, in particular to maximum likelihood estimation. We think that the information metric is a reasonable approximation to any convenient distance as far as the parametric model is a reasonable approximation to our knowledge on the studied objects.

The problem is then the following: assuming that a statistical model can not be absolutely accurate, are we concerned on not adding more noise to our results by selecting intrinsic measures of the performance?

Response to Wilfrid S. Kendall

The authors would like to take the opportunity in this rejoinder to thank twice Professor Kendall. On one hand his previous work on Riemannian barycentres and convexity were really important on the final form of several parts of the paper Oller & Corcuera (1995). Secondly, his comments and global vision on the subject connecting different research areas is already an inspiration for our future work.

The bias is a quantitative measure of the systematic error and thus should be measured in the same units as the error, the latter given by the distance. The mean square error is a quantitative measure of the impreciseness of the estimates and should be measured in the square of the distance units. Both measures are deeply related and although other connection could be used to define mean values, we believe that the Levi-Civita connection is the choice which better guarantees the intuitive meaning of both measures. Furthermore this election allows also a rather simple and natural extension to the Cramér-Rao inequality. In our opinion all other connections, as useful to give account of asymptotic estimator properties, should be regarded as other natural geometrical objects defined on the parameter space.

With respect to the interesting work of Darling (2002), we have to note that he defines intrinsic means by introducing other Riemannian metric than the information metric. This is an interesting possibility to explore but only in the case that this distance satisfies the previously-mentioned logical requirements that in our context any reasonable distance must satisfy.

Response to Steven T. Smith

The authors are totally grateful to Professor Smith for the time he has invest in the careful reading of the paper but in the extended review of the state-of-the-art of the

intrinsic estimation theory. The wide range of topics and illuminating examples covered with Professor Smith will surely help the authors in their future research work.

Concerning to the problem of selecting an appropriate distance when there are no known extrinsic reasons that force one or other selection, we cannot add much more to those reasons given before. We agree that there may exist other natural distances or indexes to be considered but again, and in our opinion, they appear of interest as far as they satisfy the previous logical requirements concerning admissible transformations.

Another interesting point involved in that question is to point out which are the basics settings determining the geometry of the parameter space: if we are only concerned to the estimation problem all the relevant information should already be incorporated in the model. The parameter space appears to be of consideration insofar it is regarded as a part of the statistical model and no isolated aspects on it yield of interest.

It has been extremely thought-provoking to the authors to follow the discussion on the tightness of the Cramér-Rao bounds. We agree that the intrinsic bounds in Theorem 3.1, (2)-(3)-(4)-(5) are not tight but if we are interested in obtaining intrinsic bounds we must do so for any estimator and not only on those which concentrate their probabilistic mass around the true parameter. Consequently we need a global analysis of the problem, which has led to the non-tightness. At any case we agree that there are interesting cases where a local analysis of the problem will be very rich. Some improved bounds, continuous on the curvature, could be obtained assuming further restriccions on the diameter of the manifold.

Observe also that the order of the approximations in any local analysis will be, in general, altered when we take expectations and this aspect should be taken into account in any risk approximations. Furthermore, it is necessary to be very careful when we develop approximations of intrinsic quantities based on coordinates point of view especially if we take expectations since the goodness of the approximation is highly dependent on the coordinate system used.

Let us finish this rejoinder hoping, like Prof. Smith, that the challenging question of making all these results closely useful to the scientific community will be attained in the future.

References

Burbea, J. & Rao, C. (1982). Entropy differential metric, distance and divergence measures in probability spaces: a unified approach. *Journal of Multivariate Analysis*, *12*, 575–596.

Čencov, N. (1982). *Statistical Decision Rules and Optimal Inference*. Providence: Trans. Math. Monographs, 53, Amer. Math. Soc. (English translation of the Russian book published in 1972, Nauka, Moscow).

Csiszar, I. (1967). Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungrica*, *2*, 299–318.

Darling, R. (2002). Intrinsic location parameter of a diffusion process. *Electronic Journal of Probability*, *7*, 23 pp. (electronic).

Heyer, H. (1982). *Theory of Statistical Experiments*. New York: Springer-Verlag (English version of Mathematische Theorie statistischer Experiments, 1973, Springer-Verlag, Berlin).

Le Cam, L. (1990). Maximum likelihood: an introduction. *International Statistical Review*, *58*, 153–171.

Oller, J. & Corcuera, J. (1995). Intrinsic analysis of statistical estimation. *Annals of Statistics*, *23*, 1562–1581.

Strasser, H. (1985). *Mathematical Theory of Statistics: Statistical Esxperiments and Asymptotic Decision Theory*. Berlin, New York: Walter de Gruyter.