

What does intrinsic mean in statistical estimation?*

Gloria García¹ and Josep M. Oller²

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 $f : \mathcal{X} \times \Theta \rightarrow \mathbb{R}$ 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 $\gamma_\xi : [0, 1] \rightarrow \Theta$, 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

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

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

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

If we set

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

it is well known that \exp_θ maps \mathfrak{D}_θ diffeomorphically onto D_θ . Moreover, if the manifold is complete the boundary of \mathfrak{D}_θ 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).

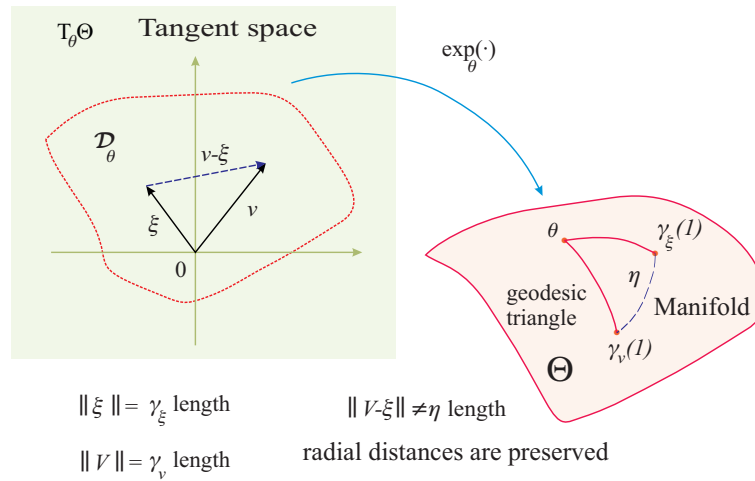


Figure 1: The exponential map

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_θ^{-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

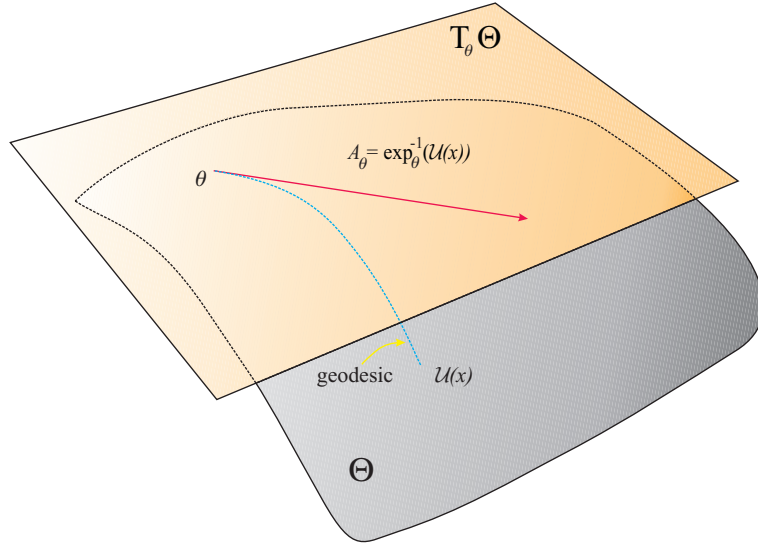


Figure 2: Estimator vector field

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

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

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

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

since $\|A(x)\|_\theta^2 = \rho^2(\mathcal{U}(x), \theta)$. 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

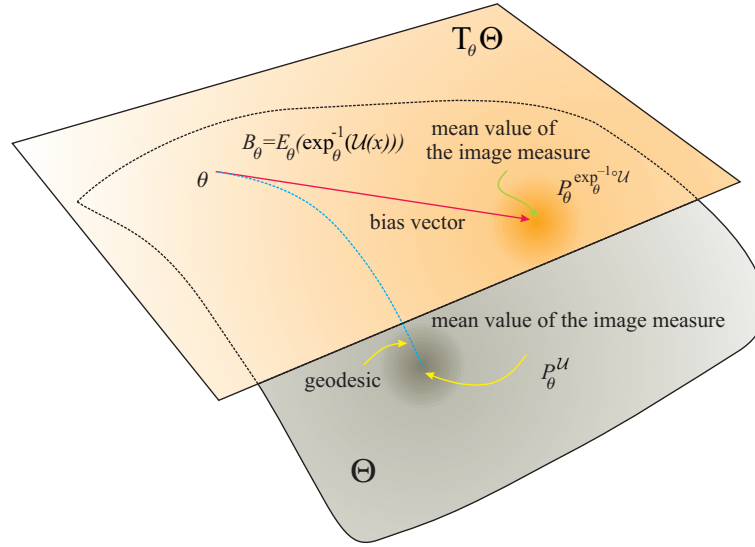


Figure 3: Bias vector field

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

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

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

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 .

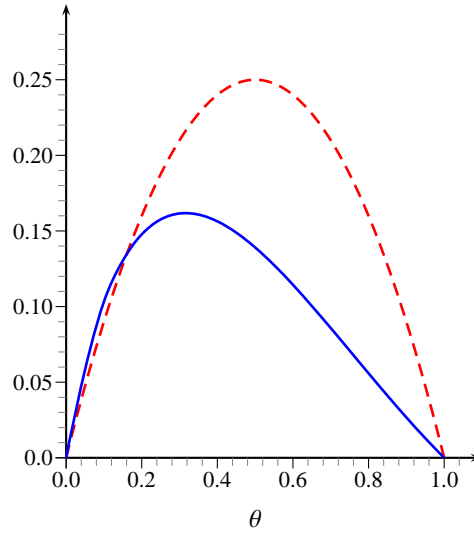


Figure 4: MSE of U (dashed line) and V (solid line).

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

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

which tends to $+\infty$ when θ_1 or θ_2 tend to 0. So $E_\theta(\rho^2(U, \theta)) = +\infty$ meanwhile $E_\theta(\rho^2(V, \theta)) < +\infty$. 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 $\{(\mathcal{X}, \mathbf{a}, \mu); \Theta; f\}$, 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 $E(\rho^2(\mathcal{U}, \theta))$ exists and the covariant derivative of $E(A)$ exists and can be obtained by differentiating under the integral sign. Then,*

1. *We have*

$$E(\rho^2(\mathcal{U}, \theta)) \geq \frac{(\operatorname{div}(B) - E(\operatorname{div}(A)))^2}{kn} + \|B\|^2,$$

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*

$$E(\rho^2(\mathcal{U}, \theta)) \geq \frac{(\operatorname{div}(B) + 1 + (n-1)\sqrt{-\mathcal{K}}\|B\|\coth(\sqrt{-\mathcal{K}}\|B\|))^2}{kn} + \|B\|^2.$$

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*

$$E(\rho^2(\mathcal{U}, \theta)) \geq \frac{(\operatorname{div}(B) + 1 + (n-1)\sqrt{\mathcal{K}}d(\Theta)\cot(\sqrt{\mathcal{K}}d(\Theta)))^2}{kn} + \|B\|^2.$$

In particular, for intrinsically unbiased estimators, we have:

4. *If all sectional Riemannian curvatures are non-positive, then*

$$E(\rho^2(\mathcal{U}, \theta)) \geq \frac{n}{k}$$

5. *If all sectional curvatures are less or equal than a positive constant \mathcal{K} and $d(\Theta) < \pi/2\sqrt{\mathcal{K}}$, then*

$$E(\rho^2(\mathcal{U}, \theta)) \geq \frac{1}{kn}$$

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

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

where v^\top 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:

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

Theorem 3.2. *The dispersion tensor S satisfies*

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

where $\text{Tr}^{i,j}$ and $G^{i,j}$ 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

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

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 $B_\theta^{\mathfrak{M}(\mathcal{U}|\mathcal{D})}$ be the bias tensor, by the Jensen inequality,

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

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, \dots, 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, \dots, 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 $(X_1, \dots, X_k) = (x_1, \dots, x_k)$ put $T = t$. since,

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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 than 3.5% for any result based on sample size $k \leq 100$.

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

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

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

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

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 $[0, 1]$.

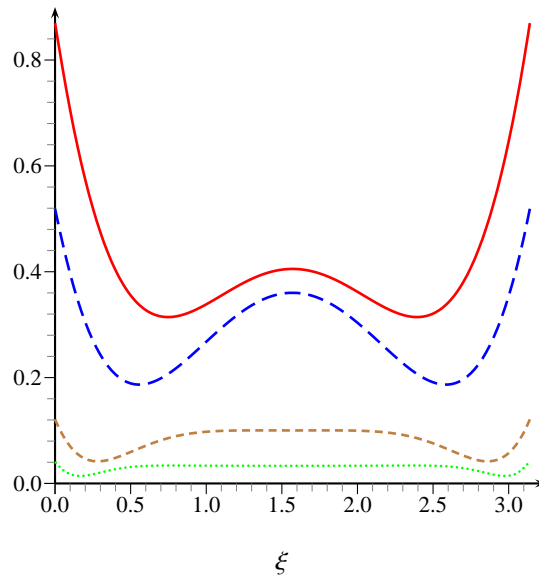


Figure 5: Riemannian risk of ξ^b , for $k = 1$ (solid line), $k = 2$ (long dashed line), $k = 10$ (short dashed line) and $k = 30$ (dotted line).

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 .

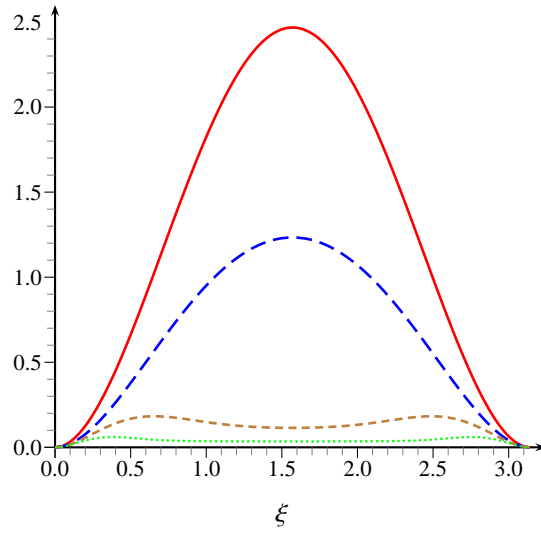


Figure 6: Riemannian risk of ξ^* , for $k = 1$ (solid line), $k = 2$ (long dashed line), $k = 10$ (short dashed line) and $k = 30$ (dotted line).

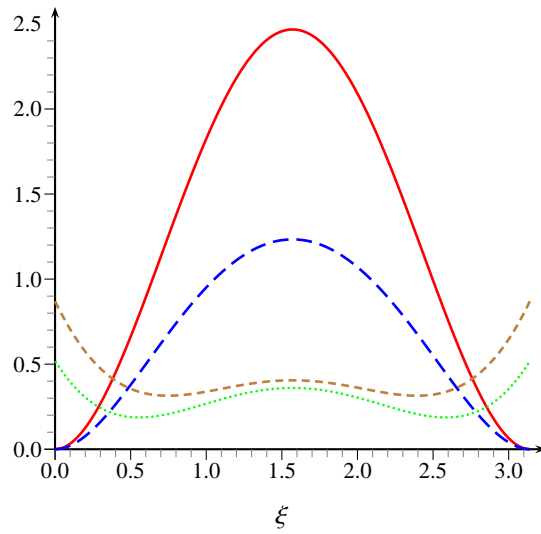


Figure 7: Riemannian risk corresponding to θ^* , for $k = 1$ (solid line), $k = 2$ (long dashed line) and corresponding to θ^b , for $k = 1$ (short dashed line) and $k = 2$ (dotted line).

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

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

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

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

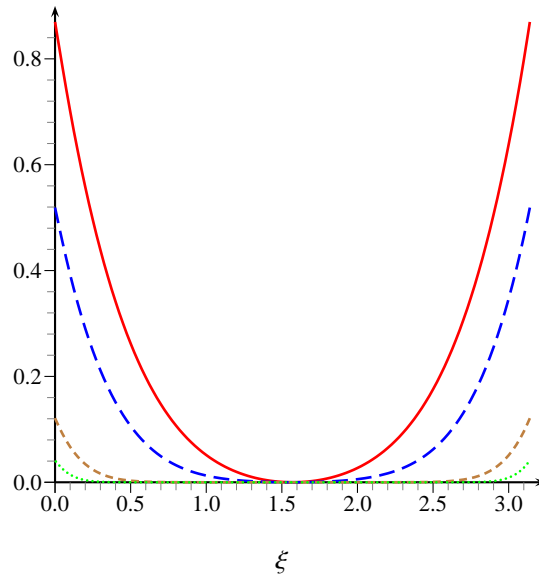


Figure 8: $\|B^b\|^2$ for $k = 1$ (solid line), $k = 2$ (long dashed line), $k = 10$ (short dashed line) and $k = 30$ (dotted line).

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, \dots, 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

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

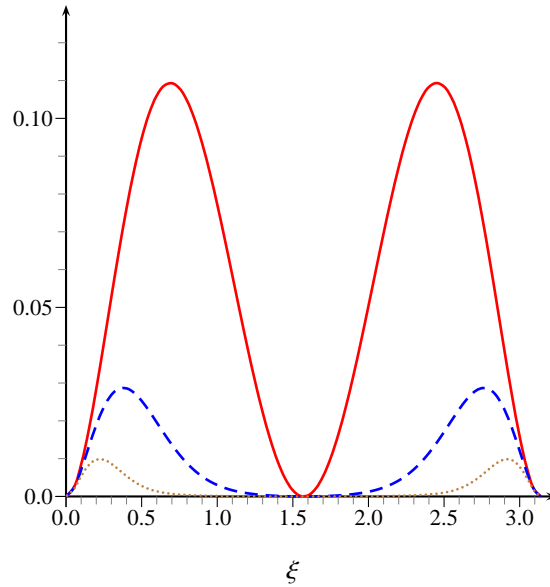


Figure 9: $\|B^*\|^2$ $k = 1, 2$ (solid line) (the same curve), $k = 10$ (dashed line) and $k = 30$ (dotted line).

We shall assume again the Jeffreys prior distribution for σ . Thus the joint density for σ and (X_1, \dots, 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 $(X_1, \dots, X_k) = (x_1, \dots, x_k)$ 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 $\pi(\cdot | s^2)$ based on the Jeffreys prior satisfies

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

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

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

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