Intrinsic Losses Based on Information Geometry and Their Applications

Yao Rong, Mengjiao Tang and Jie Zhou * (ID)

College of Mathematics, Sichuan University, Chengdu 610064, China; rongyao0618@outlook.com (Y.R.);
qingqing_tmj@126.com (M.T.)
* Correspondence: jzhou@scu.edu.cn; Tel.: +86-28-8541-5060

Abstract: One main interest of information geometry is to study the properties of statistical models that do not depend on the coordinate systems or model parametrization; thus, it may serve as an analytic tool for intrinsic inference in statistics. In this paper, under the framework of Riemannian geometry and dual geometry, we revisit two commonly-used intrinsic losses which are respectively given by the squared Rao distance and the symmetrized Kullback–Leibler divergence (or Jeffreys divergence). For an exponential family endowed with the Fisher metric and α-connections, the two loss functions are uniformly described as the energy difference along an α-geodesic path, for some $\alpha \in \{-1, 0, 1\}$. Subsequently, the two intrinsic losses are utilized to develop Bayesian analyses of covariance matrix estimation and range-spread target detection. We provide an intrinsically unbiased covariance estimator, which is verified to be asymptotically efficient in terms of the intrinsic mean square error. The decision rules deduced by the intrinsic Bayesian criterion provide a geometrical justification for the constant false alarm rate detector based on generalized likelihood ratio principle.

1. Introduction

In Bayesian analysis, the choices of a particular loss function and a featured type of priors strongly influence the resulting inference. From the standpoint of purely inference, the final results are typically required to be independent of the way that the model has been parameterized. Thus, in order to produce a Bayes estimator or action which only depends on the model assumed and the data observed, an intrinsic loss and a noninformative (or uninformative) prior are demanded. In an early paper [1], the intrinsic losses are also called the noninformative losses, since they are considered as an analogous determination of the noninformative priors. The reference [1] describes an intrinsic loss as the automated derivation from the sampling distribution without subjective inputs. Whereafter, the entropy and Hellinger losses are suggested there as two appropriate distances between distributions, and their properties are studied especially for exponential families. Some latter works by Barcelona et al. regard the additional requirements as reasonable that the intrinsic losses need be invariant under reduction to sufficient statistics [2]. A symmetric version of Kullback–Leibler (KL) divergence (also named "the intrinsic discrepancy") as such an intrinsic loss is highly recommended, and has been widely applied to develop intrinsic Bayesian analyses of hypothesis testing [3,4], point estimation [4–7], and interval estimation [2,4,7].

The invariance criteria achieved by the intrinsic losses are indeed consistent with one main topic of information geometry, which concentrates on excavating the properties of statistical manifolds that are preserved over coordinate system transformations [8]. Under certain regularity conditions, each parametric statistical model possesses a Riemannian manifold structure given by the Fisher

metric, with the parametrization identical to a coordinate system [9]. The parameters are merely labels for probability distributions, just as the coordinates for geometric objects. Moreover, the manifold structure supplied by the Fisher metric and α-connections are proven to be invariant under transformations of coordinates and reduction by sufficiency [10]. Following this line, a concept is intrinsic if it has a well-defined geometrical meaning. The Riemannian distance in this case—also named the Rao distance [11]—is taken as an intrinsic measure for the dissimilarity between two probability distributions. The reference [12] regards the squared Rao distance between the models labelled by the actual value of the parameter and its estimate as the most natural intrinsic loss, which has been utilized to seek intrinsic Bayes estimators in some concrete univariate examples. In the wider literature, statisticians also apply the Rao distance measure to stimulate the intrinsic study of nonlinear filtering [13], estimation fusion [14,15], and target tracking [16]. Besides, the intrinsic versions of the bias, mean square error (MSE) and Cramér–Rao bound (CRB) are put forward [17,18], which help conduct a systematic analysis for intrinsic estimation theory.

From the above, we can conclude that the Rao distance and the KL divergence are widely accepted as foundation quantities for the intrinsic losses. The former is well-defined on the statistical manifolds with consideration of its usual Riemannian structure, while the later is extremely famous in information theory and is also associated with the canonical divergence in Amari's dual geometries [9,19]. In this paper, we revisit two intrinsic losses that are given by the squared Rao distance and the symmetric KL divergence (also called the Jeffreys divergence), by integrating them in the framework of information geometry. For the exponential families, they are shown to be uniformly described as the energy difference along an α-geodesic path on the statistical manifold, with some $\alpha \in \{-1, 0, 1\}$. Additionally, the Jeffreys prior—as one of the most popular noninformative priors—corresponds to the Riemannian volume in this setting [20]. The elucidation of their geometrical meanings contributes to deepen our understanding and capabilities in intrinsic Bayesian analysis. Subsequently, we apply these two intrinsic losses to develop Bayesian approaches to covariance matrix estimation and range-spread target detection, which are both hot issues in radar signal processing. The sample covariance matrix (the maximum likelihood estimator (MLE) using a set of zero-mean Gaussian samples) is proven to be intrinsically biased [18]. We provide a Bayesian approach to estimate the scale factor of the sample covariance matrix, which leads to an intrinsically unbiased and asymptotically efficient covariance estimator. The detection of range-spread target is, in essence, a decision problem. Intrinsic analysis supplies a novel way to derive a consistent decision rule under different coordinate systems of measurement. The produced detectors are equivalent to a classical result based on generalized likelihood ratio principle, which will benefit an intuitive interpretation about the geometrical nature of the latter.

In the next section, we review some basic facts about information geometry and exponential family. In Section 3, we present the intrinsic Bayesian approaches to point estimation and hypothesis testing. In Section 4, some applications to covariance matrix estimation and range-spread target detection are studied in detail. Section 5 concludes.

2. The Riemannian Geometry and Dual Geometry of Exponential Family

2.1. The Fisher Metric and the α-Connections

Assume a statistical model $S = \{p(x|\xi) : \xi \in \Xi\}$ to describe the probabilistic behavior of a random variable x, with the parameter space Ξ being an open set of $\mathbb{R}^n$. Under certain regularity conditions, S can be regarded as a differentiable manifold, with $\xi = [\xi^1, \ldots, \xi^n]$ being a coordinate system. The manifold S carries the Riemannian structure induced by the Fisher information matrix $G(\xi) = [g_{ij}(\xi)]$,

$$g_{ij}(\xi) = \mathbb{E}_\xi \left[\frac{\partial \log p(x|\xi)}{\partial \xi^i} \frac{\partial \log p(x|\xi)}{\partial \xi^j} \right], \tag{1}$$

where $\mathbb{E}_{\boldsymbol{\xi}}$ signifies the expectation with respect to $p(\boldsymbol{x}|\boldsymbol{\xi})$. In this geometric setting, $\boldsymbol{G}$ plays the key role of a metric tensor [11], called the Fisher metric. Then, the infinitesimal squared distance between two closely-spaced distributions $p(\boldsymbol{x}|\boldsymbol{\xi})$ and $p(\boldsymbol{x}|\boldsymbol{\xi} + \mathrm{d}\boldsymbol{\xi})$ is defined as [21]

$$\mathrm{d}s^2 = \sum_{i,j=1}^{n} g_{ij}(\boldsymbol{\xi}) \, \mathrm{d}\xi^i \, \mathrm{d}\xi^j.$$

For a smooth manifold, the notion of affine connections permits a covariant differential calculus on its tangent bundle. A statistical manifold S possesses a one-parameter family of affine connections related to the Fisher metric, named the α-connections (see [9,22]). For an arbitrary real number α, the α-connection $\nabla^{(\alpha)}$ can be given by the Christoffel symbols of the first kind

$$\Gamma_{ij,k}^{(\alpha)}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\xi}}\left[\left(\frac{\partial^2 \log p(\boldsymbol{x}|\boldsymbol{\xi})}{\partial \xi^i \partial \xi^j} + \frac{1-\alpha}{2}\frac{\partial \log p(\boldsymbol{x}|\boldsymbol{\xi})}{\partial \xi^i}\frac{\partial \log p(\boldsymbol{x}|\boldsymbol{\xi})}{\partial \xi^j}\right)\frac{\partial \log p(\boldsymbol{x}|\boldsymbol{\xi})}{\partial \xi^k}\right], \tag{2}$$

for $i, j, k = 1, \ldots, n$. Specially, the 0-connection $\nabla^{(0)}$ is exactly the Riemannian (or Levi–Civita) connection with respect to the Fisher metric, and the exponential connection $\nabla^{(1)}$ and the mixture connection $\nabla^{(-1)}$ also have good applications in statistical analysis (e.g., [23,24]).

As detailed in [9], the geometric structure given by the Fisher metric and the α-connections obeys the following two invariance principles:

(1)   It is invariant under one-to-one reparameterizations.
(2)   It is invariant under reduction to sufficient statistics.

2.2. Geometric Structure of Exponential Family

An exponential family is a statistical model S consisting of probability densities of the form

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \exp\left\{C(\boldsymbol{x}) + \sum_{i=1}^{n} \theta^i F_i(\boldsymbol{x}) - \psi(\boldsymbol{\theta})\right\}, \tag{3}$$

where $C(\boldsymbol{x})$ and $F_i(\boldsymbol{x}), i = 1, \ldots, n$, are real-valued functions of $\boldsymbol{x}$, $\boldsymbol{\theta} = [\theta^1, \ldots, \theta^n]$ are the so-called natural parameters, and $\psi(\boldsymbol{\theta})$ corresponds to the normalization constant. Besides, the $n + 1$ functions $1, F_1(\boldsymbol{x}), \ldots, F_n(\boldsymbol{x})$ are generally required to be linearly independent [8]. The exponential families include many common statistical models, such as Gaussian, Poisson, Bernoulli, Gamma, and Dirichlet distributions.

For an exponential family S, let us investigate its geometric structure with respect to the Fisher metric g and α-connection $\nabla^{(\alpha)}$. From Equation (3), we can obtain

$$\frac{\partial \log p(\boldsymbol{x}|\boldsymbol{\theta})}{\partial \theta^i} = F_i(\boldsymbol{x}) - \frac{\partial \psi(\boldsymbol{\theta})}{\partial \theta^i} \qquad \text{and} \qquad \frac{\partial^2 \log p(\boldsymbol{x}|\boldsymbol{\theta})}{\partial \theta^i \partial \theta^j} = -\frac{\partial^2 \psi(\boldsymbol{\theta})}{\partial \theta^i \partial \theta^j}. \tag{4}$$

Note that in information geometry, the differentiation and integration of $p(\boldsymbol{x}|\boldsymbol{\theta})$ are customarily assumed to be interchangeable [9]. We can easily verify that an exponential family satisfies this assumption by reference to [25]. Thus from Equations (1) and (4), it is possible to write the components of g as

$$g_{ij}(\boldsymbol{\theta}) = -\mathbb{E}\left[\frac{\partial^2 \log p(\boldsymbol{x}|\boldsymbol{\theta})}{\partial \theta^i \partial \theta^j}\right] = \frac{\partial^2 \psi(\boldsymbol{\theta})}{\partial \theta^i \partial \theta^j}. \tag{5}$$

Further computations by Equation (2) give the expressions for $\nabla^{(\alpha)}$. Especially, we have $\Gamma_{ij,k}^{(1)} = 0$ for $i, j, k = 1, \ldots, n$.

In essence, the differential geometry with respect to the Fisher metric and the 0-connection coincides with the usual Riemannian geometry. For all exponential families, it is difficult to derive a general theory of their Riemannian geometric structures. Individually, however, the properties

of univariate and multivariate Gaussian distributions are well-studied under the framework of Riemannian geometry in [26].

Amari and Nagaoka [9,19] propose the analysis of the geometric structures of statistical models through duality. Typically, an exponential family S has a dually flat structure $(g, \nabla^{(\pm 1)})$. That is to say, S is flat with respect to the dual connections $\nabla^{(1)}$ and $\nabla^{(-1)}$. On the dual flat space $(S, g, \nabla^{(\pm 1)})$, the natural parameters $\boldsymbol{\theta}$ constitute the 1-affine coordinate system, while the (-1)-affine coordinate system can be given by the expectation parameters $\boldsymbol{\eta} = [\eta_1, \ldots, \eta_n] = \mathbb{E}[x]$. Furthermore, we can prove that the two potential functions provided by $\psi(\boldsymbol{\theta})$ and $\phi(\boldsymbol{\eta}) = \boldsymbol{\theta} \cdot \boldsymbol{\eta} - \psi(\boldsymbol{\theta})$ satisfy

$$\eta_i = \frac{\partial \psi(\boldsymbol{\theta})}{\partial \theta^i}, \quad \theta^i = \frac{\partial \phi(\boldsymbol{\eta})}{\partial \eta_i} \tag{6}$$

$$g_{ij} = \frac{\partial^2 \psi(\boldsymbol{\theta})}{\partial \theta^i \partial \theta^j}, \quad g^{ij} = \frac{\partial^2 \phi(\boldsymbol{\eta})}{\partial \eta_i \partial \eta_j},$$

$$D(p, q) = \psi(\boldsymbol{\theta}_q) + \phi(\boldsymbol{\eta}_p) - \boldsymbol{\theta}_q \cdot \boldsymbol{\eta}_p, \tag{7}$$

where g^{ij} is the (i, j)-th entry of the inverse Fisher information matrix $\boldsymbol{G}^{-1}$, $\boldsymbol{\theta}_q$ are the 1-affine coordinates of $q \in S$, $\boldsymbol{\eta}_p$ are the (-1)-affine coordinates of $p \in S$, and $D(p, q)$ denotes the KL divergence from p to q.

2.3. α-Geodesics

In the presence of an affine connection, geodesics are defined to be curves whose tangent vectors remain parallel if they are transported along it. Specifically, an α-geodesic $\gamma(t)$ is characterized by the equation

$$\frac{\nabla^{(\alpha)} \dot{\gamma}}{\mathrm{d}t}(t) = 0, \tag{8}$$

where $\dot{\gamma}$ denotes the derivative with respect to t.

It is well-known in differential geometry that the (locally) shortest curve between two points of the manifold (if it exists) is a 0-geodesic (namely the geodesic in usual Riemannian geometry; e.g., [27]). Usually, a description is used that the 0-geodesic has "constant speed", since its velocity fields have constant Riemannian norm. Under certain boundary conditions, when $\alpha = 0$, an explicit solution to Equation (8) for univariate or multivariate Gaussian distributions has been derived in [26,28,29]. However, it seems complicated to give out a general closed form of the 0-geodesics for all exponential families.

From the viewpoint of flatness, the manifold S of exponential family is dually flat with respect to $\nabla^{(\pm 1)}$. We can know from [19] that any (-1)- or 1-geodesic on S is a straight line with respect to the corresponding affine coordinate system. Thus, in terms of the $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ coordinates, the ± 1-geodesics $\gamma_{p,q}^{(1)}$ and $\gamma_{p,q}^{(-1)}$ connecting any $p, q \in S$ can be expressed by

$$\gamma_{p,q}^{(1)} : \boldsymbol{\theta}(t) = (1 - t)\boldsymbol{\theta}_p + t\boldsymbol{\theta}_q,$$

$$\gamma_{p,q}^{(-1)} : \boldsymbol{\eta}(t) = (1 - t)\boldsymbol{\eta}_p + t\boldsymbol{\eta}_q, \tag{9}$$

with the parameter $t \in [0, 1]$. Note that in Equation (9), the velocity fields $\dot{\boldsymbol{\theta}}, \dot{\boldsymbol{\eta}}$ are constant along t, but $\gamma_{p,q}^{(1)}$ and $\gamma_{p,q}^{(-1)}$ do not have constant speeds.
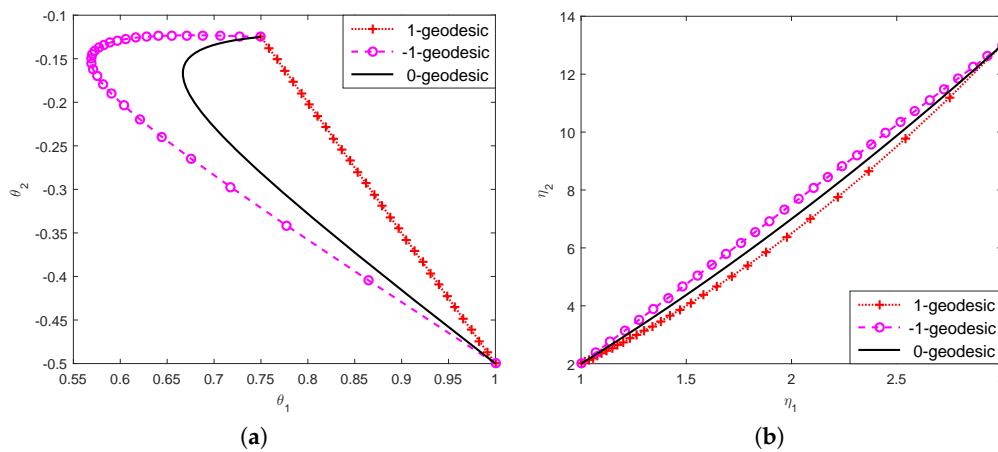
Example 1 (Univariate Gaussian Distribution)**.** *The probability density of a univariate Gaussian distribution with mean μ and variance σ^2 can be expressed as*

$$p(x|\mu, \sigma^2) = \frac{1}{2\pi\sigma} \exp\left\{ -\frac{(x - \mu)^2}{2\sigma^2} \right\} = \exp\left\{ \sum_{i=1}^{2} \theta^i F_i(x) - \psi(\boldsymbol{\theta}) \right\},$$

where $\boldsymbol{\theta} = [\theta^1, \theta^2] = [\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}]$, $F_1(x) = x$, $F_2(x) = x^2$, *and*

$$\psi(\boldsymbol{\theta}) = -\frac{(\theta^1)^2}{4\theta^2} - \frac{1}{2}\log(-\theta^2) + \frac{1}{2}\log \pi.$$

Further, we can compute the expectation parameters $\boldsymbol{\eta} = [\mathbb{E}[x], \mathbb{E}[x^2]] = [\mu, \mu^2 + \sigma^2]$. *Figure 1 illustrates the 0- and* ±1-*geodesics on the manifold of univariate Gaussian distributions.*

Figure 1. The 0- and ±1-geodesics on the manifold of univariate Gaussian distributions with given endpoints $[\mu_1, \sigma_1^2] = [1, 1]$ and $[\mu_2, \sigma_2^2] = [2, 1.44]$, under (**a**) the 1-affine coordinate system $\boldsymbol{\theta}$ and (**b**) the (-1)-affine coordinate system $\boldsymbol{\eta}$.

2.4. The Length and Energy of a Curve

The length of a piecewise smooth curve $\gamma(t) = [\xi^1(t), \ldots, \xi^n(t)], t \in [a, b]$ is defined as

$$L(\gamma) = \int_a^b \sqrt{\sum_{i,j} g_{ij} \dot{\xi}^i \dot{\xi}^j} \, \mathrm{d}t. \tag{10}$$

In some occasions, it may be more convenient to work with another related quantity termed the "energy function", which is given by the integration of the Fisher information along the curve:

$$E(\gamma) = \int_a^b \sum_{i,j} g_{ij} \dot{\xi}^i \dot{\xi}^j \, \mathrm{d}t. \tag{11}$$

Minimizing energy turns out to be equivalent to minimizing length, and both lead to the 0-geodesics on the manifold.

Remark 2. *On a manifold, both the length and energy of a curve* $\gamma(t)$ *do not change under different coordinate systems or model parameterizations, but the latter will depend on the curve parametrization* $t \mapsto \gamma(t)$ *[30].*

Now we discuss the lengths and energies of the 0- and ±1-geodesics on the manifold S of exponential family. If there exists a unique 0-geodesic connecting $p, q \in S$, the length of this 0-geodesic segment equals the well-known Rao distance between p and q [26], written as $\rho(p, q)$. From Remark 2, the energy of a curve is meaningless without specifying its parametrization. What we mainly consider in this paper are the 0- and ±1 geodesics parameterized by $[0, 1]$. Since the 0-geodesics have constant speed, it means that the integrands in Equations (10) and (11) are constant over t. Thus,

given a 0-geodesic $\gamma_{p,q}^{(0)} : [0,1] \to S$ connecting p and q, its length and energy have the following relationship [30]:

$$E\big(\gamma_{p,q}^{(0)}\big) = L^2\big(\gamma_{p,q}^{(0)}\big).$$

The ± 1-geodesics on S have explicit expressions, shown in Equation (9). Furthermore, the theorem given below describes their energy functions.

Theorem 3. *The energies of the ± 1-geodesics $\gamma_{p,q}^{(\pm 1)}$ given in Equation (9) are equal to the Jeffreys divergence $J(p,q)$ between p and q.*

**Proof.** By using Equation (5), the energy of $\gamma_{p,q}^{(1)}$ connecting $p = \boldsymbol{\theta}(0)$ and $q = \boldsymbol{\theta}(1)$ is

$$
\begin{aligned}
E\big(\gamma_{p,q}^{(1)}\big) &= \int_0^1 \sum_{i,j=1}^n g_{ij}\dot{\theta}^i\dot{\theta}^j \,\mathrm{d}t = \int_0^1 \sum_{i,j=1}^n \frac{\partial^2 \psi(\boldsymbol{\theta})}{\partial \theta^i \partial \theta^j}\frac{\mathrm{d}\theta^i}{\mathrm{d}t}\frac{\mathrm{d}\theta^j}{\mathrm{d}t} \,\mathrm{d}t \\
&= \int_0^1 \ddot{\psi}(\boldsymbol{\theta}(t)) \,\mathrm{d}t = \dot{\psi}(\boldsymbol{\theta}(t))\Big|_{t=0}^{t=1} \\
&= \sum_{i=1}^n \frac{\partial \psi(\boldsymbol{\theta}(t))}{\partial \theta^i}\frac{\mathrm{d}\theta^i}{\mathrm{d}t}\Big|_{t=0}^{t=1} \\
&= (\boldsymbol{\eta}_q - \boldsymbol{\eta}_p) \cdot (\boldsymbol{\theta}_q - \boldsymbol{\theta}_p),
\end{aligned}
$$

where the last equality above follows from Equations (6) and (9). On the other side, according to the expression of $D(p,q)$ in Equation (7), we have

$$
\begin{aligned}
J(p,q) &\triangleq D(p,q) + D(q,p) \\
&= \big(\psi(\boldsymbol{\theta}_q) + \phi(\boldsymbol{\eta}_q)\big) + \big(\psi(\boldsymbol{\theta})_p + \phi(\boldsymbol{\eta}_p)\big) - \boldsymbol{\theta}_q \cdot \boldsymbol{\eta}_p - \boldsymbol{\theta}_p \cdot \boldsymbol{\eta}_q \\
&= \boldsymbol{\theta}_q \cdot \boldsymbol{\eta}_q + \boldsymbol{\theta}_p \cdot \boldsymbol{\eta}_p - \boldsymbol{\theta}_q \cdot \boldsymbol{\eta}_p - \boldsymbol{\theta}_p \cdot \boldsymbol{\eta}_q \\
&= (\boldsymbol{\eta}_q - \boldsymbol{\eta}_p) \cdot (\boldsymbol{\theta}_q - \boldsymbol{\theta}_p).
\end{aligned}
$$

Then, $E\big(\gamma_{p,q}^{(1)}\big) = J(p,q)$. Similarly, we can prove the other half of this theorem. □

As an example, on the manifold of univariate Gaussian distributions, Figure 2 shows the energy differences along the α-geodesics parameterized by $[0,1]$, for $\alpha = -1, 0, 1$. We have known that the 0-geodesic path follows the minimum energy variation, which is also demonstrated in Figure 2. Below is a straightforward corollary of Theorem 3.
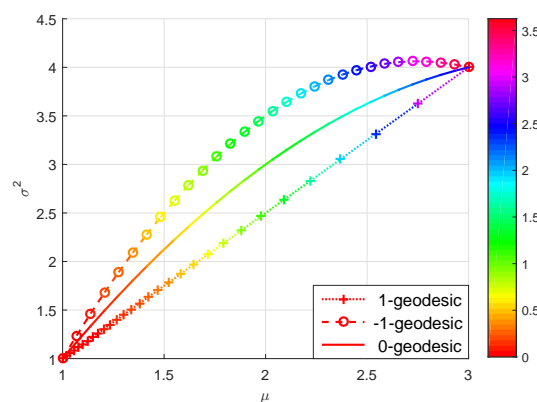
Figure 2. The energy differences along the 0- and ± 1-geodesics parameterized by $[0,1]$ on the manifold of univariate Gaussian distributions, with start point $[\mu_1, \sigma_1^2] = [1,1]$ and end point $[\mu_2, \sigma_2^2] = [3,4]$.

Corollary 4. *If p and q belong to the same exponential family, then $J(p,q) \geq \rho^2(p,q)$.*

3. Intrinsic Bayesian Analysis

3.1. Two Intrinsic Losses

Suppose that the observed data x are generated by $p(x|\xi)$, for some $\xi \in \Xi$. Let ξ_a be the actual (unknown) value of the parameter, and $\xi_g \in \Xi$ be a given value which may represent an estimator or a hypothesis. For the problems of statistical inference, we define a loss function $l(\xi_a, \xi_g)$, aiming at measuring the consequences of estimating ξ_a by ξ_g in statistical estimation [5] or judging the compatibility of ξ_g with the observations x in hypothesis testing [3].

Many conventional loss functions (such as the squared error loss, the zero-one loss) are defined to compare ξ_a and ξ_g in the parameter space, while an intrinsic loss function tends to directly measure the dissimilarity between $p(x|\xi_a)$ and $p(x|\xi_g)$. As customary, the intrinsic losses are required to be invariant under one-to-one transformations of either x or ξ (e.g. [2,5]). Let $S = \{p(x|\xi) : \xi \in \Xi\}$ form an n-dimensional manifold. From the viewpoint of information geometry, the aforementioned criteria are satisfied if a loss function has a well-defined geometrical meaning. Therefore, based on the contents of Section 2.4, we will consider the following two intrinsic losses:

(i)　Intrinsic loss based on the squared Rao distance (hereafter referred to as the Rao loss):

$$l_R(\xi_a, \xi_g) = \rho^2(\xi_a, \xi_g), \tag{12}$$

which stands for the energy difference along a 0-geodesic connecting ξ_a and ξ_g on S and parameterized by $[0, 1]$.

(ii)　Intrinsic loss based on the Jeffreys divergence (hereafter referred to as the Jeffreys loss):

$$l_J(\xi_a, \xi_g) = J(\xi_a, \xi_g), \tag{13}$$

which stands for the energy difference along a (-1)- or 1-geodesic connecting ξ_a and ξ_g on S and parameterized by $[0, 1]$ if S is an exponential family.

Since the Jeffreys divergence is a symmetrized KL divergence, it has all the properties of a metric as defined in topology except the triangle inequality property, and is thus not termed a distance. In recent years, this quantity has sometimes been used for model selection [31,32]. There exists another symmetrized version of KL divergence proposed in [2,5] for intrinsic analysis, which takes a minimum over $\{D(\xi_a, \xi_g), D(\xi_g, \xi_a)\}$ and thus copes better with the case when $p(x|\xi_g)$ and $p(x|\xi_g)$ have nested supports. Even so, we prefer to adopt the Jeffreys loss l_J in this paper since it has a better-understood geometrical interpretation.

3.2. Priors

Apart from the loss function, an appropriate choice of the prior distribution plays an equally strong role for Bayesian analysis. A particular type of priors have been developed to cope with noninformative settings, where the term "noninformative" expresses that a prior distribution in the Bayes theory is expected to have minimal effect on the posterior inference [5]. A common noninformative prior for the Bayesian paradigm is the Jeffreys prior, which is proportional to the Riemannian volume corresponding to the Fisher metric [20]. As we have learned from the geometric theory, this is an intrinsic concept and is thus not dependent on the model parametrizations.

In many early works, the Jeffreys priors are not recommended for multiparameter cases. Bernardo [33] has initiated the reference priors in Bayesian inference (also known as objective Bayesian inference), which is widely accepted for various multiparameter Bayesian problems. However, in most multidimensional occasions, the analytical evaluations related to an inference prior appear to be quite cumbersome. Thus, some numerical algorithms are designed to do the computations in objective Bayesian inference (e.g., [34,35]), which seems unfavorable for the theoretical derivations of this paper. Additionally, in Section 4, the applications of Jeffreys priors yield an appropriate scale

of the eigenvalues of sample covariance matrix, and also reproduce the classic constant false alarm ratio (CFAR) detector in radar detection theory. Hence, we believe that the Jeffreys prior may give a different consequence in intrinsic Bayesian analysis.

3.3. Intrinsic Bayesian Analysis

Let $p(\boldsymbol{\xi}|\boldsymbol{x}) \propto p(\boldsymbol{x}|\boldsymbol{\xi})p(\boldsymbol{\xi})$ be the posterior distribution with respect to the prior $p(\boldsymbol{\xi})$. From a Bayesian viewpoint, the corresponding posterior expected loss is

$$r(\boldsymbol{\xi}_g, \boldsymbol{x}) = \int_{\Xi} l(\boldsymbol{\xi}, \boldsymbol{\xi}_g)p(\boldsymbol{\xi}|\boldsymbol{x})\,\mathrm{d}\boldsymbol{\xi}.$$

Considering the invariance criteria, we take the loss function as l_R or l_J and use the Jeffreys prior π_J. This paper will discuss two aspects of intrinsic Bayesian inference: point estimation and hypothesis testing.

As formulated in [3], a hypothesis testing problem is to decide how the observed data $\boldsymbol{x}$ are compatible with the null hypothesis. Specifically, when the null hypothesis contains only one value $\boldsymbol{\xi}_0$, an intrinsic Bayesian approach to this hypothesis test is based on the positive statistic $r(\boldsymbol{\xi}_0, \boldsymbol{x})$, which suggests to reject the null hypothesis if $r(\boldsymbol{\xi}_0, \boldsymbol{x}) > \gamma$ with some threshold γ; for the composed null hypothesis case, it is suggested to take $\min_{\boldsymbol{\xi}_0 \in \Xi_0} r(\boldsymbol{\xi}_0, \boldsymbol{x})$ as the test statistic, where Ξ_0 denotes the null parameter space.

When we come to deal with a point estimation problem, the Bayes estimator of $\boldsymbol{\xi}$ is obtained by minimizing the posterior expected loss function:

$$\hat{\boldsymbol{\xi}}(\boldsymbol{x}) = \underset{\boldsymbol{\xi} \in \Xi}{\arg\min}\, r(\boldsymbol{\xi}, \boldsymbol{x}),$$

from which the obtained estimator $\hat{\boldsymbol{\xi}}(\boldsymbol{x})$ is invariant under invertible transformations of either $\boldsymbol{x}$ or $\boldsymbol{\xi}$. Besides, the property of $\hat{\boldsymbol{\xi}}(\boldsymbol{x})$ can be further analyzed using the intrinsic estimation theory.

Before finishing this subsection, it is convenient to introduce some basic concepts of intrinsic estimation theory. Fix a $\boldsymbol{\xi} \in \Xi$ and let $T_{\boldsymbol{\xi}}S$ be the tangent space at $\boldsymbol{\xi}$. Associated with a considered connection ∇, the corresponding geodesic curve with starting point $\boldsymbol{\xi}$ and initial direction $v \in T_{\boldsymbol{\xi}}S$ is denoted by $\gamma_v(t), t \in [0, 1]$. Such a curve exists when v belongs to a small neighborhood of the origin at $T_{\boldsymbol{\xi}}S$. In this case, the exponential map is defined as $\mathrm{Exp}_{\boldsymbol{\xi}}(v) = \gamma_v(1)$. Given an estimator $\hat{\boldsymbol{\xi}}(\boldsymbol{x})$, the estimator vector field is induced on S through the inverse of exponential map: $\mathcal{A}_{\boldsymbol{\xi}}(\boldsymbol{x}) = \mathrm{Exp}_{\boldsymbol{\xi}}^{-1}(\hat{\boldsymbol{\xi}}(\boldsymbol{x}))$. Further, the bias vector field is defined as

$$\mathcal{B}_{\boldsymbol{\xi}} = \mathbb{E}_{\boldsymbol{\xi}}[\mathcal{A}_{\boldsymbol{\xi}}].$$

If $\|\mathcal{B}_{\boldsymbol{\xi}}\| = 0$ for any $\boldsymbol{\xi} \in \Xi$, the estimator $\hat{\boldsymbol{\xi}}(\boldsymbol{x})$ is called intrinsically unbiased. Obviously, these definitions above are dependent on the specific choice of ∇ [18]. When a flat connection is considered, an estimator will be intrinsically unbiased if and only if it is unbiased under the corresponding affine coordinate system. Generally speaking, however, the notion of intrinsic unbiasedness is widely acknowledged only in the Riemannian case (namely, when ∇ is the Riemannian connection). Thus, in this paper, when we say that an estimator is intrinsically unbiased, we mean that this is true with regard to the 0-connection. The intrinsic MSE is defined by the mean square of the Rao distance

$$\mathrm{MSE}(\hat{\boldsymbol{\xi}}(\boldsymbol{x})) = \mathbb{E}[\rho^2(\boldsymbol{\xi}, \hat{\boldsymbol{\xi}}(\boldsymbol{x}))], \tag{14}$$

where, in the Bayesian sense, $\mathbb{E}$ denotes the expectation taken over both $\boldsymbol{x}$ and $\boldsymbol{\xi}$. An intrinsic version of the CRB gives a lower bound on the intrinsic MSE performance of any intrinsically unbiased or biased estimator. The relevant developments can be found in [17,18,36].

4. Applications

4.1. Covariance Estimation

Let $x_1, \ldots, x_n$ be random samples of size n from a zero-mean p-variate Gaussian distribution with unknown covariance matrix Σ. Consider the Jeffreys prior distribution for Σ ([37], p. 426), $\pi_J(\Sigma) = |\Sigma|^{-\frac{p+1}{2}}$. A sufficient statistic is provided by

$$S = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T,$$

which is recognized as the sample covariance matrix of the data set [38]. Assume that $n > p$; then, S is positive definite with probability one. Thus, the corresponding posterior distribution based on the Jeffreys prior is easily found to be

$$\pi(\Sigma|S) \propto p(S|\Sigma)\pi_J(\Sigma) \propto |\Sigma|^{-\frac{n+p+1}{2}} \exp\left\{ -\frac{n}{2} \operatorname{tr}(\Sigma^{-1}S) \right\},$$

which follows an inverse Wishart distribution with n degrees of freedom and scale matrix nS.

Now we restrict to consider a family of covariance estimators having the form $\Sigma_g = aS, a \in (0, +\infty)$, which contain the MLE S. The Rao distance between Σ_1 and Σ_2 is [39]

$$\rho(\Sigma_1, \Sigma_2) = \frac{1}{\sqrt{2}} \sqrt{\operatorname{tr}\left(\log^2(\Sigma_1^{-\frac{1}{2}} \Sigma_2 \Sigma_1^{-\frac{1}{2}}) \right)}, \tag{15}$$

while the Jeffreys divergence is [40]

$$J(\Sigma_1, \Sigma_2) = \frac{1}{2}\left(\operatorname{tr}(\Sigma_1^{-1}\Sigma_2) + \operatorname{tr}(\Sigma_2^{-1}\Sigma_1) - 2p \right).$$

Taking the intrinsic losses into account, we can obtain an intrinsic estimator by minimization of the posterior expected Rao loss function:

$$\hat{\Sigma}_R = \operatorname*{arg\,min}_{\Sigma_g \in \{aS|a>0\}} \int l_R(\Sigma, \Sigma_g)\pi(\Sigma|S)\,\mathrm{d}\Sigma = \operatorname*{arg\,min}_{a \in (0,+\infty)} \mathbb{E}_{\Sigma|S}\left[\rho^2(aS, \Sigma) \right], \tag{16}$$

where $\mathbb{E}_{\Sigma|S}$ signifies the expectation with respect to $\pi(\Sigma|S)$. By Equation (15), we have

$$\rho^2(aS, \Sigma) = \frac{1}{2} \operatorname{tr}\left(\log^2(a\Sigma^{-\frac{1}{2}} S \Sigma^{-\frac{1}{2}}) \right) = \frac{1}{2} \sum_{i=1}^{p} \log^2(a\lambda_i),$$

where $\lambda_i, i = 1, \ldots, p$ denote the p eigenvalues of the matrix $\Sigma^{-1}S$. In fact, it proves quite difficult to directly solve the conditional expectation in Equation (16) as a closed form. However, if we let $c = \log a$, it is easy to show that the objective function in Equation (16) is a strictly convex function with respect to c. Thus, we can resort to the Lagrange equation to seek the minimum point. Write the posterior expected Rao loss function as

$$r_R(c) = \mathbb{E}_{\Sigma|S}\left[\rho^2(aS, \Sigma) \right] = \mathbb{E}_{\Sigma|S}\left[\frac{1}{2} \sum_{i=1}^{p} (c + \log \lambda_i)^2 \right].$$

Then, let

$$\frac{\mathrm{d}r_R}{\mathrm{d}c} = \frac{\mathrm{d}}{\mathrm{d}c} \mathbb{E}_{\Sigma|S}\left[\frac{1}{2} \sum_{i=1}^{p} (c + \log \lambda_i)^2 \right] = 0.$$

By Leibniz's rule [41], we can interchange the integral and differential operators, obtaining

$$
\frac{\mathrm{d} r_R}{\mathrm{d} c} = \mathbb{E}_{\Sigma|S} \left[\frac{1}{2} \frac{\mathrm{d}}{\mathrm{d} c} \sum_{i=1}^{p} (c + \log \lambda_i)^2 \right] = \mathbb{E}_{\Sigma|S} \left[\sum_{i=1}^{p} (c + \log \lambda_i)) \right]
$$

$$
= pc + \mathbb{E}_{\Sigma|S} \left[\log |\Sigma^{-1} S| \right] = 0. \tag{17}
$$

Since $\Sigma|S \sim$ Inverse Wishart(n, nS), $S^{\frac{1}{2}} \Sigma^{-1} S^{\frac{1}{2}} |S \sim$ Wishart$(n, n^{-1} I_p)$. Combining with the fact that $|S^{\frac{1}{2}} \Sigma^{-1} S^{\frac{1}{2}}| = |\Sigma^{-1} S|$, we can calculate the conditional expectation in Equation (17) by the law of log-determinant of a Wishart matrix [42]:

$$
\mathbb{E}_{\Sigma|S} \left[\log |\Sigma^{-1} S| \right] = \sum_{i=1}^{p} \psi\left(\frac{n+1-i}{2} \right) + p \log 2 - p \log n,
$$

where $\psi(\cdot)$ is the well-known digamma function defined as $\psi(x) = \frac{\mathrm{d}}{\mathrm{d} x} \log \Gamma(x)$. Thus, from Equation (17), the intrinsic covariance estimator based on the Rao loss is $\hat{\Sigma}_R = \hat{a}_R S$, with

$$
\hat{a}_R = \exp(\hat{c}) = \frac{n}{2} \exp \left\{ -\frac{1}{p} \sum_{i=1}^{p} \psi\left(\frac{n+1-i}{2} \right) \right\}. \tag{18}
$$

Alternatively, if we consider the Jeffreys loss function l_J, then

$$
l_J(aS, \Sigma) = \frac{1}{2} \left(a \operatorname{tr}(\Sigma^{-1} S) + a^{-1} \operatorname{tr}(\Sigma S^{-1}) - 2p \right).
$$

It is easy to calculate that $\mathbb{E}_{\Sigma|S} \left[\operatorname{tr}(\Sigma^{-1} S) \right] = p$, and when $n > p + 1$,

$$
\mathbb{E}_{\Sigma|S} \left[\operatorname{tr}(\Sigma S^{-1}) \right] = \frac{np}{n - p - 1}.
$$

Therefore, the intrinsic covariance estimator based on the Jeffreys loss is $\hat{\Sigma}_J = \hat{a}_J S$, with

$$
\hat{a}_J = \sqrt{\frac{n}{n - p - 1}}.
$$

In Table 1, the scale factors $\hat{a}_R$ and $\hat{a}_J$ are evaluated for various values of n and $p = 10$.

Table 1. The scale factors of sample covariance matrix for some values of n and $p = 10$.

n	$\hat{a}_R$	$\hat{a}_J$	n	$\hat{a}_R$	$\hat{a}_J$
12	2.06698	3.46410	50	1.12585	1.13228
13	1.87853	2.54951	100	1.05867	1.06000
14	1.74997	2.16025	200	1.01114	1.02869
16	1.58326	1.78885	500	1.04598	1.01118
18	1.47864	1.60357	10^3	1.00553	1.00555
20	1.40640	1.49071	10^4	1.00055	1.00055

We now come to examine the bias and efficiency of these two intrinsic estimators.

Proposition 5. *Let* $W \sim$ *Wishart*(v, I_p) *with* $v \geq p$, *then* $\mathbb{E}[\log^k(W)] = m I_p$, *where m is a constant decided by p, k, and v.*

Proof. For a $p \times p$ unitary matrix $\boldsymbol{U}$, the random matrix $\boldsymbol{U W U}^T$ follows the same distribution with $\boldsymbol{W}$. Thus,

$$\mathbb{E}\big[\log^k(\boldsymbol{W})\big] = \mathbb{E}\big[\log^k(\boldsymbol{U W U}^T)\big].$$

Recall the important fact that $\log(\boldsymbol{A}^{-1}\boldsymbol{B A}) = \boldsymbol{A}^{-1}\log(\boldsymbol{B})\boldsymbol{A}$, where $\boldsymbol{A}$ and $\boldsymbol{B}$ are two nonsingular matrices. Thus, a further step yields

$$\mathbb{E}\big[\log^k(\boldsymbol{W})\big] = \boldsymbol{U}\mathbb{E}\big[\log^k(\boldsymbol{W})\big]\boldsymbol{U}^T.$$

Note that the equality above holds for an arbitrary unitary matrix $\boldsymbol{U}$; then, we can conclude that $\mathbb{E}\big[\log^k(\boldsymbol{W})\big]$ must be a scalar multiple of the identity matrix. □

Theorem 6. *The covariance estimator $\hat{\boldsymbol{\Sigma}}_R$ given by Equation* (18) *is intrinsically unbiased.*

Proof. For a p-dimensional positive-definite matrix $\boldsymbol{X}$, the inverse of the exponential map defined on the manifold of positive definite matrices is given by [43]

$$\exp_{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{X}) = \boldsymbol{\Sigma}^{\frac{1}{2}}\log\big(\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{X}\boldsymbol{\Sigma}^{-\frac{1}{2}}\big)\boldsymbol{\Sigma}^{\frac{1}{2}}.$$

Thus, we have

$$\begin{aligned}\mathbb{E}\big[\exp_{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\Sigma}}_R)\big] &= \mathbb{E}\big[\boldsymbol{\Sigma}^{\frac{1}{2}}\log\big(\hat{a}_R\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{S}\boldsymbol{\Sigma}^{-\frac{1}{2}}\big)\boldsymbol{\Sigma}^{\frac{1}{2}}\big]\\ &= \boldsymbol{\Sigma}^{\frac{1}{2}}\Big(\log[\hat{a}_R\boldsymbol{I}_p] + \mathbb{E}\big[\log(\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{S}\boldsymbol{\Sigma}^{-\frac{1}{2}})\big]\Big)\boldsymbol{\Sigma}^{\frac{1}{2}}.\end{aligned} \qquad (19)$$

Generally, it is difficult to directly solve the expectation of the logarithm of a Wishart matrix. However, since $n\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{S}\boldsymbol{\Sigma}^{-\frac{1}{2}} \sim \text{Wishart}(n, \boldsymbol{I}_p)$, by Proposition 5, the expectation in Equation (19) has this form

$$\mathbb{E}\big[\log(n\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{S}\boldsymbol{\Sigma}^{-\frac{1}{2}})\big] = m\boldsymbol{I}_p \qquad (20)$$

with a certain constant m. In addition, since

$$\text{tr}\left(\mathbb{E}\Big[\log[n\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{S}\boldsymbol{\Sigma}^{-\frac{1}{2}}]\Big]\right) = \mathbb{E}\Big[\log|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{S}\boldsymbol{\Sigma}^{-\frac{1}{2}}|\Big] = \sum_{i=1}^{p}\psi\left(\frac{n+1-i}{2}\right) + p\log 2,$$

we can obtain

$$m = \frac{1}{p}\sum_{i=1}^{p}\psi\left(\frac{n+1-i}{2}\right) + \log 2. \qquad (21)$$

Substituting Equations (18), (20) and (21) into Equation (19), we have $\mathbb{E}_{\boldsymbol{\Sigma}}\big[\exp_{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\Sigma}}_R)\big] = \boldsymbol{0}$, which indicates that the covariance estimator $\hat{\boldsymbol{\Sigma}}_R$ is intrinsically unbiased. □

Given a real number $a > 0$, the posterior expected Rao loss of $a\boldsymbol{S}$ can be computed by

$$\mathbb{E}_{\boldsymbol{\Sigma}|\boldsymbol{S}}\big[\rho^2(a\boldsymbol{S}, \boldsymbol{\Sigma})\big] = \mathbb{E}_{\boldsymbol{\Sigma}|\boldsymbol{S}}\left[\frac{1}{2}\text{tr}\left(\log^2(a\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{S}\boldsymbol{\Sigma}^{-\frac{1}{2}})\right)\right] = \frac{1}{2}\text{tr}\left(\mathbb{E}_{\boldsymbol{\Sigma}|\boldsymbol{S}}\left[\frac{1}{2}\log^2(a\boldsymbol{S}^{\frac{1}{2}}\boldsymbol{\Sigma}^{-1}\boldsymbol{S}^{\frac{1}{2}})\right]\right).$$

Using Proposition 5 again, we know that the right-hand side of the second equality above is constant on the sample space. Thus, the posterior expected Rao losses of $\hat{\boldsymbol{\Sigma}}_R$, $\hat{\boldsymbol{\Sigma}}_J$, and $\boldsymbol{S}$ are equal to their Bayes risks under Rao loss that average the Rao loss functions over both sample space and parameter space. Similar consequences can be deduced for the posterior expected Jeffreys losses of $\hat{\boldsymbol{\Sigma}}_R$, $\hat{\boldsymbol{\Sigma}}_J$, $\boldsymbol{S}$, and their Bayes risks under Jeffreys loss. It is still difficult to seek an explicit expression for the posterior expected Rao loss. Hence, a Monte Carlo simulation is used with 1000 trials to calculate the Bayes risks under Rao loss. The comparisons are illustrated in Figure 3.
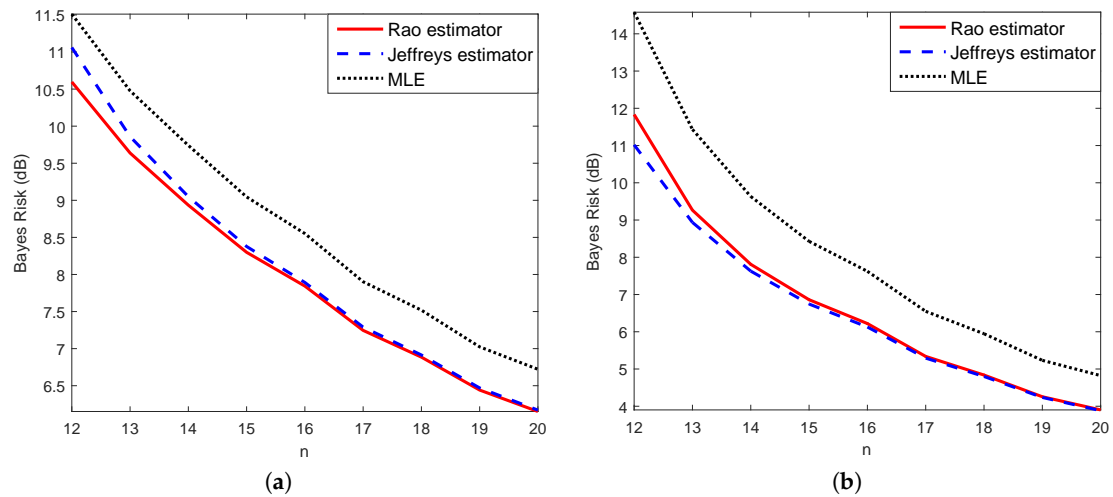
Figure 3. Comparisons of Bayes risks among $\hat{\boldsymbol{\Sigma}}_R$ (solid line), $\hat{\boldsymbol{\Sigma}}_J$ (dashed line), and S (dotted line) for $p = 10$ under (**a**) Rao loss and (**b**) Jeffreys loss.

In Figure 4, we compare the intrinsic MSEs (see Equation (14)) of the considered estimators with an intrinsic version of the CRB given in ([18], Theorem 4), which serves as a lower bound for any intrinsically unbiased covariance estimator. In fact, for our case, the intrinsic MSEs of $\hat{\boldsymbol{\Sigma}}_R, \hat{\boldsymbol{\Sigma}}_J$ and S are equal to their Bayes risks under Rao loss. Recalling that the MLE S is not intrinsically unbiased but always asymptotically efficient [18], we can conclude that the Bayes estimator $\hat{\boldsymbol{\Sigma}}_R$ under Rao loss is intrinsically unbiased and also asymptotically efficient. This property is also demonstrated in Figure 4.
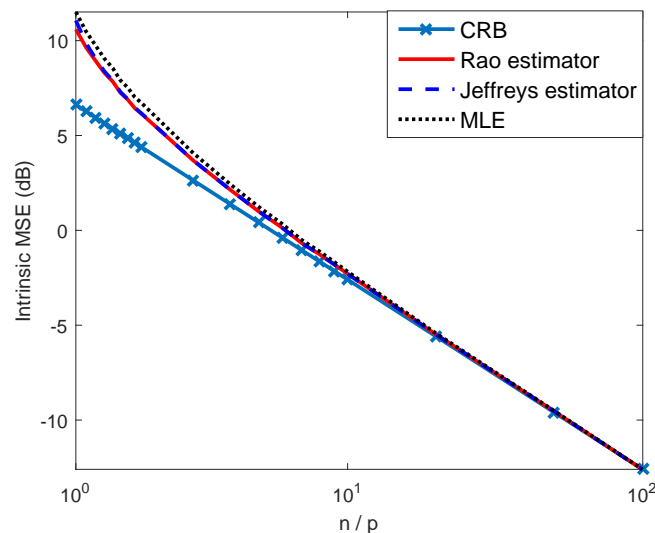
Figure 4. The intrinsic MSEs of $\hat{\boldsymbol{\Sigma}}_R$ (solid line), $\hat{\boldsymbol{\Sigma}}_J$ (dashed line), and S are compared with an intrinsic version of Cramér–Rao bound (CRB) (marked line) for $p = 10$.

4.2. Range-Spread Target Detection

The target detection in the background of Gaussian noise is a basic problem in radar signal processing. Suppose that an unknown target is spatially distributed across H range cells. Data collected from these H cells—called the primary data—consist of the possible target echoes plus Gaussian noises. As in [44], we assume that a set of secondary data coming from K cells arranged around the target are available, which serve as training samples for noise covariance estimation. Here the homogeneous noise environment is considered, which means that the secondary data are free

of signal components and share the same distribution as the noise-only part of the primary data. Therefore, the considered detection problem can be formulated as [44]

$$
\begin{aligned}
H_0 &: z_i = w_i, & i = 1, \ldots, H+K, \\
H_1 &: \begin{cases} z_i = p_i + w_i, & i = 1, \ldots, H, \\ z_i = w_i, & i = H+1, \ldots, H+K, \end{cases}
\end{aligned}
\tag{22}
$$

where $p_i, i = 1, \ldots, H$ denote the N-dimensional unknown signal vectors, and $w_i, i = 1, \ldots, H+K$, are independent and identically distributed Gaussian noises with mean zero and unknown covariance matrix M. For the brevity of notation, let us introduce the matrix forms

$$
Z_p = [z_1, \ldots, z_H], \ Z_s = [z_{H+1}, \ldots, z_{H+K}], \ P = [p_1, \ldots, p_H].
$$

Denote $Z = [Z_p, Z_s]$ and $\tilde{P} = [P, 0]$, then the matrix Z of measurements is Gaussian distributed with mean $\tilde{P}$ and covariance $I_{H+K} \otimes M$. In the absence of target, we have $P = 0$ under the null hypothesis. Therefore, the Rao distance between the assumed model and the null model is [45]

$$
\rho\left([\tilde{P}, I_{H+K} \otimes M], [0, I_{H+K} \otimes M]\right) = \sqrt{\sum_{i=1}^{H} p_i^T M^{-1} p_i}.
\tag{23}
$$

We shall adopt again the Jeffreys prior distribution for parameters (P, M) [35]:

$$
\pi_J(P, M) \propto |M|^{-\frac{N+H+1}{2}}.
$$

The joint density for Z is proportional to

$$
|M|^{-\frac{H+K}{2}} \exp\left\{ -\frac{1}{2} \operatorname{tr}\left(M^{-1}(S_1 + S_2) \right) \right\},
$$

where $S_1 = (Z_p - P)(Z_p - P)^T$, $S_2 = Z_s Z_s^T$. Thus, the corresponding posterior distribution satisfies

$$
\begin{aligned}
\pi(P, M|Z) &\propto \pi_J(P, M) p(Z|P, M) \propto |M|^{-\frac{N+H+1}{2}} \cdot |M|^{-\frac{H+K}{2}} \exp\left\{ -\frac{1}{2} \operatorname{tr}\left(M^{-1}(S_1 + S_2) \right) \right\} \\
&= |M|^{-\frac{H}{2}} \exp\left\{ -\frac{1}{2} \operatorname{tr}(M^{-1} S_1) \right\} \cdot |M|^{-\frac{H+K+N+1}{2}} \exp\left\{ -\frac{1}{2} \operatorname{tr}(M^{-1} S_2) \right\}.
\end{aligned}
$$

Since $\pi(P, M|Z) = \pi(P|M, Z)\pi(M|Z)$, we can conclude that

$$
\pi(P|M, Z) = \text{Normal}(Z_p, I_H \otimes M), \ \pi(M|Z) = \text{Inverse Wishart}(H+K, S_2).
\tag{24}
$$

Therefore, the posterior expected Rao loss of determining the null hypothesis is derived as follows:

$$
\begin{aligned}
&\mathbb{E}_{P,M|Z}\left[l_R\left([\tilde{P}, I_{H+K} \otimes M], [0, I_{H+K} \otimes M]\right) \right] \\
&= \mathbb{E}_{M|Z}\left[\mathbb{E}_{P|M,Z}\left[\sum_{i=1}^{H} p_i^T M^{-1} p_i \right] \right] = \mathbb{E}_{M|Z}\left[\sum_{i=1}^{H} z_i^T M^{-1} z_i + N \right] \\
&= N + (H+K) \sum_{i=1}^{H} z_i^T S_2^{-1} z_i.
\end{aligned}
$$

Hence, the decision rule for the compatibility of the null hypothesis with the observed data Z can be given as

$$
\text{Reject } H_0 \text{ if and only if } \sum_{i=1}^{H} z_i^T S_2^{-1} z_i > \gamma,
$$

where the threshold γ is chosen by the required false alarm probability. In fact, this result reproduces the two-step generalized likelihood ratio test [46] or Wald test [47], which is verified to have CFAR property [46].

Besides, since in this case,

$$l_J\left([\widetilde{P}, I_{H+K} \otimes M], [0, I_{H+K} \otimes M]\right) = l_R\left([\widetilde{P}, I_{H+K} \otimes M], [0, I_{H+K} \otimes M]\right),$$

we can derive an equivalent decision rule using the Jeffreys loss.

Finally, as a conclusion of this section, we will conduct a discussion about the relationship between the geometric invariance principle and the CFAR property. If we transform the random variables x to y without altering the structures of statistical model and testing problem, it is equivalent to derive a transformation on the parameter space Ξ. Let a decision statistic be an intrinsic concept in the geometrical meaning. Thus, this statistic is invariant by changing x to y, which indicates that it maintains a CFAR behavior over an orbit of the parameter space.

5. Conclusions

In this paper, the Rao loss and Jeffreys loss—which are, respectively, based on the squared Rao distance and the Jeffreys divergence—are uniformly considered in the framework of information geometry. We have elucidated their geometrical meaning based on some results of Riemannian geometry and dualistic geometry. In particular, on the manifold of an exponential family, the Rao loss and Jeffreys loss are essentially the energy differences along α-geodesic paths parameterized by $[0, 1]$, for $\alpha = -1, 0, 1$. Based on this, they certainly enjoy the invariance properties under the one-to-one transformations of random variables and model parameters. Subsequently, these two intrinsic loss functions are unitized to develop intrinsic Bayesian analysis of covariance estimation and range-spread target detection. We use the intrinsic losses to derive the scale factor of sample covariance matrix, which leads to an intrinsically unbiased and asymptotically efficient covariance estimator. On the other hand, the detectors provided by the posterior expected intrinsic losses have been proven to coincide with the classic CFAR detector in radar detection theory. In this respect, it seems to be a novel but promising approach to derive a detector having CFAR property though intrinsic analysis.

Author Contributions: Yao Rong and Mengjiao Tang put forward the original ideas and conducted the research; Jie Zhou raised the research question, reviewed this paper and provided improvement suggestions. All authors have read and approved the final manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Robert, C.P. Intrinsic losses. *Theor. Decis.* **1996**, *40*, 191–214.
2. Bernardo, J.M. Intrinsic credible regions: An objective Bayesian approach to interval estimation. *Test* **2005**, *14*, 317–384.
3. Bernardo, J.M. Nested hypothesis testing: The Bayesian reference criterion. In *Bayesian Statistics 6*; Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M., Eds.; Oxford University Press: Oxford, UK, 1999; pp. 101–130.
4. Bernardo, J.M. Integrated objective Bayesian estimation and hypothesis testing. In *Bayesian Statistics 9*; Bernardo, J.M., Bayarri, M.J., Berger, J.O., Dawid, A.P., Heckerman, D., Eds.; Oxford University Press: Oxford, UK, 2010; pp. 1–68.
5. Bernardo, J.M.; Juárez, M. Intrinsic estimation. In *Bayesian Statistics 7*; Bernardo, J.M., Bayarri, M.J., Berger, J.O., Dawid, A.P., Heckerman, D., Smith, A.F.M., West, M., Eds.; Oxford University Press: Oxford, UK, 2003; pp. 465–476.

6.　Bernardo, J.M. Intrinsic point estimation of the normal variance. In *Bayesian Statistics and Its Applications*; Upadhyay, S.K., Singhand, U., Dey, D.K., Eds.; Anamaya Publication: New Delhi, India, 2006; pp. 110–121.

7.　Bernardo, J.M.; Bernardo, J.M. Objective Bayesian point and region estimation in location-scale models. *Stat. Oper. Res. Trans.* **2007**, *31*, 3–44.

8.　Calin, O.; Udrişte, C. *Geometric Modeling in Probability and Statistics*; Springer: Basel, Switzerland, 2014.

9.　Amari, S.I.; Nagaoka, H. *Methods of Information Geometry*; American Mathematical Society: Providence, RI, USA, 2000.

10.　Čencov, N.N. *Statistical Decision Rules and Optimal Inference*; Ametrican Mathematical Society: Providence, RI, USA, 1982.

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

12.　García, G.; Oller, J.M. What does intrinsic mean in statistical estimation? *Stat. Oper. Res. Trans.* **2006**, *30*, 125–170.

13.　Darling, R.W.R. Geometrically intrinsic nonlinear recursive filters II: Foundations. Available online: https://arxiv.org/ftp/math/papers/9809/9809029.pdf (accessed on 4 August 2017).

14.　Ilea, I.; Bombrun, L.; Terebes, R.; Borda, M.; Germain, C. An M-estimator for robust centroid estimation on the manifold of covariance matrices. *IEEE Signal Process. Lett.* **2016**, *23*, 1255–1259.

15.　Tang, M.; Rong, Y.; Zhou, J. Information-geometric methods for distributed multi-sensor estimation fusion. In Proceedings of the 19th International Conference on Information Fusion, Heidelberg, Germany, 5–8 July 2016; pp. 1585–1592.

16.　Cheng, Y.; Wang, X.; Morelande, M.; Moran, B. Information geometry of target tracking sensor network. *Inf. Fusion* **2013**, *14*, 311–326.

17.　Oller, J.M.; Corcuera, J.M. Intrinsic analysis of statistical estimation. *Ann. Stat.* **1995**, *23*, 1562–1581.

18.　Smith, S.T. Covariance, subspace, and intrinsic Cramér-Rao bounds. *IEEE Trans. Signal Process.* **2005**, *53*, 1610–1630.

19.　Amari, S.I. *Information Geometry and Its Applications*; Springer: Tokyo, Japan, 2016.

20.　Jeffreys, H. An invariant form for the prior probability in estimation problems. *Proc. R. Soc. Lond. A Math. Phys. Sci.* **1946**, *186*, 453–461.

21.　Amari, S.I. Information geometry on hierarchy of probability distributions. *IEEE Trans. Inf. Theory* **2001**, *47*, 1701–1711.

22.　Amari, S.I. *Differential-Geometrical Methods in Statistics*; Vol. 28, *Lecture Notes in Statistics*; Springer: New York, NY, USA, 1985.

23.　Amari, S.I. Information geometry of the EM and em algorithms for neural networks. *Neural Netw.* **1995**, *8*, 1379–1408.

24.　Amari, S.I. Information geometry of statistical inference–an overview. In Proceedings of the IEEE Information Theory Workshop, Bangalore, India, 25 October 2002; pp. 86–89.

25.　Ay, N.; Jost, J.; Vân Lê, H.; Schwachhöfer, L. Information geometry and sufficient statistics. *Probab. Theory Related Fields* **2015**, *162*, 327–364.

26.　Skovgaard, L.T. A Riemannian geometry of the multivariate normal model. *Scand. J. Stat.* **1984**, *11*, 211–223.

27.　Postnikov, M.M. *Geometry VI: Riemannian Geometry*; Vol. 91, *Encyclopaedia of Mathematical Sciences*; Springer: Berlin/Heidelberg, Germany, 2001.

28.　Calvo, M.; Oller, J.M. An explicit solution of information geodesic equations for the multivariate normal model. *Stat. Decis.* **1991**, *9*, 119–138.

29.　Imai, T.; Takaesu, A.; Wakayama, M. Remarks on geodesics for multivariate normal models. *J. Math. Industry* **2011**, *3*, 125–130.

30.　do Carmo, M.P. *Riemannian Geometry*; Birkhäuser: Boston, MA, USA, 1992.

31.　Magnant, C.; Grivel, E.; Giremus, A.; Joseph, B.; Ratton, L. Jeffrey's divergence for state-space model comparison. *Signal Process.* **2015**, *114*, 61–74.

32.　Legrand, L.; Grivel, E. Evaluating dissimilarities between two moving-average models: A comparative study between Jeffrey's divergence and Rao distance. In Proceedings of the 24th European Signal Processing Conference, Budapest, Hungary, 29 August–2 September 2016; pp. 205–209.

33.　Bernardo, J.M. Reference posterior distributions for Bayesian inference. *J. R. Stat. Soc. Series B Stat. Methodol.* **1979**, *41*, 113–147.

34. Yang, R.; Berger, J.O. Estimation of a covariance matrix using the reference prior. *Ann. Stat.* **1994**, *22*, 1195–1211.

35. Sun, D.; Berger, J.O. Objective Bayesian analysis for the multivariate normal model. In Proceedings of Valencia / ISBA 8th World Meeting on Bayesian Statistics, Alicante, Spain, 1–6 June 2006; pp. 525–547.

36. Barrau, A.; Bonnabel, S. A note on the intrinsic Cramer-Rao bound. In *Geometric Science of Information*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 377–386.

37. Box, G.E.; Tiao, G.C. *Bayesian Inference in Statistical Analysis*; Wiley: New York, NY, USA, 1992.

38. Guerci, J.R. *Space-Time Adaptive Processing for Radar*; Artech House: Norwood, MA, USA, 2003.

39. Costa, S.I.; Santos, S.A.; Strapasson, J.E. Fisher information distance: A geometrical reading. *Discrete Appl. Math.* **2015**, *197*, 59–69.

40. Contreras-Reyes, J.E.; Arellano-Valle, R.B. Kullback–Leibler divergence measure for multivariate skew-normal distributions. *Entropy* **2012**, *14*, 1606–1626.

41. Protter, M.H.; Jr., C.B.M. *Intermediate Calculus*; Springer: New York, NY, USA, 1985.

42. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer: New York, NY, USA, 2006.

43. Moakher, M. A differential geometric approach to the geometric mean of symmetric positive-definite matrices. *SIAM J. Matrix Anal. Appl.* **2005**, *26*, 735–747.

44. Kelly, E.J. An adaptive detection algorithm. *IEEE Trans. Aerosp. Electron. Syst.* **1986**, *22*, 115–127.

45. Atkinson, C.; Mitchell, A.F. Rao's distance measure. *Sankhya* **1981**, *43*, 345–365.

46. Conte, E.; De Maio, A.; Galdi, C. CFAR detection of multidimensional signals: An invariant approach. *IEEE Trans. Signal Process.* **2003**, *51*, 142–151.

47. Liu, W.; Xie, W.; Wang, Y. Rao and Wald tests for distributed targets detection with unknown signal steering. *IEEE Signal Process. Lett.* **2013**, *20*, 1086–1089.