Normal Correlation:
An Objective Bayesian Approach

Miguel A. Juárez
University of Warwick
«m.a.juarez@warwick.ac.uk»

Abstract

In this paper we give a decision-theoretic oriented, objective Bayesian answer to the problems of point estimating and sharp hypothesis testing about the correlation coefficient of a bivariate Normal population. Under this view both problems are deemed closely related and thus a coherent answer is developed. Comparisons with frequentist results are given and an alternative interpretation of the maximum likelihood estimator is found.

Keywords: Bayesian reference criterion, intrinsic divergence, intrinsic estimator, logarithmic discrepancy, reference prior.

1 Introduction

Since the introduction in the late 19th century of the concept of correlation into the statistical analysis, a great deal of work has been done in order to estimate the correlation between measurements and to decide whether the estimated correlation is (statistically) significant. Assuming that the observed vector $\boldsymbol{x} = (x_1, x_2)$, given five parameters, follows a bivariate Normal distribution, we address the problems of estimating the correlation,ρ, between both measurements, and to decide whether the (null) hypothesis $H_0 \equiv \{\rho = \rho_0\}$ is compatible with the data.

The paper is organised as follows, Section 2 gives a brief account of some classical and objective Bayesian results on estimation and sharp hypothesis testing. The reference-intrinsic methodology, introduced by Bernardo and Rueda (2002) and Bernardo and Juárez (2003), to derive a decision rule for sharp hypothesis testing, and to provide a point estimate is presented in Section 3. Implementation of the method and comparisons with alternative approaches is carried out in Section 4, using both simulated and real data. Some final comments are given in Section 5.

2 Preliminaries

2.1 The frequentist view

The now ubiquitous law

$$f(\boldsymbol{x} \mid \boldsymbol{\mu}, \Sigma) = \frac{|\Sigma|^{-\frac{1}{2}}}{2\pi} \exp\left[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})' \, \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right], \tag{1}$$

with

$$\mathrm{E}[\boldsymbol{x}] = \boldsymbol{\mu} = \{\mu_1, \mu_2\}\,, \quad \text{and} \quad \mathrm{Var}[\boldsymbol{x}] = \Sigma^{-1} = \frac{1}{1-\rho^2}\begin{pmatrix} \lambda_1 & -\rho\lambda_1^{1/2}\lambda_2^{1/2} \\ -\rho\lambda_1^{1/2}\lambda_2^{1/2} & \lambda_2 \end{pmatrix},$$

where $\lambda_i > 0$, $i = 1, 2$, and $\rho \in (-1, 1)$, was first derived empirically in 1885 by Sir Francis Galton (Pearson and Kendall, 1970, p. 197), bringing the concept of correlation into the realm of statistical inference. Some thirty years later, Fisher (1915) showed that, given a random sample, $z = \{(x_{1i}, x_{2i})\}_{i=1}^{n}$, the sampling correlation coefficient

$$r(z) = \frac{s_{12}}{s_1 s_2}$$

where

$$s_{12} = \frac{1}{n}\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)\,,\ s_j^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_{ji} - \bar{x}_j\right)^2 \text{ and } \bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ji}\ j = 1, 2;$$

is sufficient for ρ (Jeffreys, 1961, p. 175), and that its sampling distribution is given by

$$\begin{aligned} p\,(r \mid \mu_1, \mu_2, \lambda_1, \lambda_2, \rho) &= p\,(r \mid \rho) \\ &\propto \frac{\left(1-\rho^2\right)^{\frac{n-1}{2}}\left(1-r^2\right)^{\frac{n-4}{2}}}{(1-\rho\, r)^{n-\frac{3}{2}}}\,\mathrm{F}\left(\frac{1}{2}, \frac{1}{2}, n-\frac{1}{2}, \frac{1+\rho\, r}{2}\right), \end{aligned} \tag{2}$$

where $\mathrm{F}(a, b, c, z)$ is the hypergeometric function.

Moreover, $r = r\,(z)$ is the maximum likelihood estimator (MLE) –and thus consistent. Therefore, report r as a point estimate and calculate numerically its standard error from (2).

If we are interested in testing independence between both measurements, i.e. test $H_0 \equiv \{\rho = 0\}$, it can be shown (see e.g. Lehmann, 1986) that the statistic

$$R = R\,(r) = \frac{r}{\sqrt{(1-r^2)\,/\,(n-2)}} \tag{3}$$

has, under H_0, a standard $t_{(n-2)}$ sampling distribution with $n - 2$ degrees of freedom, leading to the decision rule:

> Reject H_0, with test size α, whenever $|R| > K$, where K is the $\alpha/2$ percentile of the standard $t_{(n-2)}$ distribution.

2.2 The objective Bayesian view

In his seminal work, Jeffreys (1961) proposed

$$\pi_{J1}\,(\rho, \mu_1, \mu_2, \lambda_1, \lambda_2) \propto \lambda_1^{-1}\lambda_2^{-1} \tag{4}$$

as an objective prior for this problem, leading to the posterior

$$\pi_{J1}(\rho \mid z) \propto \frac{\left(1-\rho^2\right)^{\frac{n-1}{2}}}{(1-\rho\, r)^{n-\frac{3}{2}}}\, \mathsf{F}\left(\frac{1}{2}, \frac{1}{2}, n-\frac{1}{2}, \frac{1+\rho\, r}{2}\right), \tag{5}$$

which, omitting the constants, is identical to (2), corroborating Fisher's results on the sufficiency of r. Seeking an invariant prior, he also proposed what now we might call the Jeffreys' prior,

$$\pi_{J2}(\rho, \mu_1, \mu_2, \lambda_1, \lambda_2) \propto \lambda_1^{-1}\lambda_2^{-1}\left(1-\rho^2\right)^{-\frac{3}{2}}, \tag{6}$$

so that the posterior is of the same shape as (5) and may be written as

$$\pi_{J2}(\rho \mid z) \propto \frac{\left(1-\rho^2\right)^{\frac{n}{2}-2}}{(1-\rho\, r)^{n-\frac{3}{2}}}\, \mathsf{F}\left(\frac{1}{2}, \frac{1}{2}, n-\frac{1}{2}, \frac{1+\rho\, r}{2}\right). \tag{7}$$

Later, Lindley (1972), while studying the behaviour of posteriors derived from priors of the form,

$$\pi_L(\rho, \lambda_1, \lambda_2, \mu_1, \mu_2) \propto \lambda_1^{-1}\lambda_2^{-1}\left(1-\rho^2\right)^{c},$$

concluded that for the prior to be non-informative we must have $-2 < c \leq -1$ and recommended $c = -1$, ruling out (6).

Finally, Bayarri (1981), following the ideas in Bernardo (1979), derived the reference prior for the bivariate Normal model when the coefficient of correlation is the parameter of interest as

$$\pi(\rho, \lambda_1, \lambda_2, \mu_1, \mu_2) \propto \lambda_1^{-1}\lambda_2^{-1}\left(1-\rho^2\right)^{-1}, \tag{8}$$

confirming Lindley's intuition and setting the standard objective prior which leads to the reference posterior, for $n \geq 3$

$$\pi(\rho \mid z) \propto \frac{\left(1-\rho^2\right)^{\frac{n-3}{2}}}{(1-\rho\, r)^{n-\frac{3}{2}}}\, \mathsf{F}\left(\frac{1}{2}, \frac{1}{2}, n-\frac{1}{2}, \frac{1+\rho\, r}{2}\right). \tag{9}$$

The posterior contains all the information we have about ρ, though if we are asked about a point estimate we must propose a loss function and then minimise the posterior expected loss. The most commonly used *automatic* losses are the quadratic, the absolute and the zero-one loss functions, yielding the posterior mean, median and mode, respectively (if they exist). These can be numerically calculated from (9).

With regard to the sharp testing problem, the ordinary tool, conventional Bayes factors, implicitly assume a point mass allocated at the null value, ρ_0, which might lead to unsound results –namely, the Jeffreys-Lindley-Bartlett (JLB) paradox–, as pointed out by Lindley (1957) and Bartlett (1957). Various attempts to overcome this difficulties have been made, including intrinsic Bayes factors (Berger and Pericchi, 2001), fractional Bayes factors (O'Hagan, 1997) and neutral Bayes factors (Robert and Caron, 1996). However, they do not necessarily correspond to a Bayesian analysis for any prior and are, therefore, open to criticism.

3 Reference-Intrinsic Analysis

In this section the problems of point estimation and hypothesis testing are addressed within a decision framework and from an objective Bayesian approach which overcome these shortcomings.

3.1 The methodology

Suppose we have determined that the probabilistic behaviour of an observable $\boldsymbol{x}$ is adequately described by the parametric model

$$\mathcal{M} = \{p(\boldsymbol{x} \mid \boldsymbol{\theta}, \omega),\ \boldsymbol{x} \in X,\ \boldsymbol{\theta} \in \Theta, \omega \in \Omega\}$$

and that we are interested in making inferences about $\boldsymbol{\theta}$, with ω a nuisance parameter. Recently, Bernardo and Rueda (2002) and Bernardo and Juárez (2003) argue that the problems of point estimation and precise hypothesis testing might be posed as those of deciding which value $\boldsymbol{\theta}^* = \boldsymbol{\theta}^*(\boldsymbol{x})$ renders the best proxy, $p(\boldsymbol{x} \mid \boldsymbol{\theta}^*, \omega)$, to the assumed model given the data, and whether a given proxy model, $p(\boldsymbol{x} \mid \boldsymbol{\theta}_0, \omega)$, is consistent with the data, respectively. To this end they propose as a loss function the *intrinsic discrepancy*, defined as

$$\delta_{\boldsymbol{x}}(\boldsymbol{\theta}, \omega; \boldsymbol{\theta}_0) = \min\{k(\boldsymbol{\theta}_0 \mid \boldsymbol{\theta}, \omega), k(\boldsymbol{\theta}, \omega \mid \boldsymbol{\theta}_0)\}, \tag{10}$$

where

$$k(\boldsymbol{\theta}_0 \mid \boldsymbol{\theta}, \omega) = \min_{\omega_0 \in \Omega} \int_X p(\boldsymbol{x} \mid \boldsymbol{\theta}, \omega) \log \frac{p(\boldsymbol{x} \mid \boldsymbol{\theta}, \omega)}{p(\boldsymbol{x} \mid \boldsymbol{\theta}_0, \omega_0)}\, \mathrm{d}\boldsymbol{x}$$

and

$$k(\boldsymbol{\theta}, \omega \mid \boldsymbol{\theta}_0) = \min_{\omega_0 \in \Omega} \int_X p(\boldsymbol{x} \mid \boldsymbol{\theta}_0, \omega_0) \log \frac{p(\boldsymbol{x} \mid \boldsymbol{\theta}_0, \omega_0)}{p(\boldsymbol{x} \mid \boldsymbol{\theta}, \omega)}\, \mathrm{d}\boldsymbol{x},$$

are the corresponding Kullback-Leibler (KL) divergences.

The intrinsic discrepancy is a measure, in natural information units (*nits*), of the expected amount of information needed to discriminate between the full model $\mathcal{M}$ and a given proxy. This measure has a number of appealing properties: it is symmetric, non-negative and vanishes iff $f(\boldsymbol{x} \mid \boldsymbol{\eta}, \boldsymbol{\lambda}) = f(\boldsymbol{x} \mid \boldsymbol{\varphi}, \boldsymbol{\gamma})$, a.e. It is invariant under one-to-one transformations of the parameter of interest, under one-to-one transformations of the data and under the choice of the nuisance parameter. Further, it is additive for conditionally independent observations. If $f_1(\boldsymbol{x} \mid \boldsymbol{\psi})$ and $f_2(\boldsymbol{x} \mid \boldsymbol{\phi})$ have nested supports so that $f_1(\boldsymbol{x} \mid \psi) > 0$ iff $\boldsymbol{x} \in \mathcal{X}_1(\Psi)$, $f_2(\boldsymbol{x} \mid \boldsymbol{\phi}) > 0$ iff $\boldsymbol{x} \in \mathcal{X}_2(\Phi)$ and either $\mathcal{X}_1(\Psi) \subset \mathcal{X}_2(\Phi)$ or $\mathcal{X}_2(\Phi) \subset \mathcal{X}_1(\Psi)$, the intrinsic divergence is still well defined and reduces to the logarithmic discrepancy, viz. $\delta(\boldsymbol{\phi}; \boldsymbol{\psi}) = k(\boldsymbol{\phi} \mid \boldsymbol{\psi})$ when $\mathcal{X}_1(\Psi) \subset \mathcal{X}_2(\Phi)$ and $\delta(\boldsymbol{\phi}; \boldsymbol{\psi}) = k(\boldsymbol{\psi} \mid \boldsymbol{\phi})$ when $\mathcal{X}_2(\Phi) \subset \mathcal{X}_1(\Psi)$. For a thorough discussion see Bernardo and Rueda (2002) and Juárez (2004).

Deriving the minimum of the KL divergencies in models where the dimension of the parametric space is relatively large may prove tedious and involved. However, in some cases it is possible to interchange the optimisation and minimisation steps. We will use the following lemma, before stating the main result. All proofs are deferred to the appendix.

Lemma 1.
Let p_1 and p_2 be two probability density functions with convex support. Then, the KL divergencies, $k(p_i \mid p_j)$, $i, j = 1, 2$ are convex.

Theorem 1 (Intrinsic discrepancy in regular models).
Let $\{p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{\lambda}),\ \boldsymbol{x} \in \mathcal{X},\ \boldsymbol{\theta} \in \Theta,\ \boldsymbol{\lambda} \in \Lambda\}$ be a probabilistic model that meets the regularity conditions stated in the appendix. Then,

$$\begin{aligned}\delta^*(\boldsymbol{\theta}, \boldsymbol{\theta}_0, \boldsymbol{\lambda}) &= \inf_{\boldsymbol{\lambda}_0 \in \Lambda} \delta(\boldsymbol{\theta}, \boldsymbol{\theta}_0, \boldsymbol{\lambda}, \boldsymbol{\lambda}_0) \\ &= \min\left\{\inf_{\boldsymbol{\lambda}_0 \in \Lambda} k(\boldsymbol{\theta}, \lambda \mid \boldsymbol{\theta}_0, \boldsymbol{\lambda}_0),\ \inf_{\boldsymbol{\lambda}_0 \in \Lambda} k(\boldsymbol{\theta}_0, \boldsymbol{\lambda}_0 \mid \boldsymbol{\theta}, \boldsymbol{\lambda})\right\}.\end{aligned}$$

In particular, the exponential family meets the conditions of Theorem 1, therefore

Corlary 1 (Intrinsic discrepancy in the exponential family).
Let $\{p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{\lambda}),\ \boldsymbol{x} \in \mathcal{X},\ \boldsymbol{\theta} \in \Theta,\ \boldsymbol{\lambda} \in \Lambda\}$ belong to the exponential family; i.e. $p(\boldsymbol{x} \mid \boldsymbol{\psi}) = a(\boldsymbol{\psi}) \exp[\psi^t \boldsymbol{t}(\boldsymbol{x})]$. Then,

$$\begin{aligned}\delta^*(\boldsymbol{\theta}, \boldsymbol{\lambda}; \boldsymbol{\theta}_0) &= \inf_{\boldsymbol{\lambda}_0 \in \Lambda} \delta(\boldsymbol{\theta}, \boldsymbol{\lambda}; \boldsymbol{\theta}_0, \boldsymbol{\lambda}_0) \\ &= \min\left\{\inf_{\boldsymbol{\lambda}_0 \in \Lambda} k(\boldsymbol{\theta}, \lambda \mid \boldsymbol{\theta}_0, \boldsymbol{\lambda}_0),\ \inf_{\boldsymbol{\lambda}_0 \in \Lambda} k(\boldsymbol{\theta}_0, \boldsymbol{\lambda}_0 \mid \boldsymbol{\theta}, \boldsymbol{\lambda})\right\},\end{aligned}$$

with

$$\begin{aligned}k(\boldsymbol{\psi}_j \mid \boldsymbol{\psi}_i) &= \int p(\boldsymbol{x} \mid \boldsymbol{\psi}_i) \log \frac{p(\boldsymbol{x} \mid \boldsymbol{\psi}_i)}{p\left(\boldsymbol{x} \mid \boldsymbol{\psi}_j\right)} \, \mathrm{d}\boldsymbol{x} \\ &= M(\boldsymbol{\psi}_i) - M(\boldsymbol{\psi}_j) + \left(\boldsymbol{\psi}_j^t - \boldsymbol{\psi}_i^t\right) \nabla M(\boldsymbol{\psi}_i),\end{aligned}$$

where $\boldsymbol{\psi}_k = \{\boldsymbol{\theta}_k, \boldsymbol{\lambda}_k\}$, $k = 0, 1$, $M(\boldsymbol{\psi}) = \log a(\boldsymbol{\psi})$ y $\nabla M(\boldsymbol{\psi}) = \partial M(\boldsymbol{\psi}) / \partial \boldsymbol{\psi}$.

In our problem, the directed KL divergencies,

$$k(\rho_0 \mid \rho, \mu_1, \mu_2, \lambda_1, \lambda_2) = \frac{n}{2}\left[\log \frac{(1 - \rho_0 \rho)^2}{\left(1 - \rho_0^2\right)^2} + \log \frac{1 - \rho_0^2}{1 - \rho^2}\right],$$

and

$$k(\rho, \mu_1, \mu_2, \lambda_1, \lambda_2 \mid \rho_0) = \frac{n}{2}\left[\log \frac{(1 - \rho_0 \rho)^2}{(1 - \rho^2)^2} + \log \frac{1 - \rho^2}{1 - \rho_0^2}\right],$$

are symmetric, and hence the intrinsic discrepancy is

$$\delta(\rho;\rho_0) = \frac{n}{2}\,\log\frac{(1-\rho_0\rho)^2}{(1-\rho^2)\left(1-\rho_0^2\right)}\,. \tag{11}$$

It is readily verified that $\delta(\rho;\rho_0) \geq 0$ with equality iff $\rho = \rho_0$, and that it is a convex function of ρ_0.

In a decision problem, the parameter of interest is that which enters the loss function. Making use of this argument, Bernardo and Rueda (2002) proceed to derive the reference prior (Berger and Bernardo, 1992; Bernardo, 1979), $\pi_\delta(\boldsymbol{\theta};\omega)$, when the intrinsic discrepancy is the parameter of interest. In our case, in order to calculate the reference prior when $\delta(\rho;\rho_0)$ is the parameter of interest, note that this is a one-to-one piecewise function of ρ. Thus, given that the reference posterior is invariant under this kind of transformations (Bernardo and Smith, 1994, p. 326), we can use (8) to derive the reference posterior for ρ and then calculate the expected value of (11) under it. Obviously, (9) is the corresponding posterior.

3.2 The Bayesian Reference Criterion

The expected value of the intrinsic divergence under the reference posterior, hereafter *the intrinsic statistic*, is

$$d(\rho_0 \mid r) = \int_{-1}^{1}\frac{n}{2}\,\log\left[\frac{(1-\rho_0\rho)^2}{(1-\rho^2)\left(1-\rho_0^2\right)}\right]\pi(\rho \mid r)\,d\rho\,. \tag{12}$$

The intrinsic statistic is a measure, in natural information units (nits), of the weight of the evidence conveyed by the data against the simplification attained when acting as if $\rho = \rho_0$. Given that (11) is convex in ρ_0, $d(\rho_0 \mid r)$ is also convex. Thus, it induces the decision rule, hereafter the Bayesian reference criterion (BRC):

$$\text{Reject } \rho = \rho_0 \text{ iff } d(\rho_0 \mid r) > d^*.$$

In order to calibrate the threshold value, d^*, Bernardo and Rueda (2002) argue that the intrinsic statistic can be interpreted as the expected value of the log likelihood ratio against the simplified model, given the data. Hence, values of d^* around 2.5 would imply a ratio of $e^{2.5} \approx 12$, providing mild evidence against the null; while values around 5 ($e^5 \approx 150$) can be regarded as strong evidence against H_0; values of $d^* \geq 7.5$ ($e^{7.5} \approx 1800$) can be safely used to reject the null.

For this problem, (12) can be numerically integrated. Juárez (2004) shows that when the reference posterior is asymptotically Normal, as in this case, a good approximation even for moderate sample sizes is given by,

$$d(\theta_0 \mid \boldsymbol{x}) \approx \delta(\hat{\theta};\theta_0) + \frac{1}{2},$$

with $\hat{\theta}$, a consistent estimator of θ, for instance the MLE . In this case we obtain

$$d\,(\rho_0 \mid r) \approx \frac{n}{2}\,\log\left[\frac{(1-\rho_0\,r)^2}{(1-r^2)\left(1-\rho_0^2\right)}\right]+\frac{1}{2}. \tag{13}$$

Both (12) and (13), are depicted in Figure 1, calculated from simulated data with $\rho = 3/4$ and for different sample sizes. We can see that for small n, it is not possible to reject almost any value of the parameter and that the criterion becomes more discriminating as the sample size increases.

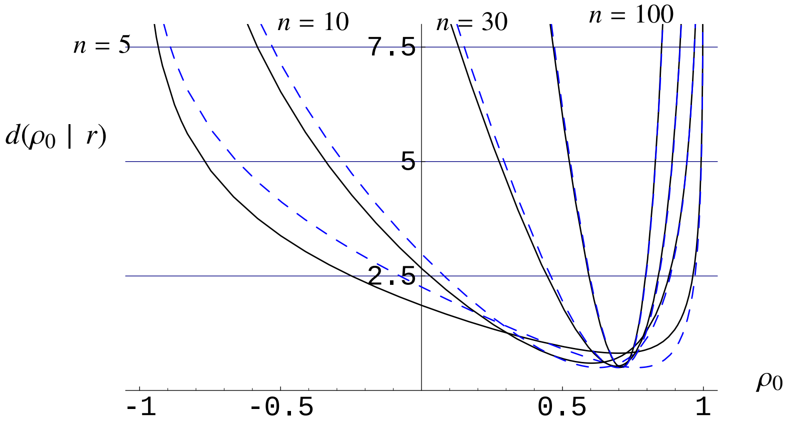


Figure 1. *The intrinsic statistic for the correlation coefficient and its approximation (dashed) for simulated values, fixing $\rho = 3/4$. According to the BRC , values of $d\,(\rho_0, r)$ above 2.5 should be considered as mild evidence against $\rho = \rho_0$, values over 5 can be think of as providing strong evidence against the null and values over 7.5 may be safely rejected.*

An important advantage of the BRC over the classical test is that with the former we are able to test any value $\rho_0 \in (-1, 1)$, in contrast with (3) which is derived only for $\rho_0 = 0$. In case we want to use a classical test for a null value different from 0, we must derive its sampling distribution or use the normal approximation, $\sqrt{n}\,(r-\rho) \xrightarrow{d} \mathrm{N}\left(0,(1-\rho^2)^{-2}\right)$, which is known to fail for small to moderate sample sizes and extreme correlation values.

3.3 The Intrinsic Estimator

A natural consequence of our decision-theoretical approach is that the obvious choice for a point estimate is the one that minimizes the intrinsic statistic. Hence, we shall say that $\boldsymbol{\theta}^*$ is *the intrinsic estimator of θ* if

$$\boldsymbol{\theta}^*(\boldsymbol{x}) = \arg\min_{\theta_0\in\Theta} d\,(\theta_0 \mid \boldsymbol{x})\,.$$

An important feature of the intrinsic estimator is that it inherits the invariance properties of the intrinsic statistic, a characteristic not shared by the most commonly used objective Bayes estimators.

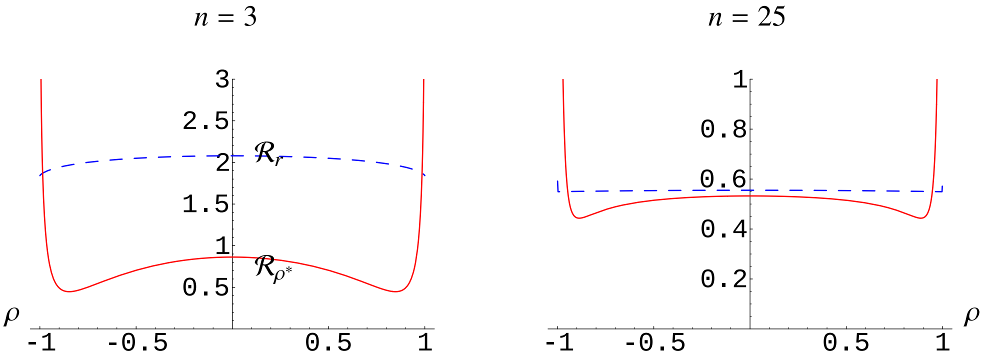
In our problem, $\rho^* = \rho^*(r)$ can be found numerically for $n \geq 3$. A good analytical approximation is (Juárez, 2004)

$$\rho^* \approx r\sqrt{\frac{n}{n+1}}.$$

One way to compare the performance of the intrinsic estimator and the MLE under homogeneous repeated sampling is to calculate their respective (intrinsic) risk functions,

$$\mathcal{R}_{\tilde{\rho}}(\rho) = \int_{-1}^{1} \delta(\rho; \tilde{\rho})\, p\,(r \mid \rho)\ \mathrm{d}\rho,$$

with $p\,(r \mid \rho)$ as in (2). Both functions are depicted in Figure 2 for the minimum sample size and for a moderate one. We can see that none of them is dominated by the other, although $\mathcal{R}_{\rho^*} < \mathcal{R}_r$ for almost all of the parametric space, and how both functions are close to each other, but at the boundaries of the sample space, as sample size grows. Indeed, Juárez (2004) shows that when the (marginal) reference posterior is asymptotically Normal, the intrinsic estimator is admissible and, if $\tilde{\theta}$ is a consistent estimator of θ, then θ^* and $\tilde{\theta}$ are asymptotically equivalent in risk, under intrinsic loss.



$n = 3$

$n = 25$

Figure 2. *Intrinsic risk function of the intrinsic estimator, $\mathcal{R}_{\rho^*}$ (solid line), and the MLE , $\mathcal{R}_r$ (dashed line), for the minimum sample size and a moderate one.*

In a number of applications, the instrument used to measure both quantities is usually the same, and so assuming $\lambda_1 = \lambda_2 = \lambda$, i.e. both measurements share the same precision, may be sensible. In this case, the logarithmic divergence (10) remains unchanged and it follows that the reference prior for this model when ρ is the parameter of interest is

$$\pi\,(\rho, \mu_1, \mu_2, \lambda) \propto \lambda^{-1}\left(1-\rho^2\right)^{-1}\ . \tag{14}$$

Combining the likelihood with (14) and integrating out the nuisance parameters $\{\mu_1, \mu_2, \lambda\}$, leads to the reference posterior

$$\pi\,(\rho \mid r) = \frac{\Gamma\left(\frac{n}{2}\right)\left(1-r^2\right)^{\frac{n-1}{2}}}{\sqrt{\pi}\,\Gamma\left(\frac{n-1}{2}\right)}\ \frac{\left(1-\rho^2\right)^{\frac{n-3}{2}}}{(1-r\rho)^{n-1}}, \tag{15}$$

with mean

$$\mathrm{E}[\rho \mid r] = r\left(1-r^2\right)^{\frac{n-1}{2}} \frac{(n-1)}{n} \mathrm{F}\left(\frac{n+1}{2}, \frac{n}{2}, \frac{n+2}{2}, r^2\right),$$

and mode

$$\mathrm{Mod}[\rho \mid r] = \frac{\sqrt{(n-3)^2 + 8r^2(n-1)} - (n-3)}{4r}.$$

It is interesting that for this case the posterior median, the maximum likelihood estimator and the intrinsic estimator coincide,

$$\rho^*(r) = \hat{\rho}(r) = \mathrm{Med}[\rho \mid r] = r;$$

this can be checked numerically. Thus suggesting an appealing interpretation of the maximum likelihood estimator as an approximation to the intrinsic estimator when both precisions are close.

4 Examples

4.1 Simulated data

In order to compare the performance under homogeneous repeated sampling of both the intrinsic estimator and the BRC with their frequentist counterparts, 5000 simulations were carried for the minimum sample size and for a moderate one. Table 1 summarises this comparison. The third and fourth columns (BRC and R) report the relative number of times when the hypothesis $\rho = 0$ was rejected under each criterion, using the comparable threshold values of $d^* = 3$ for the BRC , and $\alpha = 0.05$ for the t-test. The last column presents the relative number of times that the intrinsic estimator was closer than the MLE to the real value.

Table 1. *Comparison of the behaviour under repeated sampling of the BRC and the intrinsic estimator vs. their frequentist counterparts*

n	ρ	BRC	R	% diff.
3	$-\frac{9}{10}$	0.257	0.169	0.655
	0	0.200	0.050	1.000
	$\frac{3}{4}$	0.366	0.099	0.720
25	$-\frac{9}{10}$	1.000	1.000	0.512
	0	0.026	0.046	1.000
	$\frac{3}{4}$	0.996	0.998	0.519

As we can see from Table 1, the intrinsic estimator outmatches the MLE for the minimum sample size and their performances get closer as sample size increases. Interesting to note is the case when $\rho = 0$, where the intrinsic estimator is consistently closer to the true value than r. Further exploration of this behaviour (not shown here for brevity) suggests that this closeness characteristic of ρ^* over the MLE diminishes as the true value of the parameter moves away from 0 and the sample size increases. Regarding hypothesis testing, we confirm that the frequentist *p*-value remains constant as sample size increases, while the one corresponding to the BRC decreases. The power of both tests increases with sample size.

4.2 Heights data

Figure 3 depicts the intrinsic statistic calculated from the 1375 mother-daughter heights data recorded by Pearson and Lee (1903), with sample correlation of $r = 0.4907$. There we point out the intrinsic estimator, $\rho^* = 0.49057$ and, enclosed within the broken lines, the *non-rejection regions*, $\mathcal{R}_{d^*}$, corresponding to the threshold values: $d^* = 2.5, 5$ and 7.5. Thus, we may cast some doubts about the true correlation value between mother-daughter heights being outside the region $\mathcal{R}_{2.5} = \{0.449, 0.530\}$; we can seriously doubt that it is beyond $\mathcal{R}_5 = \{0.427, 0.549\}$ and we can be quite sure that it is not outside $\mathcal{R}_{7.5} = \{0.410, 0.563\}$. Moreover, it can be safely assumed that the true value is $\rho = 1/2$.

One of the most appealing properties of the intrinsic solutions is their invariance under piecewise one-to-one transformations. For instance, if we are interested in the coefficient of determination, ρ^2, all we have to do is to apply the corresponding transformation to the previous results. Hence, $(\rho^2)^* = (\rho^*)^2 = 0.2407$ and $\mathcal{R}_{2.5}(\rho_0^2) = \{0.202, 0.281\}$, $\mathcal{R}_5(\rho_0^2) = \{0.182, 0.301\}$ and $\mathcal{R}_{7.5}(\rho_0^2) = \{0.168, 0.317\}$.

Apparent from Figure 3 is also the fact that, given the sample size, the asymptotic approximation, $d(\rho_0 \mid r) = \delta(r; \rho_0) + 1/2$, is almost indistinguishable form the actual function. Finally, the proposed analytic approximation to the intrinsic estimator proves to be very accurate, $\rho^* = \sqrt{n/(n+1)}\, r = 0.49053$.

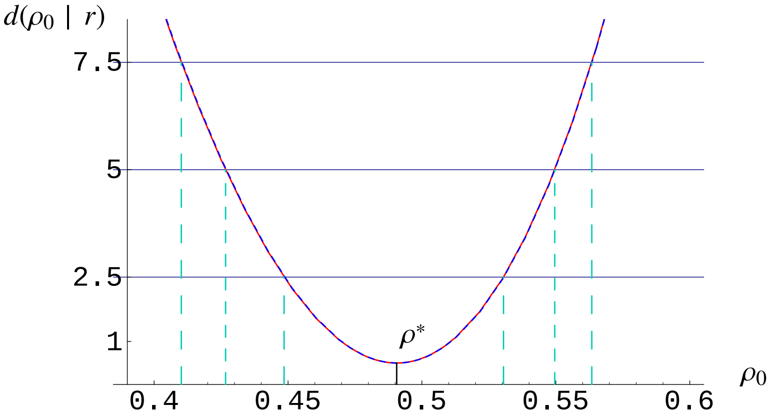
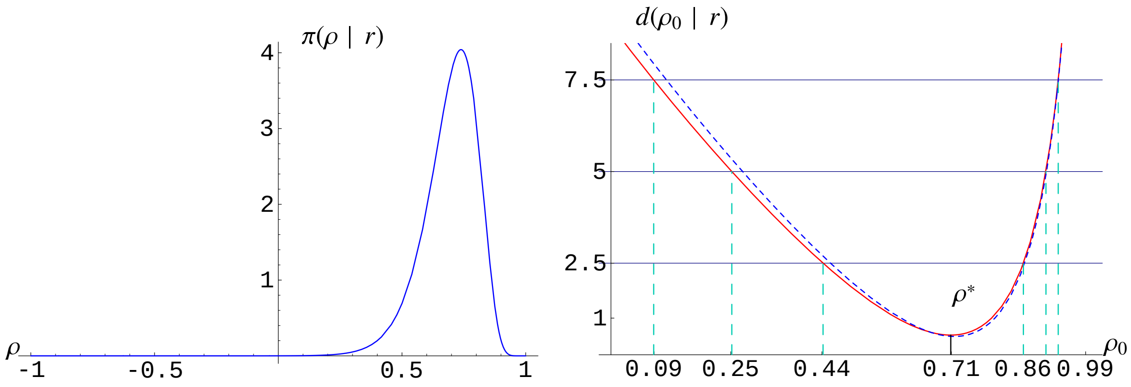


Figure 3. *The intrinsic statistic (solid line) and the asymptotic approximation (dotted line) calculated from Person and Lee's mother-daughter heights data. Also, the intrinsic estimator, ρ^* and the rejection regions corresponding to the $d^* = 2.5$, $d^* = 5$ and $d^* = 7.5$ threshold values are pointed out.*

4.3 Cancer data

Certainly, the large sample size in the previous example is responsible for the almost exact agreement between intrinsic and classical results. In order to explore the extent of this coincidence, we analyse a data set comprising the smoking ratio and the lung cancer standardized mortality ratio (SMR), recorded for males in England and Wales during 1970-1972, for each of $n = 25$ "occupation orders" or broad groups of jobs, taken from Hand *et al.* (1994, pp. 66–67). For this data set the sample correlation is $r = 0.7162$ and, unlike the heights data example, the (reference) posterior for the correlation coefficient given this data set is far from Normal, as depicted in Figure 4 (a). Nevertheless, the asymptotic approximation to the intrinsic statistic is still very accurate –Figure 4 (b)–, as well as the proposed approximation, $\rho^* \approx 0.7023$, to the intrinsic estimator, $\rho^* = 0.7087$. The (non-symmetric) non-rejection regions can again be determined for this case as $\mathcal{R}_{2.5} = \{0.442, 0.0.860\}$, $\mathcal{R}_5 = \{0.252, 0..907\}$, and $\mathcal{R}_{7.5} = \{0.089, 0.933\}$. Thus, if the basic assumption of joint normality holds for this sample, we could safely reject the hypothesis of no association between smoking ratio and lung cancer SMR. Further, we could state that the correlation is positive.



(a) *The reference posterior.*

(b) *Intrinsic statistic and the proposed approximation.*

Figure 4. *The reference posterior, $\pi(\rho \mid r)$, for the correlation between smoking ratio and lung cancer SMR on the left. The right pane shows the intrinsic statistic (solid) and the asymptotic approximation (dotted), with the intrinsic estimator as well as the non-rejection regions pointed out.*

5 Conclusions

The reference-intrinsic approach described here provides a powerful alternative to point estimation and sharp hypothesis testing, with a clear interpretation in terms of information units.

The Bayesian Reference Criterion, in contrast of the classical t-test, can be used to test any feasible point in the parametric space, with no additional effort. Its invariant properties make it readily available if we are interested in any bijective function of the parameter, as the coefficient of determination. Given the asymptotic normality of the reference posterior, its behaviour under repeated, homogeneous sampling, measured through the corresponding *p*-value for a given threshold, tends to agree with the frequentist test as sample size increases.

Unlike the most commonly utilised Bayes estimators, the intrinsic estimator is invariant under piecewise one-to-one transformations. Further, it clearly performs better than the traditional MLE , particularly in the relevant case when the correlation is near zero. We also showed that the MLE might be interpreted as an approximation to the intrinsic estimator when both precisions are close, coinciding exactly with the posterior median when they are equal.

The idea of an intrinsic interval estimate stems almost naturally from the BRC . Future research about this invariant, not necessarily symmetric regions which contain the intrinsic estimator will be reported elsewhere.

Appendix A Regularity conditions

Assume that $p\,(\boldsymbol{x} \mid \boldsymbol{\theta})$ is a probability density function and that the parametric space $\Theta \subset \mathbb{R}^k$. We say that the model is regular if

(*i*) The parametric space, Θ, is open.

(*ii*) The support $C = \{\boldsymbol{x} \in \mathcal{X} \; : \; p\,(\boldsymbol{x} \mid \boldsymbol{\theta}) > 0\}$ is the same for all $\boldsymbol{\theta} \in \Theta$.

(*iii*) For each $\boldsymbol{x} \in \mathcal{X}$, $\dfrac{\partial^2}{\partial \boldsymbol{\theta}_i \boldsymbol{\theta}_j} p\,(\boldsymbol{x} \mid \boldsymbol{\theta})$ exists and is continuous for all $i, j = 1, \ldots, k$.

(*iv*) $\int p\,(\boldsymbol{x} \mid \boldsymbol{\theta})\ \mathrm{d}\boldsymbol{x}$ can be differentiated twice w.r.t. the components of $\boldsymbol{\theta}$, under the integral.

Appendix B Proofs

B.1 Lemma 1

The set $\mathcal{S}$ is convex if for all $x, y \in \mathcal{S}$ y $\omega \in (0, 1)$ then $z = \omega\, x + (1 - \omega)\, y \in \mathcal{S}$. Further, a functional f, defined on a convex set $\mathcal{S}$, is convex if for each x e y en $\mathcal{S}$,

$$\omega\, f(x) + (1 - \omega)\, f(y) \leq f(\omega\, x + (1 - \omega)\, y)\,, \quad 0 \leq \omega \leq 1.$$

Let $\{p_i(\boldsymbol{x}), q_i(\boldsymbol{x})\}$, $i = 1, 2$ be four probability density functions. Then,

$$\omega\, p_i(\boldsymbol{x}) + (1 - \omega)\, q_i(\boldsymbol{x}) \geq 0 \qquad \text{and} \qquad \int \omega\, p_i(\boldsymbol{x}) + (1 - \omega)\, q_i(\boldsymbol{x})\ \mathrm{d}\boldsymbol{x} = 1.$$

We can write the KL divergence as

$$\begin{aligned} k(p_2 \mid p_1) &= \int p_1(\boldsymbol{x})\ \log \frac{p_1(\boldsymbol{x})}{p_2(\boldsymbol{x})}\ \mathrm{d}\boldsymbol{x} \\ &= \int p_2(\boldsymbol{x})\, g(\boldsymbol{x})\ \log g(\boldsymbol{x})\ \mathrm{d}\boldsymbol{x}\ , \end{aligned}$$

whith $g(\boldsymbol{x}) = p_1(\boldsymbol{x})/p_2(\boldsymbol{x}) > 0$, the likelihood ratio. Moreover, $\phi(t) = t\,\log t$ is convex, for all $t > 0$; i.e. let $t_1, t_2 > 0$, then, for all $\omega \in (0, 1)$, $\phi(\omega\, t_1 + (1-\omega)\, t_2) \leq \omega\, \phi(t_1) + (1-\omega)\, \phi(t_2)$. It follows that

$$\int p_2(\boldsymbol{x})\, \phi\Big(\omega\, t_1 + (1-\omega)\, t_2\Big)\, \mathrm{d}\boldsymbol{x} \leq \int p_2(\boldsymbol{x})\, \Big(\omega\, \phi(t_1) + (1-\omega)\, \phi(t_2)\Big)\, \mathrm{d}\boldsymbol{x}.$$

If we define $t_1 = t_1(\boldsymbol{x}) = p_1(\boldsymbol{x})/p_2(\boldsymbol{x})$ y $t_2 = t_2(\boldsymbol{x}) = q_1(\boldsymbol{x})/q_2(\boldsymbol{x})$, the result follows.

B.2 Theorem 1

Let $\{p\,(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{\lambda}),\ \boldsymbol{x} \in \mathcal{X},\ \boldsymbol{\theta} \in \Theta,\ \boldsymbol{\lambda} \in \Lambda\}$ be a probability model which meets the regularity conditions above and let $k_1 = k(\boldsymbol{\theta}_0, \lambda_0 \mid \boldsymbol{\theta}, \lambda)$ y $k_2 = k(\boldsymbol{\theta}, \boldsymbol{\lambda} \mid \boldsymbol{\theta}_0, \lambda_0)$ be the KL divergencies. By Lemma 1, Both functions are convex and also bounded from below by 0.

There exist two alternatives

(*i*) $k_i \leq k_j$ for all $\{\boldsymbol{\theta}, \lambda, \boldsymbol{\theta}_0, \lambda_0\} \in \Theta^2 \times \Lambda^2$ or

(*ii*) $k_i \leq k_j$ in $C \subset \Theta^2 \times \Lambda^2$ y $k_i \geq k_j$ in C'.

The proof in the first case is trivial. In the second case, and without loss of generality, assume that $k_1 \leq k_1$ en C. Define

$$\lambda^\star_C = \arg\min_{\lambda_0 \in C}\ k_1 \qquad \text{and} \qquad \lambda^*_{C'} = \arg\min_{\lambda_0 \in C'}\ k_2.$$

On the other hand, define $\boldsymbol{\lambda}^\star = \lambda^\star(\boldsymbol{\theta}, \boldsymbol{\theta}_0, \lambda) \in \Lambda$ as the value for which k_1 reaches its minimum and similarly $\lambda^* = \lambda^*(\boldsymbol{\theta}, \boldsymbol{\theta}_0, \lambda) \in \Lambda$, for k_2.

Given that the divergencies are bounded from below it must happen that if $\lambda^\star \in C$ then $\lambda^\star = \lambda^\star_C$.

On the contrary, if $\boldsymbol{\lambda}^\star \in C'$, the assumptions implies that $k(\boldsymbol{\theta}_0, \boldsymbol{\lambda}^\star \mid \boldsymbol{\theta}, \lambda) \geq k(\boldsymbol{\theta}, \boldsymbol{\lambda} \mid \boldsymbol{\theta}_0, \lambda^\star)$.

B.3 Corollary 1

Let $\{p\,(\boldsymbol{x} \mid \boldsymbol{\psi}),\ \boldsymbol{x} \in \mathcal{X},\ \boldsymbol{\psi} \in \Psi\}$ be a probability model belonging to the exponential family. It is easy to prove (Gutierrez-Peña, 1992; Robert, 1996) that

$$\begin{aligned} k(\boldsymbol{\psi}_2 \mid \boldsymbol{\psi}_1) &= \int p\,(\boldsymbol{x} \mid \boldsymbol{\psi}_1)\, \log \frac{p\,(\boldsymbol{x} \mid \boldsymbol{\psi}_1)}{p\,(\boldsymbol{x} \mid \boldsymbol{\psi}_2)}\, \mathrm{d}\boldsymbol{x} \\ &= M(\boldsymbol{\psi}_1) - M(\boldsymbol{\psi}_2) + (\boldsymbol{\psi}_2^t - \boldsymbol{\psi}_1^t)\nabla M(\boldsymbol{\psi}_1)\,, \end{aligned}$$

where $M(\boldsymbol{\psi}) = \log a(\boldsymbol{\psi})$ y $\nabla M(\boldsymbol{\psi}) = \partial M(\boldsymbol{\psi})/\partial \boldsymbol{\psi}$;.

Let $\boldsymbol{\psi} = \{\boldsymbol{\theta}, \boldsymbol{\lambda}\}$, $\Psi = \Theta \bigcup \Lambda$, be an ordered parameterisation where $\boldsymbol{\theta}$ is the parameter of interest and $\boldsymbol{\lambda}$ the nuisance parameter and let $k_1 = k(\boldsymbol{\theta}_0, \lambda_0 \mid \boldsymbol{\theta}, \lambda)$ and $k_2 = k(\boldsymbol{\theta}, \boldsymbol{\lambda} \mid \boldsymbol{\theta}_0, \lambda_0)$, the KL divergencies.

Given that $a(\boldsymbol{\theta})$ is a convex function and that $\log z$ is monotone, then $M(\boldsymbol{\theta})$ is convex. Similarly, as the KL divergencies are linear combinations of $M(\cdot)$, both functions are convex and also bounded from below by 0. Two alternatives may happen

(*i*) Either $k_i \leq k_j$ for all $\{\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\theta}_0, \boldsymbol{\lambda}_0\} \in \Theta^2 \times \Lambda^2$

(*ii*) or $k_i \leq k_j$ in $C \subset \Theta^2 \times \Lambda^2$ y $k_i \geq k_j$ in C'.

The proof is thus analogous to that of Theorem 1.

References

Bartlett, M. (1957). A comment on D. V. Lindley's statistical paradox, *Biometrika*, **44**, 533–534.

Bayarri, M. J. (1981). Inferencia bayesiana sobre el coeficiente de correlación en una población normal bivariante, *Trabajos de Estadística e Investigación Operativa*, **32** (3), 18–31.

Berger, J. O. and Bernardo, J. M. (1992). On the development of reference priors, *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.), Oxford: University Press, pp. 35–60.

Berger, J. O. and Pericchi, L. R. (2001). Objective Bayesian methods for model selection: Introduction and comparison, *Model Selection*, *Lecture Notes*, vol. 38 (P. Lahiri, ed.), Institute of Mathematical Statistics, pp. 135–207. (with discussion).

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference, *J. Roy. Statist. Soc. B*, **41** (2), 113–147.

Bernardo, J. M. and Juárez, M. A. (2003). Intrinsic estimation, *Bayesian Statistics 7* (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds.), Oxford: University Press, pp. 465–475.

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

Bernardo, J. M. and Smith, A. F. (1994). *Bayesian Theory*, Chichester: John Wiley.

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population, *Biometrika*, **10**, 507–521.

Gutierrez-Peña, E. (1992). Expected logarithmic divergence for exponential families, *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.), Oxford university press, pp. 669–674.

Hand, D. J., Daly, F., Lunn, A. D., McConway, K. J. and Ostrowski, E. (1994). *A Handbook of small data sets*, London: Chapman and Hall.

Jeffreys, H. (1961). *Theory of Probability*, Oxford: University Press, third ed.

Juárez, M. A. (2004). *Objective Bayes methods for estimation and hypothesis testing*, Unpublished PhD thesis, Universitat de Valencia.

Lehmann, E. L. (1986). *Testing Statistical Hypotheses*, New York: John Wiley, second ed.

Lindley, D. V. (1957). A statistical paradox, *Biometrika*, **44**, 187–192.

Lindley, D. V. (1972). *Introduction to Probability and Statistics from a Bayesian Viewpoint, Part 2*, Cambridge: Cambridge University Press.

O'Hagan, A. (1997). Properties of intrinsic and fractional Bayes factors, *Test*, **6**, 101–118.

Pearson, E. S. and Kendall, M. G. (1970). *Studies in the History of Statistics and Probability*, London: Griffin.

Pearson, K. and Lee, A. (1903). On the laws of inheritance in Man: I. Inheritance of Physical Characters, *Biometrika*, **2**, 357–462.

Robert, C. P. (1996). Intrinsic losses, *Theory and Decisions*, **40**, 191–214.

Robert, C. P. and Caron, N. (1996). Noninformative Bayesian testing and neutral Bayes factors, *Test*, **5**, 411–437.