CORRELATION IN A FORMAL BAYES FRAMEWORK

by

Anirban DasGupta*
Purdue University

George Casella*
Cornell University

Mohan Delampady
Purdue University and
Indian Statistical Institute

Herman Rubin*
Purdue University

William E. Strawderman*
Rutgers University

Technical Report #98-04

Department of Statistics
Purdue University
West Lafayette, IN USA

April 1999
Revised January 2000

---

* Research of these authors was supported by the National Science Foundation and the National Security Agency.

The Canadian Journal of Statistics
Vol. 28, No. ?, 2000, Pages ???-???
La revue canadienne de statistique

Correlation in a formal Bayes framework

Anirban DASGUPTA, George CASELLA, Mohan DELAMPADY,
Christian GENEST, Herman RUBIN and William E. STRAWDERMAN

Key words and phrases: Bayes; Bayes risk; BLUE; Correlation; MLE; Posterior probability; Prior; P value; Robust estimation; Unbiased estimate.
AMS 1991 subject classifications: Primary 62A15, secondary 62C05.

ABSTRACT

The authors study the marginal correlation between a parametric function and an estimate or, more generally, the marginal correlation between two functions of both the parameter and the data. They give closed form expressions for these marginal correlations and use them to derive various connections to other notions of inference such as maximum likelihood, admissibility, and unbiasedness. They also obtain a general nonparametric upper bound on Bayes risks. Their results are illustrated with examples.

RÉSUMÉ

Les auteurs étudient la corrélation marginale entre une fonction paramétrique et une estimation ou, plus généralement, la corrélation marginale entre deux fonctions dépendant à la fois d'un paramètre et des observations. Ils montrent comment calculer ces corrélations et relient ce concept à d'autres notions d'inférence telles que l'estimation à vraisemblance maximale, l'admissibilité et l'absence de biais. Ils proposent également un majorant non paramétrique pour le risque bayésien. Leurs résultats sont illustrés par des exemples.

1. INTRODUCTION

The popularity of the Pearson correlation coefficient in statistical methodology is well documented. From regression to principal components or observational studies, it is frequently the measure of choice to assess the strength of association. In this article, we present the Pearson correlation coefficient as a binding theme to connect together various approaches to statistical inference in interesting ways. However, that is not all we do. We also show that even innocuous properties of the Pearson correlation coefficient lead to useful and substantial developments in mathematical statistics, particularly Bayesian statistics.

Formally, we begin with an observable X distributed as $f(x|\theta)$ and the parameter θ distributed according to a prior π. There is then a joint probability distribution which we call P. All the developments in the article follow from consideration of the Pearson correlation coefficient between two functions $g(X, \theta)$ and

$h(X, \theta)$ under the probability distribution P. They do not necessarily have to be functions of both X and θ; it is interesting, e.g., to talk about the correlation between $g(X, \theta) = \theta$, and $h(X, \theta) = \delta(X)$, a natural estimate of θ. One can also consider the correlation between two natural estimates of θ, or some general parametric function of θ. As we shall see, there will be cases of interest in which f, g are indeed each functions of both X and θ also.

In Section 2, we give a general formula for the correlation between a parameter θ and an estimate $\delta(X)$; no particular probabilistic setup is assumed. This formula simplifies neatly for unbiased and Bayes estimates. We then show that a number of implications result from these formulas for Bayes and unbiased estimates, including a noteworthy property of Dirichlet process priors for a CDF on the real line.

In Section 3, we show various connections of the Pearson correlation coefficient to popular notions in statistical inference, such as maximum likelihood estimation, unbiasedness and robust estimation.

In Section 4, we move onto correlation between two functions of X. Examples are when one of them is an unbiased estimate, and the other a Bayes estimate, or the MLE, etc. Using certain results derived here, we obtain a general nonparametric upper bound on Bayes risk for any decision problem with squared error loss. We then relate this to the lower bound of Brown & Gajek (1990) (cf. also Sato & Akahira 1996) lower bound on Bayes risk; it turns out the two bounds coincide in the normal-normal case. Further, we check the sharpness of our upper bound in a test case that was also investigated by Brown and Gajek: the estimation of a bounded normal mean by using the Bickel-Levit prior.

Two additional examples close Section 4. One example illustrates the relation between P-values and posterior probabilities and the other explores the connection between the sample mean and the BLUE of the stationary mean of Gaussian time series. Some final brief remarks on the practical uses of our approach are made in Section 5. All proofs are given in an appendix.

The following general notation is used in the sequel: π denotes a prior for the parameter θ, m is the marginal of X, E_θ refers to the conditional expectation given θ, cov_π stands for the covariance under π of two functions of θ, ρ_π represents the correlation, and ρ_P denotes the correlation between two functions of X and θ. Similar interpretations apply to variances via the notation var_π and var_P. Also, as usual, $r(\pi, \delta)$ means the Bayes risk of the estimate $\delta(X)$, and $r(\pi)$ the Bayes risk of the Bayes estimate $\delta_\pi(X)$.

2. CORRELATION BETWEEN A PARAMETER AND AN ESTIMATE

Our first proposition gives a general correlation formula between a parameter and an estimate in a general probability space $L_2(\mathcal{X} \otimes \Theta, P)$. The formula, which simplifies considerably when applied to unbiased estimates and to the Bayes estimate corresponding to the prior induced by P, is illustrated in two special cases.

PROPOSITION 1. *Let $\delta(X)$ be any estimate of θ. Then the correlation between θ and δ equals*

$$\rho_P(\theta, \delta) = \frac{V(\pi) + \text{cov}_\pi\{\theta, b(\theta)\}}{\sqrt{V(\pi)[r(\pi, \delta) + \ \text{var}_\pi\{\theta + b(\theta)\} - E_\pi\{b^2(\theta)\}]}} \tag{1}$$

where $V(\pi) = \text{var}_\pi(\theta)$, and $b(\theta) = E_{X|\theta}\{\delta(X)\} - \theta$. In particular,

$$\rho_P(\theta, \delta) = \sqrt{\frac{V(\pi)}{r(\pi, \delta) + V(\pi)}}. \tag{2}$$

when $\delta(X)$ is an unbiased estimate of θ. For the Bayes estimate $\delta_\pi(X)$, one finds

$$\rho_P(\theta, \delta_\pi) = \sqrt{1 - \frac{r(\pi)}{V(\pi)}}. \tag{3}$$

EXAMPLE 1 (Estimation of a normal mean). Let $X_1, \ldots, X_n$ be iid $N(\theta, 1)$ with sufficient statistic $\bar{X} \sim N(\theta, 1/n)$, and let θ have a prior π in the (very large) class

$$\mathcal{C} = \{\pi\colon\ E_\pi(\theta) = 0,\ V(\pi) = 1\}.$$

Using (3) and Brown's identity on Bayes risks (Brown 1971, 1985), one finds

$$1 - \rho_P^2(\theta, \delta_\pi) = \frac{r(\pi)}{V(\pi)} = r(\pi) = \frac{1}{n} - \frac{1}{n^2} I(m), \tag{4}$$

where $m(x) = \int \sqrt{n/2\pi} e^{-n(x-\theta)^2/2} d\pi(\theta)$ and $I(\cdot)$ denotes the Fisher information functional.

Since the variance of the marginal distribution of X equals $1 + 1/n$ for any π in $\mathcal{C}$, one has

$$\inf_{\pi \in \mathcal{C}} I(m) = \frac{n}{n+1}, \tag{5}$$

as normal distributions have the Fisher information minimization property when the variance is fixed; cf. Huber (1981). From (4) and (5), therefore,

$$\sup_{\pi \in \mathcal{C}} \{1 - \rho_P^2(\theta, \delta_\pi)\} = \frac{1}{n} - \frac{1}{n(n+1)} = \frac{1}{n+1} \quad \Rightarrow \quad \inf_{\pi \in \mathcal{C}} \rho_P(\theta, \delta_\pi) = \sqrt{\frac{n}{n+1}}.$$

On the other hand, as the operator $I(\cdot)$ is convex (again, see Huber 1981), one has

$$I(m) \leq \int I\{N(\theta, 1/n)\} d\pi(\theta) = n.$$

But, if one considers an element of $\mathcal{C}$ of the form $p1(-c) + p1(c) + (1-2p)1(0)$, where $c = 1/\sqrt{2p}$ and $1(\cdot)$ denotes point mass, then for the corresponding marginal m_p, $I(m_p) \to n$ when $p \to 0$. Together, these imply

$$\inf_{\pi \in \mathcal{C}} \{1 - \rho_P^2(\theta, \delta_\pi)\} = \frac{1}{n} - \frac{1}{n^2} \sup_{\pi \in \mathcal{C}} I(m) = \frac{1}{n} - \frac{n}{n^2} = 0,$$

whence $\sup_{\pi \in \mathcal{C}} \rho_P(\theta, \delta_\pi) = 1$.

These facts together show that $\rho_P(\theta, \delta_\pi)$ converges to 1 as $n \to \infty$ *uniformly* over π in $\mathcal{C}$. This is interesting, considering that $\mathcal{C}$ is a very large class of priors. This uniform convergence would not necessarily hold true for other classes $\mathcal{C}$.

EXAMPLE 2. (A property of the Dirichlet process). Consider the problem of estimating a distribution function F nonparametrically, given a random sample $X_1, \ldots, X_n$ from F. Let π denote the Dirichlet process prior on F with parameter α, a finite measure on $\mathbb{R}$. Let F_n be the empirical distribution function. Then it is easily seen that, for any fixed x,

$$E_P\{F_n(x)F(x)\} = E_P[F(x)E_F\{F_n(x)|F(x)\}] = E_\pi\{F^2(x)\}$$

whence $\mathrm{cov}_P\{F_n(x), F(x)\} = \mathrm{var}_\pi\{F(x)\}$ and

$$\begin{aligned} \mathrm{var}_P\{F_n(x)\} &= E_\pi[\mathrm{var}\{F_n(x)|F(x)\}] + \mathrm{var}_\pi[E\{F_n(x)|F(x)\}] \\ &= \frac{1}{n}E_\pi[F(x)\{1-F(x)\}] + \mathrm{var}_\pi\{F(x)\}. \end{aligned} \tag{6}$$

Also, it is well known (cf. Ferguson 1974) that under π, $F(x)$ is distributed as Beta(α, β) with $\alpha = \alpha(-\infty, x]$ and $\beta = \alpha(I\!R) - \alpha(-\infty, x]$ for any x. Accordingly,

$$\begin{aligned} \rho_P\{F_n(x), F(x)\} &= \frac{\mathrm{var}_\pi\{F(x)\}}{\sqrt{\mathrm{var}_\pi\{F(x)\}[\mathrm{var}_\pi\{F(x)\} + \frac{1}{n}E[F(x)\{1-F(x)\}]]}} \\ &= \left[1 + \frac{1}{n}\frac{E[F(x)\{1-F(x)\}]}{\mathrm{var}_\pi\{F(x)\}}\right]^{-1/2} \\ &= \left[1 + \frac{1}{n}\{\alpha(-\infty,x] + \alpha(I\!R) - \alpha(-\infty,x]\}\right]^{-1/2} \\ &= \{1 + \alpha(I\!R)/n\}^{-1/2}, \end{aligned}$$

which is totally free of x. This property is possibly characteristic of the Dirichlet process. Note also the convergence to 1 of the above as $n \to \infty$ for any base measure α.

3. VARIOUS CONNECTIONS

Our second result highlights some of the connections which the correlation formulas of Section 2 allow to make to other established methods and concepts of statistics. In the process, we also mention a very positive but somewhat obscure property of maximum likelihood estimates (MLE). As an application, we investigate the correlation between robust estimators and the parameter they estimate.

PROPOSITION 2. *Under the joint distribution of X and θ,*

a) *If the likelihood function is unimodal for every x, then the MLE $\hat\theta$ has the property that $\rho_P(\theta, \hat\theta) \geq 0$ under any prior π.*

b) *The correlation between θ and any unbiased estimate is always nonnegative and strictly positive unless the prior is degenerate.*

c) *If θ has a UMVUE δ_U, then $\rho_P(\theta, \delta_U) \geq \rho_P(\theta, \delta)$ for any other unbiased estimate δ and the inequality is strict if δ_U is the unique UMVUE and if π gives support to the entire parameter space.*

d) *$\rho_P(\theta, \delta_\pi)$ is also always nonnegative.*

e) *If the Bayes estimate of θ is not a constant, then $\rho_P(\theta, \delta_\pi) > 0$.*

f) *Let ξ_2 be a more informative experiment than ξ_1 in the Blackwell sense; then under any prior π, $\rho_{P,\xi_2}(\theta, \delta_\pi) \geq \rho_{P,\xi_1}(\theta, \delta_\pi)$, i.e., the parameter is more correlated with a Bayes estimate under a more informative experiment.*

g) *Let X have a distribution with a density or pmf $f(x|\theta)$ which has monotone likelihood ratio in X. If $\delta(X)$ is any admissible estimate of θ, then $\rho_P\{\theta,\ \delta(X)\} \geq 0$.*

h) *In a location parameter model, i.e., if* $X \sim F(x-\theta)$ *and* $\theta \sim \pi$*, and* F*,* π *belong respectively to specified classes* $\mathcal{C}_1, \mathcal{C}_2$*, the criterion of maximizing the minimum correlation* $\inf_{F \in \mathcal{C}_1, \pi \in \mathcal{C}_2} \rho_P\ (\theta, \delta)$ *over all estimates* δ *which are unbiased under each* F *in* $\mathcal{C}$ *is equivalent to the common minimax criterion of minimizing* $\sup_{\mathcal{C}_1} \operatorname{var}_F(\delta)$.

The proof of this result is given in the appendix. As is seen there, part c) is a consequence of the following little-known property of MLE's, which we state here for the record.

PROPOSITION 3. *Let* $X|\theta$ *have a density* $f(x|\theta)$ *with respect to some dominating measure* μ*. Suppose the Cramér-Rao regularity conditions hold and the likelihood function is unimodal for every* x*. Then the MLE* $\hat{\theta} = \hat{\theta}(X)$ *has a distribution that is stochastically increasing in the parameter* θ*. In particular,* $E_\theta(\hat{\theta})$ *is increasing in* θ*, if it exists.*

EXAMPLE 3. A vast literature has now accumulated on robust estimation of a location parameter; cf., e.g., Staudte & Sheather (1990). The common criterion is the maximum variance of the asymptotic distribution of a candidate estimate δ. Here, we examine robustness from the point of view of correlation, using $\rho_P(\theta, \delta)$.

To this end, let us consider the sample mean $\bar{X}$ and select three robust estimators, namely, the median, the 12.5% trimmed mean, and the Huber M estimate with score function $\psi(x) = |x|$ if $|x| \leq k$ and k if $|x| > k$. For the latter, we use $k = 1.399$, which corresponds to the asymptotic minimax estimate for 5% contamination.

For sampling models, we take t distributions with $m > 0$ degrees of freedom, and as priors, we use t distributions as well, with degrees of freedom $\alpha > 2$. Note that these choices allow a very broad choice of tails in both the sampling model and the prior. Another point: if $m \leq 2$, the sample mean $\bar{X}$ does not have a finite variance, but this will just mean that $\rho_P(\theta, \bar{X})$ is zero if $m \leq 2$ and this fact will be included in the comparative evaluation of $\bar{X}$. We will not refer to this issue in the following again.

If δ is any of the four estimates, and if $V_F(\delta)$ denotes its variance at the sampling model F, then

$$\rho_P(\theta, \delta) = \sqrt{\frac{V(\pi)}{V(\pi) + V_F(\delta)}}, \tag{7}$$

but in three of the four cases, an exact expression simply does not exist. We thus make the natural approximation $V_F(\delta) \approx \sigma_F^2(\delta)/n$, where $\sigma_F^2(\delta)$ comes from known central limit theorems for each of these four estimates, viz.

$$\sqrt{n}\{\delta(X) - \theta\} \xrightarrow{\mathcal{L}} N(0, \sigma_F^2(\delta)).$$

Equation (7) is then replaced by

$$\rho(\alpha, m, n) = \sqrt{\frac{\frac{\alpha}{\alpha-2}}{\frac{\alpha}{\alpha-2} + \frac{\sigma^2(m,\delta)}{n}}}, \tag{8}$$

where $\sigma^2(m, \delta)$ means $\sigma_F^2(\delta)$ evaluated at the sampling model $F = t$ with m degrees of freedom. Fortunately, expressions for $\sigma^2(m, \delta)$ are available from general formulas available in Chapter 5 of Lehmann (1983):

$$\sigma^2(m,\bar{X}) = \frac{m}{m-2} \text{ for } m>2, \quad \sigma^2(m,\text{median}) = \frac{1}{4f_m^2(0)} = \frac{m}{4}B^2\left(\frac{m}{2},\frac{1}{2}\right)$$

$$\sigma^2(m,\text{trimmed mean}) = \frac{32}{9}\left[\int_0^{F_m^{-1}(.875)} x^2 f_m(x)dx + .125\{F_m^{-1}(.875)\}^2\right],$$

and

$$\sigma^2(m,\text{Huber estimate}) = \frac{2\left[\int_0^{1.399} x^2 f_m(x)dx + 1.957\{1-F_m(1.399)\}\right]}{\{2F_m(1.399)-1\}^2}.$$

In the above, $B(\cdot,\cdot)$ denotes the Beta function, f_m stands for the density of a t distribution with m degrees of freedom, F_m for its cdf, and F_m^{-1} for its quantile function.

TABLE 1: Average value of $\rho(\alpha,m,n)$ when $n=20$, α ranges from 2 to 20, and m ranges over three separate intervals.

Estimator	$(0,3]$	$[1,60]$	$[60,120]$
Mean	.307	.956	.980
Median	.915	.969	.969
Trimmed mean	.797	.977	.979
Huber estimate	.881	.978	.979

Table 1 reports the value of $\rho(\alpha,m,n)$ when $n=20$, α ranges from 2 to 20, and m ranges over three separate intervals, viz. $0<m\leq 3$ (extremely flat), $1\leq m\leq 60$ (mixed tails of great variety), and $60\leq m\leq 120$ (thin tails). On the basis of criterion (8), the median appears to be a good choice when the tail is believed to be extremely flat, while for mixed tails, the Huber estimate is a more balanced choice. For thin tails, there isn't much discrepancy, and this is actually a bit of a surprise.

4. CORRELATION BETWEEN TWO ESTIMATES

We now investigate the correlation between estimates obtained from different paradigms of inference. Our first result gives conditions under which Bayes estimates are positively correlated with maximum likelihood estimates.

PROPOSITION 4. *If the likelihood function $\ell(\theta|x)$ is unimodal for every x, then the maximum likelihood estimate $\hat{\theta}$ has the property that $\rho_P(\delta_\pi,\hat{\theta})\geq 0$ under any prior π.*

Next, we examine the correlation between Bayes and unbiased estimates.

PROPOSITION 5. *For any unbiased estimate δ of θ,*

$$\rho_P(\delta,\delta_\pi) = \frac{V(\pi)}{\sqrt{\{V(\pi)-r(\pi)\}[V(\pi)+E_\pi \mathrm{var}_\theta\{\delta(X)\}]}}.$$

This quantity is non-negative under any prior π and for any unbiased estimate, and it is strictly positive if π is nondegenerate. Furthermore, $\rho_P(\delta, \delta_\pi)$ is the largest when $\delta(X)$ is the UMVUE, if one exists.

As shown in the appendix, the above expression for $\rho_P(\delta, \delta_\pi)$ leads to the following general upper bound on Bayes risks.

PROPOSITION 6. *For any prior π, $r(\pi)$ admits the upper bound*

$$r(\pi) \leq \frac{V(\pi) \cdot E_\pi \mathrm{var}_\theta\{\delta(X)\}}{V(\pi) + E_\pi \mathrm{var}_\theta\{\delta(X)\}},$$

where $\delta(X)$ can be any unbiased estimate of θ.

Suppose, in particular, that $f(x|\theta)$ is a member of the one-parameter exponential family with θ as the expectation parameter. If the unbiased estimate $\delta(X)$ in Proposition 6 is chosen to be the UMVUE, then $\mathrm{var}_\theta\{\delta(X)\} \equiv 1/I(\theta)$ and hence in this case, one has the following special upper bound.

COROLLARY 1. *If $f(x|\theta)$ is in the one-parameter exponential family with θ as the expectation parameter, then for any prior π,*

$$r(\pi) \leq \frac{CV(\pi)}{C + V(\pi)},$$

where $C = E_\pi\{I^{-1}(\theta)\}$.

It is instructive to compare the latter upper bound with the lower bound of Brown & Gajek (1990), who showed $r(\pi) \geq C^2/(C + D)$, where

$$D = \int \frac{[\{\pi(\theta)^{-1} I(\theta)\}']^2}{\pi(\theta)} d\theta$$

for priors with a density. Our upper bound does not need a prior density.

EXAMPLE 4 Consider the problem of estimating a bounded normal mean by using the prior density $\pi(\theta) = L^{-1}\cos^2\{\theta\pi/(2L)\})$, $|\theta| \leq L$; cf. Bickel (1981), Levit (1980), and Borovkov & Sakhanienko (1980). In this case, C is exactly 1 and $V(\pi) = L^2(1/3 - 2/\pi^2)$, so that the upper bound of Corollary 1 is then given by

$$\frac{V(\pi)}{1 + V(\pi)} = \frac{L^2}{L^2 + \frac{3\pi^2}{\pi^2 - 6}}. \tag{9}$$

This is compared in Table 2 with the Borovkov-Sakhanienko (B-S) lower bound $L^2/(L^2 + \pi^2)$ and the stronger bound (B-G) of Brown & Gajek (1990). It is clear that $r(\pi)$ is close to bound (9) in this test case.

TABLE 2: Efficacy of Bound (9)

L	B-S bound	B-G bound	Exact $r(\pi)$	Bound (9)
.5	.0247	.0314	.0316	.0316
1	.0920	.1135	.1156	.1156
2	.2884	.3305	.3427	.3433
3	.4770	.5177	.5375	.5405
5	.7170	.7399	.7578	.7657
10	.9102	.9150	.9214	.9289
20	.9759	.9765	.9779	.9812

4.1. Correlation in the One-Sided Testing Problem.

The typical frequentist assessment of a null hypothesis is the P-value and the typical Bayesian assessment is the posterior probability $P(H_0|X)$. A substantial literature on the reconcilability of these two answers grew in the 1980s; cf., e.g., Berger & Sellke (1987), Casella & Berger (1987), and Oh & DasGupta (1999). Casella & Berger (1987) show that for the one-sided testing problem for a univariate $N(\theta, \sigma^2)$ mean, the frequentist and the robust Bayesian answers often coincide. For example, for testing $H_0: \theta \leq 0$ vs $H_1: \theta > 0$ when θ is a normal mean, the P-value and the minimum value of $P(H_0|\text{data})$ over all $N(0, \tau^2)$ priors are equal. Here, we show that under any $N(0, \tau^2)$ prior, and for any σ^2, the marginal correlation between the P-value and $P(H_0|X)$ is larger than $\sqrt{3/\pi}$ ($\approx$ .977205), a surprisingly uniformly high correlation.

For testing $H_0: \theta \leq 0$ versus $H_1: \theta > 0$ using $X \sim N(\theta, \sigma^2)$, consider the marginal correlation between the P-value $p(X)$ and the posterior probability of the null hypothesis $P_\pi(H_0|X)$. Then $p(X) = 1 - \Phi(X/\sigma)$ is a decreasing function of X, and $P(H_0|X)$ is also a decreasing function of X. Therefore, $\rho_P\{p(X), P(H_0|X)\} \geq 0$ for any prior π. Now consider, in particular, the class of all normal priors with mean 0 and variance τ^2. Then $\theta|X$ is normal with mean $\tau^2 X/(\sigma^2 + \tau^2)$ and variance $\sigma^2\tau^2/(\sigma^2 + \tau^2)$, the marginal distribution of X is normal with mean 0 and variance $\sigma^2 + \tau^2$, and

$$P(H_0|X) = 1 - \Phi\left(\frac{\tau^2 X}{\sigma^2 + \tau^2} \frac{\sqrt{\sigma^2 + \tau^2}}{\sigma\tau}\right) = 1 - \Phi\left(\frac{\tau X}{\sigma\sqrt{\sigma^2 + \tau^2}}\right).$$

Therefore,

$$\begin{aligned}
&\text{cov}_P\left\{p(X), P_\pi(H_0|X)\right\} \\
&= E_P\left\{\Phi\left(\frac{X}{\sigma}\right)\Phi\left(\frac{\tau X}{\sigma\sqrt{\sigma^2 + \tau^2}}\right)\right\} - E_P\left\{\Phi\left(\frac{X}{\sigma}\right)\right\} E_P\left(\frac{\tau X}{\sigma\sqrt{\sigma^2 + \tau^2}}\right) \\
&= E\left(\Phi\left(\frac{\sqrt{\sigma^2 + \tau^2}}{\sigma} Z\right)\Phi\left(\frac{\tau}{\sigma} Z\right)\right) - E\left\{\Phi\left(\frac{\sqrt{\sigma^2 + \tau^2}}{\sigma} Z\right)\right\} E\left\{\Phi\left(\frac{\tau}{\sigma} Z\right)\right\},
\end{aligned}$$

where $Z \sim N(0, 1)$.

Now notice that given independent copies Z^*, Z^{**} of Z and arbitrary constants a and b, one has $E\{\Phi(aZ)\} = P(Z^* \leq aZ) = P(Z^* - aZ \leq 0) = 1/2$ and

$$\begin{aligned} E\{\Phi(aZ)\Phi(bZ)\} &= P(Z^* - aZ \leq 0, Z^{**} - bZ \leq 0) \\ &= \frac{1}{4} + \frac{1}{2\pi}\sin^{-1}\left(\frac{ab}{\sqrt{(1+a^2)(1+b^2)}}\right), \end{aligned}$$

by a well known formula for bivariate normal distributions (cf. Anderson 1984, pp. 51-52).

Using these facts, one may then see that

$$\mathrm{cov}_P\{p(X), P(H_0|X)\} = \frac{1}{2\pi}\sin^{-1}\left(\frac{\tau}{\sqrt{\tau^2+2\sigma^2}}\right)$$

and

$$\mathrm{var}_P\{p(X)\} = \frac{1}{2\pi}\sin^{-1}\left(\frac{\sigma^2+\tau^2}{\tau^2+2\sigma^2}\right), \quad \mathrm{var}_P\{p(H_0|X)\} = \frac{1}{2\pi}\sin^{-1}\left(\frac{\tau^2}{\tau^2+\sigma^2}\right).$$

Letting $v = \tau^2/(\tau^2+\sigma^2)$, one can then conclude that

$$\rho_P\{p(X), P_\pi(H_0|X)\} = \frac{\sin^{-1}\left(\sqrt{\frac{v}{2-v}}\right)}{\sqrt{\sin^{-1}(v)\sin^{-1}\left(\frac{1}{2-v}\right)}} \equiv f(v) \quad \text{(say.)}$$

Since $\inf_{0\leq v\leq 1} f(v) = f(0) = \sqrt{3/\pi}$, the argument is complete.

4.2. Effect of Dependence.

Although independence of sample data is a common assumption in statistical analysis, many popular procedures get adversely affected by dependence among the observations. There are some exceptions; for instance, after the classic work of Grenander & Rosenblatt (1957), the sample mean $\bar{X}$ remains the standard estimate for the mean of a stationary time series although the BLUE is theoretically better. In fact, recent research has shown that $\bar{X}$ is a fine estimate for the stationary mean even under long range dependence; cf. Beran (1994) and Zang & DasGupta (1998). Below we present the marginal correlation between $\bar{X}$ and BLUE for the general stationary Gaussian case first. The general expression leads to some general bounds. We will then complete the example with the $AR(1)$ case as a specific illustration.

Suppose then $n > 1$, and let $X_1, \ldots, X_n$ be jointly normal with a common mean μ and covariance matrix Σ. The marginal variances of $\bar{X}$ and the BLUE are $V(\pi) + 1'\Sigma 1/n^2$ and $V(\pi) + 1/1'\Sigma^{-1}1$. For their marginal covariance, note that

$$\begin{aligned} E_P(\bar{X}\cdot \text{BLUE}) &= E_\pi E_\mu(\bar{X}\cdot \text{BLUE}) = E_\pi\{\mathrm{cov}_\mu(\bar{X}, \text{BLUE}) + \mu^2\} \\ &\stackrel{*}{=} E_\pi\{\mathrm{var}_\mu(\text{BLUE}) + \mu^2\} = E_\pi(\mu^2) + 1/1'\Sigma^{-1}1, \end{aligned}$$

where the relation $\stackrel{*}{=}$ is justified by Basu's theorem (Basu 1955), given that the BLUE is complete and sufficient. Hence, $\mathrm{cov}_P(\bar{X}, \text{BLUE}) = V(\pi) + 1/1'\Sigma^{-1}1$, giving

$$\rho_P(\bar{X}, \text{BLUE}) = \sqrt{\frac{V(\pi) + 1/1'\Sigma^{-1}1}{V(\pi) + 1'\Sigma 1/n^2}}. \tag{10}$$

This identity leads to a universal lower bound on $\rho_P(\bar{X}, \text{BLUE})$, free of the choice of the prior π. From the Kantorovich inequality (cf., e.g., Rao 1973, p. 74), it easily follows that $1/1'\Sigma^{-1}1 \geq (41'\Sigma 1)\lambda_1\lambda_n/\{n^2(\lambda_1+\lambda_n)^2\}$, where λ_1, λ_n are the two extreme eigenvalues of Σ. Hence, from (10),

$$\rho_P(\bar{X}, \text{BLUE}) \geq \sqrt{\frac{V(\pi) + \frac{4\lambda_1\lambda_n}{(\lambda_1+\lambda_n)^2}\ \frac{1'\Sigma 1}{n^2}}{V(\pi) + \frac{1'\Sigma 1}{n^2}}} \geq \sqrt{\frac{\frac{4\lambda_1\lambda_n}{(\lambda_1+\lambda_n)^2}\ \frac{1'\Sigma 1}{n^2}}{\frac{1'\Sigma 1}{n^2}}} = 2\sqrt{\frac{\lambda_1\lambda_n}{\lambda_1+\lambda_n}},$$

a simple global bound valid for any prior π. The broad moral of this inequality is that $\bar{X}$ and the BLUE are marginally well correlated unless the dependence is too strong.

Special correlation structures are of practical interest. Here we investigate the $AR(1)$ case with the autocovariance function $\sigma_{ij} = \rho^{|i-j|}/(1-\rho^2)$. Fortunately, Σ^{-1} is then computable in closed form: it is tridiagonal, having the two extreme elements equal to 1 in the main diagonal, the remaining elements of the main diagonal being $1+\rho^2$, and the elements in the diagonals immediately adjacent to the main diagonal being $-\rho$. Now, several lines of algebra yield $1'\Sigma^{-1}1 = n - 2\rho(n-1) + \rho^2(n-2)$ and

$$1'\Sigma 1 = \frac{n - 2\rho - n\rho^2 + 2\rho^{n+1}}{(1-\rho)^3(1+\rho)}.$$

Substitution into (10) shows that for the $AR(1)$ case,

$$\begin{aligned}
\rho_P(\bar{X}, \text{BLUE}) &\geq \sqrt{\frac{V(\pi) + \frac{1}{n-2\rho(n-1)+\rho^2(n-2)}}{V(\pi) + \frac{n-2\rho-n\rho^2+2\rho^{n+1}}{n^2(1-\rho)^3(1+\rho)}}} \\
&\geq \frac{n(1-\rho)^{\frac{3}{2}}\sqrt{1+\rho}}{\sqrt{\{n-2\rho(n-1)+\rho^2(n-2)\}\{n-2\rho-n\rho^2+2\rho^{n+1}\}}},
\end{aligned}$$

for any prior π.

Figure 1 plots this universal lower bound as a function of the autoregression parameter ρ for $n = 20$; the interesting thing is that uniformly in ρ, the lower bound is itself very high, about .95 from the plot. It reinforces conventional faith in $\bar{X}$ for this problem.

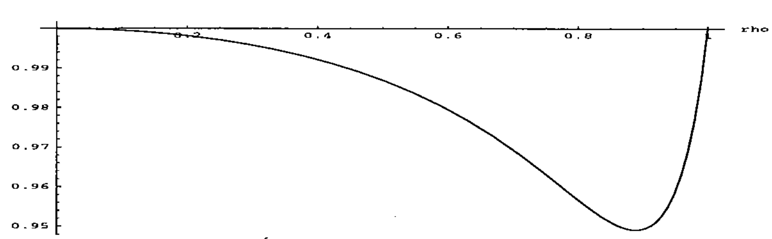

FIGURE 1. Lower Bound on $\rho_P(\bar{X}, \text{BLUE})$ under AR(1)

5. SUMMARY AND PRACTICAL DEVELOPMENTS

We have presented the Pearson correlation coefficient in a formal Bayesian framework as a theoretically useful theme to provide new insight and interpretations for established methods of statistics. Our general approach also has certain potentials for practical uses and development of actual new methodology. In Delampady *et al.* (1999), some of these ideas are developed to write proper default priors and to choose specific Bayes rules from a collection of Bayes rules.

Additional practical developments that may result from our approach include parallel theories of optimal design, choice of bandwidth in density estimation in finite samples, and choice of default models in various model selection problems. The attraction of our approach is that correlation is an omnibus measure of association, and calibrated. Whenever our approach leads to an identifiable answer, we can avoid use of a subjective criterion such as a technically convenient loss function.

APPENDIX

Proof of Proposition 1. We can assume without loss of generality that $E_\pi(\theta) = 0$. The covariance of θ and δ is then

$$\mathrm{cov}_P(\theta, \delta) = E_\pi\{\theta \cdot E_{X|\theta}\delta(X)\} = E_\pi\{\theta^2 + \theta b(\theta)\} = V(\pi) + \mathrm{cov}_\pi\{\theta, b(\theta)\},$$

while the marginal variance of δ is

$$\begin{aligned}\mathrm{var}_P\{\delta(X)\} &= E_\pi\{\mathrm{var}_\theta \delta(X)\} + \mathrm{var}_\pi[E_\theta\{\delta(X)\}] \\ &= E_\pi[E\{\delta(X) - \theta\}^2 - b^2(\theta)] + \mathrm{var}_\pi\{\theta + b(\theta)\} \\ &= r(\pi, \delta) + \mathrm{var}_\pi\{\theta + b(\theta)\} - E_\pi\{b^2(\theta)\},\end{aligned}$$

which proves (1). The latter immediately reduces to (2) when $b(\theta) \equiv 0$.

Next, let $b_\pi(\theta)$ denote $E_\theta\{\delta_\pi(X) - \theta\}$, the bias of $\delta_\pi(X)$. From DasGupta (1994), one then has

$$\begin{aligned}r(\pi) = r(\pi, \delta_\pi) &= E_P\{\delta_\pi(X) - \theta\}^2 \\ &= E_P[\{\delta_\pi(X) - \theta\}\delta_\pi(X)] - E_P[\{\delta_\pi(X) - \theta\}\theta] \\ &= E_P[\delta_\pi(X) E_{\theta|X}\{\delta_\pi(X) - \theta\}] - E_\pi[\theta E_\theta\{\delta_\pi(X) - \theta\}] \\ &= -E_\pi\{\theta b_\pi(\theta)\} = -\mathrm{cov}_\pi\{\theta, b_\pi(\theta)\},\end{aligned}$$

since $E_\pi\{b_\pi(\theta)\} = 0$. In this case, the numerator of (1) thus reduces to $V(\pi) - r(\pi) = \mathrm{var}_P\{\delta_\pi(X)\}$. Furthermore, one has

$$\mathrm{var}_\pi\{\theta + b_\pi(\theta)\} = V(\pi) + 2E_\pi\{\theta b_\pi(\theta)\} + E_\pi\{b_\pi^2(\theta)\},$$

from which the denominator of (1) also reduces to $\sqrt{V(\pi)\{V(\pi) - r(\pi)\}}$, thereby completing the derivation of (3).

Proof of Proposition 2.

a) First observe that for any estimate $\delta(x)$ with expectation $\mu(\theta) = E_\theta \delta(x)$,

$$\begin{aligned}\mathrm{cov}_P\{\theta, \delta(X)\} &= E_P\{\theta\delta(X)\} - E_\pi(\theta) E_P \delta(X) \\ &= E_\pi\{\theta\mu(\theta)\} - E_\pi(\theta) E_\pi \mu(\theta) \\ &= \mathrm{cov}_\pi\{\theta, \mu(\theta)\},\end{aligned}$$

and so $\rho_P(\theta,\delta) \geq 0$ if and only if $\rho_\pi\{\theta,\mu(\theta)\} \geq 0$. Now for the MLE, by virtue of Proposition 3, $\mu(\theta)$ is nondecreasing in θ if the likelihood function is unimodal for every x, so that $\rho_P(\theta,\hat{\theta}) \geq 0$.

b) and c) are both easily obtained from (2).

d) Consider the trivial estimate identically equal to $E_\pi(\theta)$. Its Bayes risk is $V(\pi)$ and therefore $r(\pi) \leq V(\pi)$, and so statement d) follows from (3).

e) This result follows from the same argument by observing that $r(\pi) < V(\pi)$ if the Bayes estimate is nonconstant.

f) Under ξ_2, the Bayes risk $r(\pi)$ would be smaller, and so (3) implies statement f).

g) First observe that under any prior π,

$$\begin{aligned} \mathrm{cov}_P\{\theta,\delta(X)\} &= E_P\{\theta\delta(X)\} - E_P(\theta)E_P\{\delta(X)\} \\ &= EE\{\theta\delta(X)|X\} - E_P\{\delta_\pi(X)\}E_P\{\delta(X)\} \\ &= E_P\{\delta_\pi(X)\delta(X)\} - E_P\{\delta_\pi(X)\}E_P\{\delta(X)\} \\ &= \mathrm{cov}_P\{\delta_\pi(X),\delta(X)\}. \end{aligned} \tag{11}$$

Next, from Karlin & Rubin (1956), a Bayes estimate $\delta_\pi(X)$ and an admissible estimate $\delta(X)$ are both nondecreasing functions of X under monotone likelihood ratio. Therefore, $\mathrm{cov}_P\{\delta_\pi(X),\delta(X)\} \geq 0$. Combining this with the above, statement g) follows.

h) From (2),

$$\rho_P^2(\theta,\delta) = \frac{V(\pi)}{V(\pi)+r(\pi,\delta)} = \frac{V(\pi)}{V(\pi)+\mathrm{var}_F(\delta)}$$

whence the result.

Proof of Proposition 3. We will need to show that $P_\theta\{\hat{\theta}(X) \leq c\}$ is a decreasing function of θ, for any fixed c. Consider first the case $\theta \geq c$. Then,

$$\begin{aligned} P_\theta\{\hat{\theta}(X) \leq c\} &= \int_{x:\ \hat{\theta}(x)\leq c} f(x|\theta)d\mu(x) \\ \Rightarrow \quad & \frac{d}{d\theta}P_\theta\{\hat{\theta}(X) \leq c\} = \int_{x:\theta(x)\leq c} \left\{\frac{d}{d\theta}f(x|\theta)\right\} d\mu(x). \end{aligned}$$

However, whenever x is such that $\hat{\theta}(x) \leq c$, $df(x|\theta)/d\theta \leq 0$ for any $\theta \geq c$ by the unimodality assumption (drawing a picture is helpful), and hence the integral is also non-positive, implying $dP_\theta\{\hat{\theta}(X) \leq c\}/d\theta \leq 0$, as needed. As the case $\theta \leq c$ can be treated similarly, the argument is complete.

Proof of Proposition 4. By Equation (11), $\rho_P(\delta_\pi,\hat{\theta}) \geq 0$ if and only if $\rho_P(\theta,\hat{\theta}) \geq 0$. Proposition 4 thus follows from Proposition 2.

Proof of Proposition 5. First use (11) and the assumption that $\delta(X)$ is unbiased to see that

$$\begin{aligned} \mathrm{cov}_P\{\delta_\pi(X),\delta(X)\} &= \mathrm{cov}_P\{\theta,\delta(X)\} = E_P\{\theta\delta(X)\} - E_\pi(\theta)E_P\{\delta(X)\} \\ &= E_\pi(\theta^2) - E_\pi^2(\theta) = V(\pi). \end{aligned}$$

Now $\text{var}_P\{\delta(X)\} = V(\pi) + E_\pi \text{var}_\theta\{\delta(X)\}$ and it was shown in the proof of Proposition 1 that $\text{var}_P\{\delta_\pi(X)\} = V(\pi) - r(\pi)$. Combining all this yields the desired conclusion.

Proof of Proposition 6. The proof has formal similarity to the derivation of the information inequality. Indeed,

$$\begin{aligned} \rho_P^2(\delta, \delta_\pi) \leq 1 \quad &\Rightarrow \quad V(\pi) \leq \{V(\pi) - r(\pi)\}\{V(\pi) + E_\pi \text{var}_\theta \delta(X)\} \\ &\Rightarrow \quad r(\pi)\{V(\pi) + E_\pi \text{var}_\theta \delta(X)\} \leq V(\pi) \cdot E_\pi \text{var}_\theta \delta(X), \end{aligned}$$

whence the result.

ACKNOWLEDGEMENT

We thank the Editor, the Associate Editor and the referee for their constructive suggestions and encouragement.

REFERENCES

T. W. Anderson (1984). *An Introduction to Multivariate Statistical Analysis.* Wiley, New York.

D. Basu (1955). On statistics independent of a complete sufficient statistic. *Sankhyā*, 15, 377–380.

J. Beran (1994). Statistics for Long Memory Processes, Chapman and Hall, New York.

J. Berger (1994). *Statistics for Long Memory Processes.* Chapman and Hall, New York.

J. Berger & T. M. Sellke (1987). Testing a point null hypothesis: the irreconciliability of significance levels and evidence. *Journal of the American Statistical Association*, 82, 112–122.

P. J. Bickel (1981). Minimax estimation of the mean of a normal distribution when the parameter space is restricted. *The Annals of Statistics*, 9, 1301–1309.

A. A. Borovkov & A. U. Sakhanienko (1980). On estimates of the expected quadratic risk. *Probability and Mathematical Statistics*, 1, 185–195.

L. D. Brown (1971). Admissibility, recurrent diffusions, and insoluble boundary value problems. *The Annals of Mathematical Statistics*, 42, 855–903.

L. D. Brown (1985). *Fundamentals of Exponential Families.* Institute of Mathematical Statistics Monograph, Hayward, CA.

L. D. Brown & L. Gajek (1990). Information inequalities for the Bayes risk. *The Annals of Statistics*, 18, 1578–1594.

G. Casella & R. L. Berger (1987). Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *Journal of the American Statistical Association*, 82, 106–111.

A. DasGupta (1994). *An Examination of Bayesian Inference and Methods.* Technical Report #94–16, Department of Statistics, Purdue University, West Lafayette, IN.

M. Delampady, A. DasGupta, G. Casella, H. Rubin & W. E. Strawderman (1999). *A New Approach to Default Priors and Robust Bayes Methodology.* Technical Report, Department of Statistics, Purdue University, West Lafayette, IN.

T. S. Ferguson (1974). Prior distributions on spaces of probability measures. *The Annals of Statistics*, 2, 615–629.

U. Grenander & M. Rosenblatt (1957). *Statistical Analysis of Stationary Time Series.* Wiley, New York.

P. J. Huber (1981). *Robust Statistics.* Wiley, New York.

S. Karlin & H. Rubin (1956). The theory of decision procedures for distributions with monotone likelihood ratio. *The Annals of Mathematical Statistics*, 27, 272–299.

E. L. Lehmann (1983). *Theory of Point Estimation*, Second Edition. Wiley, New York.

B. Levit (1980). Asymptotic minimax estimators of the second order. *Theory of Probability and its Applications*, 25, 552–568.

H.-S. Oh & A. DasGupta (1999). Comparison of the P-value and posterior probability. *Journal of Statistical Planning and Inference*, 76, 93–107.

C. R. Rao (1973). *Linear Statistical Inference and its Applications*, Second Edition. Wiley, New York.

M. Sato & M. Akahira (1996). An information inequality for the Bayes risk. *The Annals of Statistics*, 24, 2288–2295.

R. G. Staudte & S. J. Sheather (1990). *Robust Estimation and Testing.* Wiley, New York.

C. Zang & A. DasGupta (1998). *Bayesian inference under long range dependence.* Submitted.

Received 19 May 1999
Accepted ??? 2000

Anirban DASGUPTA: dasgupta@stat.purdue.edu
Mohan DELAMPADY: mohan@isibang.ac.in
Herman RUBIN: hrubin@stat.purdue.edu
Department of Statistics
Purdue University
West Lafayette, IN 47907-1399
USA

George CASELLA: gc15@cornell.edu
Department of Biometrics
Cornell University
Ithaca, NY 14853-7801
USA

Christian GENEST: genest@archimede.mat.ulaval.ca
Dép. de Mathématiques
et de Statistique
Université Laval
Sainte-Foy (Québec)
G1K7P4
CANADA

William E. STRAWDERMAN: straw@stat.rutgers.edu
Statistics Center
Rutgers University
New Brunswick, NJ 08903
USA