Multivariate Analysis and Its Applications
IMS Lecture Notes – Monograph Series (1994) Volume 24

SELECTION PARADOXES OF BAYESIAN INFERENCE

BY A. P. DAWID

University College London

When the inference to be made is selected after looking at the data, the classical statistical approach demands — as seems intuitively sensible — that allowance be made for the bias thus introduced. From a Bayesian viewpoint, however, no such adjustment is required, even when the Bayesian inference closely mimics the unadjusted classical one. In this paper we examine more closely this seeming inadequacy of the Bayesian approach. In particular, it is argued that conjugate priors for multivariate problems typically embody an unreasonable determinism property, at variance with the above intuition.

1. Introduction. A great deal of statistical practice involves, explicitly or implicitly, a two stage analysis of the data. At the first stage, the data are used to identify a particular parameter on which attention is to focus; the second stage then attempts to make inferences about the selected parameter.

Perhaps one of the most important intuitions which the discipline of Statistics has to offer — and perhaps one of the hardest to put across to the outside world — is the inappropriateness, in such circumstances, of a "face-value" approach, in which the second stage proceeds just as if the quantities identified at stage one had been chosen before the experiment. Instead, the selection or optimization applied at the first stage introduces a bias which needs to be allowed for — either by explicit modelling of the whole two-stage process, or by some general de-biasing technique such as cross-validation (Stone, 1974) or prequential analysis (Dawid, 1992).

In this paper we contrast the Bayesian approach to this problem with the above intuition. Since Bayesian posterior distributions are already fully conditioned on the data, the posterior distribution of any quantity is the same, whether it was chosen in advance or selected in the light of the data: that is, for a Bayesian, the face-value approach is fully valid, and no further adjustment for selection is required. This seeming paradox is all the more striking in problems where the face-value Bayesian inference closely mimics (in form, if

---

AMS 1980 Subject Classification: Primary 62H99; secondary 62F15, 62F07.

Keywords: Selection bias, conjugate prior, determinism.

not in interpretation) the face-value frequentist inference. If the latter would appear to require adjustment, why should not also the former?

EXAMPLE 1. **Treatment selection.** In an agricultural trial, p varieties are tested, with the aim of choosing that having highest mean yield. We might select, for future use, that variety i^* associated with the largest of the sample means (X_i). (Note that this need not, of course, have the largest population mean.) For simplicity we suppose $X_i \sim N(\mu_i, \sigma^2)$ independently (σ^2 known). We would then be interested in making inferences about the corresponding population mean $\mu_{i^*} = \mu^*$, say. How should such inferences take account of the optimization performed at the first stage?

The naïve or "face-value" approach, which estimates $\mu^* = \mu_{i^*}$ by $X^* = X_{i^*}$, appears inadequate: i^* is the realization of a random variable I^*, which the selection process biases towards those values of i for which X_i happens, by chance, to be in the upper tail of its distribution. Thus — as will be confirmed in Section 2 below — $X^* = X_{I^*}$ is likely to overestimate $\mu^* = \mu_{I^*}$. This is just the "resubstitution effect" which statisticians have learned to recognize and try to deal with: if a method optimized for a data-set is (naïvely) assessed on the basis of the same data, an over-optimistic view of its future performance will be obtained.

Nevertheless there is a puzzling aspect to this analysis. We would have been happy to estimate μ_{i^*} by X_{i^*}, had i^* been fixed before experimentation. Why then should we not, after observing the data and finding $I^* = i^*$, still be happy to do so? Why should the process by which we come to select a particular parameter for further attention affect the process of making inference about that parameter?

The Bayesian approach respects this alternative intuition. Suppose, for example, we use the improper prior $\pi(\boldsymbol{\mu}) \propto 1$. In the posterior we then have $\mu_i \sim N(x_i, \sigma^2)$ independently. In particular, $\mu^* = \mu_{i^*} \sim N(x^*, \sigma^2)$, and x^* could be used as a Bayesian estimate of μ^*. No account is thus taken of the selection process, and no adjustment for bias is needed. Similarly, a Bayesian $(1-\alpha)\%$ credible interval for μ^* is $x^* \pm \sigma z_{\frac{1}{2}\alpha}$, identical with an unadjusted face-value confidence interval.

EXAMPLE 2. We might alternatively select those two treatments i^* and j^* yielding the greatest sample difference $X_{i^*} - X_{j^*}$. Frequentist analysis would suggest that this overestimates $\mu_i^* - \mu_j^*$, while an improper Bayesian analysis would not.

EXAMPLE 3. The same issues arise if we use $\sum a_i^* X_i$ to estimate $\sum a_i^* \mu_i$, where $\boldsymbol{a}^*$ is chosen to maximize $\sum a_i X_i$, subject to $\sum a_i = 0$ and $\sum a_i^2 = 1$.

2. Optimization and selection: frequentist analysis

2.1. General structure. Suppose that we have a collection $\{\phi_\lambda : \lambda \epsilon \mathcal{L}\}$ of parameters of potential interest. Each ϕ_λ is a function of the full parameter θ underlying the distribution of the data X. Suppose further that for each fixed $\lambda \epsilon \mathcal{L}$ we have a "good" estimator X_λ of ϕ_λ.

In Example 1 we have $\mathcal{L} = \{1, 2, \cdots, p\}$, $\lambda = i$, $\phi_\lambda = \mu_i$ and $X_\lambda = X_i$. For example 2 we could take $\mathcal{L} = \{(i,j) : 1 \leq i \neq j \leq p\}$, $\lambda = (i,j)$, $\phi_\lambda = \mu_i - \mu_j$, $X_\lambda = X_i - X_j$; and, for Example 3, $\mathcal{L} = \{\boldsymbol{a} : \sum a_i = 0, \sum a_i^2 = 1\}$, $\lambda = \boldsymbol{a}$, $\phi_\lambda = \sum a_i \mu_i$, $X_\lambda = \sum a_i X_i$.

In an optimization problem, one focus of interest might be $\phi^{**} = \sup\{\phi_\lambda : \lambda \epsilon \mathcal{L}\}$, together with λ^{**}, the (supposed unique) value of λ achieving this (supposed possible). In so far as these are well-defined functions of θ, making inferences about them raises no special conceptual difficulties, although in practice it may be far from straightforward. In Examples 1, 2 and 3, we have, respectively, $\phi^{**} = \mu^{**} = \sup\{\mu_i\}$; $\phi^{**} = \sup\{\mu_i - \mu_j\}$; and $\phi^{**} = \sup\{\sum a_i \mu_i : \sum a_i = 0, \sum a_i^2 = 1\}$ — this last being in fact $(\sum(\mu_i - \mu_.)^2)^{\frac{1}{2}}$, achieved at $a_i^{**} = (\mu_i - \mu_.)/(\sum(\mu_i - \mu_.)^2)^{\frac{1}{2}}$, where $\mu_. = p^{-1}(\sum \mu_i)$.

The relevance of such optimized parametric quantities ϕ^{**} is limited by the fact that, without fully knowing the parameters, it is not possible to identify the optimizing value λ^{**}, so that, when we estimate ϕ^{**}, we are estimating an unachievable optimum: and it is not very useful to estimate $\sup\{\mu_i\}$, for example, if we do not know which treatment to use to achieve it. An alternative, more useful, focus of interest might be $\phi^* = \phi_{\Lambda^*}$, where the value Λ^* is suggested by the data as likely to be associated with a large parameter (for example Λ^* might index the largest sample mean). Since Λ^* is known (after seeing the data), we can in this case identify the parameter about which inference is being made. But Λ^* is itself random, and ϕ^* is thus a "data-dependent parameter". General principles of inference for such quantities are not well-developed.

2.2. Optimized parameter. Although the problem of making inference about $\phi^{**} = \sup\{\phi_\lambda\}$ is not our primary interest here, it forms a useful halfway stage on the road to our real concern: making inference about a parameter ϕ^* suggested by the data. We therefore now examine some aspects of this problem.

Suppose X_λ is unbiased for ϕ_λ, all λ. In considering estimation of ϕ^{**} and λ^{**}, we might begin by examining the "naïve" estimators

$$X^* = \sup\{X_\lambda : \lambda \epsilon \mathcal{L}\}$$

and

$$\Lambda^* = \arg\sup\{X_\lambda : \lambda \epsilon \mathcal{L}\}.$$

We then note that, since $X^* \geq X_\lambda$, all λ, $E_\theta(X^*) \geq \sup\{E_\theta(X_\lambda) : \lambda \epsilon \mathcal{L}\} = \phi^{**}$ — with equality only if $\Lambda^* = \lambda^{**}$ with probability 1. Thus X^* is indeed positively biased for ϕ^{**}, and the bias can be strong if the set $\mathcal{L}$ is rich enough. It would therefore *not* seem appropriate to use X^* to estimate ϕ^{**}, at least not without making some kind of adjustment to take this bias into account.

EXAMPLE 4. Again take $X_i \sim N(\mu_i, \sigma^2)$. Take $\lambda = \boldsymbol{a}\,(\sum a_i = 0,\ \sum a_i^2 = 1)$, $\phi_\lambda = (\sum a_i \mu_i)^2$. Then $X_\lambda = (\sum a_i X_i)^2 - \sigma^2$ is unbiased for ϕ_λ. We have $\phi^{**} = \sum(\mu_i - \mu_.)^2$, $X^* = \sum(X_i - X_.)^2 - \sigma^2$. Thus $E(X^*) = \phi^{**} + (p-2)\sigma^2$, and X^{**} is positively biased for ϕ^{**} if $p > 2$ (if $p = 2$, then, with probability 1, $\Lambda^* = \lambda^{**} = (1/\sqrt{2}, -1/\sqrt{2})$).

The above "bias" effect is not restricted to its effect on unbiased estimation, but affects other forms of inference. In particular, in many problems we have the following structure. There is a family $\mathcal{Q} = \{Q_\phi\}$ of distributions, stochastically increasing in ϕ, such that, given θ, $X_\lambda \sim Q_{\phi_\lambda}$. Moreover, the distribution of $X^* = \sup\{X_\lambda\}$ depends on θ only through $\phi^{**} = \sup\{\phi_\lambda\}$. Let Q^*_ϕ denote this distribution when $\phi^{**} = \phi$, and $\mathcal{Q}^* = \{Q^*_\phi\}$. In Example 3 we had $\phi_\lambda = \sum a_i \mu_i$, $X_\lambda = \sum a_i X_i$, $\phi^{**} = \|\boldsymbol{\mu} - \mu_. \mathbf{1}\|$, $X^* = \|\boldsymbol{X} - X_. \mathbf{1}\|$. We have the above structure with $Q_\phi = N(\phi, \sigma^2)$ and $Q^*_\phi = \{\sigma^2 \chi^{2'}_{p-1}(\phi^2/\sigma^2)\}^{\frac{1}{2}}$.

In such a case, Q^*_ϕ must be stochastically greater than Q_ϕ, all ϕ. For let g be an increasing function, and let $h(\phi) = E_\phi(g(X))$. Then h is increasing, and $E_\theta\,(g(X_\lambda)) = h(\phi_\lambda)$. Our previous bias argument now implies that $E^*_{\phi^{**}}\,(g(X^*)) = E_\theta\,(g(X^*)) > h(\phi^{**}) = E_{\phi^{**}}(g(X))$. Since g and ϕ are arbitrary, the result follows.

This property means that, under the (correct) model $\mathcal{Q}^*$ for X^* given ϕ^{**}, large values of X^* are more likely, for any ϕ^{**}, than under the (incorrect) face-value model $\mathcal{Q}$. This in turn suggests that any value of X^* can be explained by a smaller value of ϕ^{**} under $\mathcal{Q}^*$ than required by $\mathcal{Q}$, and thus that any kind of inferences made using $\mathcal{Q}$ are likely to be biased upwards (in an intuitive sense) compared with those appropriate under the correct model $\mathcal{Q}^*$. In particular, for testing $\phi^{**} = \phi_o$ against $\phi^{**} > \phi_o$, the face-value upper-tail P-value, if $X^* = x^*$, is $Q_{\phi_o}(X > x^*)$, which is smaller than $Q^*_{\phi_o}(X > x^*)$, the correct P-value. The real evidence against $\phi^{**} = \phi_o$, in favour of $\phi^{**} > \phi_o$, is thus not as strong as face-value inference would indicate. Correspondingly, face-value confidence limits will be higher than warranted.

2.3. Data-dependent parameter. Suppose again that X_λ is unbiased for ϕ_λ, all λ. We now proceed with a two-stage approach. At the first stage we use the full data X to determine the realized value λ^*, of Λ^*, which achieves $\sup\{X_\lambda\}$, and is thus likely to be associated with a large parameter ϕ_λ; and at the second stage, having thus determined that the parameter of interest is $\phi^* = \phi_{\lambda^*}$, we proceed to make inferences about it.

If we ignore the fact that λ^* was chosen as a function of X — what we

have termed the face-value approach — then we could consider using the *face-value estimator* X_{λ^*}. In terms of the full two-stage process, this would mean that we were again using the estimator $X^* = X_{\Lambda^*}$, but this time to estimate the data-dependent parameter $\phi^* = \phi_{\Lambda^*}$, rather than the optimal ϕ^{**}.

Now necessarily $\phi^* \leq \phi^{**}$, and typically this inequality is strict with probability 1. Since we have already shown that $E_\theta(X^*) > \phi^{**}$, we must therefore consider X^* as positively biased for ϕ^* (even though the meaning of this assertion is unclear when ϕ^* depends on X), and indeed still more strongly biased for estimating ϕ^* than for estimating ϕ^{**}. In particular, $E_\theta(X^* - \phi^*) > 0$ for all θ. If an estimate of ϕ^* is to be based on X^*, it thus appears that some form of downward bias correction will again be required. Similarly, other forms of face-value inference for ϕ^* based on X^* can be considered biased upwards, and in need of adjustment.

3. Bayesian inference. Let now θ have a prior distribution, and let Y_λ be the posterior expectation $E(\phi_\lambda|X)$. Define $Y^{**} = E(\phi^{**}|X)$, $Y^* = E(\phi^*|X)$. Let $Y^\dagger = \sup\{Y_\lambda\}$ be achieved at $\Lambda^\dagger$, and define $\phi^\dagger$ to be the corresponding data-dependent parameter $\phi_{\Lambda^\dagger}$. Then $Y^\dagger = E(\phi^\dagger|X)$. Y^{**}, $Y^\dagger$ and Y^* could be used as sensible Bayesian estimates of ϕ^{**}, $\phi^\dagger$ and ϕ^* respectively. It is readily seen that $Y^{**} \geq Y^\dagger \geq Y^*$. The frequentist analysis of Section 2 would lead us to expect that Y^{**}, and *a fortiori* $Y^\dagger$ and Y^*, should typically be smaller than X^*, so as to counter the effects of bias.

3.1. Improper priors. Now in many problems (such as Example 1) it is possible to choose a (necessarily improper) prior for θ such that Y_λ is unbiased for ϕ_λ, all λ. That is, we can identify Y_λ, with X_λ, and thus $\Lambda^\dagger$, $Y^\dagger$, $\phi^\dagger$ with Λ^*, X^*, ϕ^*. But then the Bayesian estimate $E(\phi^*|X) = Y^* = X_{\Lambda^*} = X^*$, which does *not* incorporate any correction for bias. Worse, since $Y^{**} > Y^*$, so far from incorporating a negative bias-correction, as seems required from the frequentist analysis, the Bayesian estimate of ϕ^{**} adjusts X^* in a *positive* direction. These properties of the improper Bayes inference seem highly undesirable, and appear to form yet another argument against the use of improper priors.

3.2. Proper priors. To see that the above problems cannot arise in quite the same form with a proper prior, we note that $0 = Y^* - E(\phi^*|X) = E(Y^* - \phi^*|X)$, so that (in the joint distribution of (X, θ)) $E(Y^* - \phi^*) = 0$. We thus can *not* have $E_\theta(Y^*) > \phi^*$ for all θ, since this would imply $E(Y^* - \phi^*) > 0$. In other words, for at least some values of the parameter, the Bayes estimate of ϕ^* does not share the positive bias associated with X^*, and can not therefore be ruled out on that score. The same holds for $Y^\dagger$ as an estimator of $\phi^\dagger$. Similarly, $E_\theta(Y^{**} - \phi^{**})$ cannot be positive for all θ (unlike $E^*_\theta(X^* - \phi^{**})$). However, even though free of logical inconsistencies, the behaviour of proper

Bayesian inference under selection can still appear unsatisfactory.

EXAMPLE 5. In Example 1, take the proper prior $\mu_i \sim N(0, \tau^2)$ independently. In the posterior, $\mu_i \sim N\left(y_i, \left(\frac{1}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1}\right)$, with $y_i = \left(\frac{\tau^2}{\tau^2+\sigma^2}\right) x_i$. Correspondingly, $\mu^* \sim N\left(y^*, \left(\frac{1}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1}\right)$ $\left(y^* = \left(\frac{\tau^2}{\tau^2+\sigma^2}\right) x^*\right)$. We see that large values of x^* are indeed shifted downwards, although the effect depends only on the size of x^*, and in no way on the selection process (for example, it is independent of the number of treatments selected from). However, if the prior variance τ^2 is large compared with the data-precision σ^2 — whether because τ^2 is large, representing "vague prior knowledge", or because σ^2 is small, being the variance σ_0^2/n of the average of a large number n of replicate observations — then y^* will be close to x^*, and one might feel that the shrinkage effect is simply not enough to counter the effect of bias.

We can also consider the alternative proper prior, under which the (μ_i) have independent Student distributions, $\mu_i \sim \tau t_\nu$. It can then be shown (Dawid, 1973) that $|y^* - x^*| \to 0$ as $x^* \to \infty$, so that, asymptotically, this Bayesian method makes no correction at all for selection bias.

3.3. The effects of prior assumptions. What is the nature of the seemingly paradoxical behaviour, under selection, of the Bayesian inferences discussed above? Do they point to a flaw in the Bayesian approach? We shall argue that this is not the case, but rather that they highlight the importance of using a prior distribution carefully chosen to represent and incorporate understanding of the problem, rather than pulled off a convenient shelf.

Consider, for Example 1, under what conditions on $\boldsymbol{\mu}$ the sampling bias in X^* is likely to be small. Clearly this will be when, for some i_0, $P(I^* = i_0)$ is close to 1, which will happen when μ_{i_0} exceeds $\max\{\mu_i : i \neq i_0\}$ by a quantity large compared with σ. Conversely, the bias will be greatest when the μ's are all equal. If the prior distribution gives very high probability to the former state of affairs, that is as much as to say that we do not expect the bias in X^* to be important, and should therefore not need to make much correction for it.

For the prior $\mu_i \sim N(0, \tau^2)$ independently, the above condition on $\boldsymbol{\mu}$ will hold with high probability when the prior variance τ^2 is large compared with the data variance σ^2, and this is just the circumstance when $Y^* = \{\tau^2/(\tau^2 + \sigma^2)\} X^*$ is close to X^*. Conversely, if τ^2 is not so large, the non-ignorable shrinkage factor $\tau^2/(\tau^2 + \sigma^2)$ can be regarded as introducing a suitable correction for bias, in this case expected to be non-negligible.

For the prior $\mu_i \sim \tau t_\nu$, when $\tau^2 \gg \sigma^2$ essentially the above discussion again applies. But, even for small τ, the long tail of the t_ν distribution means that (particularly for large p) there is a reasonable probability that the largest of the (μ_i) is much larger than the others, so that it is not surprising to find Y^*

close to X^* — especially when X^* is large, a situation which goes to confirm this structure for the (μ_i). In contrast, smaller values of X^* may require (and will get) stronger "debiasing".

The message of the above discussion is that the Bayesian solution is perfectly sensible so long as the prior is taken seriously. If a formal Bayesian analysis leads to results which appear unreasonable, the implication is that the prior distribution itself was unreasonable. All this goes to stress the importance of thinking very carefully about the prior distribution, and ensuring that one is happy with its properties and implications.

In the remainder of this paper we consider the behaviour of Bayesian inferences based on conjugate prior distributions in some multivariate problems. To the extent that this behaviour is counter-intuitive, we argue that this is because the priors used themselves embody a world-view at odds with the offended intuition. If that intuition is considered valid, such prior distributions should not, therefore, be used.

4. Discrete prediction and discrimination. Let $X_1, \cdots, X_p$ be random 0–1 predictor variables, and Y a 0–1 response variable. We have a random sample D of cases $\{(\boldsymbol{x}_i, y_i) : i = 1, \cdots, n\}$, from which we wish to construct a rule for classifying Y as 0 or 1 on the basis of observation of $\boldsymbol{X}$ alone, and estimate the associated error rate.

Let P_θ be the joint distribution of $\boldsymbol{X}$ and Y, with $\theta(\boldsymbol{x}, y) = P_\theta(\boldsymbol{X} = \boldsymbol{x}, Y = y)$. Let $\mathcal{R} = \{R_\lambda\}$ be the set of all possible classification rules. R_λ classifies $\boldsymbol{x}$ as $y_\lambda(\boldsymbol{x})$. Its correct classification rate is then $\phi_\lambda = \sum_{\boldsymbol{x}} \theta(\boldsymbol{x}, y_\lambda(\boldsymbol{x}))$. If θ is known, the optimal classification rule is: classify $\boldsymbol{x}$ as y^{**}, where $y^{**} = y^{**}(\boldsymbol{x})$ maximizes $\theta(\boldsymbol{x}, y)$ over $y = 0, 1$. The optimal correct classification rate is thus $\phi^{**} = \sum \theta(\boldsymbol{x}, y^{**}(\boldsymbol{x}))$. Of course, if θ is unknown, so is the rule y^{**}, so that ϕ^{**} is not achievable in practice.

Let $n(\boldsymbol{x}, y)$ be the frequency, in D, of $(\boldsymbol{X} = \boldsymbol{x}, Y = y)$. The naïve data-based classification rule classifies $\boldsymbol{x}$ as $y^*(\boldsymbol{x}) = y^*$, where y^* maximizes $n(\boldsymbol{x}, y)$ over y (ties, including the case $n(\boldsymbol{x}, 0) = n(\boldsymbol{x}, 1) = 0$, being broken arbitrarily: for definiteness we take $y^* = 1$ if there is a tie). If this rule is applied to future cases, its error rate will be $\phi^* = \sum_{\boldsymbol{x}} \theta(\boldsymbol{x}, y^*(\boldsymbol{x}))$. This is a data-dependent parameter, because the rule y^* is based on D.

The naïve estimate of $\theta(\boldsymbol{x}, y)$ is $\widehat{\theta}(\boldsymbol{x}, y) = n(\boldsymbol{x}, y)/n$, which, for any fixed $(\boldsymbol{x}, y)$, is unbiased. However, as discussed in Section 2, the implied estimate $\widehat{\phi} = \sum_{\boldsymbol{x}} \widehat{\theta}(\boldsymbol{x}, y^*(\boldsymbol{x}))$ will be biased upwards, whether considered as an estimate of ϕ^{**}, or of ϕ^*. Note that $\widehat{\phi}$ is just the proportion of correct classifications in D, when using the rule y^* which maximizes that proportion. Clearly this can give a very rosily optimistic view of the expected proportion ϕ^* arising when y^* is used on new cases. Indeed if, as can hold in many applications, n, though possibly large, is small compared with 2^p, we may, with high probability, have $n(\boldsymbol{x}) = 0$ or 1 for all $\boldsymbol{x}$, in which case $\widehat{\phi} = 1$.

A standard Bayesian analysis (Teather, 1974) uses the conjugate Dirichlet prior for $\theta = (\theta(\boldsymbol{x}, \boldsymbol{y}))$. If the prior parameters are $\boldsymbol{\alpha} = (\alpha(\boldsymbol{x}, \boldsymbol{y}))$ then the posterior is again Dirichlet, with parameters $\boldsymbol{\alpha} + \boldsymbol{n}$ ($\boldsymbol{n} = (n(\boldsymbol{x}, y))$). The optimal Bayes classification rule is $y^\dagger(\boldsymbol{x})$, chosen to maximize $\alpha(\boldsymbol{x}, y) + n(\boldsymbol{x}, y)$ over y. In particular, if the prior parameters are all equal we can take $y^\dagger \equiv y^*$.

The implied posterior for ϕ^* is $\beta(\alpha_1^* + r_1^*, \alpha_0^* + r_0^*)$, where $\alpha_0^* = \sum \alpha(\boldsymbol{x}, y^*(\boldsymbol{x}))$, $r_0^* = \sum n(\boldsymbol{x}, y^*(\boldsymbol{x}))$, $\alpha_1^* = \alpha_+ - \alpha_0^*$ ($\alpha_+ = \sum \alpha(\boldsymbol{x}, y)$) and $r_1^* = n - r_0^*$. In particular, the posterior expectation of ϕ^* is $(\alpha_0^* + r_0^*)/(\alpha_+ + n)$, which, if the α's are small, will be close to the (biased) face-value estimate $\widehat{\phi}$. Moreover, for large n the posterior variance will be very small. Once again this Bayesian inference appears highly misleading. Brown (1976, 1980) discusses the above difficulty and relates it to properties of the Dirichlet prior distribution. Another property of this prior, which gives some insight into the above behaviour, is studied by Fang and Dawid (1993). We consider the effect of adding further X-variables, all the time ensuring a (consistent) Dirichlet structure for the prior. We show that, under weak conditions, this implies prior probability 1 for the event that the distributions of $\boldsymbol{X}$ given Y would be such as to allow *asymptotically perfect discrimination* between the populations labelled by Y, as the number of X-variables is increased. In other words, if we take the Dirichlet structure seriously, we must believe that Y is a deterministic function of the X's. Of course, we are *a priori* uncertain of the form of this function. However, on observing data (Y and some of the X's) we shall essentially be observing some of its values (with little uncertainty), and hence shall be able to reconstruct it, in part. Further, this prior implies that, given a sample of size n, the total probability of all the $\boldsymbol{x}$-configurations not yet observed is of order n^{-1} (Brown, 1976), so that we can safely ignore such $\boldsymbol{x}$-values.

Now if the above beliefs are accepted, it is very reasonable to estimate ϕ^* by something close to $\widehat{\phi}$. When we believe in deterministic discrimination, the dangers of following random noise in the data are almost eliminated, and consequently we can ignore its biasing effects. There is thus nothing unsatisfactory about the implications of the Dirichlet prior in those situations for which it is a good description. The seeming conflict with frequentist analysis is, rather, a conflict of world views: that behind the frequentist approach taking seriously just the possibility of high residual uncertainty in Y after observing $\boldsymbol{X}$ which is essentially ruled out by the Dirichlet prior. Which world view is more appropriate must be a matter of context. Statisticians tend to work with applied problems, such as medical diagnosis and (more especially) prognosis, where it would *not* be reasonable to suppose that classification could be done almost perfectly, were only enough variables to be observed. But there are certainly problems (*e.g.* botanical classification) where the opposite is true — indeed, much of the work on pattern recognition undertaken within the artificial intelligence community seems based on archetypal applications in which

perfect classification is possible. The Dirichlet prior appears well suited to such problems, but inappropriate to express the more "statistical" problems in which residual uncertainty can never be eliminated. Other priors expressing this idea need to be developed and explored.

5. Continuous discrimination and regression. A very similar story can be told for some other common multivariate problems. Consider first the problem of discrimination between two multivariate normal distributions with common dispersion. As soon as the number of variables p considered exceeds the within-group degrees of freedom ν, it is possible to find a sample-based discriminant function yielding zero within-group sample variance, infinite sample Mahalanobis distance, and perfect sample classification. Of course, no-one would infer that this discriminant would work perfectly on future cases – the effect of bias is obvious.

When we conduct a Bayesian analysis with the usual conjugate normal-inverted Wishart prior, we can consider, for example, the discriminant function maximizing the corresponding posterior expected population Mahalanobis distance. For $p > \nu$ this will not generally separate the samples perfectly, but will tend to do so more closely for larger p: this is because, under the assumed prior distribution, the Mahalanobis distance between the populations tends almost surely to infinity as $p \to \infty$ (Dawid and Fang, 1992); and hence the same holds for its posterior expectation. This Bayesian discriminant is thus, in a sense, very close to being a naïve sample-based discriminant, and once again the conjugate Bayes approach appears to neglect the problem of bias. And, once again, this behaviour appears less unreasonable if one truly takes the conjugate prior seriously, since it implies that, with probability 1, it is possible to classify an observation perfectly on the basis of sufficiently many predictor variables. However, in many contexts this assumption will be clearly unreasonable. In that case, use of a conjugate prior can lead to highly misleading inferences.

Finally, consider the prediction of a continuous variable Y using continuous predictors $X_1, X_2, \cdots, X_p$, the joint distribution being multivariate normal. Again, for large enough p, it will be possible to find a sample-based linear predictor which exactly fits the sample data, with zero residual variation. Again, a Bayesian analogue, using a conjugate prior, will be very close to this. And yet again, this neglect of sampling bias is justifiable if the prior is taken seriously, since it implies (Dawid, 1988) that, with probability 1, the true residual uncertainty in Y decreases to 0 as $p \to \infty$: yet another instance of asymptotic determinism.

6. Discussion. We have seen that use of a conjugate prior, in a variety of multivariate problems, leads to inferences about selected parameters close to face-value sampling-theory inferences, but, worryingly, with no scope for

bias correction. We have also seen that this does not undermine the self-consistency often trumpeted as the main virtue of Bayesian inference, but is, rather, explicable in terms of an "asymptotic determinism" property implicit in the conjugate prior, under which – if it is truly believed – selection bias ceases to be a problem. The moral is that self-consistency is not enough: choice of prior distribution for Bayesian analysis is a delicate matter, which must be carefully considered in terms of the realism of its implications in the context for which it is intended.

REFERENCES

BROWN, P. J. (1976). Remarks on some statistical methods for medical diagnosis. *J. Roy. Statist. Soc. A* **139**, 104–107.

BROWN, P. J. (1980). Coherence and complexity in classification problems. *Scand. J. Statist.* **7**, 95–98.

DAWID, A. P. (1973). Posterior expectations for large observations. *Biometrika* **60**, 664–667.

DAWID, A. P. (1988). The infinite regress and its conjugate analysis (with Discussion). Bayesian Statistics 3 (Bernardo, J. M., DeGroot, M. H., Lindley, D. V. and Smith, A. F. M., Eds.). Oxford University Press, 95–110.

DAWID, A. P. (1992). Prequential data analysis. In Current Issues in Statistical Inference: Essays in Honor of D. Basu (M. Ghosh and P. K. Pathak, Eds.) *IMS Lecture Notes-Monograph Series* **17**, 113–126.

DAWID, A. P. and FANG, B. Q. (1992). Conjugate Bayes discrimination with infinitely many variables. *J Mult. Anal.* **41**, 27–42.

FANG, B. Q. and DAWID, A. P. (1993). Asymptotic properties of conjugate Bayes discrete discrimination. *J. Mult. Anal.* (to appear).

STONE, M. (1974). Cross-validatory choice and assessment of statistical predictions (with Discussion). *J. Roy. Statist. Soc. B* **36**, 111–147.

TEATHER, D. (1974). Statistical techniques for diagnosis. *J. Roy. Statist. Soc. A* **137**, 231–244.

DEPARTMENT OF STATISTICAL SCIENCE
UNIVERSITY COLLEGE LONDON
GOWER STREET, LONDON WCIE 6BT
UNITED KINGDOM