NILS L. HJORT[1]

Norwegian Computing Center

1. Introduction. Although hardly the authors' intention, these papers by Diaconis and Freedman (D & F) will probably be read by many as criticism against and pessimism about Bayesian analysis in situations with high-dimensional parameter spaces. It is perhaps also easy for the statistician scanning the papers to get an impression of "just counterexamples," which would be unfair; these and earlier papers by D & F (or F & D) contain many new important statistical ideas and also useful mathematical techniques.

I will try to be (more) positive and hope to show that thinking Bayes in semi- and nonparametric models may be a worthwhile enterprise, sometimes giving additional insight into old problems, and sometimes (dare I say often?) leading to sensible Bayes procedures that also behave agreeably in the frequentist asymptotic sense. The bulk of my comments concerns a problem that is almost as old as statistics itself, that of fitting a parametric model to a data set, and that can be attacked again with ideas underlying some of the constructions of D & F. Let $X_1, \ldots, X_n$ be a sample from some unknown distribution F with density f. Some (possibly crude) parametric family $\{F_\theta, f_\theta\colon\ \theta \in \Theta\}$ is then forced on the data. Textbooks teach us how to proceed, for example, advocating finding the maximum likelihood estimator $\hat{\theta}_{\mathrm{ML}}$, on the grounds of good asymptotic behavior, in particular, consistency. What very few textbooks tell us, however, is what $\hat{\theta}_{\mathrm{ML}}$ does when the model is wrong, i.e., there is no θ_0 with $f = f_{\theta_0}$. It is however not difficult to see that $\hat{\theta}_{\mathrm{ML}}$ still is a meaningful estimator in that it takes aim at the parameter value $\theta = \theta_1$ that minimises Kullback–Leibler "information distance"

$$I(f : f_\theta) = \int f \log(f/f_\theta)\, dx; \tag{1}$$

the log likelihood divided by n is a consistent estimate of $\int f \log f\, dx - I(f : f_\theta)$. Under appropriate conditions $\hat{\theta}_{\mathrm{ML}}$ is consistent for this "least false" parameter value. Hjort (1985a, Chapter 3) has further comments about the behavior of maximum likelihood machinery when the model is wrong.

One of the major uses of a fitted model is prediction, or probability assessments, for certain sets. Thus we could be interested in stating that approximately 90% of future Xs from a fitted normal will fall in $(\hat{\mu} - 1.645\hat{\sigma},\ \hat{\mu} + 1.645\hat{\sigma})$, or that approximately 50% of future data points from a fitted Weibull fall below $\hat{\theta}(\log 2)^{1/\hat{\alpha}}$, etc. If such statements are an important part of the statistical analysis, then there are disadvantages to using $\hat{\mu}_{\mathrm{ML}}$, $\hat{\sigma}_{\mathrm{ML}}$, resp. $\hat{\theta}_{\mathrm{ML}}$, $\hat{\alpha}_{\mathrm{ML}}$ in the case of an incorrectly specified model, and one could do better with other estimates that aimed at other versions of least false population parameters. It is the aim of the present notes to show that such least false parameters can be defined and that a suitably engineered semiparametric Bayesian setup can result in estimates that actually manage to estimate these.

[1]This work was done while the author visited Stanford University with grants from the Norwegian Computing Center and the Royal Norwegian Council for Scientific and Industrial Research.

If we acknowledge uncertainty about the chosen parametric model we should perhaps build that into a larger statistical model. A natural Bayesian approach is to give a prior density $\nu(\theta)\,d\theta$ for θ in Θ *and* some prior on the space of distributions on the sample space, centered at $\{f_\theta\colon \theta \in \Theta\}$ in some sense. Such ideas are really behind much of the work presented in D & F. One way of doing this is the following: Assume for the moment that the X_j's are univariate and write

$$X_j = F_\theta^{-1}(U_j), \qquad j = 1, \ldots, n, \tag{2}$$

where U_j has distribution $G = FF_\theta^{-1}$. That the parametric model is correct amounts to having $G = G_0 = U(0,1)$ for some θ. Uncertainty about the parametric model can therefore be modelled by a prior distribution for G (in the space of distributions on $[0,1]$), centered at G_0. An "uncertain Gaussian model," for example, is

$$X_j = \mu + \sigma Y_j, \qquad j = 1, \ldots, n, \tag{3}$$

where $Y_j = \Phi^{-1}(U_j)$ has a random distribution centered at the standard normal. To help identifiability one could restrict the space of allowable distributions for Y_j to those having zero mean and unit variance, or to those being symmetric with interquartile range $2 \times (0.674)$, etc. The latter approach would be along the lines of D & F.

The next section outlines another but related approach, still with the notorious ("herostratic" would be too harsh, even with D & F's examples) Dirichlet process prior as a building tool, and is more akin to recent work of Hani Doss (1985a, b). The method offers the possibility of building uncertainty about any parametric model into a larger semiparametric model, and allows one to specify *control sets* that may be important for later predictions based on the fitted model. The asymptotic results of Section 3 are of the same character as those of D & F and of Doss, and indeed, examples displaying "inconsistency" can be constructed. They are interpreted in a more positive light here, however. It will be seen that the resulting Bayes estimates really take aim at, and will be frequentist consistent for, completely sensible least false parameter values. Also included in Section 3 are indications of asymptotic normality results and calculations of influence functions. It emerges that the Bayes estimates, or for that matter closely related frequentist estimates, constitute robust alternatives to traditional estimates, with the advantageous capability of being flexibly tailored to any specific prediction task, and without losing much efficiency in the idealised (and unrealistic) case when the parametric model happens to be correct.

Section 4 briefly sketches some ideas for similar Bayesian semiparametric analysis in parametric survival analysis models, where the class of *beta processes* plays the natural role. Section 5 contains additional remarks.

2. Semiparametric Bayes estimation. Having (2) and (3) in mind, write for a general i.i.d. sample $X_1, \ldots, X_n$ in some X space

$$X_j = h_\theta(Y_j), \qquad j = 1, \ldots, n, \tag{4}$$

where h_θ is one-to-one on some Y space. Y_j has distribution G; if G is equal to some idealised G_0 then X_j has distribution $F_\theta = G_0 h_\theta^{-1}$ with density $f_\theta(x) = g_0(h_\theta^{-1}(x))|\partial h_\theta^{-1}(x)/\partial x|$, say.

A first construction for a prior distribution for (θ, G) could be to let $\theta \sim \nu(\theta)\,d\theta$ and G, independently, be a Dirichlet process centered at G_0. Write $G \sim \text{Dir}(kG_0)$ for such a process, with "strength of belief" parameter k and "prior guess" G_0. It turns out that θ has posterior density

$$\nu(\theta|x) = c(x) \prod_{j=1}^{n}{}^{*} f_\theta(x_j)\nu(\theta), \tag{5}$$

the * signifying that only the distinct observations are to be included. The only effect of the sophisticated extra randomness introduced by $G \sim \text{Dir}(kG_0)$ is that the likelihood is only over the distinct data points. ($c(x)$ denotes generically a function of the data $x = (x_1, \ldots, x_n)$ that gives integrated posterior density 1.)

A more fruitful approach is the following, generalising Doss' method (1985a, b). Define *m control sets* $B_1, \ldots, B_m$, constituting a measurable partition of Y space, with $G_0(B_i) = z_i$, $i = 1, \ldots, m$. Now pin down a $G \sim \text{Dir}(kG_0)$ by conditioning on $G(B_i) = z_i$, $i = 1, \ldots, m$. It can be seen that G splits into m separate and independent Dirichlet processes:

$$G = z_i G_i \quad \text{on the set } B_i, \text{ where } \quad G_i \sim \text{Dir}\big(kz_i(G_0/z_i)\big) \quad \text{on } B_i.$$

Hjort (1985b) obtains

$$\nu(\theta|x) = c(x)M(x,\theta) \prod_{j=1}^{n}{}^{*} f_\theta(x_j)\nu(\theta) \tag{6}$$

for the posterior density of θ, where

$$M(x,\theta) = \prod_{i=1}^{m} z_i^{C_i(\theta)}/\Gamma\big(kz_i + C_i(\theta)\big) \tag{7}$$

and

$$C_i(\theta) = \sum_{j=1}^{n} I\big\{X_j = h_\theta(Y_j) \in h_\theta B_i\big\} = nF_n(h_\theta B_i), \tag{8}$$

writing F_n for the usual empirical distribution of the sample. $M(x, \theta)$ is large for values of θ that make $C_i(\theta)$ close to nz_i, i.e., $F_n(h_\theta B_i)$ close to z_i, $i = 1, \ldots, m$.

EXAMPLE. Fit a normal (μ, σ^2) to data. Assume it is of interest to have approximately 25% of future data points in each of the four categories $(-\infty, \hat\mu - c\hat\sigma]$, $(\hat\mu - c\hat\sigma, \hat\mu]$, $(\hat\mu, \hat\mu + c\hat\sigma]$, and $(\hat\mu + c\hat\sigma, \infty)$, for $c = 0.674$, which would be the case for each perfectly normal underlying F, but not, for example, for even slightly skewed F, if $\hat\mu_{\text{ML}}$, $\hat\sigma_{\text{ML}}$ are used. The 25-25-25-25 goal could be important for prediction purposes, or just considered a pleasant aspect of the theoretical parametric model worth preserving to some extent for the fitted model. Define

control sets $B_1 = (-\infty, -c]$, $B_2 = (-c, 0]$, $B_3 = (0, c]$, and $B_4 = (c, \infty)$. Then

$$M(x; \mu, \sigma) = \left(\tfrac{1}{4}\right)^n \left\{ \prod_{i=1}^{4} \Gamma(k/4 + C_i(\mu, \sigma)) \right\}^{-1},$$

where

$$C_1(\mu, \sigma) = nF_n(\mu - c\sigma),\ C_2(\mu, \sigma) = nF_n(\mu - c\sigma, \mu],$$
$$C_3(\mu, \sigma) = nF_n(\mu, \mu + c\sigma],\ C_4(\mu, \sigma) = nF_n(\mu + c\sigma, \infty).$$

The posterior density (6) would in this case have two peaks, one corresponding to $M(x; \mu, \sigma)$, trying to achieve the 25-25-25-25 splitting, and one for the usual factor $\prod_{j=1}^{n} f_{\mu,\sigma}(x_j)$, which makes efforts to get the population mean and population variance correctly estimated. Since these goals coincide only in the idealised Gaussian case the Bayes estimators based on (6) try to push $\hat{\mu}_{\mathrm{ML}}, \hat{\sigma}_{\mathrm{ML}}$ so as to better achieve the stated 25-25-25-25 goal.

The theory allows multidimensional data, and extends to non-i.i.d. situations, for example regression models. Further examples are in Hjort (1985b).

3. Asymptotic behaviour of the estimates. Assume that the X_j's come from a continuous F with density f. Until D & F came along one would have expected the parts of the posterior density (6) that stem from the fixed, chosen prior distribution to be washed out by the data as n tends to infinity. $\nu(\theta)$ indeed ceases to be important even for moderate n, inviting subject-independent Jeffreys–Box–Tiao-style choices for this parametric part, but the nonparametric part $M(x, \theta)$ turns out to match the Fisherian part $\prod_{j=1}^{n} f_\theta(x_j)$ in importance.

A Stirling approximation shows that

$$\prod_{i=1}^{m} z_i^{np_i}/\Gamma(kz_i + np_i) \doteq B_n e^{-nI(p:z)} \prod_{i=1}^{m} p_i^{1/2 - kz_i},$$

where $B_n = (2\pi)^{-m/2}\exp\{n - (n + k - \frac{1}{2}m)\log n\}$ is independent of the probability vectors $p = (p_1, \ldots, p_m)$ and $z = (z_1, \ldots, z_m)$. Here $I(p : z) = \sum_{i=1}^{m} p_i \log(p_i/z_i)$ is the Kullback–Leibler distance from p to z, cf. (1), and is convex in p with a unique minimum at $p = z$. From (6) and (7) we get the approximation

$$\nu(\theta|x) \doteq c(x) e^{-nQ(F_n, \theta)} \nu(\theta) \tag{9}$$

for the posterior density, ignoring some lower-order terms, where

$$Q(F_n, \theta) = \sum_{i=1}^{m} F_n(h_\theta B_i) \log(F_n(h_\theta B_i)/z_i) - \frac{1}{n} \sum_{j=1}^{n} \log f_\theta(x_j). \tag{10}$$

The posterior density is concentrated where $Q(F_n, \theta)$ is smallest, and the Bayes estimator $\hat{\theta}$ should asymptotically behave as

$$\tilde{\theta} = \phi(F_n) = \text{the } \theta \text{ minimising } Q(F_n, \theta). \tag{11}$$

Under regularity conditions, therefore, both $\hat{\theta}$ and $\tilde{\theta}$ are frequentist consistent estimators for

$$\theta_2 = \phi(F) = \text{the } \theta \text{ minimising } Q(F, \theta), \tag{12}$$

where

$$Q(F,\theta) = \sum_{i=1}^{m} F(h_\theta B_i)\log(F(h_\theta B_i)/z_i) - \int \log f_\theta(x)\, dF(x)$$

(13)

$$= I\big((Fh_\theta B_i)_{i=1}^m : (z_i)_{i=1}^m\big) + I(f : f_\theta) - \int f \log f\, dx.$$

θ_2 enjoys interpretation as a least false parameter, and lies intuitively somewhere between the θ_1 that $\hat\theta_{\mathrm{ML}}$ aims at, discussed around (1), and a third variant θ_3 that minimises the first term in $Q(F,\theta)$, aiming at getting $(Fh_\theta B_i)_{i=1}^m$ as close to $(z_i)_{i=1}^m$ as possible.

When f really belongs to the parametric family, say $f = f_{\theta_0}$, then θ_1, θ_2, and θ_3 all coincide with the (then true) value θ_0.

When F is discrete, only $o(n)$ of $X_1,\dots,X_n$ are distinct, a.s., and the first term in $Q(F_n,\theta)$ dominates. Hence $\hat\theta$ is consistent for θ_3 in this case!

Hjort (1985b) gives conditions under which $n^{1/2}(\hat\theta - \tilde\theta) \to 0$ in probability. Limiting properties of $n^{1/2}(\hat\theta - \theta_2)$ can therefore be investigated by studying the functional ϕ above. Its influence function can be calculated. Preliminary work indicates that these are of the robust type (more cautious than for $\hat\theta_{\mathrm{ML}}$ in typical models) and that a reasonable efficiency is retained in the idealised case $f = f_{\theta_0}$. To cite but one example, $\hat\mu, \hat\sigma$ constructed as in the example of Section 2 have

$$\begin{pmatrix} n^{1/2}(\hat\mu - \mu) \\ n^{1/2}(\hat\sigma - \sigma) \end{pmatrix} \to_D N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} (1.071\sigma)^2 & 0 \\ 0 & (0.866\sigma)^2 \end{pmatrix}\right)$$

in the idealised Gaussian case. This compares well with the optimal covariance matrix $\mathrm{diag}(\sigma^2, \sigma^2/2)$.

4. Semiparametric Bayesian analysis of survival analysis models. We still need more experience in and knowledge of the consequences of Bayesian analysis of semi- and nonparametric models. Frequentist asymptotic analysis seems to have been restricted to cases where the Dirichlet process, in various disguises, has been the prior. Another testing ground could be models in survival analysis with censored data, where manageable prior processes other than the Dirichlet are available (cf. Ferguson and Phadia (1979) and Hjort (1984a)). (The classical results about maximum likelihood and Bayes analysis for i.i.d. frameworks with finitely many parameters carry over to say counting process models with censoring; see Hjort (1984b).)

Beta processes are introduced in Hjort (1984a, 1985c) as natural priors for cumulative hazard rates in nonparametric models with censoring. Suppose for example that a crude model specifies a constant rate θ for transitions from state s to state s' in a (possibly time-inhomogeneous) Markov chain. A semiparametric supermodel could structure the underlying unknown cumulative hazard $A(t)$ via $1 - dA(s) = (1 - dB(s))^\theta$, where $B(\cdot)$ is a beta process centered at $B_0(\cdot)$, $B_0(t) = t$; there is also a strength of belief parameter *function* $k(\cdot)$ to be specified. If θ is also given a prior the posterior density of θ can be worked out. Preliminary

investigation indicates that the Bayes estimate $\hat{\theta}$ converges to a value that depends on the chosen function $k(\cdot)$; this in contrast to results for the method outlined in Sections 2 and 3.

As another example, consider Cox's regression model. Imagine individual i having its own cumulative hazard $A_i(\cdot)$, and assume proportional hazards $1 - dA_i(s) = (1 - dA(s))^{\theta \exp(\beta z_i)}$, where z_i is the covariate vector, and $A(\cdot)$ is close to having unit rate $A_0(t) = t$. As above, a prior on (θ, β) can be given, and A can be taken as a beta process centered at A_0 with strength of belief parameter $k(\cdot)$. The posterior density of β can be handled. The Bayes estimate $\hat{\beta}$ is close to the usual Cox estimate for $k(\cdot)$ close to zero, and is close to the maximum likelihood estimate based on the (θ, β) model with $A = A_0$ when $k(\cdot)$ is large. The asymptotic fate of $\hat{\beta}$ is unclear for intermediate choices of $k(\cdot)$. I hope to pursue these matters later.

5. Additional remarks.

(a) D & F state that the Bayes estimates do worse than available frequentist procedures, e.g., the empirical median M_n (D & F (1986), Section 1, Remark 4). Bayes procedures that match M_n in performance can be constructed, however, if the problem is just this, i.e., estimating the true median. Let $\hat{\theta}_n$ be the Bayes estimate (posterior expectation) based on any Dirichlet process prior $\mathrm{Dir}(kF_0)$. Then $\hat{\theta}_n$ is close to the interesting estimator $\theta_n^* = \sum_{i=1}^n \binom{n-1}{i-1} (\frac{1}{2})^{n-1} x_{(i)}$, assuming $x_{(1)} < \cdots < x_{(n)}$: $\hat{\theta}_n - \theta_n^* \to_p 0$ if $k/n \to 0$, and $n^{1/2}(\hat{\theta}_n - \theta_n^*) \to_p 0$ if $k/n^{1/2} \to 0$. Also, $n^{1/2}(\theta_n^* - M_n) \to_p 0$. These statements are valid with some restrictions on the tails of the underlying continuous F.

(b) The work of D & F, and the present contribution, can be seen as an attempt to construct Bayesian robust procedures qualitatively similar to those recently worked out by Beran (1981), Millar (1981), and others: full efficiency at the parametric model and Le Cam-type robust optimality in a (shrinking) neighbourhood. The construction of Sections 2 and 3 above seems to manage this only in the not very satisfactory asymptotic framework where $k/n \to \infty$, k being the prior sample size parameter, cf. (6), where $M(x, \theta)$ is dominated by $\prod_{j=1}^n f_\theta(x_j)$ under this assumption.

(c) The semiparametric Bayes estimates constructed in Sections 2 and 3 have frequentist relatives that behave equally well (?) asymptotically. The parameter k must be specified by the user in (6) in order to compute the Bayes estimate; however (i) the asymptotics are independent of k, and (ii) it would also be possible to estimate k from the data. k large means a good fit to the parametric model.

(d) I welcome papers such as these (D & F) and Rubin (1984), discussing points of overlap and of mutual interest for Bayesianism and frequentism. I agree with Diaconis (1985) when he observes that the controversy seems to have lost its power to polarise. I also agree with a slight variation of another statement in Diaconis (1985): We should focus on the coming controversy—with those who think the computer has taken over.

REFERENCES

BERAN, R. (1981). Efficient robust estimation in parametric models. *Z. Wahrsch. verw. Gebiete* **55** 91–108.

DIACONIS, P. (1985). Bayesian statistics as honest work. *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer.* (L. M. Le Cam and R. A. Olshen, eds.) **1** 53–64. Wadsworth, Monterey, Calif.

DIACONIS, P. and FREEDMAN, D. (1986). On the consistency of Bayes estimates. *Ann. Statist.* **14** 1–26.

DOSS, H. (1985a). Bayesian nonparametric estimation of the median. I: Computation of the estimates. *Ann. Statist.* **13** 1432–1444.

DOSS, H. (1985b). Bayesian nonparametric estimation of the median. II: Asymptotic properties of the estimates. *Ann. Statist.* **13** 1445–1464.

FERGUSON, T. S. and PHADIA, E. G. (1979). Bayesian nonparametric estimation based on censored data. *Ann. Statist.* **7** 163–186.

HJORT, N. L. (1984a). Nonparametric Bayes estimators of cumulative intensities in models with censoring. Research Report, Norwegian Computing Center, Oslo.

HJORT, N. L. (1984b). Bayes estimators and asymptotic efficiency in parametric counting process models. Research Report, Norwegian Computing Center, Oslo.

HJORT, N. L. (1985a). Notes on the theory of statistical symbol recognition. Research Report, Norwegian Computing Center, Oslo.

HJORT, N. L. (1985b). Semi-parametric Bayes estimators. Unpublished.

HJORT, N. L. (1985c). Contribution to the discussion of Andersen and Borgan's "Counting process models for life history data: A review." *Scand. J. Statist.* **12** 141–150.

MILLAR, P. W. (1981). Robust estimation via minimum distance methods. *Z. Wahrsch. verw. Gebiete* **55** 73–89.

RUBIN, D. B. (1984). Bayesianly justifiable and relevant frequentist calculations for the applied statistician. *Ann. Statist.* **12** 1151–1172.

NORWEGIAN COMPUTING CENTER
P. B. 335 BLINDERN
N-0314 OSLO 3
NORWAY

WILLIAM S. KRASKER AND JOHN W. PRATT

Harvard University

This is one in a series of fascinating papers. They are easily read as painting a picture of modern Bayesianism in bad trouble but frequentism in fine shape. A larger historical perspective gives a different view, however. It should therefore be sketched, even if only impressionistically.

Time began in darkness and "inverse" probability. Then the Rev. Thomas Bayes let in some posthumous light. The postulate he identified and used was duly found arbitrary or ambiguous, and unfounded. Likewise Fisher's reference sets. Then Neyman and Pearson developed "objective" (frequentist) concepts even as Ramsey and de Finetti were proving that "subjective" Bayesianism was the only coherent theory possible. Soon (well before Pratt's 1961 and 1965 surveys) objective methods too were found arbitrary and theoretically and practically deficient even in the simplest situations (where uniformly most powerful tests are randomized).