PRIOR FEEDBACK: A BAYESIAN APPROACH TO MAXIMUM LIKELIHOOD ESTIMATION

by

Christian P. Robert
Université Paris VI and Purdue University

Technical Report #91-49C

Department of Statistics
Purdue University

August 1991

$\star$

PRIOR FEEDBACK: A BAYESIAN APPROACH TO MAXIMUM LIKELIHOOD DISTRIBUTION

by

Christian P. Robert
Université Paris VI and Purdue University

Abstract

We provide a Bayesian derivation of the maximum likelihood estimators in exponential families through a method called prior feedback. In this case, the maximum likelihood estimator can be argued to be a noninformative answer. In addition to this theoretical justification of the maximum likelihood estimator, we consider some extensions which are of more practical interest. In particular, prior feedback can provide an efficient alternative algorithm for the estimation of the parameters of a mixture.

Keywords: Maximum likelihood; conjugate priors; noninformative prior; mixture; EM algorithm.

AMS Subject Classification (1985): 62A10, 62F15, 62F10, 62F03.

1. Introduction

1.1. Maximum likelihood estimation. It is rather curious to consider that, despite the extensive use of maximum likelihood estimation in statistical inference, this method has no deep justification for small samples. Actually, apart from various asymptotic optimality arguments, the main reason for maximum likelihood estimation seems to rely on the heuristic argument of a "reasonable" criterion, which furthermore complies with the likelihood principle. Obviously, in many particular cases, the mle has been shown to be optimal in some sense (admissibility, minimaxity, ...) but other cases have been exhibited where the mle is not to be used (see, e.g., LeCam (1990)).

In this paper, we show that the mle can also be considered as a *Bayesian noninformative answer*, in a weak sense to be precised below. Apart from justifying any further maximum likelihood estimation, this result has two other implications: first, it shows that mle's can be considered as noninformative Bayes estimates when regular noninformative priors (e.g., reference priors) cannot be used. The second (and more practical) implication is that Bayesian techniques may also be used to compute mle's and come as competitors of more traditional techniques such as the *EM algorithm* (Dempster, Laird and Rubin (1977)), especially in the light of the recent progresses of *Gibbs sampling* (Gelfand and Smith (1990), Casella and George (1991)). Therefore, the connection exhibited here may be of interest to both Bayesians and likelihoodists.

The method through which we obtain the results stated above is called *prior feedback* because it uses repeatedly the observed data to update the prior in order to remove the influence of the prior input. Although this provides the mle as a formal Bayes estimator at the end of the updating process, further uses of the posterior distribution are rather limited, because it still depends on some prior parameters. However, a similar procedure can be enhanced for other parameters of interest.

1.2. Noninformative deadends. Even when they are well defined, noninformative priors are not always "foolproof" and may lead to answers that are definitely suboptimal.

Example 1. Consider $x \sim \mathcal{N}(\theta, 1)$ when the parameter of interest is e^θ, to be estimated under squared error loss. The noninformative prior is the Jeffreys prior, $\pi(\theta) = 1$, which leads to the Bayes estimator $\delta^\pi(x) = \exp(x + 1/2)$; δ^π is inadmissible and dominated by the mle, e^x. $\triangle$

Example 2. Consider $x \sim \mathcal{N}_p(\theta, I_p)$ and $\|\theta\|^2$ is to be estimated under squared error loss. The Bayes estimator is $\delta^\pi(x) = \|x\|^2 + p$, dominated by the mle, $\|x\|^2$. $\triangle$

There are also cases when noninformative priors (or more generally improper priors) cannot be defined, because the structure of the problem is somehow too weak to withstand vague priors. In such cases, noninformative substitutes have to be provided when no prior information is available.

Example 3. A mixture of two distributions,

$$x_1, \ldots, x_n \sim pf(x|\theta_1) + (1-p)f(x|\theta_2),$$

is typically a setup where improper priors cannot be used since, whatever n is, the likelihood can be decomposed into a sum of n terms, the first one being

$$p^n \prod_{i=1}^{n} f(x_i|\theta_1)$$

and thus providing no information about θ_2. Note that the same reason implies the absence of a mle when $f(x|\theta)$ is unbounded (in θ). $\triangle$

Example 4. Consider $x \sim \mathcal{N}(\theta, 1)$ and the null hypothesis $H_0: \theta = 0$ is to be tested. If we use a noninformative prior $\pi(\theta) = c$ on $H_1: \theta \neq 0$ and the weight π_0 for H_0, the posterior probability depends on c:

$$P^\pi(\theta = 0|x) = \left\{1 + c\ \frac{1-\pi_0}{\pi_0}\ \sqrt{2\pi} e^{x^2/2}\right\}^{-1}.$$

Even though $c = 1$ may appear as the most "natural" choice (since $\theta = 0$ has the same weight under the two hypotheses), there is still no indication in the usual noninformative approaches about which value of c should be used. $\triangle$

1.3. The limitations of limits. In such cases where noninformative priors cannot be defined directly, sequences of conjugate priors may sometimes provide a substitute since, actually, many noninformative priors can be written as limits of conjugate priors. However, this substitute is not necessarily of interest!

Example 3. Consider the particular mixture

$$x_1, \ldots, x_m \sim 0.4\, \mathcal{N}(\theta_1, 1) + 0.6\, \mathcal{N}(\theta_2, 1),$$

for which a mle, although not explicit, does exist. If we use the conjugate prior $\mathcal{N}(0, n)$ on both θ_1 and θ_2, the Bayes estimator of θ_1 can be written:

$$\mathbb{E}^{\pi_n}[\theta_1 | x_1, \ldots, x_m] = \sum_{k=0}^{m} \sum_{(i_\ell)} w_k^n(i_\ell)\, \frac{k}{k + 1/n}\, \overline{x}_1(i_\ell)$$

where the second sum is taken upon all partitions of $\{x_1, \ldots, x_m\}$ into two subsamples $\{x_{i_1}, \ldots, x_{i_k}\}$ and $\{x_{i_{k+1}}, \ldots, x_{i_m}\}$,

$$\begin{aligned} &\overline{x}_1(i_\ell) = \frac{1}{n} \sum_{\ell=1}^{k} x_{i\ell},\ \overline{x}_2(i_\ell) = \frac{1}{m-k} \sum_{\ell=k+1}^{m} x_{i\ell}, \\ &s_1^2(i_\ell) = \sum_{\ell=1}^{k} (x_{i\ell} - \overline{x}_1(i_\ell))^2,\ s_2^2(i_\ell) = \sum_{\ell=k+1}^{m} (x_{i\rho} - \overline{x}_2(i_\ell))^2 \end{aligned} \tag{1.1}$$

and

$$w_k^n(i_\ell) \propto \frac{(0.4)^k (0.6)^{m-k} \exp -\{s_1^2(i_\ell) + s_2^2(i_\ell)\}/2 \exp \frac{1}{2} \left(\overline{x}_1(i_\ell)^2 \frac{k}{kn+1} + \overline{x}_2^2(i_\ell) \frac{m-k}{(m-k)n+1} \right)}{\sqrt{(k + \frac{1}{n})(m - k + \frac{1}{n})}}$$

so that

$$\sum_{k=0}^{m} \sum_{(i\ell)} w_k^n(i_\ell) = 1.$$

Therefore, it follows immediately from (1.1) that the limit (in n) of the Bayes estimators is not defined since w_0^n and w_m^n go to infinity with n. A sequence of conjugate priors is thus of no use in this setting. $\triangle$

Example 4. The use of a sequence of conjugate priors for testing H_0 leads to what is known as the Jeffreys–Lindley paradox, namely the fact that the posterior probability

$P^{\pi_n}(\theta = 0|x)$ goes to 1 for every value of x when π_n is $\mathcal{N}(0, n)$ and n goes to infinity. It is often argued that this phenomenon indicates that a noninformative approach is not appropriate in a point null testing setting. (For a recent discussion of Jeffreys–Lindley paradox, see Aitkin (1991).) $\triangle$

2. The Prior Feedback Alternative

2.1. Definition and motivations. We consider a setting where conjugate priors are available for $f(x|\theta)$. This is typically the case of exponential families or of mixtures of exponential families and we will denote these conjugate priors by $\pi(\theta|x_0, \lambda)$ since, if

$$f(x|\theta) = c(x)\exp(\theta \cdot x - \psi(\theta)),$$

the conjugate priors are of the form

$$\pi(\theta|x_0, \lambda) \propto \exp(\theta \cdot x_0 - \lambda\psi(\theta)).$$

Obviously, conjugate priors are of little use in a noninformative situation since the hyperparameters x_0 and λ cannot be determined. However, if we start from an arbitrary conjugate prior $\pi(\theta|x_0, \lambda)$ and estimate $h(\theta)$, is seems reasonable to assume that the posterior expectation $\mathsf{E}^\pi[h(\theta)|x]$ is a better estimator than $\mathsf{E}^\pi[h(\theta)]$ since it is making use of the information contained in x, while $\mathsf{E}^\pi[h(\theta)]$ is essentially arbitrary. In this sense, a conjugate prior which would give $\mathsf{E}^\pi[h(\theta)|x]$ as a *prior expectation* for $h(\theta)$ would be better than the original prior. (Lemma 2 below gives a more precise meaning to this statement.)

Following this heuristic reasoning, we assume that the quantity of interest, $h(\theta)$, is such that $\mathsf{E}^\pi[h(\theta)]$ is in one–to–one correspondence with x_0. Actually, we assume in the sequel that h is strictly monotonic. (In order to avoid confusion, we denote the prior expectation as $\mathsf{E}^\pi[h(\theta)|x_0, \lambda]$ to stress the dependence on the hyperparameters.) Then, the new prior $\pi(\theta|x_1, \lambda)$, defined by the relation

$$\mathsf{E}^\pi[h(\theta)|x_1, \lambda] = \mathsf{E}^\pi[h(\theta)|x, x_0, \lambda], \tag{2.1}$$

should improve the estimation of $h(\theta)$. Naturally, the new estimator $\mathsf{E}^\pi[h(\theta)|x, x_1, \lambda]$ is still arbitrary and we can argue the same way in favor of a new prior $\pi(\theta|x_2, \lambda)$ defined

as in (2.1) with x_2 replacing x_1 and x_1 replacing x_0. This leads to the following iterative scheme: the prior distribution $\pi(\theta|x_n, \lambda)$ is defined by the implicit equation in x_n,

$$\mathsf{E}^{\pi}[h(\theta)|x_n, \lambda] = \mathsf{E}^{\pi}[h(\theta)|x, x_{n-1}, \lambda]. \tag{2.2}$$

The *prior feedback distribution* $\pi(\theta|x^*, \lambda)$ is then defined as the limit of this process, i.e. x^* satisfies the fixed point equation

$$\mathsf{E}^{\pi}[h(\theta)|x^*, \lambda] = \mathsf{E}^{\pi}[h(\theta)|x, x^*, \lambda]. \tag{2.3}$$

Note one interesting feature of (2.3): the prior feedback distribution is such that the prior expectation of $h(\theta)$ is equal to the posterior expectation of $h(\theta)$, i.e. the observation x does not modify the prior information or, conversely, the prior information contained in $\pi(\theta|x^*, \lambda)$ perfectly agrees with the information brought by x. In this sense, $\pi(\theta|x^*, \lambda)$ represents a *neutral prior*, since it somehow coincides with the sample information.

The estimator of $h(\theta)$ derived by this method, namely (2.3), has undoubtedly some undesirable properties. First, since the "prior" $\pi(\theta|x^*, \lambda)$ depends on x, the corresponding estimator is no longer Bayes and, therefore, while corresponding formally to a Bayes estimator for every x, it is likely to be suboptimal from a decision–theoretic point of view. Moreover, this approach cannot be called *noninformative* since the resulting prior still depends on the scale parameter λ. In most cases, a robustness study will therefore be necessary for assessing the sensitivity of (2.3) with respect to λ. However, we will see below that this analysis will not consider the variance going to infinity but, in the contrary, to zero. While being surprising, this phenomenon is consistent with the examples given in 1.2, where a totally noninformative approach is impossible or, at least, unsatisfactory.

Example 5. Consider again $x \sim \mathcal{N}(\theta, 1)$ when θ is the parameter of interest. Since the conjugate priors are of the form $\mathcal{N}(x_0/\lambda,\ 1/\lambda)$, the recurrence relation (2.2) can be written

$$\frac{x_n}{\lambda} = \frac{x_{n-1} + x}{\lambda + 1}$$

and leads to the limit

$$x^* = \lambda x,$$

which gives $\delta^*(x) = x$ as a prior feedback estimator, whatever λ is. $\triangle$

Example 1. For the estimation of e^θ, if $\theta \sim \mathcal{N}(x_0/\lambda,\ 1/\lambda)$, we have

$$\mathsf{E}^\pi[e^\theta|x_0,\lambda] = \exp\left\{\frac{x_0}{\lambda} + \frac{1}{2\lambda}\right\}$$

and (2.2) is then

$$\exp\left\{\frac{x_n}{\lambda} + \frac{1}{2\lambda}\right\} = \exp\left\{\frac{x_{n-1}+x}{\lambda+1} + \frac{1}{2(\lambda+1)}\right\},$$

which leads to the limit $x^* = \lambda x - \frac{1}{2}$ and the prior feedback estimator $\delta^*(x) = \exp(x)$, for every value of λ. $\triangle$

2.2 Convergence results. In the two previous examples, the estimator derived by the prior feedback method happens to be the maximum likelihood estimator, for every value of the hyperparameter λ. This independence, however, does not hold in every case and λ usually has to be large enough for the prior feedback estimator to be close to the maximum likelihood estimator. This result is rather surprising since we have to take the variance *small enough* to obtain convergence to the mle, but we can relate it to the more straightforward statement that an iterative replacement of the prior distribution by the corresponding posterior leads to a Dirac mass at the mle, since

$$\begin{aligned}\pi_n(\theta|x) &\propto \ell(\theta|x)\pi_{n-1}(\theta)\\ &\propto \ell^n(\theta|x)\pi_0(\theta).\end{aligned}$$

The convergence of the prior feedback algorithm to the maximum likelihood estimator is somehow more subtle since we keep the structure of the prior fixed (as well as the scale parameter).

First, it is possible to extend the result of Example 5 to every exponential family $\exp(\theta \cdot x - \nabla\psi(\theta))$ when the parameter of interest is the mean of x, $\nabla\psi(\theta)$. In this case,

$$\mathsf{E}^\pi[\nabla\psi(\theta)|x_0,\lambda] = x_0/\lambda \quad \text{and} \quad \mathsf{E}^\pi[\nabla\psi(\theta)|x,x_0,\lambda] = \frac{x+x_0}{\lambda+1}.$$

The recurrence relation being identical to the one in Example 5, we deduce

Lemma 1. *For every value of λ, the prior feedback estimator of the mean of an exponential family is the maximum likelihood estimator, x.*

Therefore, the estimation of the mean in exponential families does not involve some "minimal information input" in the sense that the algorithm converges for every value of λ, i.e. even for large prior variances. This is obviously related to the special nature of the mean, which is also the mode of the distribution, and of the conjugate prior, which is similar to the likelihood.

In a more general context, the following result hints at a similar behavior of the prior feedback method and supports the heuristic argument that the passage from the prior to the posterior quantity improves the estimation of $h(\theta)$.

Lemma 2. *If $f(x|\theta)$ is unimodal in θ,*

$$\mathbb{E}^{\pi}\left[(h(\theta)-h(\hat{\theta}))^2\middle|x_n,x,\lambda\right] \leq \mathbb{E}^{\pi}\left[(h(\theta)-h(\hat{\theta}))^2\middle|x_n,\lambda\right],$$

where $\hat{\theta}$ is the mle of θ.

Proof. For $\theta < \hat{\theta}$ (resp. $\theta > \hat{\theta}$), $f(x|\theta)$ is increasing (resp. decreasing) while $(h(\theta)-h(\hat{\theta}))^2$ is decreasing (resp. increasing). The result then follows from a general inequality, namely that

$$\mathsf{E}[f_1(\theta)f_2(\theta)] \leq \mathsf{E}[f_1(\theta)]\mathsf{E}[f_2(\theta)]$$

when f_1 and f_2 are of opposite variations. □□

However, despite this improvement on the average, the prior feedback method is not necessarily converging to $h(\hat{\theta})$ for every λ, as the following examples show.

Example 6. Let $x \sim \mathcal{G}a(\alpha,\theta)$. A conjugate prior distribution on θ is $\mathcal{G}a(\beta,x_0)$. If we want to estimate θ^k, the prior expectation is

$$\mathbb{E}^{\pi}[\theta^k|x_0,\beta] = \frac{\Gamma(\beta+k)}{\Gamma(\beta)}\, x_0^{-k}$$

and the recurrence relation (2.3) is then

$$x_n^{-k}\,\frac{\Gamma(\beta+k)}{\Gamma(\beta)} = (x+x_{n-1})^{-k}\,\frac{\Gamma(\alpha+\beta+k)}{\Gamma(\alpha+\beta)},$$

which gives

$$\delta^*_\beta(x) = \left(\left[\frac{\Gamma(\alpha+\beta+k)\Gamma(\beta)}{\Gamma(\beta+k)\Gamma(\alpha+\beta)} \right]^{1/k} - 1 \right)^k \frac{\Gamma(\beta+k)}{\Gamma(\beta)} \, x^{-k}$$

as a fixed point. As Figure 1 shows, this estimator is converging to the mle $(\frac{\alpha}{x})^k$ only when β is going to infinity. $\triangle$

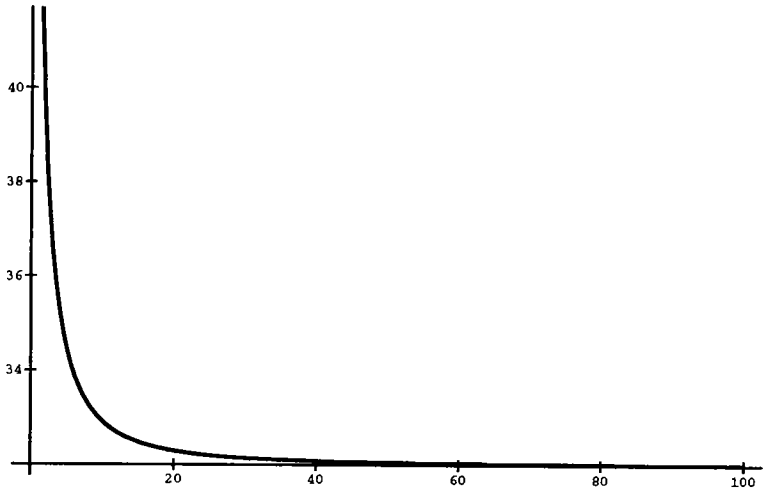

Figure 1 — Fixed point δ^*_β for $\alpha = 3$, $k = 5$ and $x = 1.5$.

Example 7. Consider again $x \sim \mathcal{G}a(\alpha, \theta)$ and $\theta \sim \mathcal{G}a(\beta, x_0)$ when the parameter of interest is $e^{-a\theta}$. The prior expectation is then

$$\mathbb{E}^\pi[e^{a\theta}|x_0, \beta] = \frac{x_0^\beta}{(a+x_0)^\beta},$$

which leads to the following recurrence relation

$$\left(\frac{x_n}{a+x_n} \right)^\beta = \left(\frac{x_{n-1}+x}{a+x_{n-1}+x} \right)^{\alpha+\beta}. \tag{2.4}$$

Obviously, the fixed point of (2.4) cannot be written analytically but, as Figure 2 shows, it is converging to $e^{-a\alpha/x}$ only when β goes to infinity. $\triangle$

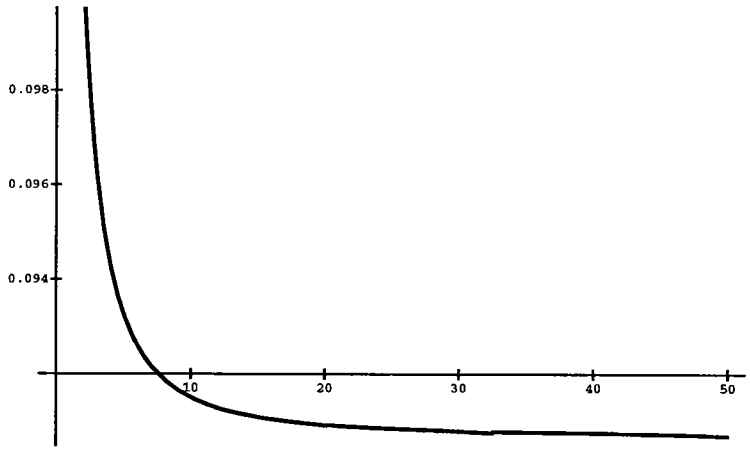

Figure 2 — Fixed point for (2.4) and $\alpha = 3,\ a = 2$ and $x = 2.5$.

The convergence of the prior feedback method to the mle is formalized in the following lemma:

Lemma 3. *The prior feedback estimator of* $h(\theta)$, $\delta^*_\lambda(x)$, *is converging to the mle* $h(\hat{\theta})$ *as* λ *goes to infinity.*

Proof. For the conjugate prior $\pi(\theta|x_0, \lambda)$, if $\xi(\theta) = \nabla\psi(\theta)$ is the mean parameterization, the mode is equal to the mean, namely $\xi^{-1}(\frac{x_0}{\lambda})$. When λ goes to infinity, this distribution converges to a Dirac point mass at 0 or at $\lim_{\lambda\to+\infty} x_0/\lambda$ when x_0 depends on λ.

Therefore, the prior feedback estimator of $h(\theta)$ converges to $\lim_{\lambda\to+\infty} h\left(\xi^{-1}(\frac{x^*}{\lambda})\right)$. Since $\mathbb{E}^\pi[h(\theta)|x^*, \lambda] = \mathbb{E}^\pi[h(\theta)|x, x^*, \lambda]$, this implies, for λ large enough, $h\left(\xi^{-1}(\frac{x^*}{\lambda})\right) \simeq h\left(\xi^{-1}(\frac{x+x^*}{\lambda+1})\right)$, i.e.

$$\frac{x^*}{\lambda} \simeq \frac{x + x^*}{\lambda + 1} \simeq x$$

and $h\left(\xi^{-1}(\frac{x^*}{\lambda})\right)$ converges to $h(\hat{\theta})$. □□

3. Extensions and Conclusions

3.1. Further examples. The results obtained in the previous section stress the link existing between maximum likelihood and conjugate priors; the mle can actually be written

as a limit of a sequence of Bayes estimators. However, from a practical point of view, this relation may appear to be rather limited since (a) it was only obtained for exponential families, where mle's are easy to compute, and (b) the function h had to be strictly monotonic. We consider below some extensions of the phenomenon in both directions.

Example 8. Let $x \sim \mathcal{N}(\theta, 1)$ and the hypothesis $H_0: \theta \leq 0$ is to be tested. A conjugate prior in this setup is $\theta \sim \mathcal{N}(x_0, \sigma^2)$, with

$$\pi(\theta \leq 0 | \sigma^2) = \Phi(-x_0/\sigma).$$

The prior feedback recurrence relation is then

$$\Phi(-x_n/\sigma) = \Phi\left(-\frac{\sigma^2 x + x_{n-1}}{\sigma\sqrt{\sigma^2+1}}\right),$$

which leads to the fixed point

$$\gamma^*(x) = \Phi\left(-\frac{\sqrt{\sigma^2+1}+1}{\sigma} x\right).$$

When σ goes to 0, $\gamma^*(x)$ goes to 0 if $x \geq 0$, 1 otherwise. This answer is also the mle, $\mathbb{I}_{[-\infty,0]}(x)$. $\triangle$

Example 9. In the more complicated case of the point null hypothesis, we consider the prior distributions

$$\pi(\theta | \pi_0, \sigma^2) = \pi_0 \mathbb{I}_0(\theta) + (1-\pi_0)\mathcal{N}(0, \sigma^2).$$

Then

$$\pi(\theta = 0 | \pi_0, \sigma^2, x) = \left[1 + \frac{1-\pi_0}{\pi_0} \frac{e^{x^2\sigma^2/2(\sigma^2+1)}}{\sqrt{\sigma^2+1}}\right]^{-1}.$$

If we actualize π_0 according to the prior feedback algorithm, the derived recurrence relation is

$$\pi_n = \left(1 + \frac{1-\pi_{n-1}}{\pi_{n-1}} \frac{e^{x^2\sigma^2/2(\sigma^2+1)}}{\sqrt{\sigma^2+1}}\right)^{-1} \tag{3.1}$$

and it is straightforward to deduce from (3.1) that π_n is converging to 1 if

$$e^{x^2\sigma^2/2(\sigma^2+1)} < \sqrt{\sigma^2+1} \tag{3.2}$$

and to 0 if $e^{x^2\sigma^2/2(\sigma^2+1)} > \sqrt{\sigma^2+1}$. Note that (3.2) is always satisfied for $|x| < 1$. This implies that, whatever σ^2 is, the prior feedback answer is to accept the null hypothesis when $|x| < 1$. This result can be related to Berger and Sellke (1987), where the authors established that the lower bound on the Bayes factors is 1 for $|x| < 1$.

Another implication of the prior feedback result is that the answer depends on σ^2 when $|x| > 1$. If we let σ^2 go to infinity, we are back to the Jeffreys–Lindley paradox, namely that we always accept H_0. If, in the contrary, we let σ^2 go to 0, it is easily seen that (3.2) is not satisfied for any given $|x| > 1$. Therefore, the prior feedback method leads to reject the null hypothesis when $|x| > 1$ if we start with σ small enough. A possible interpretation of this opposition between the two limits is that noninformative approaches lead to paradoxes in point null testing because noninformative modellings should not be undertaken in these settings. That is, for instance, the point of view of DeGroot (1972). In Caron and Robert (1991), we consider that an opposite view can be held: prior feedback indirectly provides a noninformative answer for point null tests. Actually, (3.1) is a fixed point equation if

$$e^{x^2\sigma^2/2(1+\sigma^2)} = \sqrt{\sigma^2+1}, \tag{3.3}$$

which is a null event in terms of x but determines a value of σ^2 such that prior and posterior probabilities agree. This choice of $\sigma^2(= \sigma^2(x))$ is then *neutral* in the sense mentioned above. (See Caron and Robert (1991) for more details.) $\triangle$

Example 3. Mixtures of exponential families provide an extension where Lemma 3 does not necessarily hold and where studies of convergence are much more intricated. In Robert and Soubiran (1991), we established an equivalent of Lemma 3 when only the weights of the mixture distribution are unknown. In the more general case of a mixture of two bivariate normal distributions, we were only able to show through examples that the prior feedback method was also leading to the mle when the variance parameters were decreasing.

However, this setup provides an important field of applications for the prior feedback method, if not a true noninformative alternative. Actually, available maximum likelihood algorithms, such as the *EM algorithm*, are generally performing satisfactorily except in

most difficult cases (see Robert (1991)). With the apparition of new simulations methods such as *Gibbs sampling*, it is now possible to implement very easily conjugate priors Bayesian estimation in mixture and other missing data models (see Gelfand and Smith (1990) and Diebolt and Robert (1990)) and thus propose prior feedback as a reasonable alternative for maximum likelihood estimation, especially for small samples or imbricated models where the Bayesian paradigm seems to bring a stability the other methods lack.Δ

3.2. The limitations of the prior feedback method. As we mentioned before, this method cannot claim to be fully Bayesian since the limiting distribution is determined with respect to the observation. At best, it can be classified as a more elaborate empirical Bayes method. Moreover, the fact that the limiting distribution and the associated estimator still depend on the scale parameter λ is also a major drawback since, if we let λ go to infinity, the resulting distribution cannot be used for other inferential purposes and, for a fixed λ, the approach cannot be called noninformative.

Obviously, some extensions could be proposed in order to provide an acceptable value of λ. For instance, one could look for a stabilization of the prior feedback estimator as λ goes to infinity. But the values thus obtained may be too large to be of any use; this occurs for instance for tests.

Another drawback is that this method heavily relies on the use of conjugate priors. Therefore, it excludes models where no conjugate prior exists, like *t–distributions*. However, an extension is possible when the distribution is a continuous mixture of an exponential family, since a conjugate prior exists for the mixed distribution (see Robert (1990)). For example, the normal distribution is again a "conjugate" prior in the case of the t–distribution.

Despite these negative aspects, the prior feedback method has some compelling features. On the theoretical side, it formalizes a link between the maximum likelihood estimators and the Bayesian paradigm and therefore justifies the mle on a more decision–theoretic ground. On the practical side, it may provide an alternative algorithm for the computation of the mle, even though it has no practical use for exponential families.

Acknowledgements

Part of this research was performed at the Cornell workshop on Conditional Inference sponsored by the U.S. Army Mathematical Sciences Institute and the Statistics Center, June 3–14. The author is also grateful to Steve McEachern for helpful conversations and to Jim Berger for additional financial support through NSF Grant DMS–8717799.

References

Aitkin, M, (1991). Posterior Bayes factors (with discussion). *J.R. Statist. Soc. B* **53**, 111–142.

Berger, J. and Sellke, T. (1987). Testing a point null hypothesis: the irreconciability of p–values and evidence (with discussion). *JASA* **82**, 112–122.

Caron, N. and Robert, C. (1991). Noninformative Bayesian testing and neutral Bayes factors. Rapport technique #140, L.S.T.A., Université Paris 6.

Casella, G. and George, E. (1991). Explaining the Gibbs sampler. Tech. Report 96, Statistics Research Center, University of Chicago.

DeGroot, M. (1982). Comments on Shafer (1982). *JASA* **77**, 336–338.

Dempster, Laird, N. and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J.R. Statist. Soc. B* **39**, 1–38.

Diebolt, J. and Robert, C. (1990). Estimation of finite mixture distributions through Bayesian sampling. Rapport technique #121, L.S.T.A., Université Paris 6.

Gelfand, A. and Smith, A. (1990). Sampling based approaches to calculating marginal densities. *JASA* **85**, 398–409.

LeCam, L. (1990). Maximum likelihood: an introduction. *Int. Statist. Rev.* **58**, 153–170.

Robert, C. (1990). Hidden mixtures and Bayesian sampling. Rapport technique #115, L.S.T.A., Université Paris 6.

Robert, C. (1991). Comments on Meng and Rubin's paper. *Bayesian Statistics* **4**, J. Berger, J. Bernardo, R. Dawid and A. Smith (Eds.)

Robert, C. and Soubiran, C. (1991). Estimation of a mixture model through Bayesian sampling and Prior feedback. Rapport technique #138, L.S.T.A., Université Paris 6.