Jour. Ind. Soc. Ag. Statistics
49 (Golden Jub. No.), 1996-97 : 111–118

Some Examples of Nonidentifiable Improper Posteriors[1]

Malay Ghosh
University of Florida, Gaines Ville, USA

SUMMARY

The paper introduces the notion of nonidentifiable posteriors, and brings in several examples of the same from different areas of statistics. It is pointed out also that often the impropriety of the posterior has gone unnoticed in literature.

Key words: Nonidentifiable improper posteriors, Bayesian methods, Simple location and item-response models, Generalised linear mixed effects model with spatial correlation.

1. *Introduction*

Bayesian methods are finding increasing acceptance in the theory and practice of statistics. This can partly be attributed to the fact that even with little or vague prior information, Bayesian methods can be used very effectively by employing "diffuse" or "noninformative" priors. Thus, not surprisingly, over the years, a wide variety of noninformation priors have been proposed based on diverse criteria.

One of the potential dangers with the use of such priors in that often one is led into improper posteriors. Improper posteriors can wreck havoc when one is interested in descriptive measures such as posterior means or posterior quantiles, or in inferential criteria such as credible sets. While Bayesian analysts are well-aware of this problem, in practice, the impropriety of the posterior often goes unchecked, especially for very complex models. Another cause of concern is that most Bayesian procedures, in these days, are implemented via Markov Chain Monte Carlo (MC^2) integration techniques. This requires generating samples from several conditional pdf's in order to find marginal posteriors of the parameters of interest. It may so happen that all these conditionals are proper pdf's, and yet the joint posterior pdf is improper. In other words, the impropriety of the posterior remains undetected by the MC^2 technique.

1 Research partially supported by NSF Grant Number SBR-9423996.

The impropriety of the posterior can happen in many ways, and once noninformative priors are employed, one is advised to check analytically the propriety or otherwise of posteriors by finding suitable analytical bounds of integrals under consideration. However, often there is a simple way to detect the impropriety of posteriors. If for example, the joint posterior $\pi(\theta, \phi \mid y)$ of the parameter vector (θ, ϕ) given the data y (θ, ϕ and y can be real-or-vector-valued) is expressed as $\pi(\theta, \phi \mid y) = g(\theta \mid y)$, and at least one of the components of ϕ has infinite range, then $\int \pi(\theta, \phi \mid y)\, d\phi = +\infty$ which immediately implies that $\int \pi(\theta, \phi \mid y)\, d\phi\, d\theta = +\infty$. The fact that $\pi(\theta, \phi \mid y)$ does not depend on ϕ makes the posterior nonidentifiable. The nonidentifiability is a consequence of nonidentifiable likelihoods together with nonidentifiable priors.

Often the nonidentifiability of the posterior is easier to detect by a one-to-one transformation of parameters. The fact is that if certain parameter (vector) has an improper posterior, the same is true of any one-to-one function of that parameter.

In this note, several examples of nonidentifiable improper posteriors are presented. In some of the examples, the impropriety has gone unnoticed in the literature. These examples cover a wide range of models, beginning with simple location models and dealing subsequently into more complex models such as item-response models, and generalized linear models with spatial correlation structure.

2. *Examples*

Example 1. Consider the one-way ANOVA model

$$Y_{ij} = \mu + \alpha_i + e_{ij}, (j = 1, \ldots, n_i,\ i = 1, \ldots, p) \tag{2.1}$$

where the e_{ij} are independent with some specified pdf's f_{ij}. Then, the likelihood function is given by

$$L(\mu, \alpha_1, \ldots, \alpha_p; y) = \Pi_{i=1}^{p}\ \Pi_{j=1}^{n_i}\ f_{ij}\ (y_{ij} - \mu - \alpha_i) \tag{2.2}$$

Suppose now one puts the uniform prior

$$\pi(\mu, \alpha_1, \ldots, \alpha_p) \propto 1 \tag{2.3}$$

for $(\mu, \alpha_1, \ldots, \alpha_p)$ where each component ranges over $(-\infty, \infty)$. Then, denoting the data vector by $\mathbf{y}$, the joint posterior is given by

$$\pi(\mu, \alpha_1, \ldots, \alpha_p \mid y) \propto \Pi_{i=1}^{p}\ \Pi_{j=1}^{n_i}\ f_{ij}\,(y_{ij} - \mu - \alpha_i) \tag{2.4}$$

With the one-to-one transformation $\phi = \mu$, $\theta_i = \mu + \alpha_i$ $(i = 1, \ldots, p)$, the joint posterior of $(\phi, \theta_1, \ldots, \theta_p)$ is

$$\pi(\phi, \theta_1, \ldots, \theta_p \mid y) \propto \Pi_{i=1}^{p} \; \Pi_{j=1}^{n_i} \; f_{ij}\,(y_{ij} - \theta_i) \qquad (2.5)$$

which does not depend on ϕ. Integrating with respect to ϕ over $(-\infty, \infty)$, one proves the impropriety of the posterior.

Remark 1. The impropriety of the posterior in this example is due to nonidentifiability of both the likelihood and the prior. A result is proved which provides a necessary and sufficient condition for the propriety of the posterior with a nonidentifiable likelihood. The result is implicit in Dawid [5] and in O'Hagan [7], but it is worth making it explicit due to its simplicity.

Proposition 1. Suppose the likelihood function $L(\theta, \phi; y)$ is expressible as $L(\theta, \phi; y) = g(\theta; y)$. Consider the prior $\pi(\theta, \phi)$ for (θ, ϕ). Then the posterior $\pi(\theta, \phi \mid y)$ is proper if and only if both the conditional pdf's $\pi(\theta \mid y)$ and $\pi(\phi \mid \theta)$ are proper.

Proof.

$$\begin{aligned}\pi(\theta, \phi \mid y) &\propto L(\theta, \phi; y)\,\pi(\theta, \phi) = g(\theta, y)\,\pi(\theta, \phi) \\ &= g(\theta, y)\,\pi(\theta)\,\pi(\phi \mid \theta) \propto \pi(\theta \mid y)\,\pi(\phi \mid \theta) \qquad (2.6)\end{aligned}$$

Remark 2. Example 1 also shows that if instead one views the likelihood as $L(\theta_1, \ldots, \theta_p; y)$ and put the flat prior $\pi(\theta_1, \ldots, \theta_p) \propto 1$, then the posterior $\pi(\theta_1, \ldots, \theta_p \mid y) \propto \Pi_{i=1}^{p} \; \Pi_{j=1}^{n_i} \; f_{ij}(y_{ij} - \theta_i)$ can be proper. In particular, if the errors e_{ij} are iid $N(0, \sigma^2)$ where $\sigma^2 (> 0)$ is known, then it is well-known that the posterior $\pi(\theta_1, \ldots, \theta_p \mid y) \propto \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{p} \sum_{j=1}^{n_i} (y_{ij} - \theta_i)^2\right] \propto \exp\left[-\frac{1}{2\sigma^2} \sum n_i (\bar{y}_i - \theta_i)^2\right]$ $(\bar{y}_i = n_i^{-1} \sum_{j=1}^{n_i} y_{ij}, i = 1, \ldots, p)$ is the product of p independent normal pdf's. This implies that the joint posterior of the elementary contrasts $(\alpha_1 - \alpha_p, \ldots, \alpha_{p-1} - \alpha_p) = (\theta_1 - \theta_p, \ldots, \theta_{p-1} - \theta_p)$ is also proper.

The normal linear model has been treated very extensively in Sahu and Gelfand [10]. Indeed, these authors have established in the normal case a very strong connection between the nonestimability of parameters, and corresponding impropriety of posteriors under diffuse priors.

Example 2. This example extends the results of Example 1 to item response models. These models have been used very extensively for the analysis of psychological data. As a specific example, consider ability tests or attitude tests where each individual answers a battery of questions. Let Y_{ij} denote the response of the ith individual to the jth question ($i = 1, ..., n$; $j = 1, ..., k$). Associated with the ith individual is a subject parameter β_i that expresses the capacity, ability or attitude of the individual in a given context. However, the distributions of the Y_{ij} will depend not only on the β_i, but also on some parameter α_j, where α_j represents the nature or the difficulty level of question j. Item response analysis models the distribution of Y_{ij} taking into account both β_i and α_j ($i = 1, ..., n; j = 1, ..., k$).

Consider the case when the Y_{ij} are binary random variables. This is, for example, the situation when n examinees are answering "True/False" questions. Alternately, the examinees may answer multiple choice questions, where each answer is coded as "correct" or "incorrect." Let $p_{ij} = P(Y_{ij} = 1)$ ($i = 1, \ldots, n; j = 1, \ldots, k$), where 1 denotes a correct response. The general form of a one-parameter item response model is

$$p_{ij} = F(\beta_i - \alpha_j) \tag{2.7}$$

where F is a distribution function. This is called a one parameter item response model since as a function of β_i, this has the form of a location-family distribution function with location parameter α_j. In the special case when $F(x) = \exp(x) / [1 + \exp(x)]$, that is F is the logistic distribution function, the model is the celebrated Rasch model. (Rasch [9]), while when $F(x) = \Phi(x)$, the cdf of the N(0,1) variable, the model is referred to as the probit model (Lord [8]).

Writing $\beta = (\beta_1, ..., \beta_n)$, $\alpha = (\alpha_1, ..., \alpha_k)$, the likelihood function is given by

$$L(\beta, \alpha; y) = \Pi_{i=1}^{n} \Pi_{j=1}^{k} [F^{y_{ij}}(\beta_i - \alpha_j) \bar{F}^{1-y_{ij}}(\beta_i - \alpha_j)] \tag{2.8}$$

where $\bar{F} \equiv 1 - F$. Suppose now one uses the flat prior $\pi(\beta, \alpha) \propto 1$. With the one-to-one transformation $\phi_i = \beta_i - \alpha_k$, ($i = 1, ..., n$), $\theta_{n+j} = \alpha_j - \alpha_k$ ($j = 1, ..., k-1$) and $\phi = \alpha_k$, the prior $\pi(\theta, \phi) \propto 1$, and the joint posterior of $\theta = (\theta_1, \ldots, \theta_n, \theta_{n+1}, \ldots, \theta_{n+k-1})$ and ϕ is given by

$$\pi(\theta, \phi \mid y) \propto \Pi_{i=1}^{n} \Pi_{j=1}^{k-1} [F^{y_{ij}}(\theta_i - \theta_{n+j}) \bar{F}^{1-y_{ij}}(\theta_i - \theta_{n+j})] \times \Pi_{i=1}^{n} [F^{y_{ij}}(\theta_i) \bar{F}^{1-y_{ij}}(\theta_i)] \tag{2.9}$$

which does not depend on $\phi \in (-\infty, \infty)$. Thus, the joint posterior is improper.

Remark 3. The findings of Example 1 may suggest that if the likelihood given in (2.7) is viewed only as a function of θ, then the diffuse prior $\pi(\theta) \propto 1$ may lead to a proper posterior $\pi(\theta \mid y)$. For Example 2, the result is true except on the boundary where $y_{i1} = \ldots = y_{ik} = 0$ or 1 for at least one i or $y_{1j} = \ldots = y_{nj} = 0$ or 1 for at least one j. The details are worked out in Ghosh, Ghosh and Agresti [6]. The difference between the present case, and the one in Example 1 may be attributed to the discreteness of the data in this example in contrast to the continuous data of Example 1.

Example 3. We continue with the likelihood given in Example 2, but consider instead the prior $\pi(\beta, \alpha) \propto \pi_1(\beta)\pi_2(\alpha)$ where

$$\pi_1(\beta) \propto \int_0^\infty \int_{-\infty}^\infty \Pi_{i=1}^n \, [\sigma_1^{-1} u_1(\beta_i - \mu)/\sigma_1)] \, m_1(\sigma_1) \, d\mu \, d\sigma_1 \qquad (2.10)$$

and

$$\pi_2(\alpha) \propto \int_0^\infty \int_{-\infty}^\infty \Pi_{j=1}^k \, [\sigma_2^{-1} u_2((\alpha_j - \mu)/\sigma_2)] \, m_2(\sigma_2) \, d\mu \, d\sigma_2 \qquad (2.11)$$

u_1, u_2, m_1 and m_2 being proper or improper pdf's. Clearly, both $\Pi_1(\beta)$ and $\Pi_2(\alpha)$ are hierarchical priors. For example, conditional on μ and $\sigma_1, \beta_1, \ldots, \beta_n$ are iid with common location-scale pdf $\sigma_1^{-1} u_1((\beta - \mu)/\sigma_1)$. Similarly, conditional on μ and σ_2, $\alpha_1, \ldots, \alpha_k$ are iid with common location-scale pdf $\sigma_2^{-1} u_2((\alpha - \mu)/\sigma_2)$. Marginally μ, σ_1, and σ_2 are mutually independent with $\mu \sim$ uniform $(-\infty, \infty)$, while σ_1 and σ_2 have pdf's $m_1(\sigma_1)$ and $m_2(\sigma_2)$. To recognize the nonidentifiability of the posterior, it is convenient to write the posterior as

$$\begin{aligned} \pi(\beta, \alpha, \mu, \sigma_1, \sigma_2 \mid y) \propto\ & \Pi_{i=1}^m \Pi_{j=1}^k \, [F^{y_{ij}}(\beta_i - \alpha_j) \bar{F}^{1-y_{ij}}(\beta_i - \alpha_j)] \\ & \times \Pi_{i=1}^n \, [\sigma_1^{-1} u_1((\beta_i - \mu)/\sigma_1)] \\ & \times \Pi_{j=1}^k \, [\sigma_2^{-1} u_2((\alpha_j - \mu)/\sigma_2)] \, m_1(\sigma_1) \, m_2(\sigma_2) \qquad (2.12) \end{aligned}$$

With the one-to-one transformation $\theta_i = \beta_i - \mu$, $(i = 1, \ldots, n)$, $\theta_{n+j} = \alpha_j - \mu$, $(j = 1, \ldots, k)$, $\theta_{n+k+m} = \sigma_m$, $(m = 1, 2)$, and $\phi = \mu$, the transformed posterior is given by

$$\pi(\theta, \phi \mid y) \propto \Pi_{i=1}^{n} \Pi_{j=1}^{k} [F^{y_{ij}}(\theta_i - \theta_{n+j}) \bar{F}^{1-y_{ij}}(\theta_i - \theta_{n+j})]$$
$$\times \Pi_{i=1}^{n} [\theta_{n+k+1}^{-1} u_1(\theta_i / \theta_{n+k+1})] \Pi_{j=1}^{k} [\theta_{n+k+2}^{-1} u_2(\theta_{n+j} / \theta_{n+k+2})]$$
$$\times m_1(\theta_{n+k+1}) m_2(\theta_{n+k+2}) \quad (2.13)$$

Since this posterior does not depend on $\phi \in (-\infty, \infty)$, its nonidentifiability and impropriety follow immediately.

Example 4. Finally, consider a generalized linear mixed effects model with some spatial correlation structure. We begin with the generalized linear model where conditional on $\theta_1, \ldots, \theta_n$, $Y_1, \ldots, Y_n$ are independent with pdf's

$$f(y_i \mid \theta_i) = \exp[\phi^{-1}(y_i \theta_i - \Psi(\theta_i))] + c(y_i; \phi)] \quad (2.14)$$

where $\theta_1, \ldots, \theta_n$ are unknown, but $\phi\ (> 0)$ is known. Important special cases are the Bernoulli model with probabilities of success p_i or the Poisson (λ_i) models. In each case $\phi = 1$. In the Bernoulli case $\theta_i = \log(p_i / (1 - p_i))$, while in the Poisson case $\theta_i = \log \lambda_i$.

Next consider the spatial mixed effects model

$$\theta_i = \mu + \mathbf{x}_i^T \mathbf{b} + u_i + v_i \quad (2.15)$$

where μ is the unknown general effect, $\mathbf{x}_i$, $(p \times 1)$ are the known design vectors, and $\mathbf{b}$ is the unknown regression vector. It is assumed that the matrix $(1, \mathbf{x}_1, \ldots, \mathbf{x}_p)$ has rank $= p + 1 < n$. The u_i and the v_i are mutually independent with the v_i iid $N(0, \sigma_v^2)$, while $u_1, \ldots, u_n$ have joint pdf

$$f(u_1, \ldots, u_n) = (\sigma_u^2)^{-\frac{1}{2}n} \exp[-\sum_{i} \sum_{\sim j} w_{ij}\, g(u_i - u_j) / (2\sigma_u^2)] \quad (2.16)$$

where g is a function symmetric about zero. The summation is over all pairs $i \sim j$ that are deemed as neighbors, and the w_{ij} are a corresponding set of specified positive weights.

To complete the Bayesian hierarchical model, it is assumed that $\mu, \mathbf{b}, \sigma_u^2$, and σ_v^2 are mutually independent with $\mu \sim \text{uniform}(-\infty, \infty)$, $\mathbf{b} \sim \text{uniform}(R^p)$, $\sigma_u^{-2} \sim \text{Gamma}(\frac{1}{2}a, \frac{1}{2}g)$, and $\sigma_v^{-2} \sim \text{Gamma}(\frac{1}{2}c, \frac{1}{2}d)$. [A random variable $Z \sim \text{Gamma}(\alpha, \beta)$ if it has pdf $f(z) \propto \exp(-\alpha z) z^{\beta - 1}$].

The prior given in (2.16) is the celebrated *pairwise difference prior* considered by several authors. Among others, this appears in Clayton and Kaldor [4], Besag, York and Mollie [2], Bernardinelli and Montomoli [1], and Besag, Green, Higdon and Mengersen [3]. The first three papers were primarily concerned with disease maps, and used some version of the Poisson likelihood slightly different from the present one. Our main contention is to show, however, is that the inclusion of the intercept term b_0 along with the pairwise difference prior given in (2.16) creates nonidentifiability in the model leading thereby to a nonidentifiable improper posterior.

To this end, first write the joint posterior of $\theta, \mathbf{u}, \mu, \mathbf{b}, \sigma_u^2$ and σ_v^2 as

$$\begin{aligned}\pi(\theta, \mathbf{u}, \mu, \mathbf{b}, \sigma_u^2, \sigma_v^2 \mid \mathbf{y}) \propto\ & \exp\Big[\sum_1^n \{\theta_i y_i - \Psi(\theta_i)\}\Big] \\ & \times \exp\Big[-\sum_{i=1}^n (\theta_i - \mu - \mathbf{x}_i^T \mathbf{b} - u_i)^2 / (2\sigma_v^2)\Big] \\ & \times (\sigma_v^2)^{-\frac{1}{2}n} \exp\Big[-\sum_{i \sim j}\sum w_{ij} g(u_i - u_j)/(2\sigma_u^2)\Big] (\sigma_u^2)^{-\frac{1}{2}n} \\ & \times \exp(-a/(2\sigma_u^2))(\sigma_u^2)^{-\frac{1}{2}g-1} \exp(-c/2\sigma_v^2))(\sigma_v^2)^{-\frac{1}{2}d-1}\end{aligned}$$

With the one-to-one transformation $z_i = \mu + u_i$ $(i = 1, \ldots, n)$, the joint posterior of $\theta, \mathbf{z} = (z_1, \ldots, z_n), \mu, \mathbf{b}, \sigma_u^2$ and σ_v^2 is given by

$$\begin{aligned}\pi(\theta, \mu, \mathbf{b}, \sigma_u^2, \sigma_v^2 \mid \mathbf{y}) \propto\ & \exp\Big[\sum_1^n \{\theta_i y_i - \Psi(\theta_i)\}\Big] \\ & \times \exp\Big[-\sum_{i=1}^n (\theta_i - \mathbf{x}_i^T \mathbf{b} - z_i)^2 / (2\sigma_v^2)\Big] \\ & \times (\sigma_v^2)^{-\frac{1}{2}n} \exp\Big[-\sum_{i \sim j}\sum w_{ij}\, g(z_i - z_j)/(2\sigma_u^2)\Big] (\sigma_u^2)^{-\frac{1}{2}n} \\ & \times \exp(-a/(2\sigma_u^2))(\sigma_u^2)^{-\frac{1}{2}g-1} \exp(-c/(2\sigma_v^2))\,(\sigma_v^2)^{-\frac{1}{2}d-1} \qquad (2.17)\end{aligned}$$

which does not depend on $\mu \in (-\infty, \infty)$. This leads to a nonidentifiable improper posterior.

Inspite of the easily recognizable impropriety of the above posterior, it appears that its impropriety has gone unnoticed in the literature prior to Besag *et al.* [3]. It is easy to fix the impropriety of this posterior by removing the intercept term. The details are omitted. Besag *et al.* [3] recognize the impropriety of the posterior, but claim that the posterior of θ is still proper. This, however, does not seem to be the case.

REFERENCES

[1] Bernardinelli, L. and Montomoli, C., 1992. Empirical Bayes versus fully Bayesian analysis of geographical variation in disease risk. *Statistics in Medicine,* **11**, 983-1007.

[2] Besag, J., York, J.C. and Mollie, A., 1991. Bayesian image restoration with two applications in spatial statistics (with discussion). *Ann. Inst. Statist. Math.,* **43**, 1-59.

[3] Besag, J., Green, P., Higdon, D. and Mengersen, K., 1995. Bayesian computations and stochastic systems (with discussion). *Statist. Sci.,* **10**, 1-66.

[4] Clayton, D.G. and Kaldor, J., 1987. Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics,* **43**, 671-681.

[5] Dawid, A.P., 1979. Conditional independence in statistical theory (with discussion). *J. Roy. Statist. Soc.,* **B41**, 1-31.

[6] Ghosh, M., Ghosh, A. and Agresti, A., 1996. *Bayesian estimation of item response models using improper priors.* Technical Report 510. Department of Statistics, University of Florida, USA.

[7] O'Hagan, A., 1994. *Kendall's Advanced Theory of Statistics.* **2B**, John Wiley and Sons, New York.

[8] Lord, F.M., 1953. An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. *Psychometrika,* **18**, 57-75.

[9] Rasch, G., 1961. On general laws, and the meaning of measurement in psychology. *Proc. of the 4th Berkeley Symposium on Mathematical Statistics,* **4**, 321-334, University of California Press.

[10] Sahu, S.K. and Gelfand, A.E., 1995. On propriety of posteriors and Bayesian identifiability in generalized linear models. Preprint.