LIKELIHOOD IMPUTATION

Li-Chun Zhang, Statistics Norway
Kongensgt 6, P.B. 8131 Dep., Oslo 0033, Norway (E-mail: lcz@ssb.no)

Key Words: Incomplete data, Latent structure models, Residual log-likelihood.

1. Introduction

Denote by $\{f(x;\theta), \theta \in \Theta\}$ some parametric model of X with parameter θ, both may possibly be vector-valued. It is said to have a *latent structure* if the complete data x can only be observed through a statistic of it, i.e. $y = Y(x)$, in which case x is said to be *latent* and y its *manifestation*. (Notice that the term "latent" refers here to the entire complete data set instead of merely its unobserved part.) In particular, $\{f_1(y;\theta), \theta \in \Theta\}$ denotes the model marginalized towards y, with the the *marginal* model function $f_1(y;\theta)$. A latent structure is *ignorable* if Y is sufficient for θ under the *complete* model, i.e. $f(x;\theta|y) = f(x|y)$, where $f(x;\theta|y)$ is the conditional probability mass (or density) function of x given $Y = y$. Nonignorable latent structures arise from all genuine incomplete-data situations, of which sample survey with nonresponse and/or measurement errors constitutes one of the many familiar as well as important instances.

Given the data y under the latent structure model, an *imputation* is any latent x^* satisfying $y = Y(x^*)$. Notice that the superscript has been used to underline that x^* is not truly observed. Moreover, denote by Ω^* the set of all possible imputations given y. Now imputations may be desirable due to the general interest in fixing up the data base for public uses, or because one wishes to study the systematic difference between the observed marginal model and the latent complete model as well as the sensitivity of the model towards the latent structure, *etc.*.

From a parametric inference point of view, an imputation is said to be *likelihood consistent* with its manifestation only if the latent structure is ignorable, in which case x^* yields the identical inference as y. Otherwise, and more generally, they lead to different inferences. We propose *to choose the imputation such that this difference is minimized.* To be explicit, denote by $T_1(Y)$ some summary statistics of inference based on the observation, and $T(X^*)$ the corresponding statistics had X^* been available. Suppose the difference in T_1 and T can in some well defined sense be measured by $D(T_1, T)$. The x^* which minimizes $D(t_1, T)$ conditional to $Y = y$ can then be considered optimal for that inferential purpose. For instance, suppose $T_1(y) = \hat{\theta}$ is the maximum likelihood estimate (m.l.e.) based on y and $f_1(y;\theta)$, and $T(x^*)$ the m.l.e. $\hat{\theta}^*$ based on x^* and $f(x;\theta)$. Suppose their difference is measured by a distance function, say, $D(\hat{\theta}, \hat{\theta}^*) = ||\hat{\theta} - \hat{\theta}^*||_2$. The optimal imputation is then x^* which yields the closest $\hat{\theta}^*$ to $\hat{\theta}$. For $f(x;\theta)$ from the exponential families of distributions, this x^* can be generated by the final E-step of the EM algorithm, based on which $\hat{\theta}^*$ coincides with $\hat{\theta}$.

Such optimality of the imputation rests clearly on the inferential purpose; and an imputation which is optimal in one sense can be poor in others. However, since all the information contained in y and x^* are summarized in their respective likelihoods, denoted by $L_1(\theta;y)$ and $L(\theta;x^*)$, we are prompted to consider the case where (T_1, T) are simply set to be these likelihoods themselves. Now the difference between the information contents of L_1 and L is, above all, reflected in the variation of the *residual likelihood* in θ, i.e.

$$\begin{aligned} L_r(\theta;x^*|y) &= L(\theta;x^*)/L_1(\theta;y) \\ \propto \quad f(x^*;\theta|y) &= f(x^*;\theta)/f_1(y;\theta), \end{aligned}$$

since L_r contains all the information which lies in x^* yet outside of y. Notice that in case x^* is likelihood consistent with y, L_r becomes a constant of θ. The likelihood imputation developed below aims therefore at making the residual likelihood as flat as possible, in which sense inference based on the imputed data is made as close as possible to that based on the observed data. Clearly, though, variants of the likelihood imputation can be derived in the similar spirit should one choose to focus on other summary statistics of inference.

In contrast, standard model-assistant single imputation method opt for the conditional mean of X averaged over its estimated conditional distribution $f(x;\hat{\theta}|y)$ (Greenlees, Reece, and Zieschang, 1982; Bjørnstad and Walsøe, 1991) which, however, does not directly address the difference in information contents between an imputation and its manifestation. For instance, in case the conditional-mean imputation coincides with the final E-step of the EM algorithm, it carries no other information than what $\hat{\theta}$ does, which is why the E-step remains only the auxiliary part of the algorithm and is largely forgotten afterwards.

Moreover, although we shall be focusing on the single imputation here, our approach does also have implications for multiple imputation (Rubin, 1976, 1987; Schenker and Welsh, 1988), under which inferences based on m imputations (for not too large m) can be combined in such ways that the results are consistent to the various orders with those directly based on the observations. Compared to single imputation, this method is particularly successful in assessing the accuracy of parameter estimation — see however Efron (1994) for a nonparametric Bootstrap approach and our discussion of the concept of *effective sample size* below. In any case, since each randomization almost surely generates different set of multiple imputations, it is still legitimate to ask which one of them we should choose.

Section 2 defines the likelihood imputation. We explain its properties and implementations. Section 3 uses two examples for illustration, sometimes in comparison with the conditional-mean single imputation. Section 4 draws attention to several potential subjects for future study. Finally, we refer more details and cases to a fuller version of the current presentation (Zhang, 1997).

2. Likelihood imputation

2.1 The definition

We measure the variation of the residual log-likelihood over some parameter region Θ_h by the *(uniform) δ-value* of x^*, i.e.

$$\delta(x^*) = \int_{\Theta_h} l_r^2/c_h \; d\theta - (\int_{\Theta_h} l_r/c_h \; d\theta)^2,$$

where $l_r = \log L_r$ and $c_h = \int_{\Theta_h} d\theta$ the Lebesgue measure of Θ_h chosen. Notice that the δ-value is thus invariant towards proportional observed likelihoods provided so is Θ_h; and it vanishes in case x^* is likelihood consistent with y. Notice also that the idea of assessing the proportionality between two functions through the variation in their log-difference has otherwise been adopted for the Gibbs stopper (Tanner and Wong, 1987; Wei and Tanner, 1990; Ritter and Tanner, 1992). Usually, we form Θ_h in the same way as we draw confidence regions for θ, i.e.

$$\Theta_h(\beta) = \{\theta : \; l_1(\hat{\theta};y) - l_1(\theta;y) \le \chi^2_\beta(d_\theta)/2\}$$

where d_θ denotes the number of free parameters and χ^2_β the β-quantile of the χ^2-distribution. Appealing to the asymptotic χ^2-distribution of the log-likelihood ratio statistic, $\Theta_h(\beta)$ is an approximate $100\beta\%$ confidence region of the parameter and will often be shorthanded as Θ_β.

Given observation $Y = y$, Ω^* becomes a linearly ordered set induced by the δ-value; and the *likelihood imputation* is such that no other imputation has a smaller δ-value. What is essential here, however, is the idea of minimizing the difference in information contents between an imputation and its manifestation, regardless of the actual quantification of this difference.

Suppose it is legitimate to exchange the integrations and the differentiations involved, the likelihood imputation satisfies the *(uniform) δ-equation*, i.e.

$$Cov_h(l_r, l_r') = 0$$

for $x^* \in \Omega^*$ and $l_r' = \partial l_r/\partial x$, w.r.t. the uniform distribution c_h^{-1} over $\Theta_h(\beta)$, i.e. the inverse of the Lebesgue measure of $\Theta_h(\beta)$. Observe that l_r on the likelihood imputation has the geometric interpretation of being orthogonal to its derivatives in the corresponding L^2-space.

Unlike the conditional-mean imputation, denoted by $\hat{x}^* = E[X;\hat{\theta}|y]$, the likelihood imputation depends on an entire high-likelihood region instead of one of its interior points. It nevertheless results into consistent parameter estimator with considerable generality. We outline briefly the basic arguments. Denote by $Q_j(\theta)$ the quadratic form of θ at some square matrix j, i.e. $Q_j(\theta) = \theta^T j\theta$. Heuristically, suppose that the following quadratic log-likelihoods apply to some neighbourhood of $\hat{\theta}$, i.e.

$$l(\theta; x^*) = l(\hat{\theta}^*; x^*) - \frac{1}{2}Q_{\hat{\jmath}}(\theta - \hat{\theta}^*)$$
$$l_1(\theta; y) = l_1(\hat{\theta}; y) - \frac{1}{2}Q_{\hat{\jmath}_1}(\theta - \hat{\theta}),$$

where $\hat{\jmath} = j(\hat{\theta})$ and $\hat{\jmath}_1 = j_1(\hat{\theta})$ are the respective observed information matrices derived from l and l_1 both, however, evaluated at $\hat{\theta}$. The symmetry of $\Theta_h(\beta)$ around $\theta = \hat{\theta}$ would imply that the δ-equation is solved at the x^* which yields $\hat{\theta}^* = \hat{\theta}$, i.e. the center of $\Theta_h(\beta)$. (One needs only to reflect over the scalar θ and realize that the general case is essentially the same.) Asymptotically, the error terms of these quadratic expansions are arbitrarily small with a probability tending to unity (e.g. when $\hat{\theta}$ attains asymptotic normality), in which case the probability that the δ-equation is solved at some $\hat{\theta}^*$ arbitrarily close to $\hat{\theta}$ also tends to unity, so that $\hat{\theta}^*$ based on the likelihood imputation is a consistent estimator of θ.

2.2 The implementation

The integrals involved in the δ-value and δ-equation may well turn out to be difficult to handle analytically, in which case they can be evaluated by the Monte Carlo method (Hammersley and Handscomb, 1964; Rubenstein, 1981) under which it is not necessary to calculate the constant c_h. More explicitly, let a *grid* be a set of randomly generated parameter values $\{\theta_1, ..., \theta_d\} \in \Theta_h$, the simplest unbiased Monte Carlo estimate of $\delta(x^*)$ is given by $\sum_{i=1}^{d}[l_r(\theta_i) - \bar{l}_r]^2/(d-1)$, where $\bar{l}_r = \sum_{i=1}^{d} l_r(\theta_i)/d$. The other integrals involved can similarly be approximated by their Monte Carlo estimates evaluated at the *same* grid points. The resultant solution of the δ-equation is called the *Monte Carlo likelihood imputation* w.r.t. $\{\theta_1, ..., \theta_d\}$.

There is consequently a question on the size of the grid, especially with high dimensional parameter. Practically, one could run the algorithm repeatedly for grids with increasing sizes till the equilibrium is reached — the assessment of which can be based on a test on the equal 'variance' of the residual log-likelihoods from different runs. To reduce the dimension of the δ-equation, one may apply the method to the sufficient statistics of x^* instead, provided the post randomization is feasible. In particular, when $f(x;\theta)$ belongs to the exponential families of distributions, the Monte Carlo likelihood imputations obtained from different runs can be pooled to form a new imputation with reduced δ-value. Due to the linearity of $l(\theta; x^*)$ in the sufficient statistics of x^*, the residual log-likelihood on the pooled imputation coincides with the pooled residual log-likelihood, whose δ-value is smaller than the pooled δ-value unless the residual log-likelihoods are pairwise perfectly 'correlated'. The size of the grid can therefore as well be controlled by the 'correlations' between the residual log-likelihoods from different runs. It is also clear that multiple likelihood imputations do not occur in case of exponential $f(x;\theta)$, though the possibility should always be entertained otherwise.

Meanwhile, the minimization of the δ-value is often simplified provided the residual likelihood can be factorized into

$$\prod_{j=1}^{k} L_{r_j}(\theta; x_j|y_j) \propto \prod_{j=1}^{k} f(x_j; \theta|y_j),$$

where $(y_1, ..., y_k) = (Y_1(x_1), ..., Y_k(x_k))$ and both x_j and y_j, $1 \leq j \leq k$, can be vectors themselves. That is, conditional to y, the complete data can be partitioned into independent $x_1, ..., x_k$. The δ-value is then given as

$$\sum_{i,j=1}^{k} Cov_h(l_{r_i}, l_{r_j}) = \sum_{i,j=1}^{k} \delta(x_i^*, x_j^*),$$

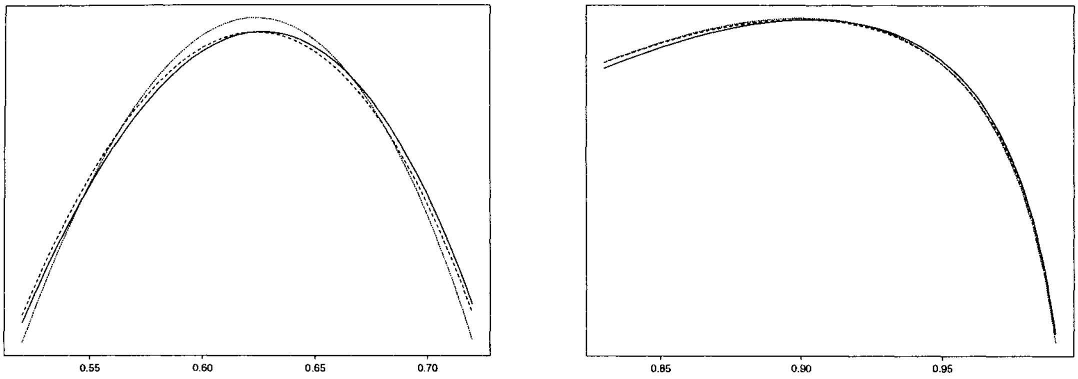
where $\delta(x_i^*, x_j^*)$ is called the *(i,j)-th cross δ-value* for $1 \leq i,j \leq k$. Clearly, therefore, we can minimize the cross δ-values involving x_i^* at given x_j^* $(j \neq i)$ and iterate. This breaks the δ-equation into smaller pieces, and generates a sequence of imputations decreasing in their δ-values by construction.

3. Examples

3.1 Genetic linkage model (Rao, 1965, p 368-9)

Suppose $n = 197$ animals (y) are divided into four categories, with counts $(y_1, ..., y_4)$. Augment the data to obtain the latent $x = (x_1, ..., x_5)$ such that $y_1 = x_1 + x_2$, $y_2 = x_3$, $y_3 = x_4$ and $y_4 = x_5$, which are multinomially distributed according to the

Figure 1: Genetic linkage model. The solid observed log-likelihood $l_1(\theta; y)$ against the dotted imputed complete log-likelihood $l(\theta; x^*)$ and the dashed effective (imputed) complete log-likelihood $(m/n)l(\theta; x^*)$ after vertical shifts. (The left plot based on $y = (125, 18, 20, 34)$ and the right one $y = (14, 0, 1, 5)$.)



probability vector $p_x = (1/2, \theta/4, (1-\theta)/4, (1-\theta)/4, \theta/4)$.

Suppose first $y = (125, 18, 20, 34)$. We search through $\Omega^* = \{x_1^*; 0 < x_1^* < y_1\}$ over $\Theta_{0.95} = (0.52, 0.72)$, i.e. for any grid evenly spread out over $\Theta_{0.95}$ we calculate the Monte Carlo δ-value for each $x_1^* \in \Omega^*$. The corresponding Monte Carlo likelihood imputation is given by the x_1^* with the smallest Monte Carlo δ-value. The result settles already for small grid sizes (say, $d = 5$), which yields the likelihood imputation $x_1^* = 96$ and the m.l.e. based on this $\hat{\theta}^* = 0.6238$. The EM-estimate in this case is $\hat{\theta} = 0.6268$ (Dempster et al., 1977) where the E-step treats the data as if they were continuous. The procedure was similarly carried out for another more extreme data set $y = (14, 0, 1, 5)$ over $\Theta_h = (0.833, 0.950)$ which is the range of $\hat{\theta}^*$ conditional to y — the sample is too small for the asymptotic results on Θ_β to apply. The Monte Carlo likelihood imputation settles on $x_1^* = 10$, giving $\hat{\theta}^* = 0.9$ as compared to the EM-estimate $\hat{\theta} = 0.903$. (See Figure 1 — the scaling factor m/n is discussed in Sec. 4.)

Alternatively, treating the data as if they were continuous and solving for the δ-equation, we obtain the likelihood imputation as $x_2^* = Cov_h[y_1 \log(2+\theta), \log\theta]/Var_h(\log\theta)$. Taking the nearest integers, we have $[x_2^*] = 29$ for $y = (125, 18, 20, 34)$ and $[x_2^*] = 4$ for $y = (14, 0, 1, 5)$, which are the same as above.

Meanwhile, conditional to $Y_1 = y_1$, X_1^* can be considered as, say, the total success among y_1 i.i.d. Bernoulli trials, denoted by $(Z_1^*, ..., Z_{y_1}^*)$ and $x_1^* = \sum_{i=1}^{y_1} z_i^*$, to which the iterative algorithm applies. That is, at each iteration, we minimize $\sum_i \delta(z_i^*, z_j^*)$ for $j = 1, ..., y_1$ in succession. For instance, let $y = (125, 18, 20, 34)$ and set $z_1^* = \cdots = z_{125}^* = 0$. The algorithm returns $z_j^* = 1$ for j up to 96, upon which it converges since it returns zero for z_{97}^* which is the same as its initial value. Setting $z_1^* = \cdots = z_{125}^* = 1$, the algorithm returns $z_j^* = 0$ up to $j = 29$ at which convergence is reached. Indeed, due to the binary domain of Z_i^*, the convergence can be reached within the first iteration for any initial values of $(z_1^*, ..., z_{y_1}^*)$.

3.2 Linear regression with right-censored data

Schmee and Hahn (1979) studied the motorette data (Table 1) to which they fitted a linear regression model, i.e.

$$x_i = \beta_0 + \beta_1 v_i + \sigma\epsilon_i \quad i = 1, ..., 40,$$

where $\epsilon_i \sim N(0, 1)$ and (temperature$+273.2)v_i = 1000$. Denote by θ the parameter vector $(\beta_0, \beta_1, \sigma)$, whose residual log-likelihood l_r is

$$-\sum_{j=1}^{23}[\log\sigma + \log H_j + (x_j^* - \mu_j)^2/(2\sigma^2)]$$

Table 1: The motorette data (Schmee and Hahn, 1979). Each row corresponds to different temperatures at which the motorettes were tested, which shows the time to failure on the $\log_{10}$-scale, where a star-superscript indicates that the observation was right-censored.

3.907*	3.907*	3.907*	3.907*	3.907*	3.907*	3.907*	3.907*	3.907*	3.907*
3.246	3.443	3.537	3.549	3.577	3.687	3.716	3.736*	3.736*	3.736*
2.611	2.611	3.128	3.128	3.158	3.225*	3.225*	3.225*	3.225*	3.225*
2.611	2.611	2.702	2.702	2.702	2.723*	2.723*	2.723*	2.723*	2.723*
Monte Carlo likelihood imputation over the 95% confidence region.									
4.091*	4.094*	4.139*	4.157*	4.165*	4.234*	4.358*	4.410*	4.419*	4.516*
3.246	3.443	3.537	3.549	3.577	3.687	3.716	3.836*	3.877*	4.070*
2.611	2.611	3.128	3.128	3.158	3.225*	3.270*	3.427*	3.721*	3.842*
2.611	2.611	2.702	2.702	2.702	2.729*	2.854*	2.887*	3.088*	3.174*

where, for the j-th right-censored observation, $\mu_j = \beta_0 + \beta_1 v_j$ and $H_j = 1 - \Phi\{(y_j - \mu_j)/\sigma\}$. This allows us to solve the δ-equation iteratively.

Data augmentation (Tanner, 1993, p 65), with $m = 100$ chains and running for 10 iterations, has been used to locate the approximate 95% confidence region $\Theta_{0.95}$. (We are not particularly bothered with the convergence of the data augmentation, i.e. the accuracy in $\Theta_{0.95}$.) Evaluating the cross δ-values at $d = 1000$ points, we obtained one Monte Carlo likelihood imputation (Table 1), based on which $\hat{\theta}^* = (-6.017, 4.315, 0.268)$ and $\hat{\mu}^*_\nu = (4.179, 3.719, 3.299, 2.732)$ for $\nu = (150, 170, 190, 220)$. The corresponding EM-estimates (Tanner, 1993, p 43) are (-6.019, 4.311, 0.259) and (4.164, 3.707, 3.284, 2.719). For comparison, the Monte Carlo likelihood imputation generated over the approximate $\Theta_{0.50}$, gives $\hat{\theta}^* = (-6.071, 4.335, 0.272)$ and $\hat{\mu}^*_\nu = (4.172, 3.710, 3.288, 2.719)$, which seems to indicate the robustness of the likelihood imputation towards the choice of Θ_β.

It is worth noticing that the variation in the imputed values here results directly from a minimization procedure. This happens because the likelihood imputation aims at the 'right' latent likelihood, i.e. the 'right' latent sufficient statistics, which in turn requires variability in the imputed data at fixed temperatures. In contrast, the estimated conditional means are (4.239, 3.933, 3.455, 2.928) depending on the value of v_i but is the same at a given temperature, which is not only unnatural but also misleading — giving $\hat{\sigma}^* = 0.225$.

4. Discussion

An *imputor*, i.e. a function of the observation Y taking range in the sample space consistent with Y, can be said to be *likelihood consistent* if it yields likelihood consistent imputations; it is so *asymptotically* if the standardized, imputed residual log-likelihoods, i.e. l_r/n, converge to a constant independent of the parameter in probability. In case the likelihood imputor fails to satisfy asymptotic likelihood consistency, the *imputed residual information* $\hat{j}^*_r = j(\hat{\theta}^*) - j_1(\hat{\theta}^*)$ does not vanish in probability, which in turn indicates the amount of information that has been imputed. Now, restoring public data bases using single imputation very much depends on how successful we can summarize this imputed information in a concise, robust manner. The δ-value can only be used comparatively; whereas $\hat{j}^*_r$ apparently too dependent on the parameterization.

Under the repeated sampling, improvements can always be achieved by applying a concept of *effective sample size* in connection with the likelihood imputation. In short, instead of the actual sample size n, we use the effective sample size, say, m and base inference on $(m/n)l(\theta; x^*)$ instead of $l(\theta; x^*)$. This implies to minimize the variation of the effective residual log-likelihood $(m/n)l(\theta; x^*) - l_1(\theta; y)$ instead of the proper l_r. Notice the difference resulted from minimizing m posterior to the likelihood imputation or simultaneously with it. Notice also that the extent of the improvement is nevertheless not independent of the model assumed, i.e. whether it leads to favourable shapes of the log-likelihoods. This is a criticism, within the relevance of restoring public data bases, shared by all model-assistant imputation techniques — single as well as multiple. The genetic linkage model again provides a simple illustration.

Example: Genetic linkage model, cont'd. Suppose $y = (125, 18, 20, 34)$. The standard deviation of $\hat{\theta}$ is 5.146×10^{-2} derived from $\hat{j}_1$, and 4.805×10^{-2} from $\hat{j}^*$ at $x_1^* = 96$. Minimizing $Var_h[(m/n)l(\theta; x^*) - l_1(\theta; y)]$ towards m/n with x_1^* fixed at 96 gives $m/n = 0.871$, i.e. $[m] = 172$. Treat this as the 'effective' sample size and derive the adjusted standard deviation of $\hat{\theta}$ from $(m/n)l(\theta; x^*)$ gives us 5.148×10^{-2}. Whereas minimizing towards m/n and x^* simultaneously gives $[x_1^*] = 95$ and $m/n = 0.865$, i.e. $[m] = 170$, and the adjusted standard deviation 5.151×10^{-2}.

Suppose $y = (14, 0, 1, 5)$ instead. The standard deviation of $\hat{\theta}$ derived from $\hat{j}_1$ is 9.363×10^{-2}, and 9.232×10^{-2} from $\hat{j}^*$ with $x_1^* = 10$. Minimizing $Var_h[(m/n)l(\theta; x^*) - l_1(\theta; y)]$ towards m/n with x_1^* fixed at 10 gives $m/n = 0.9999$, leading practically to no adjustment; whereas minimizing towards m/n and x^* simultaneously gives us $[x_1^*] = 10$ and $m/n = 0.967$, i.e. $[m] = 19$, with the adjusted standard deviation 9.389×10^{-2}. $\diamondsuit$

The conditional-mean imputation carries little information on the model in addition to what is summarized in $\hat{\theta}$, whereas the likelihood imputation depends on an entire high-likelihood region so that a series of likelihood imputations constructed over variously chosen likelihood regions are able to cast light on the likelihood as a whole; and we outline briefly one way in which such information can be gained.

The latent structure entails loss of degree of freedom in data, so that restrictions on the dimension of θ sometimes becomes necessary. Since, technically, the likelihood imputation can be generated regardless of such restrictions, one may compare, at least around the local maxima of l_1, the full-model likelihood imputation with the restricted ones, even when the full paramater is not identifiable. In case they differ considerably from each other, which basically indicates that the quadratic approximation to l_1 no longer remains appropriate under the full model, reconsideration of the restrictions made would seem prudent.

The kind of sensitivity analysis described here is only made possible by the likelihood imputation. Its potentials in this area, however, have by no means been exhausted, especially once we start to consider alternative quantifications of the difference in information contents with other perhaps more specific inferential purposes on mind.

References

Bjørnstad, J.F. and Walsøe, H.K. (1991). Predictive likelihood in nonresponse problems. In *American Statistical Association 1991 Proceedings of the section on Survey Research Methods*, pp. 152–6.

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

Efron, B. (1994). Missing data, imputation, and the Bootstrap (with discussion). *J. Am. Statist. Assoc.*, **89**, 463–79.

Greenlees, J.S., Reece, W.S., and Zieschang, K.Y. (1982). Imputation of missing values. *J. Am. Statist. Assoc.*, **77**, 251–61.

Hammersley, J.M. and Handscomb, D.C. (1964). *Monte Carlo Methods.* London: Chapman and Hall.

Ritter, C. and Tanner, M.A. (1992). The Gibbs stopper and the greedy Gibbs sampler. *J. Am. Statist. Assoc.*, **87**, 861–8.

Rubenstein, R. (1981). *Simulation and the Monte Carlo Method.* New York: Wiley.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, **63**(3), 581–92.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys.* New York: Wiley.

Schenker, N. and Welsh, A.H. (1988). Asymptotic results for multiple imputation. *Ann. Statist.*, **16**, 1550–66.

Schmee, J. and Hahn, G.J. (1979). A simple method for regression analysis with censored data. *Technometrics*, **21**, 417–32.

Tanner, M.A. (1993). *Tools for Statistical Inference* (2nd edn). Springer-Verlag.

Tanner, M.A. and Wong, W.H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *J. Am. Statist. Assoc.*, **82**, 528–50.

Wei, G.C.G. and Tanner, M.A. (1990). Posterior computations for censored regression data. *J. Am. Statist. Assoc.*, **85**, 829–39.

Zhang, L.-C. (1997). Likelihood imputation. *Scand. J. Statist*, To appear.