Fraction of Missing Information and Convergence Rate of Data Augmentation *

Jun S. Liu

Department of Statistics, Harvard University, Cambridge, MA 02138

1 Introduction

The Gibbs sampler and other MCMC methods (Gelfand and Smith 1990, Smith and Roberts 1993, Tanner and Wong 1987), which become popular recently in statistical analysis with complicated models, are no more than some devices for generating random samples from an analytically intractable target distribution. The basic idea underlying all these methods is to construct a Markov chain with the target distribution as its equilibrium distribution. The methods differ only in the use of Markov transition functions. For example, the transition function for the Gibbs sampler with systematic scan can be expressed as a product of a sequence of conditional distributions (Smith and Roberts 1993, Liu, Wong and Kong 1994b); while the transition function for a Metropolis-Hastings algorithm consists of a "proposed" transition and a "thinning down" device (Metropolis et al. 1953, Hastings 1970, Smith and Roberts 1993). Many theoretical work has emerged in understanding convergence properties of the MCMC methods. See, for example, Geman and Geman (1984), Gelman and Rubin (1992), Geyer (1992), Liu, et al. (1994a,b), Liu (1992, 1994), Mykland, Tierney and Yu (1993), Roberts (1992), Roberts and Polson (1994), Rosenthal (1993a,b), Schervish and Carlin (1993), Tierney (1991), just to start a list. Here, by taking a slightly different angle to look at the convergence problem, we investigate relationships among various concepts in describing a Gibbs sampler and the associated Bayesian missing data problem: the rate of convergence, sample autocorrelations, and the fraction of missing information.

* The author thanks Alan Zaslavsky and Yingnian Wu for helpful discussions and computing assistance. This work is partly supported by NSF grant DMS 94-04344 and Milton Fund of Harvard University.

We distinguish two different situations for the Gibbs sampler: Data Augmentation which refers to a Gibbs sampler with only two iterative components (see Tanner and Wong 1987 for its original version, and Liu et al. 1994a for structural study), and the general Gibbs sampler (Gelfand and Smith 1990). A reason for doing this is that the two component case provides us some extra structure that a general Gibbs sampler does not possess, and the analysis of this simple case can suggest some useful methods for dealing with more general ones.

By making use of covariance structures of Data Augmentation established in Liu et al. (1994a,b), we find that the convergence rate of the induced Markov chain can be characterized by the *maximal fraction of missing information*, which is closely related to the work of Meng and Rubin (1992) for the EM algorithms. Conversely, because of this characterization, we can use autocorrelations of a stationary Gibbs sampling sequence to estimate the fraction of missing information of any quantity of interest, which is useful for deciding how many multiple imputations will be provided.

This article is arranged as follows. We review the concept of fraction of missing information in Section 2. In Section 3, we present structures and several connections for Data Augmentation. A generalization to the general Gibbs sampler is contained in Section 4. A graphical method for comparing different schemes, using the relationships found in Sections 3 and 4, is described in Section 5. In Section 6, we analyze an example for match-making in "broken regression" (DeGroot, Feder, and Goel 1971).

2 The Fraction of Missing Information

The concept of fraction of missing information was first introduced together with the so-called *missing*

information principle by Orchard and Woodbury (1972). It is later proved to be an important concept for studying the EM algorithms (Dempster, Laird and Rubin 1977). Specifically, Louis (1982) presented a method for finding the observed information, and Meng (1991) and Meng and Rubin (1993) systematically explored the concept and used it to characterize the rate of convergence for the EM and the ECM algorithms.

To introduce the fraction of missing information conveniently, we let Θ denote the parameter vector in our model, let Y denote the observed part of an imaginary complete data set, and let Z denote the missing part. A simple identity underlying the missing information principle and the EM algorithms is

$$\begin{aligned}\log[p(\Theta \mid Y)] &= \log[p(\Theta \mid Y, Z)] \\ &- \log[p(Z \mid \Theta, Y)] + \log[p(Z \mid Y)],\end{aligned}$$

which implies

$$\begin{aligned}-\frac{\partial^2 \log p(\Theta \mid Y)}{\partial \Theta^2} &= -\frac{\partial^2 \log p(\Theta \mid Y, Z)}{\partial \Theta^2} \\ &+ \frac{\partial^2 \log p(Z \mid \Theta, Y)}{\partial \Theta^2}.\end{aligned}$$

Integrating out the missing data Z with respect to $p(Z|\Theta, Y)$, we arrive at the following missing information principle

$$\begin{aligned}\text{Observed Information} &= \text{Complete Information} \\ &- \text{Missing Information.}\end{aligned}$$

Denoting each term by I_{obs}, I_{com}, and I_{mis}, respectively, we can define the *fraction of missing information* as

$$\gamma_{_L} = \frac{I_{mis}(\Theta)}{I_{com}(\Theta)} = 1 - \frac{I_{obs}(\Theta)}{I_{com}(\Theta)},$$

where the I functions are evaluated at the true parameter value. When Θ is a 1-dim parameter, the above quantity is well defined. Otherwise, the above definition takes a matrix form. Meng (1991) used the largest eigenvalue of the missing fraction matrix $I_{mis}^{-1}(\Theta) I_{com}(\Theta)$ to characterize the convergence rate of the EM algorithm.

Now let us take a Bayesian viewpoint. Suppose a prior distribution $p_0(\Theta)$ is given, and we are interested in $h \equiv h(\Theta)$ (one can view this as a way of eliminating nuisance parameters). If one can impute the missing data, i.e., draw samples $Z^{(1)}, \ldots, Z^{(m)}$ from the predictive distribution $p(Z|Y)$, then the posterior distribution of h, $p(h|Y)$, can be approximated by

$$p(h \mid y) \approx \frac{1}{m}\{p(h|Y, Z^{(1)}) + \cdots p(h|Y, Z^{(m)})\}.$$

For example, $Z^{(1)}, \cdots, Z^{(m)}$ can be draws from an iterative sampling scheme. When using the above multiple imputation type of approximations, *the fraction of missing information* is usually important for one to understand the impact of the missing data on the estimation of h. Also, it is important for one to decide how many imputations should be provided. As Rubin (1987) advocated, m can be chosen as small as 3 to 5 for estimating posterior mean of h. Of course, in this case, the fraction of missing information with respect to h can not be too high.

The fraction of missing information in the Bayesian framework can be easily defined as (Rubin 1987)

$$\begin{aligned}\gamma_{_B} &= \frac{\text{var}\{E(h \mid Y, Z) \mid Y\}}{\text{var}(h \mid Y)} \\ &= 1 - \frac{E\{\text{var}(h \mid Y, Z) \mid Y\}}{\text{var}(h \mid Y)}\end{aligned}$$

which can be explained as the extra variation caused by missing Z.

Note that in large sample and when h=θ, since $\text{var}(h|Y) \approx 1/I_{obs}$ and $E\{\text{var}(h|Y, Z)\} \approx 1/I_{com}$, the two definitions of the fraction of missing information, $\gamma_{_B}$ and $\gamma_{_L}$, are equivalent.

3 Structures for Data Augmentation

We call a special situation of the Gibbs sampler *Data Augmentation* if there are only two components for iterative sampling (Liu et al. 1994). We use Θ and Z to denote the respective components in Data Augmentation to emphasize its connection with Bayesian missing data problems.

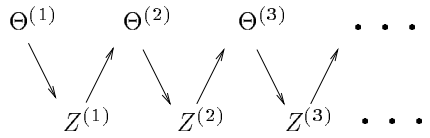
Let $\Theta^{(1)}, Z^{(1)}, \Theta^{(2)}, Z^{(2)}, \ldots,$ be consecutive draws from a stationary Data Augmentation. In other words, we assume that $\Theta^{(1)}$ is drawn from the target distribution $p(\Theta|Y, Z)$. In the following, since everything is conditioned on Y, we will omit it in all expressions. For example, when we write $E\{h(\Theta)|Z\}$, it actually means $E\{h(\Theta)|Y, Z\}$.

Consider two consecutive draws from Data Augmentation, we find that

$$E(h^{(k)}h^{(k+1)}) = E\{E(h^{(k)}h^{(k+1)} \mid Z^{(k)})\} \qquad (1)$$

$$= E\{E(h^{(k)} \mid Z^{(k)})E(h^{(k+1)} \mid Z^{(k)})\}$$
$$= E\{E^2(h \mid Z)\},$$

where the first equality follows from an elementary fact that $E(A) = E[E(A|B)]$; the second and third equalities follow from the fact that $\Theta^{(k)}$ and $\Theta^{(k+1)}$ are conditionally independent and identically distributed given $Z^{(k)}$. These facts can be illustrated by the following diagram:



From the diagram, we observe that $\Theta^{(1)}$ connects with $\Theta^{(2)}$ through $Z^{(1)}$, and, from the definition of the scheme, $(\Theta^{(1)}, Z^{(1)})$ and $(\Theta^{(2)}, Z^{(1)})$ have the same joint distribution when the chain is stationary. These two properties only hold for Data Augmentation, not for the general Gibbs sampler. However, this type of dependence graph can be applied to a general Gibbs sampler and provide useful intuition. In Section 5 we will illustrate how to use these diagrams to compare different schemes.

As a consequence of (2), we have the following identity

$$\mathrm{cov}\{h(\Theta^{(k)}), h(\Theta^{(k+1)})\} = \mathrm{var}[E\{h(\Theta) \mid Z\}]$$

The formula implies that the correlation coefficient between the two consecutive h's are

$$\rho(h^{(k)}, h^{(k+1)}) = \gamma_B.$$

An intuition of this is that the higher the fraction of missing information, the more "sticky" the sample outputs from Data Augmentation, and vice versa. The extra variance caused by the missing data, $\mathrm{var}\{E(h|Z)\}$, can then be estimated as

$$\hat{v}_{mis} = \frac{1}{m-1}\sum_{k=1}^{m-1} h^{(k)}h^{(k+1)} - (\bar{h}_m)^2.$$

If, on the other hand, $g(Z) = E(h|Z)$ is easy to compute, one may also approximate $\mathrm{var}\{E(h|Z)\}$ by

$$\tilde{v}_{mis} = \sum_{i=1}^{m}(g^{(i)} - \bar{g}_m)^2/(m-1),$$

where $g^{(i)} = E(h|Z^{(i)})$ and $\bar{g}_m = (g^{(1)} + \cdots + g^{(m)})/m$. This is a variation of Rao-Blackwellization (Gelfand and Smith 1990, Liu et al. 1994a).

Intuitively, it seems that the latter estimation is better. For example,

$$\mathrm{var}\{h^{(1)}h^{(2)}\} = E\{(h^{(1)}h^{(2)})^2\} - [E\{E^2(h|Z)\}]^2$$
$$= E\{E^2(h^2 \mid Z)\} - [E\{E^2(h|Z)\}]^2,$$

while

$$\mathrm{var}(g^2) = E\{E^4(h \mid Z)\} - [E\{E^2(h \mid Z)\}]^2.$$

Hence, by the Cauchy-Schwarz inequality, we have

$$\mathrm{var}(g^2) \le \mathrm{var}\{h^{(1)}h^{(2)}\}.$$

Furthermore, by Theorem 3.1 of Liu et al. (1994)

$$\mathrm{cov}\{(g^{(1)})^2, (g^{(k+1)})^2\}$$
$$= \mathrm{var}\{E(\cdots E[E\{g^2(Z)|\Theta\}|Z]\cdots)\}$$

where the right hand side has k expectation signs. Also, we notice that

$$E\{g^2(Z)|\Theta\} = E\{E[g(Z)h(\Theta)|Z]|\Theta\}.$$

For $\hat{v}_{mis}$, we let $f(\Theta) = E[E\{h(\Theta)|Z\}|\Theta]$, which is just $E(h^{(2)}|\Theta^{(1)})$. Then we have

$$\mathrm{cov}(h^{(1)}h^{(2)}, h^{(k+1)}h^{(k+2)})$$
$$= \mathrm{cov}(h^{(2)}f^{(2)}, h^{(k+1)}h^{(k+2)})$$

which, for the same reason as above, has the following expression

$$\mathrm{var}\{E(\cdots E[E\{h(\Theta)f(\Theta)|Z\}|\Theta]\cdots)\}$$

where there are $k-1$ expectation signs on the right hand side. However

$$E\{h(\Theta)f(\Theta)|Z\} = E\{E[h(\Theta)g(Z) \mid \Theta] \mid Z\}$$

If we compare the expression of lag-k autocovariance for the $(g^{(i)})^2$ sequence with that for the $h^{(i)}h^{(i+1)}$ sequence, we find that the former always has one more conditional expectation sign than the latter. However since the orders of the conditionings are different, there is no clear comparison between the two except for the case when lag=1, in which case, the autocovariance for the latter expression is always greater than or equal to the former.

The following analogy is helpful for understanding the above discussion. Consider two scenarios: (i) a vector **a** is projected to vector **b** and then to vector **c**; (ii) **a** is directly projected to **c**. How do we compare the length of the projections? Apparently, if the three vectors are in the same plane and **b**

lies between **a** and **c**, the latter projection is smaller than the former one. But in most other cases, the former is smaller than the latter. This corresponds to comparing $\mathrm{var}[E\{E(X|Y)|Z\}]$ and $\mathrm{var}\{E(X|Z)\}$.

For any two random variables U and V, we define the *maximal correlation* between them as

$$R(U,V) = \sup_{\mathrm{var}\{t(U)\}=\mathrm{var}\{s(V)\}=1} corr\{t(U), s(V)\}.$$

It is well understood that for a reversible stationary Markov chain $X^{(1)}, X^{(2)}, \ldots$, the maximal correlations between two consecutive states, $R(X^{(k)}, X^{(k+1)})$, is equal to λ, where $1-\lambda$ is the so-called "spectral gap." See Liu et al. (1994a,b) for more references. For discrete case, λ is just the magnitude of the second largest eigenvalue (in absolute value). For nonreversible chain, the scaled long-range maximal correlation is equal to λ (Liu et a. 1994b). That is,

$$\lim_{k\to\infty} \{R(X^{(1)}, X^{(k+1)})\}^{1/k} = \lambda.$$

It is shown in Liu et al. (1994a) that the maximal correlation between two consecutive draws of Data Augmentation, $R(\Theta^{(k)}, \Theta^{(k+1)})$ is the intrinsic rate of convergence of the scheme, and is equal to $R^2(\Theta, Z)$.

On the other hand, under mild conditions (see Csàki and Fischer 1960), there exists a pair of functions $h_0(\Theta)$ and $g_0(Z)$ with unit variance such that $\mathrm{corr}(h_0, g_0)$=$R(\Theta, Z)$ (denoted as R later), and

$$E\{g_0(Z) \mid \Theta\} = R\, h_0(\Theta) \quad (2)$$
$$E\{h_0(\Theta) \mid Z\} = R\, g_0(Z) \quad (3)$$

Therefore, h_0 suffers the *maximal fraction of missing information*

$$\gamma_{B}(h_0) = \mathrm{var}\{E(h_0|Z)\}/\mathrm{var}(h_0) = R^2,$$

and the maximal fraction of missing information is equal to the rate of convergence of Data Augmentation. If a function h is correlated with h_0 (with respect to π), then

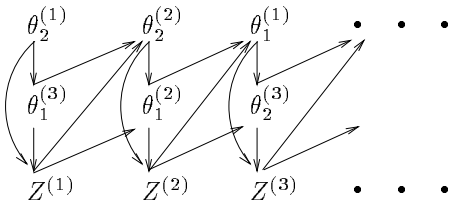
$$\{corr(h^{(1)}, h^{(k+1)})\}^{1/k} \to \lambda$$

as k goes to infinity. This follows from spectral decomposition of h (Liu 1991, Garen and Smith 1994, Roberts 1992). It suggests that the *maximal fraction of missing information* can be estimated by the output sequence of the Gibbs sampler.

4 Missing Information in the General Gibbs Sampler

We now turn our attention to the general Gibbs sampler with systematic scan. There are two situations commonly encountered in practice. We shall discuss them in the order of increasing complexity.

Case 1. $\Theta = (\theta_1, \theta_2)$, $Z = Z$. That is, given Θ, Z can be drawn directly; but θ_1 must be drawn conditional on both θ_2 and Z, and θ_2 must be drawn conditional on θ_1 and Z. Note that this can be generalized obviously. The following diagram illustrates the sampler:



Hence,

$$\begin{aligned} & \mathrm{cov}\{h(\theta_1^{(1)}), h(\theta_1^{(2)})\} \\ = \;& E\{h(\theta_1^{(1)})h(\theta_1^{(2)})\} - E\{h(\theta_1)^2\} \\ = \;& \mathrm{var}[E\{h(\theta_1) \mid \theta_2, Z\}] \end{aligned}$$

which implies that lag-1 autocorrelation of the h sequence is in general not its fraction of missing information with respect to Z, but is a quantity that reflects dependency between θ_1 and (θ_2, Z). Note that

$$\mathrm{var}[E\{h(\theta_1) \mid \theta_2, Z\}] \geq \mathrm{var}[E\{h(\theta_1)|Z\}].$$

Another way around is to design a function $g(Z)$ and to estimate the maximal correlation between Θ and Z from it. For example, if it happens that we know g_0 in (2) and (3), then by Lemma 4 of Liu (1994),

$$\begin{aligned} \mathrm{cov}\{g_0(Z^{(k)}), g_0(Z^{(k+1)})\} &= \mathrm{var}[E\{g_0(Z) \mid \Theta\}] \\ &= R^2\, \mathrm{var}\{h_0(\Theta)\}. \end{aligned}$$

Here R^2 is the maximal fraction of missing information and is an upper bound for $\gamma_{B}(h)$. This duality provides us the following scheme for obtaining an estimate of the maximal fraction of missing information.

Step 1. Design a function $g(Z)$. Usually this can just be a linear function (e.g., see Liu 1991).

Step 2. Estimate lag-k autocorrelation r_k for the g sequence for $k = 1, 2, \ldots$, after the chain converges, and fit the exponential model

$$r_k = c\rho^k.$$

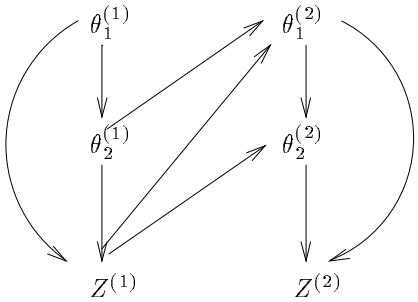
Garren and Smith (1994) provided refined methods. The fitted value $\hat{\rho}$ is an estimate of $R(\Theta, Z)$.

Case 2. $\Theta = (\theta_1, \theta_2)$ and $Z = (z_1, z_2)$. This is the case where the fraction of missing information can not be estimated from the sample autocorrelations. The maximal fraction of missing information can be extracted from long range autocorrelations by the same reason as explained in Case 1.

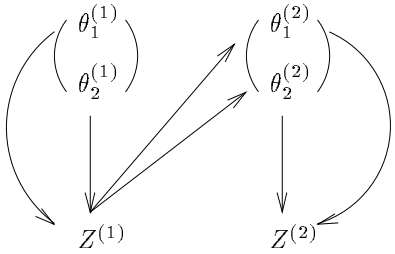
5 Compare Schemes via Diagrams

In running a Gibbs sampler or a more general MCMC algorithm, one usually has flexibilities in designing sampling schemes. As with many iterative methods , we are usually faced with a dilemma: we either have to sacrifice computational ease for iterative simulation in exchange for fast convergence, or have to suffer slow convergence in exchange for computational simplicity. Only in some rare situations as explored in Liu (1994) be we satisfied in both ways. Specifically, when the Bayesian predictive distribution is simple, one can use the *predictive updated* version to improve convergence without sacrificing computational simplicity. Liu et al. (1994a) and Liu (1994) provided some theoretical arguments based on operator theory. Here we use diagrams to illustrate autocorrelation structures. We hope that the analysis in this section can shed light on more complicated general situations.

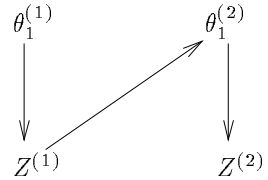
For the sake of simple argument, suppose the sampler involves three components (θ_1, θ_2, Z) and each component is visited in turn: $\theta_1 \rightarrow \theta_2 \rightarrow Z$. The following diagram shows dependency between two consecutive iterations. For example, $\theta_1^{(2)}$ is generated by a draw from $\pi(\theta_1|\theta_2^{(1)}, Z)$, which is illustrated in the diagram by two arrows connecting $\theta_2^{(1)}$ and Z with $\theta_1^{(2)}$. Other arrows have similar implications. This diagram shows that the two consecutive states depend on each other via the connection between $(\theta_1^{(1)}, Z^{(1)})$ and $(\theta_1^{(2)}, \theta_2^{(2)})$ as illustrated by three arrows in the middle or the diagram.



Next diagram illustrates a *grouping* scheme, where it is assumed that given Z, (θ_1, θ_2) can be drawn together. The diagram illustrates that dependency between two consecutive states is via the connection between $Z^{(1)}$ and $(\theta_1^{(2)}, \theta_2^{(2)})$, where only two arrows are used for this connection. Compared with the above diagram for the original sampler, dependency between the two consecutive states for *grouping* is weaker.



Our final diagram represents the *collapsing* scheme, in which we assume that θ_2 can be theoretically integrated out so that the sampler is applied only to the two remaining components. In this diagram, the only connection between two consecutive states is that between $Z^{(1)}$ and $\theta_1^{(2)}$. Only one arrow is used, which indicates the weakest correlation among the three schemes.



We expect that this type of analysis can be generalized to other situations to help one design efficient sampling schemes.

6 An Example: Broken Regression

Suppose x_i, $i = 1, \ldots, 100$, are i.i.d. normal with variance τ^2; and $y_i = \alpha + \beta x_i + \epsilon_i$, where the ϵ_i are i.i.d. from $N(0, \sigma^2)$. It is a standard regression problem if we observe (x_i, y_i) for $i = 1, \cdots, 100$. Suppose, however, the pairing information is somehow lost and we can only observe u_i, $i = 1, \ldots, 100$, a random shuffle of the y_i. The problem is no longer trivial. This can also be viewed as a special case of file matching problem. DeGroot et al. (1971) studied this problem with an objective to maximize the number of correct matches. We are interested in estimating β and the corresponding fraction of missing information (for not knowing the matching).

Let Q be the permutation that produces the u_i from the y_i. The main difficulty is that Q is missing. Let $\Theta = (\alpha, \beta)$ and $U = (u_1, \ldots, u_{100})$. With a prior distribution on Θ, Data Augmentation can be applied if we can (a) draw Q from $p(Q|\Theta, U)$ and (b) draw Θ from $p(\Theta|Q, U)$. Step (b) is simple since it only involves multivariate t-distribution. Step (a) is nontrivial. As was implemented in a preliminary report of Y. Wu (Dept. of Statist., Harvard U.), step (a) can be accommodated by a "Metropolized shuffling" scheme. Roughly speaking, a random shuffling scheme is employed that provides us a Markov chain on the space of all permutations. Based on this chain, we can apply Metropolis-Hastings rejection rule to achieve our target distribution $p(Q|\Theta, U)$. In our simulation, we used switch shuffling (randomly draw two cards and switch them). Within each iteration (i.e., a cycle of Steps (a) and (b)), 500 Metropolized shuffles were conducted, since, as theory suggested, $O(n \log(n))$ steps are needed to shuffle n cards uniformly.

We simulated a data set with $\tau^2 = 1$, $\sigma^2 = 1$, and $\alpha = 0$. Assuming that $\alpha = 0$ is known, we used a flat prior for β. Figure 1 illustrates our results. Panel(1,1) shows the posterior distribution of β, where the x's were simulated from $N(0, 1)$ and the true β was zero. As indicated, its variance is 0.12, considerably larger than 0.01, the complete-data posterior variance of β. Panel(1,2) shows the autocorrelations among the β's. The fraction of missing information can be estimated as $\hat{\gamma}_B = 0.924$ from the autocorrelation plot. As theory in Sections 2 and 3 indicated,

$$(1 - \gamma_B)\text{var}(\beta \mid U) = E\{\text{var}(\beta \mid U, Q)\}$$

where the RHS is average complete-data variance. This identity was experimentally confirmed since $(1 - 0.923) \times 0.12 = 0.009$ which is close to the theoretical value 0.01. Panel(2,1) is the same posterior distribution, but the x's were simulated from $N(0, 1)$ and the true β=0. With the x's far from origin, both the posterior variance, 0.021, and the fraction of missing information, 0.619, were considerably smaller. In Panel(3,1), the x were simulated from $N(1, 1)$ and the true $\beta = 1$. It seems to suggest that the fraction of missing information is not related to the true value of β, but is very sensitive to $\sum x_i^2$.

An intuitive solution of the problem is to sort both the x and the u first and then do a regression on the sorted data. But this procedure overestimates β and does not provide proper inference. The above Bayesian method we employed, however, is unbiased (with flat prior) and supplies proper variance estimation. When $\sum x_i^2$ is extremely large, the sorting method (essentially any method) works well, implying that the matching information is unimportant for the inference of β. This, together with the foregoing simulation study, suggests a conjecture that the fraction of missing information for β monotonely decreases as $\sum x_i^2$ increases.

References

Csàki, P., and Fischer, J.H.(1960), "Contributions to the problem of maximal correlation," *Matematikao Kotato Intezet, Kozlemenyei*, **5**, 325-337.

DeGroot, M.H., Feder, P.I., and Goel, P.K. (1971), "Matchmaking," *Ann. Math. Statist.*, **42**, 578-593.

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

Gelfand, A.E. and Smith, A.F.M. (1990), "Sampling-based approaches to calculating marginal densities," *J. Amer. Statist. Assoc.*, **85**, 398–409.

Gelman, A. and Rubin, D.B. (1992), "Inference from iterative simulation using multiple sequences (with discussion)," *Statist. Sci.*, **7**, 457-511.

Geman, S. and Geman, D. (1984), "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images," *IEEE Trans. on Pattn Anal. and Mach. Intell.*, **6**, 721–741.

Geyer, C.J. (1992), "Practical Markov chain Monte Carlo", *Statist. Sci.*, **7**, 473-483.

Hastings, W.K. (1970), "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, **57**, 97-109.

Liu, J.S. (1991), "Correlation structure and convergence rate of the Gibbs sampler," Ph.D. *Thesis*, Dept. of Statist., U. of Chicago.

Liu, J.S. (1992), "Metropolized independent sampling scheme with comparisons to rejection sampling and importance sampling," *Tech. Rep.*, Stat. Dept., Harvard U. To appear in *Statistics and Computing*.

Liu, J.S. (1994), "The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem," *J. Amer. Statist. Assoc.*, **89**, in press.

Liu, J.S., Wong, W.H. and Kong, A. (1994a), "Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes," *Biometrika*, **81**, 27-40.

Liu, J.S., Wong, W.H. and Kong, A. (1994b), "Covariance structure and convergence rate of the Gibbs Sampler with various scans", *J. Roy. Statist. Soc.*, Ser. B **55**, in press.

Meng, X. (1991), "Towards complete results for some incomplete-data problems," Ph.D. *Thesis*, Dept. of Statist., Harvard U.

Meng, X. and Rubin, D.B. (1993), "Maximum likelihood estimation via the ECM algorithm: A general framework," *Biometrika*, 267-278.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. (1953), "Equations of state calculations by fast computing machines," *J. Chem. Phys.*, **21**, 1087-1091.

Mykland, P., Tierney, L. and Yu, B. (1992), "Regeneration in Markov chain samplers," *Tech. Rep.*, Dept of Statist., U. of Chicago.

Orchard, T. and Woodbury, M.A. (1972), "A missing information principle: theory and applications," In *Proc. of the 6th Berkeley Symposium on Math. Stat. and Prob.*, 697-715.

Roberts, G.O. and Polson, N.G. (1994), "On the geometric convergence of the Gibbs sampler," *J. Roy. Statist. Soc.*, Ser. B **55**, 377-384.

Roberts, G.O. (1992), "Convergence diagnostics of the Gibbs sampler," In *Bayesian Statistics* **4**, eds. J. Bernardo, J.O. berger, A.P. Dawid and A.F.M. Smith, 763-773, Oxford University Press.

Rosenthal, J.S. (1993a), "Rates of convergence for Data Augmentation on finite sample spaces," *Ann. Appl. Prob.* **3**, 319-339.

Rosenthal, J.S. (1993b), "Minorization conditions and convergence rates for Markov chain Monte Carlo," *Tech. Rep.*, Dept. of Statist., U. of Toronto.

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

Schervish, M.J., and Carlin, B.P. (1992), "On the convergence of successive substitution sampling," *J. Comp. Graph. Statist.* **1**, 111-127.

Garren, S.T. and Smith, R.L. (1994), "Convergence diagnostics for Markov chain samplers," *Tech. Rep.*, Dept. of Statist., U. of North Carolina.

Smith, A.F.M., and Roberts, G.O. (1993), "Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods (with discussion)," *J. Roy. Statist. Soc.*, Ser. B, **55**, 3-23.

Tanner, M.A. and Wong, W.H. (1987), "The calculation of posterior distributions by data augmentation (with discussion)," *J. Amer. Statist. Assoc.*, **82**, 528–550.

Tierney, L. (1991), "Markov chains for exploring posterior distributions", *Tech. Rep. 560*, School of Statistics, University of Minnesota.

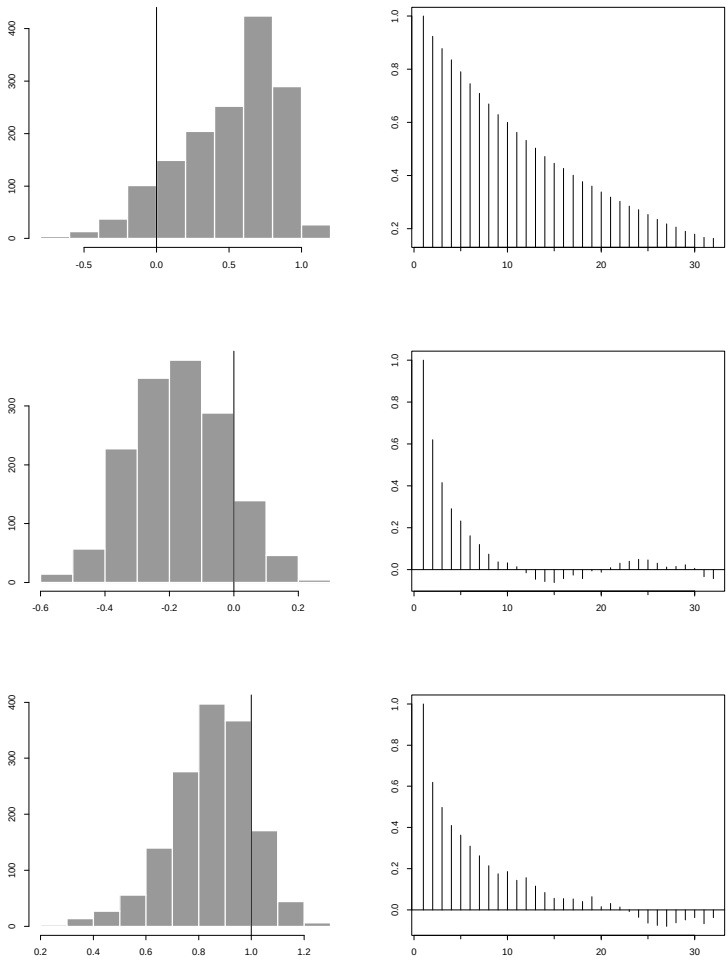


Figure 1: Estimation in broken regression