Efficient Exact Sampling Procedures for Bayesian Generalized Linear Models *

Kjell Nygren †

*Keywords: Bayesian Statistics, Log-concave likelihood functions, Generalized Linear Models, Normal Regression, Poisson Regression, Likelihood Subgradient Densities

†Kjell Nygren is Senior Manager - Marketing Analytics, IMS Health, 960 Harvest Drive, Building A, Blue Bell, PA 19422 (email: knygren@us.imshealth.com)

Abstract

Bayesian statistics is built around the concept of posterior probability and requires the integration of complex high-dimensional integrals. Except for the simplest of models, there are no reliable analytical or numerical integration techniques for such models. Current estimation procedures typically make use of Markov Chain Monte Carlo techniques that rely on Markov Chains which converge to draws from the densities of interests. Unfortunately, there are few known results regarding the rate of convergence for Markov Chains. This makes it hard to judge when a Markov Chain has been run long enough. In this paper, we derive efficient exact sampling procedures for a large class of Bayesian Generalized linear models. This should facilitate the development of Bayesian softwares that do not require convergence diagnostics on the part of users.

Our first set of procedures are for models with normal and multivariate data. We show that such models have the same posterior densities as fictitious models with log-concave likelihood functions. This result allows us to make use of Likelihood-subgradient densities and accept-reject procedures in order to produce samples for such models. The resulting procedures are easy to implement and should be efficient as long as the prior is relatively weak and centered not too far from the center of the data. We provide explicit sampling procedures for a model of a normal population, a model of a multivariate normal population, and a regression model with normal error terms. The same type of procedures should be applicable to a large class of other models as well.

Our second set of results concerns models with multivariate normal priors and log-concave likelihood functions. We show that mixtures of restricted likelihood subgradient densities can be used together with accept-reject procedures in order to produce

exact samples from posterior densities. This result allows us to improve on the acceptance rates that would result from the use of a single likelihood subgradient density. We show explicitly how to sample from a mixture of restricted multivariate normal likelihood subgradient densities. The efficiency of the resulting procedure depends on the number of restricted likelihood subgradient densities and their corresponding positioning. For the case of a multivariate normal prior, we show how to compute a function which can be used to compare the acceptance rates of any two mixtures of restricted likelihood subgradient densities. This function should be useful in choosing the positioning of restricted likelihood subgradient densities. The sampling method is illustrated for models with Poisson data and canonical link functions. The same type of procedures can also be used for any model with a multivariate normal prior and a log-concave likelihood function.

1 Introduction

Bayesian statistics is built around the concept of posterior probability. First proposed by Bayes (1763) and Laplace (1785, 1810) in the 18th century, calculations of posterior probabilities requires integration of complex high-dimensional integrals. Except for the simplest of models, there are no reliable analytical or numerical integration techniques available for such models.

Monte Carlo simulation (see e.g., Metropolis and Ulam (1949), Eckhardt (1987) and Fishman (1999)) offers a possible alternative integration technique. The simplest forms of Monte Carlo simulation involves generating independent and identically distributed draws from the densities of interest. Most of these exact sampling procedures make use of either the inverse transform method or von Neumann's (1951) accept-reject procedure. In Bayesian statistics, exact Monte Carlo sampling procedures have largely been limited to models with natural conjugate priors (Raiffa and Schlaifer, 1961). Models with natural conjugate priors have posterior densities of the same functional form as the prior density. The adaptive rejection sampling Algorithm (Gilks and Wild 1992, Gilks 1992) and Adaptive Rejection Metropolis Algorithms (Gilks, Best and Tan 1995) make efficient use of von Neumann's accept-reject procedure in the univariate case and has been implemented as part of Gibbs sampling procedures (Geman and Geman 1984, Gelfand and Smith 1990). These algorithms gradually construct an improved approximation to the full conditional posterior density by using tangents to the full conditional posterior density.

Generalized linear models with canonical links are frequently used Bayesian models. Several papers have suggested inexact sampling procedures for such models (Dellaportas and

Smith (1993), Breslow and Clayton (1993), Gamerman (1997)). These inexact sampling procedures make use of Markov Chain Monte Carlo techniques based on the Metropolis-Hastings algorithm (Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller 1953, Hastings 1970). The Metropolis-Hastings Algorithm makes use of a Markov Chain that converges to a sample from the posterior density when run long enough. Bounds on the rate at which the Metropolis-Hastings algorithm converges are unavailable for most models. Hence it is not known how long the suggested inexact sampling procedures need to be run before the Markov Chain gets close enough to the desired density. In a recent paper, Nygren (2003) introduced likelihood subgradient densities and showed how such densities could be used with accept-reject procedures in order to generate exact samples from posterior densities. Likelihood subgradient densities are constructed from prior densities by subtracting a linear term from the log of the density. The gradient of the linear term is a subgradient for the negative of the log-likelihood function at some point $\overline{\mathbf{x}}$. Such subgradients exist at every point for log-concave likelihood functions.

In this paper, we build on the results in Nygren (2003) in order to construct efficient sampling procedures for Bayesian Generalized Linear models with non-conjugate priors. Our first set of results concerns models with normal or multivariate normal data and independent priors for regression and variance parameters. Such models do not have log-concave likelihood functions and hence the results in Nygren (2003) are not directly applicable. Our sampling procedure instead makes use of the remarkable property that such models have associated fictitious models with (i) the same posterior density and (ii) a log-concave likelihood function. The results in Nygren (2003) are then applicable to the fictitious models. This useful property is used in order to develop efficient sampling procedures for a model of a normal population,

a model of a multivariate normal population, and a Bayesian regression model with normal error terms. The acceptance rates for the resulting sampling procedures are related to the strengths of the priors in the respective models. For weak priors, the procedures should be very efficient. For very strong priors, the efficiency of the procedures declines. While these first set of results are presented for models with normal and multivariate normal priors, the same procedures could be used for more general prior specifications as long as the prior densities are bounded above.

Our second set of results make use of mixtures of restricted likelihood subgradient densities in order to improve on the exact sampling procedures presented in Nygren (2003). The key property of a Likelihood Subgradient density that enable its use in accept-reject procedures is that a multiple of the density bounds the posterior density from above. The smallest such multiple of the density moreover equals the posterior density at the point for which the gradient of the subtracted term is a subgradient for the negative of the log-likelihood function. As a result, the smallest such multiple is a good approximation for the posterior density in a neigborhood of that point, but a poorer approximation further from that point. By using mixtures of restricted likelihood subgradient densities, we are able to construct better general approximations to the posterior density than what results from a single likelihood subgradient density. The key theorem in this paper shows that mixtures of restricted likelihood subgradient densities can be used to generate samples from posterior densities. We also demonstrate how to construct a mixture of likelihood subgradient densities from a class of restricted multivariate normal likelihood subgradient densities for which sampling procedures are known. The method is then illustrated using the Poisson regression model. The same method can also be used to generate samples for any model

with a log-concave likelihood function and a multivariate normal prior. Such models include the logit, multinomial-logit, and conditional-logit models.

The rest of this paper is organized as follows. In section 2, we derive exact sampling procedures for models with Normal and Multivariate Normal data. Section 3 introduces mixtures of likelihood subgradient densities and shows that such densities can be used with accept-reject procedures in order to generate samples from their corresponding posterior densities. It also demonstrates how to sample from such mixtures of restricted likelihood subgradient densities for a class of mixtures of restricted multivariate normal likelihood subgradient densities. In section 4, we show how to use mixtures of likelihood subgradient densities in generalized linear models with Poisson data. Some concluding discussion is contained in section 5. Section 6, finally, contains the proof of the theorem related to mixtures of restricted likelihood subgradient densities.

2 Generalized Linear Models with Normal and Multivariate Normal Data

In Nygren (2003), we showed how Likelihood subgradient densities could be used to generate samples from posterior densities. Models for normal data with independent priors for the variance and mean parameters do not have likelihood subgradient densities. In this section, we show that such models have the same posterior densities as fictitious models for which likelihood subgradient densities exist. We make use of this property in order to derive efficient exact sampling procedure for models with Normal and Multivariate normal data.

2.1 Models for Normal Populations

We first consider a Bayesian model of a normal population with independent priors for the mean and variance parameters. Specifically, we consider the following:

$$\sigma^2 \sim Inverse - Gamma(n_0/2, n_0 Var_0/2)$$

$$\beta \sim Normal(\mu, \sigma_0^2)$$

$$y_i \sim Normal(\beta, \sigma^2), i = 1, \ldots, n_1.$$

Here $\mathbf{y}$ represents an observed data vector. We note that this is a non-conjugate model. In order to develop an exact sampling procedure for this model, we introduce the following fictitious model (Model A):

$$\sigma^2 \sim Inverse - Gamma((n_0 + n_1 - 1)/2, (n_0 Var_0 + \sum_{i=1}^{n1}(y_i - \overline{y})^2)/2)$$

$$\beta \sim Normal(\overline{y}, \sigma^2/n_1)$$

$$\mu \sim Normal(\beta, \sigma_0^2)$$

where $\overline{y} := \sum_{i=1}^{n_1} y_i/n_1$, and μ is now interpreted as the data. We now establish the following important relationship between the Bayesian Normal Population Model and Model A.

Claim 1. *Fictitious model A has the same posterior density as the Bayesian Normal Population Model.*

Proof of Claim 1: It suffices to show that the log-posterior density for the two models have the same terms involving β and σ^2. Starting with the log-posterior from the Bayesian Normal population model, we have (where K is a constant that does not depend on β or σ^2)

$$
\begin{aligned}
&[-((n_0/2)-1)\ln(\sigma^2) - (1/2)n_0 Var_0/\sigma^2] \\
&\quad +[-(1/2)\textstyle\sum_{i=1}^{n_1}((y_i-\beta)^2/\sigma^2 + \ln(\sigma^2))] \\
&\qquad +[-(1/2)(\beta-\mu)^2/\sigma_0^2] + K \;=\; [-(((n_0+n_1)/2)-1)]\ln(\sigma^2) \\
&\qquad\qquad +[-(1/2)n_0 Var_0]/\sigma^2 \\
&\qquad\qquad +[-(1/2)\textstyle\sum_{i=1}^{n_1}((y_i-\overline{y})-(\beta-\overline{y}))^2]/\sigma^2 \\
&\qquad\qquad +[-(1/2)(\beta-\mu)^2/\sigma_0^2] + K \\
&\qquad\qquad = \; [-(((n_0+n_1)/2)-1)]\ln(\sigma^2) \\
&\qquad\qquad\quad +[-(1/2)(n_0 Var_0 + \textstyle\sum_{i=1}^{n_1}(y_i-\overline{y})^2]/\sigma^2 \\
&\qquad\qquad\quad +[-(1/2)\textstyle\sum_{i=1}^{n_1}((\beta-\overline{y})^2 \\
&\qquad\qquad\quad -2(y_i-\overline{y})(\beta-\overline{y}))]/\sigma^2 \\
&\qquad\qquad\quad +[-(1/2)(\beta-\mu)^2/\sigma_0^2] + K \\
&\qquad\qquad = \; [-(((n_0+n_1-1)/2)-1)]\ln(\sigma^2) \\
&\qquad\qquad\quad -[(1/2)(n_0 Var_0 + \textstyle\sum_{i=1}^{n_1}(y_i-\overline{y})^2)]/\sigma^2 \\
&\qquad\qquad\quad -[(1/2)\ln(\sigma^2/n_1) + (1/2)(\beta-\overline{y})^2(n_1/\sigma^2)] \\
&\qquad\qquad\quad -[(1/2)(\beta-\mu)^2/\sigma_0^2] + K - (1/2)\ln(n_1)
\end{aligned}
$$

which is recognized as the log-posterior for the fictitious model. Q.E.D.

We note that the likelihood function for fictitious model A obtains its maximum at μ. Hence it follows from Fact 1 in Nygren (2003) that the prior of the fictitious model can be used with an accept-reject procedure to generate a sample from the Posterior density.

We thus have the following exact sampling procedure for the Bayesian Normal Population Model.

Repeat Until Acceptance:

(i) Generate $\sigma^2 \sim Inverse - Gamma((n_0 + n_1 - 1)/2, (n_0 Var_0 + \sum_{i=1}^{n1}(y_i - \overline{y})^2)/2)$.

(ii) Generate $\beta \sim Normal(\overline{y}, \sigma^2/n_1)$.

(iii) Generate $U \sim Uniform(0, 1)$.

(iv) Accept if $ln(U) \leq -(1/2)(\beta - \mu)^2(1/\sigma_0^2)$.

We note that this procedure should have reasonable acceptance rates as long as the prior is relatively weak and centered not too far from the center of the support of the fictitious prior.

An important generalization of the Bayesian Normal Population model is to a multivariate normal population. We now consider the following Bayesian Multivariate-normal population model:

$$\mathbf{\Sigma} \sim Inverse - Wishart(n_0 \mathbf{Var_0}, n_0)$$

$$\beta \sim Multivariate - Normal(\mu, \mathbf{\Sigma_0})$$

$$\mathbf{y_i} \sim Multivariate - Normal(\beta, \mathbf{\Sigma}), i = 1, \ldots, n_1.$$

where again $\mathbf{y}$ represents an observed data vector.

As in the univariate case, we now introduce a fictitious model (Model B):

$$\mathbf{\Sigma} \sim Inverse - Wishart((n_0 \mathbf{Var_0} + \sum_{i=1}^{n1}(\mathbf{y_i} - \overline{\mathbf{y}})^T(\mathbf{y_i} - \overline{\mathbf{y}})), n_0 + n_1 - 1)$$

$$\beta \sim Multivariate - Normal(\overline{\mathbf{y}}, (1/n_1)\mathbf{\Sigma})$$

$$\mu \sim Multivariate - Normal(\beta, \mathbf{\Sigma_0})$$

where $\overline{\mathbf{y}} := \sum_{i=1}^{n_1}(1/n_1)\mathbf{y_i}$,and μ is interpreted as the data. We now show the following equivalence result.

Claim 2. *Fictitious model B has the same posterior density as the Bayesian Multivariate-Normal Population Model.*

Proof of Claim 2: It suffices to show that the log-posterior density for the two models have the same terms involving β and $\mathbf{\Sigma}$. Starting with the log-posterior from the p-dimensional Bayesian Multivariate-Normal population model, we have (where again K is a constant that does not depend on β or $\mathbf{\Sigma}$)

$$
\begin{aligned}
& & [-((n_0-p-1)/2)\ln(|\mathbf{\Sigma}|)] & \\
& & -[(1/2)trace(n_0\mathbf{Var_0}\mathbf{\Sigma^{-1}})] & \\
& & -[(1/2)\textstyle\sum_{i=1}^{n_1}(\mathbf{y_i}-\beta)^T\mathbf{\Sigma^{-1}}(\mathbf{y_i}-\beta)] & \\
& & -[(1/2)\textstyle\sum_{i=1}^{n_1}\ln(|\mathbf{\Sigma}|)] & \\
-[(1/2)(\beta-\mu)^T\mathbf{\Sigma_0^{-1}}(\beta-\mu)]+K & = & [-((n_0+n_1-p-1)/2)\ln(|\mathbf{\Sigma}|)] & \\
& & -[(1/2)trace(n_0\mathbf{Var_0}\mathbf{\Sigma^{-1}})] & \\
& & -[(1/2)\textstyle\sum_{i=1}^{n_1}(\mathbf{y_i}-\overline{\mathbf{y}})^T\mathbf{\Sigma^{-1}}(\mathbf{y_i}-\overline{\mathbf{y}})] & \\
& & -[(1/2)\textstyle\sum_{i=1}^{n_1}(\beta-\overline{\mathbf{y}})^T\mathbf{\Sigma^{-1}}((\beta-\overline{\mathbf{y}}) & \\
& & -2(\mathbf{y_i}-\overline{\mathbf{y}}))] & \\
& & -[(1/2)(\beta-\mu)^T\mathbf{\Sigma_0^{-1}}(\beta-\mu)]+K & \\
& = & [-((n_0+(n_1-1)-p-1)/2)\ln(|\mathbf{\Sigma}|)] & \\
& & -[(1/2)trace(n_0\mathbf{Var_0}\mathbf{\Sigma^{-1}})] & \\
& & -[(1/2)\textstyle\sum_{i=1}^{n_1}(\mathbf{y_i}-\overline{\mathbf{y}})^T\mathbf{\Sigma^{-1}}(\mathbf{y_i}-\overline{\mathbf{y}})] & \\
& & -[(1/2)(\beta-\overline{\mathbf{y}})^Tn_1\mathbf{\Sigma^{-1}}(\beta-\overline{\mathbf{y}})] & \\
& & +[(1/2)\ln(|(1/n_1)\mathbf{\Sigma}|) & \\
& & -[(1/2)(\beta-\mu)^T\mathbf{\Sigma_0^{-1}}(\beta-\mu)]+K+(p/2)\ln(n_1) & \\
& = & [-((n_0+(n_1-1)-p-1)/2)\ln(|\mathbf{\Sigma}|)] & \\
& & -[(1/2)trace((n_0\mathbf{Var_0} & \\
& & +\textstyle\sum_{i=1}^{n_1}(\mathbf{y_i}-\overline{\mathbf{y}})^T(\mathbf{y_i}-\overline{\mathbf{y}}))\mathbf{\Sigma^{-1}})] & \\
& & -[(1/2)\textstyle\sum_{i=1}^{n_1}(\mathbf{y_i}-\overline{\mathbf{y}})^T\mathbf{\Sigma^{-1}}(\mathbf{y_i}-\overline{\mathbf{y}})] & \\
& & -[(1/2)(\beta-\overline{\mathbf{y}})^Tn_1\mathbf{\Sigma^{-1}}(\beta-\overline{\mathbf{y}})] & \\
& & +[(1/2)\ln(|(1/n_1)\mathbf{\Sigma}|) & \\
& & -[(1/2)(\beta-\mu)^T\mathbf{\Sigma_0^{-1}}(\beta-\mu)]+K+(p/2)\ln(n_1) &
\end{aligned}
$$

which is recognized as the log-posterior for the fictitious model. Q.E.D.

We note that the likelihood function for fictitious model B obtains its maximum at vector μ. Combining Claim 2 with Fact 1 in Nygren (2003), we thus have the following exact sampling procedure for the Bayesian Multivariate Normal Population Model.

Repeat Until Acceptance:

(i) Generate $\mathbf{\Sigma} \sim Inverse - Wishart((n_0\mathbf{Var_0} + \sum_{i=1}^{n1}(\mathbf{y_i} - \overline{\mathbf{y}})^T(\mathbf{y_i} - \overline{\mathbf{y}})), n_0 + n_1 - 1)$.

(ii) Generate $\beta \sim Multivariate - Normal(\overline{\mathbf{y}}, (1/n_1)\mathbf{\Sigma})$.

(iii) Generate $U \sim Uniform(0, 1)$.

(iv) Accept if $ln(U) \leq -(1/2)(\beta - \mu)^T\mathbf{\Sigma_0^{-1}}(\beta - \mu)$.

We note that this procedure again should have reasonable acceptance rates as long as the prior is relatively weak and centered not too far from the center of the support of the fictitious prior.

2.2 Bayesian Regression with Normal Error terms

The classical linear regression model is perhaps the most important model in all of statistics. In this section, we derive efficient sampling procedures for a Bayesian regression model with normal error terms and independent priors for the variance and regression parameters. More specifically, we consider the following model

$$\sigma^2 \sim Inverse - Gamma(n_0/2, n_0Var_0/2)$$

$$\beta \sim Multivariate - Normal(\mu, \mathbf{\Sigma_0})$$

$$y_i \sim Normal(\mathbf{X_i}\beta, \sigma^2), i = 1, \ldots, n_1.$$

Here $\mathbf{y}$ represents an observed data vector. We note that this is a non-conjugate model. Before developing the general sampling procedure, we first consider two important special cases of this Bayesian Normal regression model.

If $\mathbf{X}^T\mathbf{X}$ has full rank m, then we can introduce the following fictitious model (Model C):

$$\sigma^2 \sim Inverse - Gamma((n_0 + n_1 - m)/2, (n_0 Var_0 + \sum_{i=1}^{n1}(y_i - \mathbf{X_i}\bar{b})^2)/2)$$

$$\beta \sim Multivariate - Normal(\bar{b}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$$

$$\mu \sim Normal(\beta, \mathbf{\Sigma_0})$$

where $\bar{b} := (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$, and μ is now interpreted as the data. In the case of full rank, we can now show the following equivalence result.

Claim 3. *If $\mathbf{X}^T\mathbf{X}$ has full rank, then fictitious model C has the same posterior density as the Normal Regression Model.*

Proof of Claim 3: It suffices to show that the log-posterior density for the two models have the same terms involving β and σ^2. Starting with the log-posterior from the m-dimensional Bayesian Regression model, we have (where $\mathbf{b} := (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ and K is a constant that does not depend on β or σ^2)

$$
\begin{aligned}
&-[((n_0/2)-1)\ln(\sigma^2)] \\
&-[(1/2)n_0 Var_0/\sigma^2] \\
&-[(1/2\sigma^2)(\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta)] \\
&-[(1/2)n_1\ln(\sigma^2)] \\
&-[(1/2)(\beta-\mu)^T{\boldsymbol{\Sigma_0}}^{-1}(\beta-\mu)] \\
+K &= [-(((n_0+n_1)/2)-1)\ln(\sigma^2)] \\
&\quad -[(1/2)n_0 Var_0/\sigma^2] \\
&\quad -[(1/2\sigma^2)((\mathbf{y}-\mathbf{X}b)-(\mathbf{X}\beta-\mathbf{X}b))^T((\mathbf{y}-\mathbf{X}b)-(\mathbf{X}\beta-\mathbf{X}b))] \\
&\quad -[(1/2)(\beta-\mu)^T{\boldsymbol{\Sigma_0}}^{-1}(\beta-\mu)]+K \\
&= [-(((n_0+n_1)/2)-1)\ln(\sigma^2)] \\
&\quad -[(1/2)n_0 Var_0/\sigma^2] \\
&\quad -[(1/2\sigma^2)(\mathbf{y}-\mathbf{X}b)^T(\mathbf{y}-\mathbf{X}b)] \\
&\quad -[(1/2\sigma^2)(\beta-b)^T\mathbf{X}^T\mathbf{X}(\beta-b)] \\
&\quad -[(1/2)(\beta-\mu)^T{\boldsymbol{\Sigma_0}}^{-1}(\beta-\mu)]+K \\
&= [-(((n_0+n_1-m)/2)-1)\ln(\sigma^2)] \\
&\quad -[(1/2)n_0 Var_0/\sigma^2] \\
&\quad -[(1/2\sigma^2)(\mathbf{y}-\mathbf{X}b)^T(\mathbf{y}-\mathbf{X}b)] \\
&\quad -[(1/2)\ln(|\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}|+(1/2\sigma^2)(\beta-b)^T\mathbf{X}^T\mathbf{X}(\beta-b)] \\
&\quad -[(1/2)(\beta-\mu)^T{\boldsymbol{\Sigma_0}}^{-1}(\beta-\mu)]+K+[(1/2)\ln(|(\mathbf{X}^T\mathbf{X})^{-1}|)] \\
&= [-(((n_0+n_1-m)/2)-1)\ln(\sigma^2)] \\
&\quad -[(1/2)(n_0 Var_0+(\mathbf{y}-\mathbf{X}b)^T(\mathbf{y}-\mathbf{X}b))/\sigma^2] \\
&\quad -[(1/2)\ln(|\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}|+(1/2\sigma^2)(\beta-b)^T\mathbf{X}^T\mathbf{X}(\beta-b)] \\
&\quad -[(1/2)(\beta-\mu)^T{\boldsymbol{\Sigma_0}}^{-1}(\beta-\mu)]+K+[(1/2)\ln(|(\mathbf{X}^T\mathbf{X})^{-1}|)]
\end{aligned}
$$

which is recognized as the log-posterior for the fictitious model. Q.E.D.

We note that the likelihood function for fictitious model C obtains its maximum at vector μ. Combining Claim 3 with Fact 1 in Nygren (2003), we thus have the following exact sampling procedure for the m-dimensional Full-Rank Bayesian Regression model with normal error terms.

Repeat Until Acceptance:

(i) Generate $\sigma^2 \sim Inverse - Gamma((n_0 + n_1 - m)/2, (n_0 Var_0 + \sum_{i=1}^{n1}(y_i - \mathbf{X_i}\bar{b})^2)/2)$.

(ii) Generate $\beta \sim Multivariate - Normal(\bar{b}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$.

(iii) Generate $U \sim Uniform(0, 1)$.

(iv) Accept if $ln(U) \leq -(1/2)(\beta - \mu)^T \mathbf{\Sigma_0^{-1}}(\beta - \mu)$.

We note that this procedure again should have reasonable acceptance rates as long as the prior is relatively weak and centered not too far from the center of the support of the fictitious prior.

Let us now introduce some additional notation and state a couple of facts.

Definition 1. *The Precision* **Prec** *of a multivariate-normal distribution is the inverse of the Variance-Covariance Matrix.*

Denote by β_1 the first m_1 elements and by β_2 the last $m - m_1$ elements of an m-dimensional multivariate normal vector. In the following, we make use of corresponding sub-matrices of the variance-covariance and precision matrices. The following two facts are straightforward to verify.

Fact 1. *The marginal distribution for* β_1 *is*

$$\beta_1 \sim Multivariate - Normal(\mu_1, Var_{11})$$

Fact 2. *The conditional distribution for* β_2 *is*

$$\beta_2|\beta_1 \sim Multivariate - Normal(\mu_2 - \mathbf{Prec_{22}}^{-1}\mathbf{Prec_{21}}(\beta_1 - \mu_1), \mathbf{Prec_{22}}^{-1})$$

Let us now consider the following Partitioned Bayesian regression where only the elements of β_1 enter as independent variables:

$$\sigma^2 \sim Inverse - Gamma(n_0/2, n_0 Var_0/2)$$

$$\beta_1 \sim Multivariate - Normal(\mu_1, \mathbf{\Sigma_{0,11}})$$

$$\beta_2 \sim Multivariate - Normal(\mu_2 - \mathbf{Prec_{22}}^{-1}\mathbf{Prec_{21}}(\beta_1 - \mu_1), \mathbf{Prec_{22}}^{-1})$$

$$y_i \sim Normal(\mathbf{X1_i}\beta_1, \sigma^2), i = 1, \ldots, n_1.$$

If $\mathbf{X1}^T\mathbf{X1}$ has rank m_1 then it is straightforward to verify that the following procedure can be used to generate a sample from the posterior density of this model:

(i) Generate a sample for (β_1, σ^2) using the procedure for the Full-rank Bayesian regression model above

(ii) Sample $\beta_2 \sim Multivariate - Normal(\mu_2 - \mathbf{Prec_{22}}^{-1}\mathbf{Prec_{21}}(\beta_1 - \mu_1), \mathbf{Prec_{22}}^{-1})$.

In order to handle the general case of a Bayesian regression of less than full rank, we will make use of a singular value decomposition for the design matrix. The following two facts are well known (see e.g., The Numerical Algorithms Group, Chapter f01, section 2 (2002)).

Fact 3. *A real valued n by m Matrix $\mathbf{X}$ of rank $r \leq k = \min(n, m)$ can be factored as the singular value decomposition*

$$\mathbf{X} = \mathbf{QDP}^T$$

where $\mathbf{Q}$ is an n by n orthogonal matrix, $\mathbf{P}$ is an m by m orthogonal matrix and D is an n by m diagonal matrix with non-negative elements. The first k columns of $\mathbf{Q}$ and $\mathbf{P}$ are the left- and right-hand singular vectors of $\mathbf{A}$ respectively and the k diagonal elements of $\mathbf{D}$ are the singular values of $\mathbf{A}$. $\mathbf{D}$ may be chosen so that

$$d_1 \geq d_2 \geq \ldots \geq d_k \geq 0$$

and in this case case, if rank($\mathbf{X}$)=r then

$$d_1 \geq d_2 \geq \ldots \geq d_r > 0, d_{r+1} = \ldots = d_k = 0.$$

Fact 4. *If $\tilde{\mathbf{Q}}$ and $\tilde{\mathbf{P}}$ consists of the first r columns of $\mathbf{Q}$ and $\mathbf{P}$ respectively and $\mathbf{S}$ is an r by r diagonal matrix with diagonal elements $d_1, d_2, \ldots, d_r$ then $\mathbf{X}$ is given by*

$$\mathbf{X} = \tilde{\mathbf{Q}}\mathbf{S}\tilde{\mathbf{P}}^{\mathbf{T}}.$$

The properties in the next two facts follow from properties of Orthogonal matrices.

Fact 5. *Let $\tilde{\mathbf{X}} := \mathbf{QD}$, $\tilde{\beta} := \mathbf{P}^T\beta$, and $\tilde{\mathbf{X1}} = \tilde{\mathbf{Q}}\mathbf{S}$, then*

(i) $\mathbf{P}^T$ is a non-singular m by m matrix,

(ii) $\tilde{\mathbf{X1}}$ is an n by r matrix with rank r,

(iii) the first r elements of the vector $\tilde{\beta}$ are given by $\tilde{\beta}_1 := \tilde{\mathbf{P}}^T\beta$, and

(iv) $\mathbf{X}\beta = \tilde{\mathbf{X}}\tilde{\beta} = \tilde{\mathbf{X1}}\tilde{\beta}_1$.

Fact 6. *If β has a multivariate normal distribution with mean vector μ and Precision matrix* **Prec** *then $\tilde{\beta}$ has a multivariate normal distribution with mean vector $\tilde{\mu} := \mathbf{P}^T\mu$ and Precision matrix* $\tilde{\mathbf{Prec}} := ((\mathbf{P}^T)^{-1})^T\mathbf{Prec}(\mathbf{P}^T)^{-1} = \mathbf{P}^T\mathbf{PrecP}$.

We now express the general rank Bayesian regression in terms of $\tilde{\beta}$. The above facts implies that the resulting regression is the following partitioned regression model:

$$\sigma^2 \sim Inverse - Gamma(n_0/2, n_0 Var_0/2)$$

$$\tilde{\beta}_1 \sim Multivariate - Normal(\tilde{\mu}_1, \tilde{\mathbf{\Sigma}}_{\mathbf{0,11}})$$

$$\tilde{\beta}_2 \sim Multivariate - Normal(\tilde{\mu}_2 - \tilde{\mathbf{Prec}}_{\mathbf{22}}^{-1}\tilde{\mathbf{Prec}}_{\mathbf{21}}(\tilde{\beta}_1 - \tilde{\mu}_1), \tilde{\mathbf{Prec}}_{\mathbf{22}}^{-1})$$

$$y_i \sim Normal(\tilde{\mathbf{X}}\mathbf{1_i}\tilde{\beta}_1, \sigma^2), i = 1, \ldots, n_1.$$

Making use of this ability to rewrite the model into a partitioned regression provides us with the following exact sampling procedure for the general rank Bayesian regression model:

(i) Generate a sample for $(\tilde{\beta}, \sigma^2)$ using the procedure for the Partitioned Bayesian regression model above

(ii) Compute $\beta = (\mathbf{P}^T)^{-1}\tilde{\beta} = \mathbf{P}\tilde{\beta}$.

2.3 Discussion

This section derived exact sampling procedures for Bayesian Models with normal and multivariate normal data. The efficiency of the procedures are related to the strength and positioning of the priors. With the exception of models with very strong or badly centered

priors, the procedures presented should be efficient. The underlying procedures should also be easily generalizable to large classes of other models with normal and multivariate normal data. We also note that in most of the models, the normal and multivariate normal priors could be replaced with other priors. The crucial property that more general priors would need to satisfy is to have known upper bounds. For such more general priors, the accept condition would then be modified so as to accept a proposed draw whenever the value of the uniform variable is less than the value of the prior divided by that upper bound. The only case in which such more general priors could not be easily used is for the regression model of less than full rank. Our procedure for that model makes explicit use of the multivariate normality of the associated prior.

3 Mixtures of Restricted Generalized Likelihood Subgradient Densities

In Nygren (2003), we introduced likelihood subgradient densities and generalized likelihood subgradient densities. We also showed how such densities could be used in exact sampling procedures. In the examples section, we showed explicitly how they could be used for Bayesian Poisson and Conditional logit regression models. For a model with multivariate normal data, we showed that the acceptance rate for the procedure depended on the amount of information contained in the data relative to the amount of information contained in the prior. More specifically, the acceptance rate was high when the prior contained a lot of information and low when it contained only small amounts of information relative to the

prior. This property is likely to make the acceptance rate unreasonably low for many models. In this section, we introduce mixtures of Generalized Likelihood subgradient densities. We also describe how such densities can be used to obtain dramatic improvements in acceptance rates over those obtained with generalized likelihood subgradient densities. The sampling procedures are illustrated using Mixtures of a subclass of restricted multivariate normal likelihood subgradient densities.

3.1 General Results

We first recall the following definition of a Generalized likelihood-subgradient density from Nygren (2003).

Definition 2. *A probability density function $q(.)$ is a Generalized Likelihood-Subgradient probability density for a posterior distribution $\pi(.|\boldsymbol{y})$ with prior distribution $\pi(.)$ and likelihood function $f(\boldsymbol{y}|.)$ at a point $\overline{\boldsymbol{x}} \in \boldsymbol{X}$ if there exists a subgradient $\boldsymbol{c}(\overline{\boldsymbol{x}})$ for the negative of the log of a function h at $\overline{\boldsymbol{x}}$, such that:*

(i) h bounds $f(\boldsymbol{y}|.)$ from above;

(ii) $MGF(\boldsymbol{c}(\overline{\boldsymbol{x}})) := \int_{\boldsymbol{x} \in \boldsymbol{X}} e(-\boldsymbol{c}(\overline{\boldsymbol{x}})^T \boldsymbol{x}) \pi(\boldsymbol{x}) d\boldsymbol{x}$ is finite; and

(iii) $\forall \boldsymbol{x} \in \boldsymbol{X} : q(\boldsymbol{x}) = exp(-\boldsymbol{c}(\overline{\boldsymbol{x}})^T \boldsymbol{x}) \pi(\boldsymbol{x}) / MGF(\boldsymbol{c}(\overline{\boldsymbol{x}}))$.

We now show that Likelihood Subgradient densities satisfy the following key property.

Fact 7. *If $q(.)$ is a Generalized Likelihood-Subgradient probability density for a posterior distribution $\pi(.|\boldsymbol{y})$ with prior distribution $\pi(.)$ and likelihood function $f(\boldsymbol{y}|.)$ at a point $\overline{\boldsymbol{x}} \in \boldsymbol{X}$ then*

$$\frac{MGF(c(\overline{\boldsymbol{x}}))h(\overline{\boldsymbol{x}})}{exp(-c(\overline{\boldsymbol{x}})^T \overline{\boldsymbol{x}}) f(\mathbf{y})} q(\mathbf{x}) \geq \pi(\boldsymbol{x}|\boldsymbol{y}), \ \forall \mathbf{x} \in \mathbf{X}.$$

Proof of Fact 7: Since $h(.)$ bounds $f(\mathbf{y}|.)$ from above and $\mathbf{c}(\overline{\mathbf{x}})$ is a subgradient for $-\ln(h(.))$ at $\overline{\mathbf{x}}$, it follows that for every $\mathbf{x} \in \mathbf{X}$,

$$
\begin{aligned}
&-\mathbf{c}(\overline{\mathbf{x}})^T\overline{\mathbf{x}} - \ln(h(\overline{\mathbf{x}})) + \\
&\ln(f(\mathbf{y}|\mathbf{x})) + \mathbf{c}(\overline{\mathbf{x}})^T\mathbf{x} \;\leq\; -\mathbf{c}(\overline{\mathbf{x}})^T\overline{\mathbf{x}} - \ln(h(\overline{\mathbf{x}})) + \ln(h(\mathbf{x})) + \mathbf{c}(\overline{\mathbf{x}})^T\mathbf{x} \\
&\qquad\qquad \leq\; -\mathbf{c}(\overline{\mathbf{x}})^T\overline{\mathbf{x}} - \ln(h(\overline{\mathbf{x}})) + [\ln(h(\overline{\mathbf{x}})) - \mathbf{c}(\overline{\mathbf{x}})^T(\mathbf{x} - \overline{\mathbf{x}})] + \mathbf{c}(\overline{\mathbf{x}})^T\mathbf{x} \\
&\qquad\qquad =\; 0.
\end{aligned}
$$

We then have

$$
\begin{aligned}
-\mathbf{c}(\overline{\mathbf{x}})^T\overline{\mathbf{x}} - \ln(h(\overline{\mathbf{x}})) + \ln(f(\mathbf{y}|\mathbf{x})) + \mathbf{c}(\overline{\mathbf{x}})^T\mathbf{x} &\leq 0 \\
&\Updownarrow \\
(f(\mathbf{y}|\mathbf{x})exp(-\mathbf{c}(\overline{\mathbf{x}})^T\overline{\mathbf{x}}))/(h(\overline{\mathbf{x}})exp(-\mathbf{c}(\overline{\mathbf{x}})^T\mathbf{x})) &\leq 1 \\
&\Updownarrow \\
h(\overline{\mathbf{x}})exp(-\mathbf{c}(\overline{\mathbf{x}})^T\mathbf{x})/exp(-\mathbf{c}(\overline{\mathbf{x}})^T\overline{\mathbf{x}}) &\geq f(\mathbf{y}|\mathbf{x}) \\
&\Updownarrow \\
\frac{MGF(c(\overline{\mathbf{X}}))h(\overline{\mathbf{X}})}{exp(-c(\overline{\mathbf{X}})^T\overline{\mathbf{X}})f(\mathbf{y})}\frac{exp(-\mathbf{c}(\overline{\mathbf{x}})^T\mathbf{x})\pi(\mathbf{X})}{MGF(c(\overline{\mathbf{X}}))} &\geq \frac{f(\mathbf{y}|\mathbf{x})\pi(\mathbf{x})}{f(\mathbf{y})} \\
&\Updownarrow \\
\frac{MGF(c(\overline{\mathbf{X}}))h(\overline{\mathbf{X}})}{exp(-c(\overline{\mathbf{X}})^T\overline{\mathbf{X}})f(\mathbf{y})}q(\mathbf{x}) &\geq \pi(\mathbf{x}|\mathbf{y})
\end{aligned}
$$

Q.E.D.

When $h(.)$ is chosen to coincide with $f(\mathbf{y}|.)$, the above expression holds with equality at point $\overline{\mathbf{x}}$. A multiple of a Generalized likelihood subgradient density thus bounds the posterior density from above and moreover obtains a tangency at the point $\overline{\mathbf{x}}$ if it is a Likelihood subgradient density. The multiple of the Likelihood Subgradient density is therefore a good approximation of the posterior density for points close to $\overline{\mathbf{x}}$. It is, however, a poorer approximation for points further from the the point $\overline{\mathbf{x}}$. How quickly the approximation becomes

poorer depends on the strength of the prior relative to the amount of information contained in the data. The approximation remains close far from $\overline{\mathbf{x}}$ if the prior is strong, but becomes poor quickly if the prior is weak.

In the following, we show how the use of mixtures of generalized likelihood subgradient densities will allow us to improve on the approximation of the posterior that results from a single Generalize Likelihood subgradient density. We define restricted Generalized Likelihood subgradient densities as follows.

Definition 3. *A probability density $\tilde{q}(.)$ is a restricted Generalized Likelihood-Subgradient density if there exists a Generalized Likelihood Subgradient density $q(.)$ and a subset A of X such that*

$$\tilde{q}(x) = \begin{cases} \frac{q(x)}{\int_{a \in A} q(a)dx} & \text{if } x \in A \\ 0 & \text{otherwise.} \end{cases}$$

Our improved approximation for the posterior density can now be constructed using a partition of X into K subsets $A_1, A_2, \ldots, A_K$ and associated restricted Generalized-Likelihood subgradient densities. We define a mixture Generalized Likelihood-Subgradient density as follows.

Definition 4. *A probability density $r(.)$ is a mixture Generalized Likelihood-Subgradient density if there exists a finite partition of X into K subsets $A_1, A_2, \ldots, A_K$, a bounding function h as above, and an associated set of restricted likelihood subgradient densities $\tilde{q}_1, \tilde{q}_2, \ldots, \tilde{q}_K$ such that*

$$\forall \mathbf{x} \in \mathbf{X} : r(\mathbf{x}) = \sum_{k=1}^{K} p_k \tilde{q}_k(\mathbf{x})$$

where

$$
\begin{aligned}
p_l &:= (\tfrac{MGF(c(\overline{\boldsymbol{x}}_l))h(\overline{\boldsymbol{x}}_l)}{exp(-c(\overline{\boldsymbol{x}}_l)^T\overline{\boldsymbol{x}}_l)f(\mathbf{y})}\int_{x\in A_l} q_l(\mathbf{x})dx)/(\sum_{k=1}^K \tfrac{MGF(c(\overline{\boldsymbol{x}}_k))h(\overline{\boldsymbol{x}}_k)}{exp(-c(\overline{\boldsymbol{x}}_k)^T\overline{\boldsymbol{x}}_k)f(\mathbf{y})}\int_{x\in A_k} q_k(\mathbf{x})dx) \\
&= (\tfrac{MGF(c(\overline{\boldsymbol{x}}_l))h(\overline{\boldsymbol{x}}_l)}{exp(-c(\overline{\boldsymbol{x}}_l)^T\overline{\boldsymbol{x}}_l)}\int_{x\in A_l} q_l(\mathbf{x})dx)/(\sum_{k=1}^K \tfrac{MGF(c(\overline{\boldsymbol{x}}_k))h(\overline{\boldsymbol{x}}_k)}{exp(-c(\overline{\boldsymbol{x}}_k)^T\overline{\boldsymbol{x}}_k)}\int_{x\in A_k} q_k(\mathbf{x})dx).
\end{aligned}
$$

Our main theorem now shows that such Likelihood-Subgradient densities can be used with accept-reject procedures in order to generate a sample from the corresponding posterior density.

Theorem 1. *Let $r(.)$ be a mixture Generalized Likelihood-Subgradient probability density for a posterior distribution $\pi(.|\boldsymbol{y})$, defined on a set $\boldsymbol{X}$, with prior distribution $\pi(.)$ and likelihood function $f(\boldsymbol{y}|.)$. Let $\boldsymbol{Z}$ denote a random vector with probability density function $r(.)$ and let U be a random variable with a uniform distribution on the interval $I := [0,1]$. If $U(i) \leq exp(-\boldsymbol{c}(\overline{\boldsymbol{x}}_{l(\mathbf{Z}(\mathbf{x}))})\overline{\boldsymbol{x}}_{l(\mathbf{Z}(\mathbf{x}))})f(\boldsymbol{y}|\boldsymbol{Z(x)})/exp(-\boldsymbol{c}(\overline{\boldsymbol{x}}_{l(\mathbf{Z}(\mathbf{x}))})\boldsymbol{Z(x)})h(\overline{\boldsymbol{x}}_{l(\mathbf{Z}(\mathbf{x}))})$, then $\boldsymbol{Z(x)}$ has the probability density function $\pi(.|\boldsymbol{y})$. Moreover, the expected number of draws for $\boldsymbol{Z}$ and U required before $U(i) \leq exp(-\boldsymbol{c}(\overline{\boldsymbol{x}}_{l(\mathbf{Z}(\mathbf{x}))})\overline{\boldsymbol{x}}_{l(\mathbf{Z}(\mathbf{x}))})f(\boldsymbol{y}|\boldsymbol{Z(x)})/exp(-\boldsymbol{c}(\overline{\boldsymbol{x}}_{l(\mathbf{Z}(\mathbf{x}))})\boldsymbol{Z(x)})h(\overline{\boldsymbol{x}}_{l(\mathbf{Z}(\mathbf{x}))})$ is given by*

$$
\begin{aligned}
a &:= \sum_{k=1}^K \tfrac{MGF(c(\overline{\boldsymbol{x}}_k))h(\overline{\boldsymbol{x}}_k)}{exp(-c(\overline{\boldsymbol{x}}_k)^T\overline{\boldsymbol{x}}_k)f(\mathbf{y})}\int_{x\in A_k} q_k(\mathbf{x})dx \\
&= (\sum_{l=1}^K \int_{x\in A_l} q_l(\mathbf{x})dx)(\sum_{k=1}^K(\tfrac{MGF(c(\overline{\boldsymbol{x}}_k))f(\mathbf{y}|\overline{\boldsymbol{x}}_k)}{exp(-c(\overline{\boldsymbol{x}}_k)^T\overline{\boldsymbol{x}}_k)f(\mathbf{y})})(\tfrac{\int_{x\in A_k} q_k(\mathbf{x})dx}{\sum_{l=1}^K \int_{x\in A_l} q_l(\mathbf{x})dx})).
\end{aligned}
$$

The reason this result holds is that a multiple of a Mixture Generalized likelihood subgradient density bounds the posterior density from above. If the bounding function h coincides with the likelihood function, then the multiple of the Mixture Generalized likelihood subgradient density coincides with the posterior density at each or the K points at which the underlying Likelihood subgradient densities obtain their tangencies. As a result, Mixture Generalized likelihood subgradient densities are capable of providing significantly better approximations

to the posterior density than those obtained through the use of a single Generalized likelihood subgradient density.

3.2 Mixtures of Restricted Multivariate Normal Likelihood Subgradient Densities

Likelihood subgradient densities are constructed from prior densities by subtracting a linear term. In the case of a multivariate normal prior, the corresponding likelihood subgradient densities are always multivariate normal. Hence simple sampling procedures are available for such likelihood subgradient densities. Restricted likelihood subgradient densities in general may be challenging to sample from. This can be the case even if simple procedures are available for the underlying likelihood subgradient density (e.g., if it is multivariate normal). In this section, we show that there is an important subclass of restricted multivariate normal densities for which simple sampling procedures are available. For mixtures of restricted multivariate normal densities from within that class, there are therefore also simple sampling procedures available.

We first note the following important fact.

Fact 8 (Anderson, 1958, p.19)**.** *Let* $\mathbf{\Sigma}$ *be the variance covariance matrix of a multivariate normal distribution. Then there exists uniquely a non-singular lower triangular matrix* $\mathbf{d}$ *such that* $\mathbf{\Sigma} = \mathbf{d}\mathbf{d}^T$.

The matrix $\mathbf{d}$ is known as the Cholesky decomposition. This fact now allows us to consider the properties of a transformed variable $\tilde{\beta} := \mathbf{d}^{-1}\beta$.

Fact 9. *Suppose* β *has a multivariate normal distribution with mean vector* μ*, variance-*

covariance matrix Σ*, and Cholesky decomposition matrix d. Then* $\tilde{\beta} := \mathbf{d}^{-1}\beta$ *is multivariate normal with mean vector* $\tilde{\mu} = \mathbf{d}^{-1}\mu$*. Moreover, the variance-covariance matrix for* $\tilde{\beta}$ *is the identity matrix.*

Let us now consider restricted multivariate normal distributions for which

$$A := \{\beta \in \mathbf{R}^m | \underline{a} \leq \mathbf{d}^{-1}\beta \leq \overline{a}\}$$

where we allow for the possibility that some elements of $\underline{a}$ are $-\infty$ and some elements of $\overline{a}$ are $+\infty$. The independence of each element of vector $\tilde{\beta}$ makes it possible to use the following procedure to generate a sample from such a restricted multivariate normal density.

(i) Generate each element of the vector $\tilde{\beta}$ from a restricted normal, where the jth element is generated from a restricted normal with mean parameter $\tilde{\mu}_j$, variance parameter 1, lower bound $\underline{a}_j$ and upper bound $\overline{a}_j$.
(ii) Compute $\beta = \mathbf{d}\tilde{\beta}$.

We note that the first step in this generation procedure can be implemented using the inverse transform method. Efficient implementation requires fast functions capable of computing the Cummulative and Inverse Cummulative Density functions of a standard normal. We also note that this makes generation from mixture likelihood subgradient densities based on restricted likelihood subgradient densities of this type straightforward.

Denote by $F\underline{a}_j$ and $F\overline{a}_j$ the cummulative density of the jth element of vector $\tilde{\beta}$ at $\underline{a}_j$ and $\overline{a}_j$. In the notation of our earlier subsection, we then have

$$\int_{x \in A} q(x)dx = \prod_{j=1}^{m} (F\overline{a}_j - F\underline{a}_j).$$

In Nygren (2003), we also noted that if the prior is multivariate normal, then

$$MGF(\mathbf{c}(\overline{\mathbf{x}})) = exp(-\mathbf{c}(\overline{\mathbf{x}})^T\mu_0 + \frac{1}{2}\mathbf{c}^T(\overline{\mathbf{x}})\Sigma_0\mathbf{c}(\overline{\mathbf{x}})).$$

For the kinds of restricted multivariate normal densities considered here, it is thus possible to compute both the probability of each restricted multivariate normal as well as the following numerically

$$\begin{aligned} g &:= \ln(\textstyle\sum_{k=1}^{K} \frac{MGF(c(\overline{\mathbf{X}_k}))h(\overline{\mathbf{X}_k})}{exp(-c(\overline{\mathbf{X}_k})^T\overline{\mathbf{X}_k})} \int_{x\in A_k} q_k(\mathbf{x})dx) \\ &= \ln(a) + \ln(f(\mathbf{y})). \end{aligned}$$

This last expression can be used in order to minimize the expected number of draws before acceptance over a class of Mixture Likelihood subgradient densities. A brief discussion of how this can be done is contained in the next section.

4 Generalized Linear Models with Log-Concave Likelihood Functions and Multivariate Normal Priors

4.1 A Univariate Model with Poisson Data

We first illustrate the ideas of the previous section with a univariate model. The model we consider is the following:

$$\beta \sim Normal(\mu, \sigma_0^2)$$

$$y_i \sim Poisson(exp(\beta)), i = 1, \ldots, N$$

where $\mathbf{y}$ represents an observed data vector. We note that this is a non-conjugate model.

Before proceeding to our estimation procedure, we will first examine the likelihood function and the corresponding gradient. The log-likelihood function for this model is given by

$$\begin{aligned} LL(\beta) &:= \sum_{i=1}^{N}(y_i\beta - exp(\beta) - \ln(y_i!)) \\ &= N(\overline{y}\beta - exp(\beta)) - \sum_{i=1}^{N}\ln(y_i!). \end{aligned}$$

where $\overline{y} = (1/N)\sum_{i=1}^{N} y_i$. The gradient of the log-likelihood function is given by

$$\frac{\partial LL(\beta)}{\partial \beta} = N(\overline{y} - exp(\beta)).$$

The unique subgradient is given by

$$c(\beta) = N(exp(\beta) - \overline{y}).$$

If we pick a point $\overline{\beta} \in \mathbf{R}$, then there is a unique likelihood subgradient density at $\overline{\beta}$. More specifically, it is the following density

$$Normal(\mu - \sigma_0^2 N(exp(\overline{\beta}) - \overline{y}), \sigma_0^2).$$

If used by itself in an accept-reject procedure, a draw from this density should be accepted whenever

$$\ln(U) \leq N(\exp(\overline{\beta})\beta - \exp(\beta)) - N(\exp(\overline{\beta})\overline{\beta} - \exp(\overline{\beta})).$$

If β had the above normal density, then $\tilde{\beta}$ would be the following normal

$$Normal((1/\sigma_0)(\mu - \sigma_0^2 N(exp(\overline{\beta}) - \overline{y})), 1).$$

If this density was restricted to values between $\underline{a}$ and $\overline{a}$, then the cummulative density at the lower and upper bounds respectively would correspond to the cummulative density of a

standard normal density at $\underline{a}-(1/\sigma_0)(\mu-\sigma_0^2 N(exp(\overline{\beta})-\overline{y}))$ and $\overline{a}-(1/\sigma_0)(\mu-\sigma_0^2 N(exp(\overline{\beta})-\overline{y}))$.

Suppose now that we wanted to simulate from the posterior density using a mixture of three restricted likelihood subgradient densities at the points $\sigma_0\tilde{\beta}_1$, $\sigma_0\tilde{\beta}_2$, and $\sigma_0\tilde{\beta}_3$. We could then partition the real number line into three segment $(-\infty, \overline{a}_1], (\underline{a}_2, \overline{a}_2], (\underline{a}_3, +\infty]$, where $\overline{a}_1 = \underline{a}_2$ and $\overline{a}_2 = \underline{a}_3$. Moreover, we could position $\tilde{\beta}_1$, $\tilde{\beta}_2$, $\tilde{\beta}_3$, $\overline{a}_1$, and $\overline{a}_2$ so as to minimize g in section 3.2.

4.2 The Poisson Regression Model

We now consider the following Poisson Regression model

$$\beta \sim Multivariate - Normal(\mu, \mathbf{\Sigma_0})$$

$$y_i \sim Poisson(exp(\mathbf{X_i}\beta)), i = 1, \ldots, N.$$

where $\mathbf{y}$ is a vector of observed data. For the Poisson regression model, the log likelihood function is

$$\ln L = \mathbf{y}^T\mathbf{X}\beta - \sum_{i=1}^{N}[exp(\mathbf{X}_i\beta) + \ln y_i!]$$

which is known to be a concave function. The gradient for the log-likelihood function is

$$\frac{\partial \ln L}{\partial \beta} = \mathbf{X}^T(\mathbf{y} - exp(\mathbf{X}\beta))$$

and the unique subgradient is given by

$$\mathbf{c}(\beta) = \mathbf{X}^T(exp(\mathbf{X}\beta) - \mathbf{y}).$$

If we pick a point $\overline{\beta} \in \mathbf{R}^m$, then there is a unique likelihood subgradient density at $\overline{\beta}$. More specifically, it is the following density

$$Multivariate - Normal(\mu - \Sigma_0 \mathbf{X}^T (exp(\mathbf{X}\beta) - \mathbf{y}), \Sigma_0).$$

If used by itself in an accept-reject procedure, a draw from this density should be accepted whenever

$$\ln(U) \quad \leq \quad [exp(\mathbf{X}_i\overline{\beta})^T \mathbf{X}\beta - \textstyle\sum_{i=1}^N exp(\mathbf{X}_i\beta)] - [exp(\mathbf{X}_i\overline{\beta})^T \mathbf{X}\overline{\beta} - \textstyle\sum_{i=1}^N exp(\mathbf{X}_i\overline{\beta})].$$

If β had the above Multivariate-Normal density, then $\tilde{\beta}$ would have the following Multivariate-Normal density

$$Multivariate - Normal(\mathbf{d}^{-1}(\mu - \Sigma_0 \mathbf{X}^T (exp(\mathbf{X}\beta) - \mathbf{y})), \mathbf{I}).$$

The following is a possible approach to sampling for this density:

(i) Partition each dimension for $\tilde{\beta}$ as in the univariate case. In other words, dimension j is partitioned into l_j intervals $(-\infty, \overline{a}_{1,j}), (\underline{a}_{2,j}, \overline{a}_{2,j}), \ldots, (\underline{a}_{l_j,j}, +\infty)$ with corresponding points $\tilde{\beta}^*_{1,j}, \tilde{\beta}^*_{2,j}, \ldots, \tilde{\beta}^*_{l_j,j}$ where $\overline{a}_{k,j} = \underline{a}_{k+1,j}$.

(ii) Form a partition of the overall space by crossing the partitions for the individual dimensions, yielding a partition with $l := \prod_{j=1}^m l_j$ elements. Associated with the kth element of this partition we would then have a unique corresponding vector $\tilde{\beta}^*_k$ and an associated restricted likelihood subgradient density at $\mathbf{d}\tilde{\beta}^*_k$.

(iii) Construct the mixture likelihood subgradient density from these restricted likelihood subgradient densities.

(iv) Use the mixture likelihood subgradient density with the corresponding accept-reject procedure in order to generate a sample from the posterior density. The accept-reject procedure can make use of the squeeze method implementation used by Nygren (2003) in the Poisson regression example.

We note that the partitions and points in (i) can be positioned so as to minimize g in section 3.2.

4.3 Discussion

In this section, we illustrated how mixture likelihood subgradient densities can be used in order to generate samples from a univariate model with Poisson data and from a Poisson regression model. The same methods are generally applicable to any model with a multivariate normal prior and a log-concave likelihood function. These type of models include the logit, multinomial-logit and conditional logit models.

5 Concluding Discussion

This paper has developed efficient exact sampling procedures for Bayesian Generalized linear models. The sampling procedures for models with normal and multivariate data should be easy to implement. Those models also allow for more general prior specifications than those used in this paper. The sampling procedures for models with multivariate normal priors and log-concave likelihood functions make use of mixture likelihood subgradient densities. These procedures are more complex to implement and requires choosing the number of restricted likelihood subgradient densities to employ as well as careful positioning of the restricted

likelihood subgradient densities. The latter should, however, be a fairly straightforward optimization problem that can make use of the function g derived in section 3.2. Studies on the number of restricted likelihood subgradient densities needed for reasonable acceptance rates in practice would be of great value here.

Our hope is that the procedures presented here will lead to the development of easy to use Bayesian software that do not require users to assess the convergence of Markov Chains. Such convergence diagnostics are challenging to perform even for sophisticated users and the need for them presents a major obstacle to the use of Bayesian methods among less sophisticated users. Hopefully, this paper will contribute to making basic Bayesian data analysis as simple as traditional frequentist data analysis. The present author is currently working on developing an excel add-in tool that implements the models presented in this paper. This tool uses compiled C++ code in the background. The tool and corresponding C++ code will be made available through the present author's web page.

The development of efficient sampling procedures for more general models than those presented here would be highly desirable. A class of models that would be natural to consider are hierarchical random effects models. It may be possible to develop efficient exact sampling procedures for such models using similar approaches to those employed here. The results presented here should also make it possible to implement such models using two block Gibbs samplers. Hence the development of tight bounds on the rate of convergence for two block Gibbs samplers would also be particularly useful. Promising results in this direction already exists (Hobert and Geyer 1998, Jones and Hobert 2003). Results in these areas may enable the development of Bayesian softwares that can handle complex models without requiring convergence diagnostics on the part of the user.

6 Proofs:

Proof of Theorem 1: Random variables $\mathbf{Z}$ and U have the joint density function defined by

$$v(U(i), \mathbf{Z(x)}) = r(x) = p_{l(\mathbf{Z(x)})}\tilde{q}(\mathbf{Z(x)}),\ U(i) \in [0,1],\ \mathbf{Z(x)} \in \mathbf{X}.$$

where $\mathbf{Z}(\mathbf{x})$ is an element of $A_{l(\mathbf{Z(x)})}$. We need to show that the conditional density function for $\mathbf{Z}$

$$v_{\mathbf{Z}}(.|U(i) \leq \frac{exp(-\mathbf{c}(\overline{\mathbf{x}}_{l(\mathbf{Z(x)})})\overline{\mathbf{x}}_{l(\mathbf{Z(x)})})}{h(\overline{\mathbf{x}}_{l(\mathbf{Z(x)})})}\frac{f(\mathbf{y}|\mathbf{Z(x)})}{exp(-\mathbf{c}(\overline{\mathbf{x}}_{l(\mathbf{Z(x)})})\mathbf{Z(x)})})$$

satisfies the property that

$$v_{\mathbf{Z}}(.|U(i) \leq \frac{exp(-\mathbf{c}(\overline{\mathbf{x}}_{l(\mathbf{Z(x)})})\overline{\mathbf{x}}_{l(\mathbf{Z(x)})})}{h(\overline{\mathbf{x}}_{l(\mathbf{Z(x)})})}\frac{f(\mathbf{y}|\mathbf{Z(x)})}{exp(-\mathbf{c}(\overline{\mathbf{x}}_{l(\mathbf{Z(x)})})\mathbf{Z(x)})}) = \pi(\mathbf{Z(x)}|\mathbf{y})\ \forall \mathbf{x} \in \mathbf{X}.$$

Set $d(\mathbf{x}) = exp(-\mathbf{c}(\overline{\mathbf{x}}_{l(\mathbf{x})})^T\overline{\mathbf{x}}_{l(\mathbf{x})})f(\mathbf{y}|\mathbf{x})/exp(-\mathbf{c}(\overline{\mathbf{x}}_{l(\mathbf{x})})^T\mathbf{x})h(\overline{\mathbf{x}}_{l(\mathbf{x})})$. Since $h(.)$ bounds $f(\mathbf{y}|.)$ from above and $\mathbf{c}(\overline{\mathbf{x}}_{l(\mathbf{x})})$ is a subgradient for $-\ln(h(.))$ at $\overline{\mathbf{x}}_{l(\mathbf{x})}$, it follows that for every $\mathbf{x} \in \mathbf{X}$,

$$\begin{aligned}
\ln(d(\mathbf{x})) &= -\mathbf{c}(\overline{\mathbf{x}}_{l(\mathbf{x})})^T\overline{\mathbf{x}}_{l(\mathbf{x})} - \ln(h(\overline{\mathbf{x}}_{l(\mathbf{x})})) + \ln(f(\mathbf{y}|\mathbf{x})) + \mathbf{c}(\overline{\mathbf{x}}_{l(\mathbf{x})})^T\mathbf{x} \\
&\leq -\mathbf{c}(\overline{\mathbf{x}}_{l(\mathbf{x})})^T\overline{\mathbf{x}}_{l(\mathbf{x})} - \ln(h(\overline{\mathbf{x}}_{l(\mathbf{x})})) + \ln(h(\mathbf{x})) + \mathbf{c}(\overline{\mathbf{x}}_{l(\mathbf{x})})^T\mathbf{x} \\
&\leq -\mathbf{c}(\overline{\mathbf{x}}_{l(\mathbf{x})})^T\overline{\mathbf{x}}_{l(\mathbf{x})} - \ln(h(\overline{\mathbf{x}}_{l(\mathbf{x})})) + \\
&\quad [\ln(h(\overline{\mathbf{x}}_{l(\mathbf{x})})) - \mathbf{c}(\overline{\mathbf{x}}_{l(\mathbf{x})})^T(\mathbf{x} - \overline{\mathbf{x}}_{l(\mathbf{x})})] + \mathbf{c}(\overline{\mathbf{x}}_{l(\mathbf{x})})^T\mathbf{x} \\
&= 0.
\end{aligned}$$

Hence $d(\mathbf{x}) \leq 1$ for every $\mathbf{x} \in \mathbf{X}$. To simplify our notation below, we define $\mathbf{x}^* := \mathbf{Z(x)}$. It then follows that

$$
\begin{aligned}
v_{\mathbf{Z}}(\mathbf{x}^*|U(i) \leq d(\mathbf{x}^*)) &= (\textstyle\int_0^{d(\mathbf{X}^*)} p_{l(\mathbf{X}^*)}\tilde{q}_{l(\mathbf{X}^*)}(\mathbf{x}^*)dU)/ \\
&\quad (\textstyle\int_{\mathbf{X}\in\mathbf{X}}\int_0^{d(\mathbf{X})} p_{l(\mathbf{X})}\tilde{q}_{l(\mathbf{X})}(\mathbf{x})dU) \\
&= ((\tfrac{MGF(c(\overline{\mathbf{X}}_{l(\mathbf{X}^*)}))h(\overline{\mathbf{X}}_{l(\mathbf{X}^*)})}{exp(-c(\overline{\mathbf{X}}_{l(\mathbf{X}^*)})^T\overline{\mathbf{X}}_{l(\mathbf{X}^*)})f(\mathbf{y})}) \textstyle\int_0^{d(\mathbf{X}^*)} q_{l(\mathbf{X}^*)}(\mathbf{x}^*)dU)/ \\
&\quad (\textstyle\int_{\mathbf{X}\in\mathbf{X}}(\tfrac{MGF(c(\overline{\mathbf{X}}_{l(\mathbf{X})}))h(\overline{\mathbf{X}}_{l(\mathbf{X})})}{exp(-c(\overline{\mathbf{X}}_{l(\mathbf{X})})^T\overline{\mathbf{X}}_{l(\mathbf{X})})f(\mathbf{y})}) \textstyle\int_0^{d(\mathbf{X})} q_{l(\mathbf{X})}(\mathbf{x})dU) \\
&= (\tfrac{MGF(c(\overline{\mathbf{X}}_{l(\mathbf{X}^*)}))h(\overline{\mathbf{X}}_{l(\mathbf{X}^*)})}{exp(-c(\overline{\mathbf{X}}_{l(\mathbf{X}^*)})^T\overline{\mathbf{X}}_{l(\mathbf{X}^*)})f(\mathbf{y})}) \textstyle\int_0^{d(\mathbf{X}^*)} \tfrac{exp(-\mathbf{c}(\overline{\mathbf{X}}_{l(\mathbf{X}^*)})^T\mathbf{X}^*)\pi(\mathbf{X}^*)}{MGF(\mathbf{c}(\overline{\mathbf{X}}_{l(\mathbf{X}^*)}))}dU/ \\
&\quad \textstyle\int_{\mathbf{X}\in\mathbf{X}}(\tfrac{MGF(c(\overline{\mathbf{X}}_{l(\mathbf{X})}))h(\overline{\mathbf{X}}_{l(\mathbf{X})})}{exp(-c(\overline{\mathbf{X}}_{l(\mathbf{X})})^T\overline{\mathbf{X}}_{l(\mathbf{X})})f(\mathbf{y})}) \textstyle\int_0^{d(\mathbf{X})} \tfrac{exp(-\mathbf{c}(\overline{\mathbf{X}}_{l(\mathbf{X})})^T\mathbf{X})\pi(\mathbf{X})}{MGF(\mathbf{c}(\overline{\mathbf{X}}_{l(\mathbf{X})}))}dUdX \\
&= (\tfrac{h(\overline{\mathbf{X}}_{l(\mathbf{X}^*)})}{exp(-c(\overline{\mathbf{X}}_{l(\mathbf{X}^*)})^T\overline{\mathbf{X}}_{l(\mathbf{X}^*)})f(\mathbf{y})})d(\mathbf{x}^*)exp(-\mathbf{c}(\overline{\mathbf{X}}_{l(\mathbf{X}^*)})^T\mathbf{x}^*)\pi(\mathbf{x}^*)/ \\
&\quad \textstyle\int_{\mathbf{X}\in\mathbf{X}}(\tfrac{h(\overline{\mathbf{X}}_{l(\mathbf{X})})}{exp(-c(\overline{\mathbf{X}}_{l(\mathbf{X})})^T\overline{\mathbf{X}}_{l(\mathbf{X})})f(\mathbf{y})})d(\mathbf{x})exp(-\mathbf{c}(\overline{\mathbf{X}}_{l(\mathbf{X})})^T\mathbf{x})\pi(\mathbf{x})d\mathbf{x} \\
&= (\tfrac{h(\overline{\mathbf{X}}_{l(\mathbf{X}^*)})}{exp(-c(\overline{\mathbf{X}}_{l(\mathbf{X}^*)})^T\overline{\mathbf{X}}_{l(\mathbf{X}^*)})})(\tfrac{exp(-\mathbf{c}(\overline{\mathbf{X}}_{l(\mathbf{X}^*)})^T\overline{\mathbf{X}}_{l(\mathbf{X}^*)})f(\mathbf{y}|\mathbf{X}^*)}{h(\overline{\mathbf{X}}_{l(\mathbf{X}^*)})exp(-\mathbf{c}(\overline{\mathbf{X}}_{l(\mathbf{X}^*)})^T\mathbf{X}^*)})(\tfrac{exp(-\mathbf{c}(\overline{\mathbf{X}}_{l(\mathbf{X}^*)})^T\mathbf{X}^*)\pi(\mathbf{X}^*)}{f(y)})/ \\
&\quad \textstyle\int_{\mathbf{X}\in\mathbf{X}}(\tfrac{h(\overline{\mathbf{X}}_{l(\mathbf{X})})}{exp(-c(\overline{\mathbf{X}}_{l(\mathbf{X})})^T\overline{\mathbf{X}}_{l(\mathbf{X})})})(\tfrac{exp(-\mathbf{c}(\overline{\mathbf{X}}_{l(\mathbf{X})})^T\overline{\mathbf{X}}_{l(\mathbf{X})})f(\mathbf{y}|\mathbf{X})}{h(\overline{\mathbf{X}}_{l(\mathbf{X})})exp(-\mathbf{c}(\overline{\mathbf{X}}_{l(\mathbf{X})})^T\mathbf{X})})(\tfrac{exp(-\mathbf{c}(\overline{\mathbf{X}}_{l(\mathbf{X})})^T\mathbf{X})\pi(\mathbf{X})}{f(y)})d\mathbf{x} \\
&= \tfrac{f(y|\mathbf{X}^*)\pi(\mathbf{X}^*)/f(\mathbf{y})}{\textstyle\int_{\mathbf{X}\in\mathbf{X}} f(\mathbf{y}|\mathbf{X})\pi(\mathbf{X})/f(\mathbf{y})d\mathbf{X}} \\
&= f(\mathbf{y}|\mathbf{x}^*)\pi(\mathbf{x}^*)/f(\mathbf{y}) \\
&= \pi(\mathbf{x}^*|\mathbf{y})
\end{aligned}
$$

establishing the equivalence of the two distributions. To see that the moreover statement holds, we first note that the probability of generating a sample from $\pi(.|\mathbf{y})$ using a single draw is given by

$$
\begin{aligned}
\int_{\mathbf{X}\in\mathbf{X}}\int_0^{d(\mathbf{X})} r(\mathbf{x})dud\mathbf{x} &= \int_{\mathbf{X}\in\mathbf{X}}\int_0^{d(\mathbf{X})} p_{l(\mathbf{x})}\tilde{q}_{l(\mathbf{x})}(\mathbf{x})dud\mathbf{x} \\
&= (1/\sum_{k=1}^{K}\frac{MGF(c(\overline{\mathbf{X}_k}))h(\overline{\mathbf{X}_k})}{exp(-c(\overline{\mathbf{X}_k})^T\overline{\mathbf{X}_k})f(\mathbf{y})}\int_{x\in A_k} q_k(\mathbf{x})dx)* \\
&\quad (\int_{\mathbf{X}\in\mathbf{X}}\int_0^{d(\mathbf{X})}\frac{MGF(c(\overline{\mathbf{x}}_{l(\mathbf{x})}))h(\overline{\mathbf{x}}_{l(\mathbf{x})})}{exp(-c(\overline{\mathbf{x}}_{l(\mathbf{x})})\overline{\mathbf{x}}_{l(\mathbf{x})})f(\mathbf{y})}q_{l(\mathbf{x})}(\mathbf{x})dud\mathbf{x}) \\
&= (1/\sum_{k=1}^{K}\frac{MGF(c(\overline{\mathbf{X}_k}))h(\overline{\mathbf{X}_k})}{exp(-c(\overline{\mathbf{X}_k})^T\overline{\mathbf{X}_k})f(\mathbf{y})}\int_{x\in A_k} q_k(\mathbf{x})dx)* \\
&\quad (\int_{\mathbf{X}\in\mathbf{X}} d(\mathbf{x})\frac{MGF(c(\overline{\mathbf{x}}_{l(\mathbf{x})}))h(\overline{\mathbf{x}}_{l(\mathbf{x})})}{exp(-c(\overline{\mathbf{x}}_{l(\mathbf{x})})\overline{\mathbf{x}}_{l(\mathbf{x})})f(\mathbf{y})}\frac{exp(-\mathbf{c}(\overline{\mathbf{X}}_{l(\mathbf{x})})^T\mathbf{x})\pi(\mathbf{X})}{MGF(\mathbf{c}(\overline{\mathbf{X}}_{l(\mathbf{x})}))}d\mathbf{x}) \\
&= (1/\sum_{k=1}^{K}\frac{MGF(c(\overline{\mathbf{X}_k}))h(\overline{\mathbf{X}_k})}{exp(-c(\overline{\mathbf{X}_k})^T\overline{\mathbf{X}_k})f(\mathbf{y})}\int_{x\in A_k} q_k(\mathbf{x})dx)* \\
&\quad (\int_{\mathbf{X}\in\mathbf{X}}(\frac{exp(-\mathbf{c}(\overline{\mathbf{X}}_{l(\mathbf{x})})^T\overline{\mathbf{X}}_{l(\mathbf{x})})f(\mathbf{y}|\mathbf{X})}{h(\overline{\mathbf{X}}_{l(\mathbf{x})})exp(-\mathbf{c}(\overline{\mathbf{X}}_{l(\mathbf{x})})^T\mathbf{X})})(\frac{h(\overline{\mathbf{x}}_{l(\mathbf{x})})exp(-\mathbf{c}(\overline{\mathbf{X}}_{l(\mathbf{x})})^T\mathbf{X})\pi(\mathbf{X})}{exp(-c(\overline{\mathbf{x}}_{l(\mathbf{x})})\overline{\mathbf{x}}_{l(\mathbf{x})})f(\mathbf{y})})d\mathbf{x}) \\
&= (1/\sum_{k=1}^{K}\frac{MGF(c(\overline{\mathbf{X}_k}))h(\overline{\mathbf{X}_k})}{exp(-c(\overline{\mathbf{X}_k})^T\overline{\mathbf{X}_k})f(\mathbf{y})}\int_{x\in A_k} q_k(\mathbf{x})dx)* \\
&\quad (\int_{\mathbf{X}\in\mathbf{X}} f(\mathbf{y}|\mathbf{x})\pi(\mathbf{x})/f(\mathbf{y})d\mathbf{x}) \\
&= (1/\sum_{k=1}^{K}\frac{MGF(c(\overline{\mathbf{X}_k}))h(\overline{\mathbf{X}_k})}{exp(-c(\overline{\mathbf{X}_k})^T\overline{\mathbf{X}_k})f(\mathbf{y})}\int_{x\in A_k} q_k(\mathbf{x})dx).
\end{aligned}
$$

Given the independence of the draws, it follows from the properties of the geometric distribution that the expected number of draws before success is given by

$$
\begin{aligned}
a^* &= (1/(1/\sum_{k=1}^{K}\frac{MGF(c(\overline{\mathbf{X}_k}))h(\overline{\mathbf{X}_k})}{exp(-c(\overline{\mathbf{X}_k})^T\overline{\mathbf{X}_k})f(\mathbf{y})}\int_{x\in A_k} q_k(\mathbf{x})dx)) \\
&= \sum_{k=1}^{K}\frac{MGF(c(\overline{\mathbf{X}_k}))h(\overline{\mathbf{X}_k})}{exp(-c(\overline{\mathbf{X}_k})^T\overline{\mathbf{X}_k})f(\mathbf{y})}\int_{x\in A_k} q_k(\mathbf{x})dx \\
&= (\sum_{l=1}^{K}\int_{x\in A_l} q_l(\mathbf{x})dx)(\sum_{k=1}^{K}(\frac{MGF(c(\overline{\mathbf{X}_k}))f(\mathbf{y}|\overline{\mathbf{X}_k})}{exp(-c(\overline{\mathbf{X}_k})^T\overline{\mathbf{X}_k})f(\mathbf{y})})(\frac{\int_{x\in A_k} q_k(\mathbf{x})dx}{\sum_{l=1}^{K}\int_{x\in A_l} q_l(\mathbf{x})dx})).
\end{aligned}
$$

Q.E.D.

7 References

Anderson, T.W. (1958). *An introduction to Multivariate Statistical Analysis*, Wiley, New York.

Bayes, T. (1763), "An essay towards solving a problem in the doctrine of chances," *Philosophical Transactions of the Royal Society*, 330-418. Reprinted, with biographical note by G.A. Barnard, *Biometrica*, 45, 293-315 (1958).

Breslow, N. E., and Clayton, D. G. (1993), "Approximate inference in generalized linear mixed models," *Journal of the American Statistical Association*, 88, 9-25.

Dellaportas, P., and Smith, A.F.M. (1993), "Bayesian Inference for generalized linear and proportional hazard models via gibbs sampling," *Applied Statistics*, 42, Issue 3, 443-459.

Eckhardt, R. (1987), "Stan Ulam, John von Neumann, and the Monte Carlo method," *Los Alamos Science*, Special Issue 15, 131-137.

Fishman, G.S. (1999), *Monte Carlo: Concepts, Algorithms, and Applications*, Springer-Verlag, New York, NY.

Gamerman, D. (1997), "Sampling from the posterior distribution in generalized linear mixed models," *Statistics and Computing*, 7, 57-68.

Gelfand, A.E., and Smith, A.F.M. (1990), "Sample-based approaches to calculating marginal densities," *Journal of the American Statistical Association*, 85, 398-409.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (2000), *Bayesian Data Analysis*, CRC Press LLC, Boca Raton, Florida.

Geman, S., and Geman, D. (1984), "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, PAMI-6, 721-740.

Gilks, W.R. (1992), "Derivative-free adaptive rejection sampling for Gibbs sampling," *Bayesian Statistics*, 4 (eds. Bernardo, J., Berger, J., Dawid, A.P., and Smith, A.F.M.), Oxford University Press.

Gilks, W.R. and Wild, P. (1992), "Adaptive rejection sampling for Gibbs sampling," *Applied Statistics*, 41, pp. 337-348.

Gilks, W.R., Best, N.G. and Tan, K.K.C. (1995), "Adaptive rejection Metropolis sampling," *Applied Statistics,* 44, pp. 455-472.

Hastings, W.K. (1970), "Monte Carlo sampling methods using Markov chains and their applications," *Biometrica*, 57, 92-109.

Hobert, J.P., and Geyer, C.J. (1998), "Geometric Ergodicity of Gibbs and Block Gibbs Samplers for a Hierarchical Random Effects Model," *Journal of Multivariate Analysis*, 67, 414-430.

Jones, G.L., and Hobert, J.P. (2003), "Sufficient Burn-in for Gibbs Samplers for a Hierarchical Random Effects Model," Accepted for Publication in *The Annals of Statistics.*

Laplace, P-S. (1785), "Memoire sur les formules qui sont fonctions de tres grands nombres," In *Memoires de l'Academie Royale des Sciences.*

—- (1810), "Memoire sur les formules qui sont fonctions de tres grands nombres et sur leurs applications aux probabilities," In *Memoires de l'Academie des Sciences de Paris.*

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. (1953), "Equations of state calculations by fast computing machines," *Journal of Chemical Physics*,

21, 1087-1092.

Metropolis, N., and Ulam, S. (1949), "The Monte Carlo Method," *Journal of the American Statistical Association*, 44(247), 335-341.

The Numerical Algorithms Group Ltd(2002), *The Nag C Library Manual, Mark 7*, Oxford, UK.

Nygren, K. (2003), "Exact Sampling from Posterior Densities Using Likelihood Subgradient Densities with Accept-Reject Procedures", submitted to *Journal or the American Statistical Association*.

Raiffa, H., and Schlaifer, R. (1961), *Applied Statistical Decision Theory*, Division of Research, Graduate School of Business Administration, Harvard University, Boston.

von Neumann, J. (1951), "Various techniques used in connection with random digits," *Monte Carlo Method*, Applied Mathematics Series 12, National Bureau of Standards, Washington, D.C.