# BAYESIAN MULTIVARIATE LINEAR REGRESSION WITH APPLICATION TO CHANGEPOINT MODELS IN HYDROMETEOROLOGICAL VARIABLES. PART I. MODEL DEVELOPMENT.

*Rapport de recherche No R-838* *Décembre 2005*

# Bayesian Multivariate Linear Regression with Application to Changepoint Models in Hydrometeorological Variables. Part I. Model Development

Jérôme J.J. Asselin[1], Taha B.M.J. Ouarda[2*], Seidou Ousmane

*(1) Centre Hospitalier de l'Université de Montréal, 3875 St-Urbain, 3e étage, Montréal (QC), Canada H2W 1V1*

*(2) INRS-ETE, Chair in Statistical Hydrology/Canada Research Chair on the Estimation of Hydrological Variables, 490 rue de la Couronne, Quebec (QC), Canada G1K 9A9*

**Research report R-838**

**Submitted to Water Resources Research**
**December 2005**

***Corresponding author:** Tel: (418) 654-3842
Fax: (418) 654-2600
Email: taha_ouarda@ete.inrs.ca

**Abstract**

Multivariate linear regression is one of the most popular modeling tools in hydrology and climate sciences for explaining the link between key variables. Piecewise linear regression is not always appropriate since the relationship may experiment sudden changes due to climatic, environmental or anthropogenic perturbations. To address this issue, a practical and general approach to the Bayesian analysis of the multivariate regression model is developed. The approach allows simultaneous single changepoint detection in a multivariate sample, and can account for missing data in the response variables and/or in the explicative variables. It also improves on the changepoint detection model of *Rasmussen* [2001] by allowing a more flexible thus more realistic prior specification for the existence of a change, the date of change as well as for the regression parameters. The estimation of all unknown parameters is achieved by Monte Carlo Markov Chain simulations. The paper also discusses the generalization of the approach to a broader class of switching models, including multiple changepoint models and segmented multivariate regression. Applications of the methodology are presented in a companion paper.

**Keywords:** Bayesian analysis, multivariate regression, changepoint, segmented multivariate regression, switching model, hydrology.

## 1. Introduction

Due to the growing evidence of climate change, the common assumption of stationarity of hydrologic phenomena no longer holds. Several recently published works point out shifts

or trend changes in hydrologic time series [e.g. *Salinger,* 2005; *Woo et Thorne,* 2003; *Burn et Elnur,* 2002]. Possible reasons of change in statistical characteristics of observed data series include natural or anthropogenic actions on the physical environment (deforestation, construction of hydraulic structures, pollution, etc.), and modifications in measurement equipment or protocol.

To deal with these non-stationary data sets, changepoint analysis in hydrologic time series is regularly revisited using various assumptions on the data model, on the parameter that exhibits a change as well as on the type of change. Most of the published methodologies use classical statistical methods to detect changes in slopes or intercept of linear regression models [*Solow*, 1987; *Easterling and Peterson*, 1995; *Vincent*, 1998; *Lund and Reeves*, 2002; *Wang*, 2003]. Other curve fitting methods are used in some rare cases [e.g. *Sagarin and Micheli,* 2001; *Bowman et al.*, 2004].

The changepoint problem was also addressed in Bayesian statistics: *Gelfand et al.* [1990] discussed Bayesian analysis of a variety of normal data models, including regression and ANOVA-type structures, where they allowed for unequal variances. *Barry and Hartigan* [1993] developed a Bayesian analysis for a multiple changepoint problem, where a normal random variable that can be added anytime to the mean of the series, but only with a certain probability. *Stephens* [1994] implemented Bayesian analysis of a multiple changepoint problem where the number of changepoints is assumed known, but the times of occurrence of the changepoints remain unknown. Other authors emphasized on the single changepoint problem. We cite for example *Carlin et al.* [1992] who applied a three-

stage hierarchical Bayesian analysis to a simple linear changepoint model for normal data where a single changepoint occurs on the regression and variance parameters. *Perreault et al.* [2000a; 2000b] gave Bayesian analyses of several changepoint models of univariate normal data. All of these authors implemented their analyses using Gibbs sampling. *Rasmussen* [2001] considered a single changepoint in a simple linear regression model with noninformative priors and derived the exact analytical posterior distribution of the regression parameters. His model assumes that the changepoint occurred with certainty, and does not allow a clear diagnosis of the existence of the change. *Perreault et al.* [2000c] developed an exact analytical Bayesian analysis of a changepoint in the mean of a series of multivariate normal random variables.

The model presented in this paper allows simultaneous changepoint analysis of several time series, each time series being modeled as a linear combination of a set of explicative variables. It generalises the model of *Rasmussen* [2001] to cases where there is more than one response variable, to cases where the changepoint does not occur with certainty and to cases where informative priors on the regression parameters are required. It also improves on the models of *Perreault et al.* [2000a, b, c] which are all special cases of the model presented in this paper. Unfortunately, the solution is no longer analytic and inference is performed using Monte-Carlo Markov Chain simulation.

The outline of this paper is as follows: the general changepoint model is presented in Section 2. Section 3 presents the Monte-Carlo Markov Chain integration methods that will be used in this paper, with special attention to Gibbs sampling. Section 4 presents the

technical developments that enable MCMC sampling of the parameters of the multivariate regression model. The methodology is adapted in Section 5 to account for missing data in the response variables and/or in the design matrices. The methodology is then generalized to segmented multivariate regression in Section 6. Based on this generalization, the final remarks in Section 7 discuss the wide scope of potential applications of the procedure described in this work. The method described in this paper provides a practical approach to highly complicated problems, such as switching models or segmented regression.

## 2. The general changepoint model

*Carlin et al.* [1992] present a general formulation of the changepoint problem. Let $\mathbf{Y} = (\mathbf{Y}_1, ..., \mathbf{Y}_n)'$ be a sequence of random vectors which displays a changepoint at an unknown time of change $\tau \in \{1, ..., n\}$. The simplest formulation for the changepoint model assumes that the random vector $\mathbf{Y}_t$ has a probability density function (pdf) $f$ for $t = 1, ..., \tau$ and pdf $g$ for $t = \tau + 1, ..., n$. The case $\tau = n$ stands for the absence of a changepoint. It follows that the likelihood of $\tau$ is

$$L(\tau \mid \mathbf{Y}) = \prod_{t=1}^{\tau} f(\mathbf{Y}_t) \prod_{t=\tau+1}^{n} g(\mathbf{Y}_t) \qquad [1]$$

with $\prod_{t=n+1}^{n} g(.) = 1$, if $\tau = n$. For known pdf's $f$ and $g$, the maximum likelihood estimate (MLE) for $\tau$ can be directly obtained by calculating the likelihood [1] for each $\tau \in \{1, 2, ..., n\}$. When we consider parametric families $f(\cdot \mid \xi)$ and $g(\cdot \mid \zeta)$, the likelihood becomes

$$L(\tau,\xi,\zeta \mid \mathbf{Y}) = \prod_{t=1}^{\tau} f(\mathbf{Y}_t \mid \xi) \prod_{t=\tau+1}^{n} g(\mathbf{Y}_t \mid \zeta)$$

Obtaining maximum likelihood estimates of $\tau$, $\xi$ and $\zeta$ can be challenging for problems involving a large number of unknown parameters, even when using numerical methods as opposed to mathematical inference. A Bayesian formulation of the changepoint problem gives an alternate approach to inferring on the parameters. Assuming a prior $p(\tau,\xi,\zeta)$ for the parameters, the joint distribution of data and parameters is

$$L(\tau,\xi,\zeta \mid \mathbf{Y}) p(\tau,\xi,\zeta) \qquad [2]$$

which is proportional to the joint posterior distribution of $\tau$, $\xi$ and $\zeta$. Obtaining the exact posterior marginal distribution of the parameter $\tau$ requires the integration of [2] with respect to $\xi$ and $\zeta$. However, this might not be practical in high-dimensional problems. In such cases, we prefer to approximate the posterior distribution using Markov chain Monte Carlo methods as will be discussed later.

Our stategy in this paper is to redesign the changepoint model into a multivariate regression model, so that normal theory can be used and applied whenever possible. This greatly simplifies analytical developments. The changepoint model (redesigned later in this section) assumes that the $(r \times 1)$ vector $\mathbf{Y}_t$ is related to the $(r \times m^*)$ matrix $\mathbf{X}_t$ by

$$\mathbf{Y}_t = \mathbf{X}_t \boldsymbol{\theta}_t^{(\tau)} + \boldsymbol{\upsilon}_t \qquad [3a]$$

where

$$\boldsymbol{\theta}_t^{(\tau)} = \begin{cases} \boldsymbol{\beta}_1^*, & 1 \le t \le \tau \\ \boldsymbol{\beta}_2^*, & \tau < t \le n \end{cases} \qquad [3b]$$

under the constraints

$$\boldsymbol{\beta}_1^* = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_0)' \text{ and } \boldsymbol{\beta}_2^* = (\boldsymbol{\beta}_2, \boldsymbol{\beta}_0)' \qquad [3c]$$

The dimensions of the vectors $\boldsymbol{\theta}_t^{(\tau)}$, $\boldsymbol{\beta}_1^*$, $\boldsymbol{\beta}^*$, $\boldsymbol{\beta}_0$, $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ are respectively $(m^* \times 1)$, $(m^* \times 1)$, $(m^* \times 1)$, $(m_0^* \times 1)$, $(m_1^* \times 1)$ and $(m_1^* \times 1)$. Equation [3.c] implies that $m^* = m_0^* + m_1^*$. It is also assumed that error terms $\{\boldsymbol{\upsilon}_t\}$ are independent and identically distributed following $N[0, \Sigma_y]$.

Model [3a] assumes a changepoint in the vector $\boldsymbol{\theta}_t^{(\tau)}$ from the subvector $\boldsymbol{\beta}_1$ to the subvector $\boldsymbol{\beta}_2$. The subvector $\boldsymbol{\beta}_0$ is assumed to remain part of $\boldsymbol{\theta}_t^{(\tau)}$ throughout the observation series. This feature allows to model, as a special case, a changepoint in the intercept parameter.

By defining $\boldsymbol{\theta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_0)'$ and

$$\boldsymbol{\Delta}_t^{(\tau)} = \begin{pmatrix} \delta_t^{(\tau)} \mathbf{I}_{m_1^*} & (1-\delta_t^{(\tau)}) \mathbf{I}_{m_1^*} & 0 \\ 0 & 0 & \mathbf{I}_{m_0^*} \end{pmatrix}$$

where $\mathbf{I}_{m_0^*}$ and $\mathbf{I}_{m_1^*}$ are the identity matrixes of dimension $m_0^*$ and $m_1^*$, and

$$\delta_t^{(\tau)} = \begin{cases} 1, & t \le \tau \\ 0, & t > \tau \end{cases}$$

model [3a] can be written in a simpler form as

$$\mathbf{Y}_t = \mathbf{X}_t \boldsymbol{\Delta}_t^{(\tau)} \boldsymbol{\theta} + \boldsymbol{\upsilon}_t \qquad [4]$$

Hence, with the knowledge of the time $t$ of the changepoint, the changepoint structure can be redesigned as a single multivariate regression equation. The general model

$$\mathbf{Y}_t = \mathbf{F}_t\boldsymbol{\theta} + \boldsymbol{\upsilon}_t \qquad [5]$$

where $\mathbf{F}_t$ is any $(r \times m)$ design matrix, will be studied in Section 4. Models [3] and [4] correspond to the special case $\mathbf{F}_t = \mathbf{X}_t\boldsymbol{\Delta}_t^{(\tau)}$. Note that when there is a continuity constraint at the changepoint, the expressions of $\boldsymbol{\theta}$ and $\boldsymbol{\Delta}_t^{(\tau)}$ are different. These expressions are given in Appendix 1 for the case of two linear relationships before and after the changepoint, and the case of a linear relationship followed by a constant mean, assuming that the mean is continuous at the changepoint.

## 3. Monte Carlo Markov Chain

To make inference on a parameter of a Bayesian model, it will be necessary to integrate the joint posterior probability with respect to all the other parameters. Except in very simple cases where the solution is analytical, this integration is carried out through computer simulation. The idea of studying the stochastic properties of a random variable through computer simulation is not recent (see *Metropolis and Ulam*, 1949). Contributions from *Metropolis et al.* [1953] and *Hastings* [1970] led to a general method nowadays referred to as the Metropolis-Hastings algorithm. When all conditional distributions are known, Gibbs sampling [*Geman and Geman*, 1984] is preferred to the Metropolis-Hastings algorithm because it leads to less numerical problems. The power of the Metropolis-Hastings algorithm and the Gibbs sampler is undeniable. They allow Bayesian analysis of highly complicated models even when exact closed-form solutions are theoretically impossible to obtain.

### *3.1 The Metropolis-Hastings Algorithm*

This is an algorithm that allows us to simulate from any distribution for which the pdf $p(\cdot)$ is known up to a multiplicative constant: there is no need to know the normalizing constant since the algorithm depends on the pdf only through ratios of the form $p(\alpha_1)/p(\alpha_2)$, where $\alpha_1$ and $\alpha_2$ are sample points. A comprehensive introduction to the Metropolis-Hastings algorithm is presented in *Chib and Greenberg* [1995]. A summary of the algorithm is presented herein. First define a candidate generating distribution $q(\alpha,\alpha^*)$ which selects a random value $\alpha^*$ from any given starting point $\alpha$. Assuming a starting point $\alpha^{(1)}$, repeat the following for $j = 1,...,N$.

a. Generate a candidate $\alpha^*$ from $q(\alpha^{(j)},\cdot)$ and $u$ from the uniform U(0,1) distribution.

b. Calculate $a(\alpha^{(j)},\alpha^*)$ by

$$a(\alpha,\alpha^*) = \begin{cases} \min\left[\dfrac{p(\alpha^*)q(\alpha^*,\alpha)}{p(\alpha)q(\alpha,\alpha^*)},1\right], & \text{if } p(\alpha)q(\alpha,\alpha^*) > 0 \\ 1, & \text{otherwise} \end{cases}$$

c. Define

$$\alpha^{(j+1)} = \begin{cases} \alpha^*, & \text{if } u \le a(\alpha^{(j)},\alpha^*) \\ \alpha^{(j)}, & \text{otherwise} \end{cases}$$

Under the reasonably general regularity conditions of irreducibility and aperiodicity, the process converges to the target density $p(\cdot)$. These conditions mean that, if $\alpha_1^*$ and $\alpha_2^*$ are any possible values of the random structure $(\alpha_1,...,\alpha_m)$, it must be possible to move from

$\alpha_1^*$ to $\alpha_2^*$ in a finite number of iterations and the number of iterations required for such move is not necessarily a multiple of some integer. The first generated values should be discarded to allow the series to reach convergence to the target density.

### *3.2 Gibbs sampling*

Research in Bayesian analysis using Gibbs sampling is exploding nowadays. Gibbs sampling is a method that allows simulation of a multivariate distribution for which all conditional distributions are known. The procedure was first introduced by *Geman and Geman* [1984]. Like the Metropolis-Hastings algorithm (see above), Gibbs sampling can be used to estimate the joint posterior distribution of parameters. The method ultimately generates random variables from the joint posterior distribution. *Casella and George* [1992] gave a comprehensive introduction of the Gibbs sampler. See also the pioneering papers of *Geman and Geman* [1984], *Tanner and Wong* [1987], and *Gelfand and Smith* [1990]. To summarize the method, suppose we want to generate from the joint distribution of $m$ potentially multivariate random variables $\alpha_1,...,\alpha_m$. Gibbs sampling consists of sequentially updating the sampled values of these variables by generating from $p(\alpha_1 \mid \{\alpha_j, j \neq 1\})$, $p(\alpha_2 \mid \{\alpha_j, j \neq 2\})$, ..., $p(\alpha_m \mid \{\alpha_j, j \neq m\})$. This sequence is then repeated until the appropriate sample size is reached. The first generated values (or burn-in) of Gibbs sampling should be discarded to allow the process to reach convergence to the joint distribution. See *Ritter and Tanner* [1992] for some solutions to burn-in issues and problems with the generating distributions. The process converges to the joint distribution of $\{\alpha_j\}$ under the regularity conditions of irreducibility and aperiodicity stated in the

preceding paragraph, Gibbs sampling is an important special case of the Metropolis-Hastings algorithm. Because Gibbs sampling does not require the choice of a candidate distribution to sample from, it is usually preferred to the Metropolis-Hastings algorithm when the full conditional distributions are available and easy to sample from.

## 4. Posterior distributions of parameters

In this section, we list the posterior distributions needed to implement the Gibbs sampler for model [5]. In Section 4.1, under a normal prior, the conditional posterior distribution of $\boldsymbol{\theta}$ given $\Sigma_y$ will be provided. In Section 4.2, we obtain the conditional posterior distribution of $\Sigma_y$ given $\boldsymbol{\theta}$. There will be no restriction on the prior for $\Sigma_y$, but conjugate priors will also be considered, namely the inverse-Wishart prior and the case of independent data. These conditional distributions are useful for performing Gibbs sampling from the joint posterior of $\boldsymbol{\theta}$ and $\Sigma_y$.

To simplify the developments, an approach similar to the one proposed by *Gelman et al.* [1995] is adopted: model [5] is expressed into the equivalent univariate multiple regression representation by stacking the observed $\mathbf{Y}_t$'s in a single vector $\mathbf{Y}^v$. Hence, we define

$$\mathbf{Y}^v = (\mathbf{Y}'_1, \mathbf{Y}'_2, \ldots, \mathbf{Y}'_n)'$$
$$\mathbf{F} = \left(\mathbf{F}'_1, \ldots, \mathbf{F}'_n\right)$$
$$\boldsymbol{\upsilon}^v = (\boldsymbol{\upsilon}'_1, \boldsymbol{\upsilon}'_2, \ldots, \boldsymbol{\upsilon}'_n)'$$

where $\mathbf{Y}^v$ is the $(nr \times 1)$ vector of observations, $\mathbf{F}$ is the $(nr \times m)$ matrix of explanatory variables, and $\boldsymbol{\upsilon}^v$ is the $(nr \times 1)$ multivariate normal vector of residuals with zero mean.

The covariance structure of $\upsilon^{v}$ assumes independence over time, that is,

$$\mathrm{Var}(\upsilon^{v}) = \mathbf{I}_n \otimes \Sigma_y$$

where $\otimes$ is the kronecker product operator. For instance, if $\mathbf{A} = (a_{ij})$ and $\mathbf{B} = (b_{ij})$ then

$$\mathbf{A} \otimes \mathbf{B} = \left(a_{ij}\mathbf{B}\right)$$

Model [*5*] is then simply expressed as the univariate regression model

$$\mathbf{Y}^{v} = \mathbf{F}\boldsymbol{\theta} + \upsilon^{v}$$

Under the assumption of normality of the residual vector $\upsilon^{v}$, it follows that

$$\mathbf{Y}^{v} \mid \mathbf{F}, \boldsymbol{\theta}, \Sigma_y \sim \mathrm{N}[\mathbf{F}\boldsymbol{\theta}, \mathbf{I}_n \otimes \Sigma_y] \qquad [7]$$

The prior distributional assumptions for the Bayesian context are defined in the following section of the paper.

### 4.1 Conditional Posterior of $\theta$ given $\Sigma_y$

Under model [*5*] with normal prior

$$\boldsymbol{\theta} \mid \mathbf{F}, \Sigma_y \sim \mathrm{N}[\boldsymbol{\theta}_0, \Sigma_\theta],$$

conditional inference of $\mathbf{Y}^{v}$ and $\boldsymbol{\theta}$ given the design matrix $\mathbf{F}$ and the variance matrix $\Sigma_y$ is a consequence of

$$\begin{pmatrix} \mathbf{Y}^{v} \\ \boldsymbol{\theta} \end{pmatrix} \mid \mathbf{F}, \Sigma_y \sim \mathrm{N}\left[\begin{pmatrix} \mathbf{f} \\ \boldsymbol{\theta}_0 \end{pmatrix}, \begin{pmatrix} \mathbf{Q} & \mathbf{S}' \\ \mathbf{S} & \Sigma_\theta \end{pmatrix}\right] \qquad [8]$$

where

$$\mathbf{f} = \mathbf{F}\boldsymbol{\theta}_0$$

$$\mathbf{Q} = \mathbf{I}_n \otimes \Sigma_y + \mathbf{F}\Sigma_\theta\mathbf{F}' \qquad [9]$$

$$\mathbf{S} = \Sigma_\theta \mathbf{F}'$$

From [8], it follows from normal theory that

$$\boldsymbol{\theta} \mid \mathbf{Y}, \mathbf{F}, \Sigma_y \square \ \mathrm{N}[\mathbf{m}, \mathbf{C}]$$

where

$$\mathbf{m} = \boldsymbol{\theta}_0 + \mathbf{S}\mathbf{Q}^{-1}(\mathbf{Y}^{\mathrm{v}} - \mathbf{f})$$

$$\mathbf{C} = \Sigma_\theta - \mathbf{S}\mathbf{Q}^{-1}\mathbf{S}' \qquad [10a]$$

$$= (\Sigma_\theta^{-1} + \mathbf{F}'(\mathbf{I}_n \otimes \Sigma_y^{-1})\mathbf{F})^{-1} \qquad [10b]$$

If a proper prior for $\boldsymbol{\theta}$ is selected, [10a] is well defined. However, if $|\Sigma_\theta| \to \infty$, then [10a] may be computationally undefined, so [10b] should be used when $\Sigma_\theta^{-1}$ is easily obtained and $|\Sigma_\theta^{-1}|$ is finite. Since $\mathbf{Q}$ is $(nr \times nr)$, $\mathbf{Q}^{-1}$ should be calculated as

$$\mathbf{Q}^{-1} = (\mathbf{I}_n \otimes \Sigma_y^{-1}) - (\mathbf{I}_n \otimes \Sigma_y^{-1})\mathbf{F}\mathbf{C}\mathbf{F}'(\mathbf{I}_n \otimes \Sigma_y^{-1}) \qquad [11]$$

with $\mathbf{C}$ obtained from [10b], rather than by directly inverting [9]. The advantage of [11] is that $\Sigma_y^{-1}$ is only $(r \times r)$ and $\mathbf{C}$ is only $(m \times m)$ in contrast to the $(nr \times nr)$ matrix $\mathbf{Q}^{-1}$.

A sampling method of the multivariate normal is given in Appendix 2.

### 4.2 Conditional Posterior of $\Sigma_y$ given $\boldsymbol{\theta}$

We shall make use of the convenient notation that, for any parameter $\zeta$, $p(\zeta)$ denotes the pdf of $\zeta$. In general, we have

$$p(\Sigma_y \mid \mathbf{Y},\mathbf{F},\boldsymbol{\theta}) \propto p(\Sigma_y \mid \mathbf{F},\boldsymbol{\theta}) p(\mathbf{Y}^{\mathrm{v}} \mid \mathbf{F},\boldsymbol{\theta},\Sigma_y)$$
$$\propto p(\Sigma_y \mid \mathbf{F},\boldsymbol{\theta}) \prod_{t=1}^{n} p(\mathbf{Y}_t \mid \mathbf{F},\boldsymbol{\theta},\Sigma_y)$$
$$\propto p(\Sigma_y \mid \mathbf{F},\boldsymbol{\theta}) \mid \Sigma_y \mid^{-n/2} \exp(-\mathrm{tr}(n\hat{\Sigma}_y \Sigma_y^{-1})/2)$$

where

$$\hat{\Sigma}_y = n^{-1} \sum_{t=1}^{n} \upsilon_t \upsilon_t', \upsilon_t = \mathbf{Y}_t - \mathbf{F}_t \boldsymbol{\theta} \qquad [12]$$

Hence, under model [12] and the assumption of inverse-Wishart prior

$$\Sigma_y \mid \mathbf{F},\boldsymbol{\theta} \sim \mathrm{W}_{\upsilon}^{-1}(\Lambda_y)$$

the conditional posterior distribution of $\Sigma_y$ is

$$\Sigma_y \mid \mathbf{Y},\mathbf{F},\boldsymbol{\theta} \sim \mathrm{W}_{\upsilon+n}^{-1}(\Lambda_y + \upsilon\hat{\Sigma}_y)$$

A description of the inverse-Wishart is presented with a sampling algorithm in Appendix 4. The noninformative case corresponds to $v \to -1$ and $|\Lambda_y| \to 0$ [*Gelman et al.*, 1995]. However, the sample size $n$ must be sufficiently large in order to ensure a proper posterior. A careful approach would be to consider reasonably vague but proper priors.

In the case when $\Sigma_y = \delta^2 \Gamma_y$, where $\Gamma_y$ is a known positive definite matrix, a conjugate inverse-gamma prior

$$\delta^2 \mid \mathbf{F},\boldsymbol{\theta} \sim \mathrm{G}^{-1}(a,b)$$

is assumed (see Appendix 3 for a description of the inverse-gamma distribution.) The corresponding conditional posterior for $\delta^2$ is

$$\delta^2 \mid \mathbf{Y},\mathbf{F},\boldsymbol{\theta} \sim \mathrm{G}^{-1}\left(a + \frac{nr}{2}, b + \frac{1}{2}\mathrm{tr}(n\hat{\Sigma}_y \Gamma_y^{-1})\right)$$

The noninformative case corresponds to imposing $a \to 0$ and $b \to 0$. Again, this improper prior should be used only if the sample size $n$ is sufficiently large. Finally, the important case of independent response variables corresponds to $\Gamma_y = \mathbf{I}_r$. In such case, we have

$$\delta^2 \mid \mathbf{Y}, \mathbf{F}, \boldsymbol{\theta} \square \mathrm{G}^{-1}\left(a + \frac{nr}{2}, b + \frac{1}{2}\mathrm{tr}(n\hat{\Sigma}_y)\right)$$

### 4.3. Changepoint Inference

In this section, we focus on the changepoint inference part of the problem. As noted in the introduction, any changepoint in a regression model can be modelled by a plain regression model conditioned on the time of changepoint. It is then a matter of "rewriting" the design matrices $\{\mathbf{X}_t\}$ as a single matrix $\mathbf{F}$ given $\tau$ and obtaining a conditional posterior for the time of changepoint.

For any prior $p(\tau \mid \{\mathbf{X}_t\}, \boldsymbol{\theta}, \Sigma_y)$ under model [5] with $\mathbf{F}_t = \mathbf{X}_t \Delta_t^{(\tau)}$ depending on $\tau$, we have

$$\begin{aligned} p(\tau \mid \mathbf{Y}, \{\mathbf{X}_t\}, \boldsymbol{\theta}, \Sigma_y) &\propto p(\mathbf{F}, \boldsymbol{\theta}, \Sigma_y \mid \mathbf{Y}) \\ &\propto p(\mathbf{F}, \boldsymbol{\theta}, \Sigma_y) p(\mathbf{Y}^{\mathrm{v}} \mid \mathbf{F}, \boldsymbol{\theta}, \Sigma_y) \\ &\propto p(\mathbf{F}, \boldsymbol{\theta}, \Sigma_y) \mid \Sigma_y \mid^{-n/2} \exp(-\mathrm{tr}(n\hat{\Sigma}_y \Sigma_y^{-1})/2) \end{aligned} \qquad [13]$$

where $\hat{\Sigma}_y$ is obtained from [12]. This result can be used to sample $\tau$ under any prior assumption on $\boldsymbol{\theta}$, $\Sigma_y$ and the missing values. Equation [13] is the "regression" version of [2]: it is the exact posterior density of all unknown parameters. Hence, this equation would remain valid for any structure built in $\mathbf{F}$. This feature will be further exploited in

Section 7. The use of the Metropolis-Hastings algorithm with [13] provides a general method to generate from the joint posterior of $\{\mathbf{F}, \boldsymbol{\theta}, \Sigma_y\}$, although this may be computationally difficult in practice, which explains why direct Gibbs sampling with conjugate priors is often preferred.

Although Gibbs sampling of $\tau$ from [13] is always possible (provided that the regularity conditions of Section 3.1 are satisfied), it is possible to do better under further prior assumptions. In Section 4.1, we have assumed a normal prior for $\boldsymbol{\theta}$. With this additional assumption, we can integrate [13] with respect to $\boldsymbol{\theta}$ and we have

$$p(\tau \mid \mathbf{Y}, \{\mathbf{X}_t\}, \Sigma_y) \propto p(\mathbf{F}, \Sigma_y \mid \mathbf{Y})$$

$$\propto p(\mathbf{F}, \Sigma_y) p(\mathbf{Y}^v \mid \mathbf{F}, \Sigma_y) \qquad [14]$$

where

$$\mathbf{Y}^v \mid \mathbf{F}, \Sigma_y \square \mathrm{N}[\mathbf{f}, \mathbf{Q}]$$

is directly obtained from [8]. Since the parameters $\tau$ and $\boldsymbol{\theta}$ may be strongly dependent, the use of [14] as opposed to [13] has the desirable feature of reducing the dependencies in the series of Gibbs samplers. Therefore, the use of [14] would improve mixing and would speed up convergence to the joint posterior of all parameters. Ideally, we should integrate [14] with respect to $\Sigma_y$ as well, but our prior assumptions render this task very difficult. *Perreault et al.* [2000c] performed successfully a similar integration under a simpler model with more restraining priors.

When choosing the prior for $\tau$, since the particular event $\tau = n$ stands for the absence of a changepoint, it might be appropriate to place more or less prior probability mass on this event, depending on the question of interest or on the prior knowledge of the data. In their application example, *Carlin et al.* [1992] used the discrete uniform on $\{1,2,...,n\}$ as a prior pmf for $\tau$.

## 5. Extension to Missing Data

In practice, the data set could contain missing values. Bayesian methods cope with this problem elegantly by simply replacing the missing values by unknown parameters that are updated in the Gibbs sampling routine the same way it is done for the parameters of interest. In order to update a missing value through Gibbs sampling, we need its conditional distribution given all other parameters and data. This section presents the conditional distributions that allow Gibbs sampling of missing values in $\mathbf{Y}^{v}$ or $\mathbf{F}$.

The case where missing values are present in $\mathbf{Y}^{v}$ is examined. For any given matrix (or vector) $\mathbf{a}$, let $\mathbf{a}_{(u)}$ be the matrix (or vector) composed of the values in $\mathbf{a}$ corresponding to the set of indices $U$. Hence, define $\mathbf{Y}^{v}_{(M)}$ to be the vector of missing values in $\mathbf{Y}^{v}$, where $M$ is the set of indices corresponding to the missing values in $\mathbf{Y}^{v}$, and define $\mathbf{Y}^{v}_{(O)}$ to represent the vector of observed values in $\mathbf{Y}^{v}$, where $O$ is the set of indices corresponding to the observed values.

From [7], the posterior distribution of the missing values in $\mathbf{Y}^{v}$ is

$$\mathbf{Y}^{v}_{(M)} \mid \mathbf{Y}^{v}_{(O)}, \mathbf{F}, \Sigma_y \sim N(\tilde{\mathbf{f}}_{(M)}, \tilde{\mathbf{Q}}_{(M\times M)})$$

where

$$\tilde{\mathbf{f}}_{(M)} = \mathbf{f}_{(M)} + \mathbf{Q}_{(M\times O)}(\mathbf{Q}_{(O\times O)})^{-1}(\mathbf{Y}^{v}_{(O)} - \mathbf{f}_{(O)})$$

and

$$\tilde{\mathbf{Q}}_{(M\times M)} = \mathbf{Q}_{(M\times M)} - \mathbf{Q}_{(M\times O)}(\mathbf{Q}_{(O\times O)})^{-1}\mathbf{Q}_{(O\times M)}$$

If $\mathbf{F}$ have missing data, it can also be generated by Gibbs sampling. With this approach, the model for $\mathbf{F}$ cannot be ignored and prior distributional assumptions on $\mathbf{F}$ must be considered. For instance, in the case of model [4], the prior must account for the changepoint structure $\mathbf{F}_t = \mathbf{X}_t \boldsymbol{\Delta}_t^{(\tau)}$. We present a solution for this special case.

Define $\mathbf{X}_t^{v}$ by stacking the columns of $\mathbf{X}_t$ into a single column vector. Inference on missing data will be obtained from the conditional joint distribution of $\mathbf{Y}_t$ and $\mathbf{X}_t^{v}$. These two components are related by model [4], which can also be written as

$$\mathbf{Y}_t = (\boldsymbol{\theta}_t^{(\tau)'} \otimes \mathbf{I}_r)\mathbf{X}_t^{v} + \boldsymbol{\upsilon}_t \qquad [15]$$

Let $\{\mathbf{X}_1^{v}, \ldots, \mathbf{X}_n^{v}\}$ be independent with

$$\mathbf{X}_t^{v} \mid \mu_x, \Omega_x \sim N[\mu_x, \Omega_x], \; t = 1, \ldots, n$$

and define for a given time $t$ when some values are missing

$$\mathbf{Z}_t = \begin{pmatrix} \mathbf{Y}_t \\ \mathbf{X}_t^{v} \end{pmatrix}$$

From [15], it follows that

$$\mathbf{Z}_t \mid \tau, \boldsymbol{\theta}, \Sigma_y, \mu_x, \Omega_x \sim \mathrm{N}[\mathbf{g}_t, \mathbf{R}_t]$$

where

$$\mathbf{g}_t = \begin{pmatrix} (\boldsymbol{\theta}_t^{(\tau)} \otimes \mathbf{I}_r)\mu_x \\ m_x \end{pmatrix}$$

$$\mathbf{R}_t = \begin{pmatrix} \Sigma_y + (\boldsymbol{\theta}_t^{(\tau)} \otimes \mathbf{I}_r)\Omega_x(\boldsymbol{\theta}_t^{(\tau)} \otimes \mathbf{I}_r) & (\boldsymbol{\theta}_t^{(\tau)'} \otimes \mathbf{I}_r)\Omega_x \\ \Omega_x(\boldsymbol{\theta}_t^{(\tau)} \otimes \mathbf{I}_r) & \Omega_x \end{pmatrix}$$

From model assumptions, $\{\mathbf{Z}_1, ..., \mathbf{Z}_n\}$ are independent. Hence, the posterior distribution of the missing values in $\mathbf{Z}_t$ is directly obtained by normal theory, that is,

$$\mathbf{Z}_{t(M)} \mid \mathbf{Z}_{t(O)}, \tau, \boldsymbol{\theta}, \Sigma_y, \mu_x, \Omega_x \sim \mathrm{N}[\tilde{\mathbf{g}}_{t(M)}, \tilde{\mathbf{R}}_{t(M\times M)}]$$

where

$$\tilde{\mathbf{g}}_{t(M)} = \mathbf{g}_{t(M)} + \mathbf{R}_{t(M\times O)}(\mathbf{R}_{t(O\times O)})^{-1}(\mathbf{Z}_{t(O)} - \mathbf{g}_{t(O)}) \qquad [16a]$$

$$\tilde{\mathbf{R}}_{t(M\times M)} = \mathbf{R}_{t(M\times M)} - \mathbf{R}_{t(M\times O)}(\mathbf{R}_{t(O\times O)})^{-1}\mathbf{R}_{t(O\times M)} \qquad [16b]$$

However, $(\mathbf{R}_{t(O\times O)})^{-1}$ may not be strictly positive definite, but only nonnegative definite. This will happen if, for example, an intercept parameter is part of model [5]. It can be shown that (16) remains valid if the *g*-inverse (generalized inverse) $(\mathbf{R}_{t(O\times O)})^{-}$ is used instead of $(\mathbf{R}_{t(O\times O)})^{-1}$. The *g*-inverse of a matrix $\mathbf{A}$ is denoted by $\mathbf{A}^{-}$ and can be calculated by $\mathbf{A}^{-} = \Gamma\Lambda^{-1}\Gamma'$, where $\Gamma$ is a column orthonormal matrix of eigenvectors corresponding to the $s$ non-zero eigenvalues $\Lambda = \mathrm{diag}(\lambda_1, ..., \lambda_s)$ of $\mathbf{A}$. A more general definition of the *g*-inverse is reviewed in *Mardia et al.* [1979].

To provide an estimation tool for the parameters $\mu_x$ and $\Omega_x$, we consider the conjugate normal-inverse-Wishart prior

$$\Omega_x \sim W_{\upsilon_0}^{-1}(\mathbf{V}_0) \qquad [17a]$$

$$\mu_x | \Omega_x \sim N[\mu_0, \Omega_x/k_0] \qquad [17b]$$

From *Gelman et al.* [1995], the conditional posterior is

$$\Omega_x | \{\mathbf{X}_t\} \sim W_{\upsilon_n}^{-1}(\mathbf{V}_n)$$
$$\mu_x | \{\mathbf{X}_t\}, \Omega_x \sim N[\mu_n, \Omega_x/k_n]$$

where

$$\mu_n = \frac{1}{k_n}(k_0\mu_0 + n\bar{\mathbf{X}})$$
$$k_n = k_0 + n$$
$$\upsilon_n = \upsilon_0 + n$$
$$\mathbf{V}_n = \mathbf{V}_0 + \mathbf{S} + \frac{k_0 n}{k_n}(\bar{\mathbf{X}} - \mu_0)(\bar{\mathbf{X}} - \mu_0)'$$

and

$$\bar{\mathbf{X}} = \frac{1}{n}\sum_{t=1}^{n} \mathbf{X}_t^{\mathrm{v}}$$
$$\mathbf{S} = \sum_{t=1}^{n} \mathbf{X}_t^{\mathrm{v}}\mathbf{X}_t^{\mathrm{v}} - n\bar{\mathbf{X}}\bar{\mathbf{X}}'$$

We can easily obtain samples from this joint posterior by first sampling $\Omega_x | \{\mathbf{X}_t\}$ and then sampling $\mu_x | \{\mathbf{X}_t\}, \Omega_x$. See Appendices A.1 and A.3 for sampling from multivariate normal and inverse-Wishart distributions. The noninformative multivariate Jeffreys' prior density for $\{\mu_x, \Omega_x\}$ is

$$[\mu_x, \Omega_x] \propto |\Omega_x|^{-(rm^*+1)/2}$$

This is the limiting case of the normal-inverse-Wishart prior in [*17*] when $k_0 \to 0$, $\upsilon_0 \to -1$ and $|\mathbf{V}_0| \to 0$. The posterior distribution $\{\mu_x, \Omega_x | \{\mathbf{X}_t\}\}$ for this case can be written as

$$\begin{aligned} \Omega_x | \{\mathbf{X}_t\} &\sim \mathrm{W}_{n-1}^{-1}(\mathbf{S}) \\ \mu_x | \{\mathbf{X}_t\}, \Omega_x &\sim \mathrm{N}[\bar{\mathbf{X}}, \Omega_x/n] \end{aligned}$$

in which we find the sample estimators of the parameters $\mu_x$ and $\Omega_x$.

## 6. Further Generalizations

This paper presented a method of analysis for a single changepoint in a multivariate regression model. However, the same idea can be exploited for more complicated models, including a multiple changepoint model or the more general segmented multivariate regression. The latter is formally described as

$$\mathbf{Y}_t = \mathbf{X}_t \boldsymbol{\beta}_i + \boldsymbol{\upsilon}_t, \text{ if } x_t \in ]\alpha_{i-1}, \alpha_i]$$

for $t = 1, \ldots, n$ and $i \in \{1, \ldots, l\}$. The $(r \times 1)$ observations $\{\mathbf{Y}_t\}$ are modeled as piecewise regressions depending on the covariates $\{x_t\}$. $\{\mathbf{X}_t\}$ are $(r \times m)$ design matrices, $\{\boldsymbol{\beta}_i\}$ are $(m \times 1)$ regression parameters, and $\{\boldsymbol{\upsilon}_t\}$ are $(r \times 1)$ residual vectors. A natural approach for the analysis of this model is to obtain estimation equations for the $\beta_i$'s separately. However, this problem is greatly simplified if the model is written as its equivalent multivariate regression form. Define the $(1 \times l)$ row vector

$$\delta^{(t)} = (\overbrace{0,0,...,0,1}^{i},0,...,0), \text{ if } x_t \in (\alpha_{i-1},\alpha_i]$$

With $\boldsymbol{\theta}' = (\boldsymbol{\beta}'_1,...,\boldsymbol{\beta}'_l)$, the segmented multivariate regression is equivalent to

$$\mathbf{Y}_t = \mathbf{X}_t(\delta^{(t)} \otimes \mathbf{I}_m)\boldsymbol{\theta} + \boldsymbol{\upsilon}_t$$

Under the appropriate assumptions on the residuals $\{\upsilon_t\}$, results of Sections 4 and 5 are immediately applicable. With $\mathbf{F}_t = \mathbf{X}_t(\delta^{(t)} \otimes \mathbf{I}_m)$, the conditional posterior [13] (or [14] if $\boldsymbol{\theta}$ has a normal prior) can be used to obtain the conditional posterior of the parameters $\{\alpha_i\}$ and perform their Gibbs sampling. There is no doubt that the same idea can be used to obtain a practical solution for a wide variety of switching models, a few of which being shown in Appendix 1.

## 7. Conclusions

This paper has provided a simple implementation of Bayesian analysis for multivariate regression via Gibbs sampling. The method was extended to the inclusion of missing values and to the inclusion of a changepoint structure in the model. An attractive feature of the approach presented in this paper is that it can be applied to cases that can't be analysed with recently published changepoint detection methodologies such as *Rasmussen* [2001] and *Perreault et al.* [2000a,b,c]: it can readily be applied to cases where the changepoint simultaneously occurs in several response variables, to cases where the change does not occur with certainty and to cases where informative priors are appropriate. Several applications that highlight these features are presented in a companion paper.

An interesting future development would be to relax the assumption of constant residual variance over time and the one of normality. A potential approach for this would be to introduce dependencies in the variance evolution over time, hence allowing for variable variance estimation. The scope of possibilities for the developed approach goes beyond the analysis of the single changepoint problem. The case of multiple changepoints could also be easily treated using the idea in Section 6. Potential applications of this model include not only changepoint models, but also other switching models such as segmented multivariate regression or shifting-level models. Such generality is made possible by the fact that the design matrices $\{\mathbf{F}_t\}$ can be structured in accordance to such models. This opens the door for a practical approach to analyzing these models and applying them in the field of water resources.

## 8. Acknowledgements

The financial support provided by the Natural Sciences and Engineering Research Council of Canada (NSERC), the Northern Study Center (CEN) of Laval University and the Ouranos consortium is gratefully acknowledged.

## References

Barry, D., and Hartigan, J. A. (1993). A Bayesian Analysis for Change Point Problems. *J. Amer. Stat. Assoc.* **88**:309-319.

Bowman, A.W., Pope, A. and Ismail, B. (2004). Detecting discontinuities in nonparametric regression curves and surfaces. Research report, Department of statistics, University of Glasgow. [http://www.stats.gla.ac.uk/%7Eadrian/research-reports/discontinuity-paper.ps]

Burn, D.H. and Hag Elnur, M.A., (2002). Detection of hydrologic trends and variability. *Journal of Hydrology* **255**: 107-122.

Carlin, B. P., Gelfand, A. E., and Smith, A. F. M. (1992). Hierarchical Bayesian analysis of changepoint problems. *Applied Statistics* **41**: 389-405.

Casella, G. and George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician* **46:** 167-174.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician* **49:**327-335.

Easterling., D.R. and Peterson., T.C. (1992) Techniques for detecting and adjusting for artificial discontinuities in climatological time series: a review. *Proc. of the Fifth International Meeting on Statistical Climatology, 22-26 june 1996, Toronto, Ontario, Canada.*

Gelfand, A. E., Hills, S. E., Racine-Poon, A., and Smith, A. F. M. (1990). Illustration of Bayesian Inference in Normal Data Models Using Gibbs Sampling. *J. Amer. Stat. Assoc.*,**85**: 972-985.

Gelfand, A. E., and Smith, A. F. M. (1990). Sampling-Based Approaches to Calculating Marginal Densities. *J. Amer. Stat. Assoc.* **85:** 398-409.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian data analysis*. Chapman & Hall, London.

Geman, S., and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6** : 721-741.

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

Lund R. and Reeves J. (2002) Detection of undocumented changepoints: A revision of the two-phase regression model. *J. Climate* **15**, 2547-2554.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate Analysis*. Academic Press, New York.

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**:1087-1092.

Metropolis, N., and Ulam S. (1949). The Monte Carlo method. *J. Amer. Stat. Assoc.*, **44**:335-341.

Perreault, L., Bernier, J., Bobée, B., and Parent, E. (2000a). Bayesian change-point analysis in hydrometeorological time series 1. Part 1. The normal model revisited. *J. of Hydrology* **235**: 221-241.

Perreault, L., Bernier, J., Bobée, B., and Parent, E. (2000b). Bayesian change-point analysis in hydrometeorological time series 2. Part 2. Comparison of change-point models and forecasting. *J. of Hydrology* **235**: 242-263.

Perreault, L., Parent, É., Bernier, J., and Bobée, B. (2000c). Retrospective multivariate Bayesian change-point analysis: A simultaneous single change in the mean of several hydrological sequences. *Stochastic Environmental Research and Risk Assessment*,**14**: 243-261.

Rasmussen, P. (2001). Bayesian estimation of change points using the general linear model. *Water Resources Research* **37**:2723-2731.

Ritter, C., and Tanner, M. A. (1992). Facilitating the Gibbs Sampler: The Gibbs Stopper and the Griddy-Gibbs Sampler. *J. Amer. Stat. Assoc.* **87**:861-868.

Sagarin R and Micheli F (2001). Climate change in nontraditional data sets. *Science* **294** : 811.

Salinger, M. (2005). Climate Variability and Change: Past, Present and Future – An Overview. *Climatic Change* **70**: 9-29

Solow A.R. (1987) Testing for climate change: an application of the two-phase regression model. *J. Climate Appl. Met.* **26**, 1401-1405.

Stephens, D. A. (1994). Bayesian Retrospective Multiple-changepoint Identification. *Applied Statistics* **43**: 159-178.

Tanner, M. A., and Wong, W. (1987). The Calculation of Posterior Distributions by Data Augmentation (with discussion). *J. Amer. Stat. Assoc.*, **82**: 528-550.

Vincent L.A. (1998) A technique for the identification of inhomogeneities in Canadian temperature series. *J. Climate* **11**, 1094-1105.

Wang X.L. (2003) Comments on 'Detection of Undocumented Changepoints: A revision of the Two-Phase regression model'. *J. Climate* **16**, 3383-3385.

Woo, M. et Thorne, R. (2003). Comment on 'Detection of hydrologic trends and variability' by Burn, D.H. and Hag Elnur, M.A., 2002. Journal of Hydrology 255, 107-122. *Journal of hydrology* **277**:150-160.

## Appendix 1: Design matrix F when there is a continuity constraints at the changepoint

When there is a continuity constraint at the changepoint, the expression of the design matrix F is slightly different of that presented at Section 2. We give here its expression for two practical cases.

### A1.1 Continuity constraint at the changepoint

$$Y_t = \begin{cases} X_t\beta_1' + \nu_t \; if \; t \le \tau \\ (X_t - X_\tau)\beta_2' + X_\tau\beta_1' + \nu_t \; if \; t > \tau \end{cases}$$

$$X_t^* = (X_t, X_\tau)$$

$$\theta = (\beta_1, \beta_2)$$

$$\Delta_t^\tau = \begin{pmatrix} (\delta_t)^\tau I_m & \left(1-(\delta_t)^\tau\right) I_m \\ \left(1-(\delta_t)^\tau\right) I_m & -\left(1-(\delta_t)^\tau\right) I_m \end{pmatrix}$$

$$F_t = X_t^* \Delta_t^\tau$$

Note that in this special case, if $\mathbf{X}$ has a column with constant values, the coefficient of the first element of $\boldsymbol{\beta}_2$ is always null, thus this parameter should not be updated in the MCMC computations.

**A.1.2 linear relationship before the change, constant mean after the change, and continuity of the mean model at the changepoint**

$$Y_t = \begin{cases} X_t\beta' + \nu_t \; if \; t \le \tau \\ X_\tau\beta' + \nu_t \; if \; t > \tau \end{cases}$$

$$X_t^* = (X_t, X_\tau)$$

$$\Delta_t^\tau = \begin{pmatrix} (\delta_t)^\tau I_{m^*} \\ \left(1-(\delta_t)^\tau\right) I_{m^*} \end{pmatrix}$$

$$\theta = \beta'$$

$$F_t = X_t^* \Delta_t^\tau$$

$$Y_t = F_t\theta + \nu_t$$

**Appendix 2: Sampling from a Multivariate Normal**

The target distribution is a $p$-dimension normal $\mathrm{N}(\mu, \mathbf{S})$ with density

$$f(\mathbf{x}) = \frac{1}{(2p)^{p/2}} |\mathbf{S}|^{-1/2} \exp\left( -\frac{1}{2}(\mathbf{x}-\mu)'\mathbf{S}^{-1}(\mathbf{x}-\mu) \right)$$

1. Obtain a matrix $\mathbf{A}$ such that $\mathbf{A}\mathbf{A}' = \mathbf{S}$ (using the eigenvalue decomposition, for example).

2. Generate $\mathbf{z} = (z_1, ..., z_p)'$, a vector consisting of $p$ independent draws from a standard normal distribution.

3. $\mathbf{x} = \mu + \mathbf{A}\mathbf{z} \square \mathrm{N}(\mu, \mathbf{S})$

## Appendix 3. Sampling from an Inverse-Gamma

The target distribution is $G^{-1}(a,b)$, $a>0$, $b>0$, with density

$$f(\theta)=\frac{b^a}{\Gamma(a)}\theta^{-(a+1)}e^{-b/\theta},\theta>0$$

1. Simulate $\gamma$ from a $G(a,b)$ distribution with density

$$f(x)=\frac{b^a}{\Gamma(a)}x^{a-1}e^{-bx},x>0.$$

2. $\theta=1/\gamma \square G^{-1}(a,b)$. The expectation and variance are respectively

$$E(\theta)=b/(a-1) \text{ for } a>1$$

and

$$\text{Var}(\theta)=\frac{b^2}{(a-1)^2(a-2)} \text{ for } a>2.$$

## Appendix 4. Sampling from an Inverse-Wishart

The target distribution is $W_\upsilon^{-1}[\mathbf{S}]$ where $\mathbf{S}$ is a $(k \times k)$ symmetric positive definite matrix and $\upsilon \in \{k, k+1, k+2, ...\}$. The density function is

$$f(\Sigma) = \left( 2^{\upsilon k/2} p^{k(k-1)/4} \prod_{i=1}^{k} \Gamma\left( \frac{\upsilon + 1 - i}{2} \right) \right)^{-1}$$
$$|\mathbf{S}|^{\upsilon/2} |\Sigma|^{-(\upsilon+k+1)/2} \exp\left( -\frac{1}{2} \mathrm{tr}(\mathbf{S}\Sigma^{-1}) \right)$$

where $\Sigma$ is symmetric and positive definite. The following algorithm was first proposed by *Odell and Feiveson* [1966].

1. Simulate $\alpha_1, \cdots, \alpha_n$, $\upsilon$ independent samples from a $k$-dimension multivariate normal distribution $\mathrm{N}(0, \mathbf{I})$. Define the $(k \times n)$ matrix $\alpha = (\alpha_1, \cdots, \alpha_n)$.

2. Obtain a matrix $\mathbf{A}$ such that $\mathbf{AA}' = \mathbf{S}$ (using the eigenvalue decomposition, for example).

3. $\Sigma = \mathbf{A}(\alpha\alpha)^{-1}\mathbf{A}' \square\ W_\upsilon^{-1}(\mathbf{S})$.

The expectation of an Inverse-Wishart is

$$\mathrm{E}(\Sigma) = \mathbf{S}/(\upsilon - k - 1) \text{ for } \upsilon \geq k + 2$$