Sampling from Gaussian Markov random fields conditioned on linear constraints

D. P. Simpson[1] I. W. Turner[2] A. N. Pettitt[3]

(Received 31 August 2006; revised 7 April 2008)

Abstract

Gaussian Markov random fields (GMRFs) are important modeling tools in statistics. They are often utilised to model spatially structured uncertainty, seasonal variation and other trends in the data. These last two examples of GMRFs are part of a larger class of GMRFs conditioned on linear constraints. Performing Monte Carlo Markov Chain inference on these models requires a large number of samples from GMRFs conditioned on linear constraints. Therefore it is vital to have fast and efficient methods for performing these samples. This article presents three Krylov subspace methods for sampling from a GMRF conditioned on linear constraints based on solving a Karush–Kuhn–Tucker, or saddle point, system.

See http://anziamj.austms.org.au/ojs/index.php/ANZIAMJ/article/view/131 for this article, Published July 17, 2008. ISSN 1446-8735

Contents

1 Introduction

Gaussian Markov random fields (GMRFs) are used in the statistical modeling of a variety of phenomena. They can be used to model structured spatial effects, seasonal variation, and other data trends. A large list of applications of GMRFs can be found in the monograph by Rue and Held [1].

A GMRF is defined by considering a cloud of points $\{\mathbf{s}_i\}_{i=1}^n$ in $\mathbb{R}^d$ and defining a Gaussian random variable x_i, $i = 1, \ldots, n$ at each point. These random variables are referred to as a Gaussian Markov random field. The joint distribution of the GMRF has the probability density function

$$\pi(\mathbf{x}|\mathbf{A}, \mathbf{b}) \propto \exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{b}^T\mathbf{x}\right), \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{n \times n}$ is a symmetric positive semi-definite 'precision' matrix and the mean $\boldsymbol{\mu}$ is given by $\mathbf{A}\boldsymbol{\mu} = \mathbf{b}$, that is, for invertible $\mathbf{A}$, $\mathbf{x}$ is a normally distributed random vector with mean $\mathbf{A}^{-1}\mathbf{b}$ and covariance matrix $\mathbf{A}^{-1}$ (written $\mathbf{x} \sim N(\mathbf{A}^{-1}\mathbf{b}, \mathbf{A}^{-1})$). In the applications to be considered $\mathbf{A}$ is large, sparse and symmetric positive definite.

A class of GMRFs with singular 'precision' matrices, known as intrinsic GMRFs, are used in statistical modeling to remove trend components in data [1]. Let $\tilde{\mathbf{A}}$ be a symmetric positive semi-definite matrix with nullity$(\tilde{\mathbf{A}}) = k$, let $\{\mathbf{n}_1, \mathbf{n}_2, \ldots, \mathbf{n}_k\}$ be an orthonormal basis for $\mathcal{N}(\tilde{\mathbf{A}})$ and let $\mathbf{N} = [\mathbf{n}_1, \mathbf{n}_2, \ldots, \mathbf{n}_k]$. Then, for any $\mathbf{y} \in \mathcal{N}(\tilde{\mathbf{A}})$, the improper density $\pi(\mathbf{x}|\tilde{\mathbf{A}}, \mathbf{0})$ satisfies $\pi(\mathbf{x}|\tilde{\mathbf{A}}, \mathbf{0}) = \pi(\mathbf{x} + y|\tilde{\mathbf{A}}, \mathbf{0})$; that is, the zero-mean improper GMRF specified by 'precision' matrix $\tilde{\mathbf{A}}$ is invariant to the addition of vectors in the nullspace of $\tilde{\mathbf{A}}$. A common example is the first order random walk on a line which has $\mathbf{N} = [\mathbf{e}]$, where $\mathbf{e}$ is a vector of ones, and is, therefore, invariant to constant shifts in all components [1, Chapter 3].

A more general form of these densities is found using the following argument. Let $\mathbf{y}$ be a zero-mean GMRF with non-singular precision matrix $\tilde{\mathbf{A}} + \alpha\mathbf{N}\mathbf{N}^T$, where $\alpha > 0$, then the density of $\mathbf{y}|\mathbf{N}^T\mathbf{y} = \mathbf{0}$ is normally distributed with mean $\mathbf{0}$ and 'precision' $\tilde{\mathbf{A}}$. It follows that these intrinsic GMRFs are a simple case of GMRFs conditioned on linear constraints, which are denoted $\mathbf{x}|\mathbf{B}\mathbf{x} = \mathbf{c}$, where $\mathbf{x}$ is a proper GMRF with symmetric, positive definite precision matrix $\mathbf{A}$ and $\mathbf{B} \in \mathbb{R}^{k \times n}$ has rank k [2]. In most applications the number of constraints k is much smaller than the number of data points n.

2 Sampling from a GMRF conditioned on linear constraints

A sample from a GMRF conditioned on linear constraints $\tilde{\mathbf{x}}$ is calculated from an unconditional sample $\mathbf{x}$ from $N(\mathbf{0}, \mathbf{A}^{-1})$ using the update formula

$$\tilde{\mathbf{x}} = \mathbf{x} - \mathbf{A}^{-1}\mathbf{B}^T \left(\mathbf{B}\mathbf{A}^{-1}\mathbf{B}^T\right)^{-1} (\mathbf{B}\mathbf{x} - \mathbf{c}), \tag{2}$$

where $\mathbf{B} \in \mathbb{R}^{k\times n}$ and $\mathbf{c} \in \mathbb{R}^k$. This is referred to as 'conditioning by Kriging' by Rue [2]. Writing equation (2) as $\tilde{\mathbf{x}} = \mathbf{x} - \boldsymbol{\delta x}$, the update can be found as the solution to the linear system [3, cf. dual-Kriging equations]

$$\begin{pmatrix} \mathbf{A} & \mathbf{B}^T \\ \mathbf{B} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{\delta x} \\ \mathbf{y} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{B}\mathbf{x} - \mathbf{c} \end{pmatrix}. \tag{3}$$

For the remainder of this article we denote this system as $\boldsymbol{\Lambda}\mathbf{v} = \mathbf{b}$ and refer to it as the Karush–Kuhn–Tucker (KKT) system, after the linear system derived from the KKT conditions in constrained optimisation [4].

Before considering methods for computing $\boldsymbol{\delta x}$, it is necessary to outline methods for sampling from an unconditional GMRF. The standard method for sampling from an unconditional GMRF uses the Cholesky decomposition $\mathbf{A}$ to transform a vector of independently and identically distributed standard normal variables [2]. Due to the size and sparsity of A, a Krylov subspace method can be used to efficiently approximate the action of the inverse square root of $\mathbf{A}$ on a vector of i.i.d. standard normals. This method for computing unconditional samples was investigated by Simpson et al. [5]. Implementation details for both of these sampling methods can be found in the papers by Simpson et al. [5] and Rue [2]. The following proposition summarises these two methods for sampling from an unconditional GMRF.

Proposition 1 *Let* $\mathbf{x} \sim N(\mathbf{0}, \mathbf{A}^{-1})$ *be a proper, zero-mean* GMRF *and let* $\mathbf{z}$ *be a vector of i.i.d. standard normal variables. Let* $\mathbf{A} = \mathbf{L}\mathbf{L}^T$ *be the*

Cholesky decomposition of $\mathbf{A}$*. Then*

$$\mathbf{x}_1 = \mathbf{A}^{-1/2}\mathbf{z} \quad \textit{and} \quad \mathbf{x}_2 = \mathbf{L}^{-T}\mathbf{z} \tag{4}$$

are (dependent) samples from $\mathbf{x}$*.*

3 Computing the correction

A method for computing the update $\boldsymbol{\delta x}$ using (2) was presented by Rue [2]. This method used the Cholesky decomposition that had already been computed during the unconditional sampling to solve the matrix equation

$$\mathbf{AX} = \mathbf{B}^T. \tag{5}$$

The update was then calculated using the formula $\boldsymbol{\delta x} = \mathbf{X}(\mathbf{BX})^{-1}\mathbf{z}$. This corresponds to a segregated method for solving the KKT system (3) [4]. As $\mathbf{A}$ and, therefore, $\mathbf{\Lambda}$ are large and sparse, it is natural to consider the use of Krylov subspace methods to solve for the correction. This approach is a natural extension of work presented by Simpson et al. [5].

The methods for solving the KKT system (3) are divided into two main classes: segregated methods that solve for $\mathbf{y}$ and $\boldsymbol{\delta x}$ separately, investigated in Sections 3.1 and 3.2; and coupled methods that solve for $\mathbf{y}$ and $\boldsymbol{\delta x}$ jointly, investigated in Section 3.3. Benzi, Golub and Liesen [4] surveyed methods for solving the KKT system. The main objective here is to explore the performance of Krylov subspace techniques for approximating the corrections. Two segregated methods and one coupled method are outlined in the following sections.

3.1 Segregated method 1: A multiple Krylov subspace approach

The first method for solving (5) is a direct extension of the method presented by Rue [2] to Krylov subspaces. This approach solves k linear systems $\mathbf{A}\mathbf{X}_{*i} = \mathbf{b}_i$ using Krylov subspace methods, where $\mathbf{X}_{*i}$ is the ith column of $\mathbf{X}$ and $\mathbf{B}^T = [\mathbf{b}_1, \ldots, \mathbf{b}_k]$. From the definition of the Frobenius norm, if the residual in the solution to each linear system satisfies $\|\mathbf{r}_m^{(i)}\|_2 \leq \epsilon$, the 2-norm of the residual satisfies $\|\mathbf{R}\|_2 = \|\mathbf{B}^T - \mathbf{A}\mathbf{X}\|_2 \leq \|\mathbf{R}\|_F \leq \sqrt{k}\epsilon$.

Before deriving an exact expression for ϵ, we first need to recall some basic facts about Krylov subspaces. The Krylov subspace of $\mathbf{A}$ generated by $\mathbf{b}$ is defined as $\mathcal{K}_m(\mathbf{A}, \mathbf{b}) = \text{span}\{\mathbf{b}, \mathbf{A}\mathbf{b}, \mathbf{A}^2\mathbf{b}, \ldots, \mathbf{A}^{m-1}\mathbf{b}\}$. This basis for $\mathcal{K}_m(\mathbf{A}, \mathbf{b})$ given in the definition is usually a poor basis for numerical computations and an orthonormal basis $\{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_m\}$ is preferred. This orthonormal basis is computed for symmetric $\mathbf{A}$ using the Lanczos decomposition

$$\mathbf{A}\mathbf{Q}_m = \mathbf{Q}_m\mathbf{T}_m + \beta_m\mathbf{q}_{m+1}\mathbf{e}_m^T, \tag{6}$$

where $\mathbf{Q}_m = [\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_m]$, $\mathbf{T}_m = \mathbf{Q}_m^T\mathbf{A}\mathbf{Q}_m$ is a symmetric tridiagonal matrix, $\mathbf{e}_m$ is the mth vector in the canonical basis for $\mathbb{R}^m$ and $\mathbf{Q}_m^T\mathbf{q}_{m+1} = \mathbf{0}$ [6]. This is essentially a partial orthogonal reduction of $\mathbf{A}$ to a tridiagonal form.

An approximate solution to $\mathbf{A}\mathbf{x} = \mathbf{b}$ is $\mathbf{x}_m = \mathbf{Q}_m\mathbf{y}$, where $\mathbf{y}$ is the solution to $\mathbf{T}_m\mathbf{y} = \|\mathbf{b}\|_2\mathbf{e}_1$. This particular choice of $\mathbf{y}$ leads to the Conjugate Gradient method for symmetric positive definite $\mathbf{A}$ and the approximate solution $\mathbf{x}_m = \mathbf{Q}_m\mathbf{y}$ is optimal in the norm induced by the $\mathbf{A}$-inner product on $\mathbb{R}^n$ [6]. The residual satisfies $\|\mathbf{r}_m\|_2 = \|\mathbf{b} - \mathbf{A}\mathbf{Q}_m\mathbf{y}\|_2 = \beta_m|\mathbf{e}_m^T\mathbf{y}|$ and, therefore, $\epsilon = \max_i\|\mathbf{b}^{(i)} - \mathbf{A}\mathbf{x}_m^{(i)}\|_2 = \max_i \beta_m^{(i)}|\mathbf{e}_m^T\mathbf{y}^{(i)}|$. It is possible to calculate the norm of the residual while building the Lanczos decomposition at little additional cost and, therefore, the subspace size m can be chosen to ensure ϵ is less than a prescribed tolerance [6].

Algorithm 1 summarises this method for calculating the correction.

Algorithm 1: A sequential Krylov subspace method for calculating the correction to a zero-mean GMRF conditioned on linear constraints.

Input: The size of the GMRF n, the precision matrix A, the constraint matrix B^T and a tolerance ϵ.

Output: The constraint correction δx and $X = A^{-1}B^T$.

1 **for** $i = 1, 2, \ldots, k$ **do**

2 Solve $AX_{*i} = b_i$ using the preconditioned conjugate gradient method until $\|r_m^{(i)}\| \leq \epsilon/\sqrt{k}$.

3 **end**

4 Form $S = BX$ and solve $Sw = z$.

5 Set $\delta x = Xw$.

3.2 Segregated method 2: a band Lanczos approach

The second method for approximating the solution to (5) exploits the fact that $k \ll n$ and builds a block Krylov subspace $\mathcal{K}_m(\mathbf{A}, \mathbf{B}^T)$. This approach uses more storage than method 1, but our numerical experiments show that it uses fewer matrix–vector products. Analogously to the single vector case, one computes an orthogonal basis for $\mathcal{K}_m(\mathbf{A}, \mathbf{B}^T)$ using a block Lanczos decomposition of the form

$$\mathbf{A}\mathbf{U}_m = \mathbf{U}_m\mathbf{T}_m + \mathbf{V}_{m+1}\mathbf{T}_{m+1,m}\mathbf{E}_m^T,$$

where $\mathbf{U}_m = [\mathbf{V}_1, \ldots, \mathbf{V}_m]$ is an orthonormal for $\mathcal{K}_m(\mathbf{A}, \mathbf{B}^T)$, $\mathbf{B}^T = \mathbf{V}_1\mathbf{W}$ is the truncated QR-factorisation of $\mathbf{B}^T$, $\mathbf{T}_m = \mathbf{U}_m^T\mathbf{A}\mathbf{U}_m \in \mathbb{R}^{mk\times mk}$ is a symmetric matrix with bandwidth $k+1$, $\mathbf{U}_m^T\mathbf{V}_{m+1} = \mathbf{0}$, and $\mathbf{E}_m$ is the last k columns the n-dimensional identity matrix [6]. Generalising the result in Section 3.1, the residual $\mathbf{R}_m = \mathbf{B}^T - \mathbf{A}\mathbf{X}_m$ in the approximation $\mathbf{X}_m = \mathbf{U}_m\mathbf{Y}$ where $\mathbf{Y}$ is the solution to $\mathbf{T}_m\mathbf{Y} = \left(\mathbf{W}^T, \mathbf{0}^T\right)^T$ satisfies

$$\|\mathbf{B}^T - \mathbf{A}\mathbf{U}_m\mathbf{Y}\|_2 = \|\mathbf{T}_{m+1,m}\mathbf{E}_m^T\mathbf{Y}\|_2 \leq \|\mathbf{T}_{m+1,m}\|_2\|\mathbf{E}_m^T\mathbf{Y}\|_2.$$

The extension to a restarted method is trivial and is outlined in Algorithm 2.

Algorithm 2: A block-Lanczos method (Ruhe's variant) for calculating the correction to a zero-mean GMRF conditioned on linear constraints.

Input: The size of the GMRF n, the precision matrix A and the constraint matrix B^T.

Output: The constraint correction δx and $X = A^{-1}B^T$.

Set $R = B^T$.
repeat
  Compute QR-decomposition $R = QW$.
  Use Ruhe's variant of the Block Lanczos Method [6] to form $AU_m = U_m T_m + V_{m+1} T_{m+1,m} E_m^T$.
  Calculate $Y = T_m^{-1} \begin{pmatrix} W \\ 0 \end{pmatrix}$.
  Set $X = X + U_m Y$.
  Set $R = V_{m+1} T_{m+1,m} E_m^T Y$.
until *convergence criterion is met [6]*;
Form $S = BX$ and solve $Sw = z$.
Set $\delta x = Xw$.

3.3 Method 3: a coupled approach

The assumption that $\text{rank}(\mathbf{B}) = k$ is sufficient for the KKT matrix $\mathbf{\Lambda}$ to be nonsingular, although it is not positive definite [4]. Therefore, we consider an iterative method for solving $\mathbf{\Lambda v} = \mathbf{b}$. Recalling that $k \ll n$, the product of the KKT matrix $\mathbf{\Lambda}$ with a vector is computed in, at most, $C(\mathbf{A}) + 2kn$ flops, where $C(\mathbf{A})$ is the cost of the matrix vector product involving $\mathbf{A}$. Due to the inexpensive matrix-vector products, the augmented linear system can be solved using a Krylov subspace method. Unfortunately, as $\mathbf{\Lambda}$ is not positive definite, the method described in Section 3.1 which is known as the

full orthogonalisation method (FOM) is no longer equivalent to the conjugate gradient method and is, therefore, not necessarily the solution from $\mathcal{K}_m(\mathbf{\Lambda}, \mathbf{b})$ of minimum norm. Despite this, the approximate solution generated using FOM is usually quite good. However, if we choose $\mathbf{y}$ to be the least squares solution to the overdetermined linear system

$$\begin{pmatrix} \mathbf{T}_m \\ \beta_m \mathbf{e}_m^T \end{pmatrix} \mathbf{y} = \|\mathbf{b}\|_2 \mathbf{e}_1 \,,$$

where $\mathbf{\Lambda Q}_m = \mathbf{Q}_m \mathbf{T}_m + \beta_m \mathbf{q}_{m+1} \mathbf{e}_m^T$, the resulting approximation $\mathbf{b}_m = \mathbf{Q}_m \mathbf{y}$ is optimal. This method, known as MINRES [7], is used in the case study.

3.4 Preconditioning

The segregated methods (Algorithms 1 and 2) are easily preconditioned using any symmetric preconditioner for $\mathbf{A}$ [6]. In the tests reported in the next section we used the incomplete Cholesky decomposition, denoted in Table 1 by IC(0), which resulted in significant speedup over the unpreconditined system. However, preconditioning coupled methods is a much more difficult proposition. Benzi, Golub and Liesen [4] gave a nice survey of preconditioners for the KKT system. If multiple samples are required, the cost of building a high quality preconditioner is negligible as it is averaged out over the number of samples. An alternative preconditioning strategy might use the spectral information generated in the previous subspaces to build an adaptive preconditioner in the spirit of Burrage and Erhel [8]. These preconditioning methods will be investigated in future research.

4 Case study

For this example we simulated 1000 pseudo-random points $\{\mathbf{s}_j\}_{i=1}^{1000}$ in the unit square and used the precision matrix

$$A_{ij} = \begin{cases} 1 + |\phi| \sum_k \chi_{\{\|\mathbf{s}_k - \mathbf{s}_i\| < \delta\}}\,, & i = j\,, \\ -\phi \chi_{\{\|\mathbf{s}_j - \mathbf{s}_i\| < \delta\}}\,, & i \neq j\,, \end{cases}$$

where χ_A is the set indicator function [9]. The constraint matrix $\mathbf{B} \in \mathbb{R}^{10\times 1000}$ was generated randomly and $\mathbf{c}$ was a random vector in the range of $\mathbf{B}$.

The performance of the three methods is described in Table 1. In the table, m is the dimension of the Krylov subspace and r is the number of restarts performed. The importance of preconditioning the segregated methods cannot be overstated: the unpreconditioned methods require, at the very least, one an a half times the number of matrix-vector products as their preconditioned counterparts. The most efficient method tested (in terms of matrix-vector products) is the coupled method 3. Even without preconditioning, this method converges with $m = 100$. Of the segregated methods, it appears that, if if it is not restarted, the band-Lanczos method 2 is the fastest of the two investigated. However, when the band-Lanczos method 2 is restarted, the number of matrix vector products required for convergence increases rapidly. Conversely, the multiple Krylov subspace method 1 appears to require the storage of only a small number of vectors at the cost of increasing the number of matrix-vector products.

5 Conclusion

We presented three methods for sampling from a Gaussian Markov random field conditioned on linear constraints. The question to ask is *which method is best*? As is often the case with numerical methods, there is no simple

TABLE 1: Results of the three methods for calculating the correction applied to the case study.

Method	Parameters	M	M-V Prod	Update error
1	$m = 40$	IC(0)	400	4e-7
1	$m = 70$	None	700	2e-7
2	$m = 150\,,\ r = 0$	IC(0)	150	3e-7
2	$m = 50\,,\ r = 17$	IC(0)	850	7e-7
2	$m = 70\,,\ r = 7$	IC(0)	490	2e-7
2	$m = 250\,,\ r = 0$	None	250	3e-7
2	$m = 50\,,\ r = 44$	None	2200	7e-7
2	$m = 70\,,\ r = 24$	None	1680	1e-7
3	$m = 100$	None	100	5e-7

answer. If one wishes to compute multiple, accurate samples, the segregated methods perform best as $\mathbf{X}$ can be approximated to high accuracy. This is then used to calculate additional samples for little additional cost [2]. When selecting a segregated method, one must decide whether less storage or fewer matrix-vector products are preferred and chose the method accordingly. If accuracy is not important, for example if the samples are then thresholded [9], then it may be cheaper to use the coupled method—especially if a good preconditioner is available.

Acknowledgements: We thank Professor Gene Golub for pointing out the link between the Schur complement reduction method and the KKT system.

References

[1] H. Rue and L. Held. *Gaussian Markov Random Fields: Theory and Applications*. Chapman and Hall/CRC, 2005. USA. C1042, C1043

[2] H. Rue. Fast Sampling of Gaussian Markov Random Fields. *J. R. Statist. Soc. B*, 63:325–338, 2001. doi:10.1111/1467-9868.00288 C1043, C1044, C1045, C1046, C1051

[3] N. Cressie. Reply to Wahba. *The American Statistician*, 44(3):256–258, 1990. C1044

[4] M. Benzi, G. Golub, and J. Liesen. Numerical solutions of saddle point problems. *Acta Numerica*, 14:1–137, 2005. doi:10.1017/S0962492904000212 C1044, C1045, C1048, C1049

[5] D. P. Simpson, I. W. Turner, and A. N. Pettitt. Fast sampling from Gaussian Markov random field using Krylov subspace approaches, *Scandinavian Journal of Statistics*, Submitted, 2008. C1044, C1045

[6] Y. Saad. *Iterative Methods for Sparse Linear Systems*. PWS Publishing Company, Boston, 1993. C1046, C1047, C1048, C1049

[7] C. C. Paige and M. A. Saunders. Solution of sparse indefinite systems of linear equations. *SIAM J. Numerical Analysis*, 12:617–629, 1975. doi:10.1137/0712047 C1049

[8] K. Burrage and J. Erhel. On the performance of various adaptive preconditioned GMRES strategies. *Numerical Linear Algebra with Applications*, 5(2):101–121, 1998. doi:0.1002/(SICI)1099-1506(199803/04)5:2¡101::AID-NLA127¿3.0.CO;2-1 C1049

[9] A. N. Pettitt, I. S. Weir, and A. G. Hart. A conditional autoregressive Gaussian process for irregularly spaced multivariate data with application to modelling large sets of binary data. *Statistics and Computing*, 12, 2002. doi:10.1023/A:1020792130229 C1050, C1051

Author addresses

1. **D. P. Simpson**, School of Mathematical Sciences, Queensland University of Technology, Brisbane, AUSTRALIA.
mailto:dp.simpson@student.qut.edu.au

2. **I. W. Turner**, School of Mathematical Sciences, Queensland University of Technology, Brisbane, AUSTRALIA.

3. **A. N. Pettitt**, School of Mathematical Sciences, Queensland University of Technology, Brisbane, AUSTRALIA.