arXiv:1107.2205v1 [stat.ME] 12 Jul 2011

Sequential Monte Carlo EM for multivariate probit models

BY GIUSI MOFFA

Institut für funktionelle Genomik, Universität Regensburg,
Josef Engertstraße 9, 93053 Regensburg, Germany
giusi.moffa@klinik.uni-regensburg.de

AND JACK KUIPERS

Institut für theoretische Physik, Universität Regensburg, D-93040 Regensburg, Germany
jack.kuipers@physik.uni-regensburg.de

SUMMARY

A Monte Carlo EM algorithm is considered for the maximum likelihood estimation of multivariate probit models. To sample from truncated multivariate normals we introduce a sequential Monte Carlo approach, while to improve the efficiency in driving the sample particles to the truncation region Student t distributions are invoked before taking their limit to a normal. After the initial sampling, a sequential Monte Carlo step can be performed to shift to new parameter values, recycling the samples and so reducing the computational cost. We discuss the identifiability issue and show that the invariance of the likelihood provides the means to ensure that constrained and unconstrained maximization are equivalent. Finally, for the multivariate probit model we derive a simple iterative procedure for either maximization which takes effectively no computational time. Applying our method to the widely used Six Cities dataset we find parameters which improve the maximum likelihood compared to other approaches.

Some key words: Maximum likelihood, Multivariate probit, Monte Carlo EM, adaptive sequential Monte Carlo

1. INTRODUCTION

Multivariate probit models, originally introduced by Ashford & Sowden (1970) for the bivariate case, are particularly useful tools to capture some of the correlation structure of binary, and more generally multinomial, response variables (McCulloch, 1994; McCulloch & Rossi, 1994; Bock & Gibbons, 1996; Chib & Greenberg, 1998; Natarajan et al., 2000; Imai & van Dyk, 2005). Inference for such models is typically computationally involved and often still impracticable in high dimensions. To mitigate these difficulties, Varin & Czado (2010) recently proposed a pseudo-likelihood approach as a surrogate for a full likelihood analysis. Similar pairwise likelihood approaches were also previously proposed by Kuk & Nott (2000) and Renard et al. (2004).

Due to the data augmentation nature of the problem, the estimation maximization (EM) algorithm (Dempster et al., 1977) is typically employed for maximizing the likelihood as its iterative procedure is usually more attractive than classical numerical optimization schemes. Each iteration consists of an estimation (E) step and a maximization (M) step and both should ideally be easy to implement. For cases in which the E step is analytically intractable, Wei & Tanner (1990) introduced a Monte Carlo version of the EM algorithm. Sampling from the truncated normal distributions involved is often based on Markov chain Monte Carlo (MCMC) methods and the Gibbs sampler in particular (see e.g. Geweke, 1991). As a different option we employ a sequential Monte Carlo (SMC) sampler (Del Moral et al., 2006) instead. Though originally introduced in dynamical scenarios (Gordon et al., 1993; Kitagawa, 1996; Liu & Chen, 1998; Doucet et al., 2001) as a more general alternative to the well known Kalman filter (Kalman, 1960), SMC algorithms can also be used in static inference (see e.g. Chopin, 2002) where artificial dynamics are introduced. When the target is a truncated multivariate normal, as in our case, an obvious sequence of distributions is obtained by gradually shifting the truncation region to the desired position. Since normal distributions decay very quickly in the tails, we propose to use flatter Student t distributions to drive the SMC particles more efficiently towards the end region, and only then take the appropriate limit to recover the required truncated multivariate normal.

The main difficulty in the M step rests with the computational complexity of standard numerical optimization over large parameter spaces, for which Meng & Rubin (1993) suggested a conditional maximization approach. A simple extension of their method allows us to define an iterative procedure to further maximize the likelihood at each M step. Though the likelihood converges, there is no guarantee that the parameters converge to a point (Wu, 1983). Restrictions to the parameter space have then been introduced to treat the identifiability issue where the data does not determine the parameters uniquely (McCulloch & Rossi, 1994; Bock & Gibbons, 1996), raising the problem of constrained maximization, normally significantly more difficult than unconstrained. However, the constraints are necessarily artificial and we show that the two maximizations can be made identical, and how they can be easily computed. Finally we validate our methods by comparison with previous approaches (Chib & Greenberg, 1998; Craig, 2008).

2. MULTIVARIATE PROBIT MODEL

2·1. *Notation*

Following the formulation in Chib & Greenberg (1998), denote by y^j a binary vector corresponding to the jth observation of a response variable Y^j with p components. Let x_i^j be a size k_i column vector containing the covariates associated to its ith component and define $X^j \triangleq \text{diag}((x_1^j)^{\mathrm{T}}, \ldots, (x_p^j)^{\mathrm{T}})$ as a $p \times k$ block diagonal matrix, with $k = \sum_{i=1}^{p} k_i$. A multivariate probit model with parameters $\beta \in \mathbb{R}^k$ and Σ, a $p \times p$ covariance matrix, can be specified

$$\mathrm{pr}\{Y^j = y^j \mid X^j, \beta, \Sigma\} = \int_{A_1^j} \cdots \int_{A_p^j} \phi_p(z^j; X^j\beta, \Sigma)\, \mathrm{d}z^j, \quad A_i^j = \begin{cases} (0, \infty) & \text{if } y_i^j = 1, \\ (-\infty, 0] & \text{if } y_i^j = 0, \end{cases} \tag{1}$$

where ϕ_p is the density function of a multivariate normal random variable with mean $\mu = X^j\beta$ and covariance matrix Σ. The vector of regression coefficient is $\beta = (\beta_1^{\mathrm{T}}, \ldots, \beta_p^{\mathrm{T}})^{\mathrm{T}}$, with each subvector $\beta_i \in \mathbb{R}^{k_i}$ corresponding to the ith component of the response variable. Naturally the situation where the β_i are all identical is a special case.

The probit model can also be understood in terms of a latent variable construction, where the observations are actually obtained from a sample of multivariate Gaussian vectors $\{z^1, \ldots, z^N\}$ from random variables $Z \sim \mathcal{N}(X^j\beta, \Sigma)$ as $y_i^j = I_{z<0}(z_i^j)$, with I the indicator function.

The covariance matrix Σ is a crucial parameter for the multivariate probit model and indirectly accounts for any dependence among the components of the response variable. The identity matrix corresponds to the assumption of independence and the model reduces to a collection of one dimensional cases, for which β can be easily estimated and used as starting point for more elaborate inference strategies. An alternative starting covariance matrix can be obtained (Emrich & Piedmonte, 1991) by pairwise approximations, which are likely however to lead to non positive definite matrices. ‘Bending’ techniques as suggested by Hayes & Hill (1981) are then necessary to ensure the positivity of the eigenvalues.

2·2. *Monte Carlo EM*

An EM algorithm (Dempster et al., 1977) allows us to build a sequence $\{\psi^m\}$ of estimated parameters such that the likelihood is non decreasing. In terms of the complete (Y, Z) and conditional $(Z \mid Y, \psi^m)$ missing data distributions for a given estimate ψ^m at iteration m and observed data Y, the log-likelihood is

$$l(\psi \mid Y) = \log(\mathrm{pr}\{Y \mid \psi\}) = Q(\psi, \psi^m) - H(\psi, \psi^m),$$

$$Q(\psi, \psi^m) = E_{Z|Y,\psi^m}\left[\log(\mathrm{pr}\{Y, Z \mid \psi\})\right], \quad H(\psi, \psi^m) = E_{Z|Y,\psi^m}\left[\log(\mathrm{pr}\{Z \mid Y, \psi\}))\right].$$

Having the difference of two logs means that the argument of each is only defined up to the same multiplicative factor. Jensen's inequality implies that $H(\psi, \psi^m) \leq H(\psi^m, \psi^m)$, so that the likelihood is certainly increased at each step if $Q(\psi^{m+1}, \psi^m) \geq Q(\psi^m, \psi^m)$, leading to a generalized EM. Ideally we wish to set ψ^{m+1} to the value of ψ which maximizes $Q(\psi, \psi^m)$, as required by the actual EM.

For the multivariate probit model, in terms of the latent variables $Z^j \sim \mathcal{N}(X^j\beta, \Sigma)$ and letting $\psi = (\beta, \Sigma)$ be the parameter vector, the complete data log-likelihood function is

$$\log(\mathrm{pr}\{Y, Z \mid \psi\}) = \sum_{j=1}^{N} \log\left[I_{A^j}(z^j)\phi(z^j; X^j\beta, \Sigma)\right].$$

Using the cyclicity of the trace and ignoring some normalizing constants, the corresponding $Q(\psi, \psi^m)$ function (Chib & Greenberg, 1998) can be written as

$$Q(\psi, \psi^m) = -\frac{N}{2}\left[\log|\Sigma| + \mathrm{tr}\left\{\Sigma^{-1}\frac{1}{N}\sum_{j=1}^{N} E_{Z^j|Y^j,\psi^m}\left\{(Z^j - X^j\beta)(Z^j - X^j\beta)^{\mathrm{T}}\right\}\right\}\right]. \quad (2)$$

The second term of (2) is analytically intractable since it involves expectations with respect to high dimensional truncated multivariate Gaussian densities. In a Monte Carlo EM approach (Wei & Tanner, 1990) the expectations can be approximated as

$$E_{Z^j|Y^j,\psi^m}\left\{(Z^j - X^j\beta)(Z^j - X^j\beta)^{\mathrm{T}}\right\} \simeq \sum_{k=1}^{M} W^{j(k)}(Z^{j(k)} - X^j\beta)(Z^{j(k)} - X^j\beta)^{\mathrm{T}}, \quad (3)$$

over a weighted sample $\{W^{j(k)}, Z^{j(k)}\}_{k=1}^{M}$, possibly approximated, from $\pi(z^j \mid y^j, \psi^m) = \mathrm{TMN}(A^j, X^j\beta, \Sigma)$, a multivariate normal distribution truncated to the domain A^j.

3. SMC AND THE E STEP

3·1. *Sequential Monte Carlo for truncated multivariate normals*

Sequential Monte Carlo samplers (Del Moral et al., 2006) are a class of iterative algorithms to produce weighted sample approximations from a sequence $\{\pi_n\}$ of distributions of interest where the normalizing constant C_n need not be known, $\pi_n = \gamma_n/C_n$. For a given probability distribution π, one obtains a collection of weighted samples $\{W^{(k)}, Z^{(k)}\}$ such that $E_\pi(h(Z)) \simeq \sum_{k=1}^{M} W^{(k)}h(Z^{(k)})$, where M is the number of particles and h a function of interest. In a static scenario the main purpose is to obtain such an approximation from the last element of the targeted sequence.

In order to control for the degeneracy of the sample, resampling (see Douc et al., 2005, for a review of resampling schemes) is typically performed when the effective sample size (ESS), as defined by Liu & Chen (1998): $\mathrm{ESS}^{-1} = \sum_{k=1}^{M} (W_n^{(k)})^2$, falls below a given threshold $\mathrm{ESS}^* = rM$ (with $0 < r < 1$). The move from the target π_{n-1} to the target π_n is achieved by means of a transition kernel K_n, so that $Z_n^{(k)} \sim K_n(Z_{n-1}^{(k)}, \cdot)$, and updating the normalized weights

$$W_n^{(k)} \propto W_{n-1}^{(k)} \tilde{w}_n^{(k)}, \quad \tilde{w}_n(Z_{n-1}^{(k)}, Z_n^{(k)}) = \frac{\gamma_n(Z_n^{(k)}) L_{n-1}(Z_n^{(k)}, Z_{n-1}^{(k)})}{\gamma_{n-1}(Z_{n-1}^{(k)}) K_n(Z_{n-1}^{(k)}, Z_n^{(k)})}, \quad k = 1, \ldots, M.$$

The quantity L_{n-1} in the expression for the incremental weights $\tilde{w}_n^{(k)}$ is a backward kernel introduced by Del Moral et al. (2006) to address computational issues. A typical choice for K_n is given by MCMC kernels with π_n as an invariant distribution and in particular we adopt a random walk Metropolis Hastings kernel. The samples at a given iteration n are obtained by moving each particle k to a new location $Z_n^{(k)} = Y^k \sim \mathcal{N}(Z_{n-1}^{(k)}, \Sigma_n)$ with probability $\alpha^k = 1 \wedge \rho^k$ and leaving it unchanged otherwise, with $\rho^k = \pi_n(Y^k)/\pi_n(Z_{n-1}^{(k)})$. The covariance matrix $\Sigma_n = \kappa \widehat{\Sigma}_\pi$ in the random walk proposal is a scaled version of an approximation $\widehat{\Sigma}_\pi$ (typically obtained from the previously simulated sample) of the target covariance matrix. As extensively investigated in the MCMC literature (for example the original paper of Gilks et al., 1998; Haario et al., 2001; Atchadé & Rosenthal, 2005, or the more recent review of Andrieu & Thoms, 2008) the scaling factor κ can be adaptively tuned by monitoring the average empirical acceptance probability $\hat{\alpha}_n$ at iteration n. For the Metropolis Hastings transition kernel, this can be evaluated as $\hat{\alpha}_n = \sum_{k=1}^{M} W_n^{(k)} (1 \wedge \pi_n(Y_n^{(k)})/\pi_n(Z_{n-1}^{(k)}))$. Adaptation of the transition kernel specifically within SMC has recently been considered by Jasra et al. (2011).

3·2. *Multivariate normals via Students*

Since the probability of the random walk Metropolis to move towards the tails of a Gaussian distribution decreases exponentially, a SMC method involving normals may be highly inefficient in moving samples towards regions of low probability. To achieve higher rates of acceptance in the tails we suggest starting with a flatter distribution: the multivariate (of dimension p) Student t distribution $\mathcal{T}(\nu, \mu, \Sigma)$ with degree of freedom ν, mean vector μ and covariance matrix Σ, which can be defined (Nadarajah & Kotz, 2005) as

$$f(z) = \frac{\Gamma(\frac{\nu+p}{2})}{\Gamma(\frac{\nu}{2})(\pi\nu)^{p/2}|\Sigma|^{1/2}} \left[1 + \frac{1}{\nu}(z-\mu)^{\mathrm{T}}\Sigma^{-1}(z-\mu)\right]^{-\frac{\nu+p}{2}}. \tag{4}$$

Replacing the ν in the denominator inside the square brackets by $(\nu - 2)$, and correspondingly changing the normalization factor, would provide the Student distribution with a covariance of Σ. As it stands, the distribution in (4) actually has a covariance of $\nu\Sigma/(\nu - 2)$ which further increases the acceptance in the tails. Once in the region of low probability we allow the degree of freedom to grow to infinity ($\nu \to \infty$) so the distribution approaches a p-variate Gaussian with the same mean and covariance matrix Σ.

To sample in the region of interest A, we define a sequence of target distributions $\{\pi_n\}_0^T$ such that the first target is an unconstrained multivariate Student and the last one is the same distribution truncated to A. Quite naturally the intermediate distributions are defined in terms of intermediate target domains $\{A_n\}_0^T$, included in each other $A_{k+1} \subset A_k$, with $A_T \equiv A$ and $A_0 \equiv \mathbb{R}^p$. The local target π_n at iteration n of the SMC algorithm is then

$$\pi_n(z) = \frac{\gamma_n(z)}{C_n}, \qquad \gamma_n(z) = \left[1 + \frac{1}{\nu}(z-\mu)^{\mathrm{T}}\Sigma^{-1}(z-\mu)\right]^{-\frac{\nu+p}{2}} I_{A_n}(z),$$

where C_n is a normalizing constant which can be estimated (Del Moral et al., 2006) from

$$\widehat{C}_n = C_0 \prod_{i=1}^{n} \widehat{\frac{C_i}{C_{i-1}}}, \qquad \widehat{\frac{C_i}{C_{i-1}}} = \sum_{k=1}^{M} W_{i-1}^{(k)} \tilde{w}_i(Z_{i-1}^{(k)}, Z_i^{(k)}),$$

and C_0 follows from (4). This ultimately allows us to obtain the probabilities of the regions in (1) and hence the likelihood for the probit model.

After reaching the required region, we define another sequence of target distributions starting from the truncated Student and increasing the degree of freedom ν until it is large enough that we can replace the Student with the desired truncated multivariate normal. One could also vary both the truncation region and the degree of freedom concurrently in the sequence of target distributions, but since the main reason for introducing the flatter Student distribution is to aid moving to regions of low probability we chose this two-step approach.

3·3. *Adaptive approach to artificial dynamics*

Other than for tuning the transition kernel K_n, adaptive strategies can also be used to define the artificial dynamics leading to the distribution of interest π_T. We do not address the problem of finding the optimal path linking an initial measure π_0 to the target π_T on the space of distributions in the sense of Gelman & Meng (1998), who actually deal with this issue in relation to Monte Carlo sampling methods for the evaluation of ratios of normalizing constants. Here we assume instead that the functional form of the intermediate distribution is given and can be described in terms of a parameter θ. An adaptive strategy to move from π_0 to π_T is one that does not require the sampling points $\{\theta_n\}$ defining the intermediate targets $\{\pi_n\}$ to be fixed a priori, but allows us to determine them dynamically on the basis of the local difficulty of the problem.

Adaptation can be achieved by controlling some statistics related to the performance of the algorithm and evolving with the parameter θ, and the ESS introduced in subsection 3·1 is an ideal quantity to monitor. Theoretically we wish to solve

$$\mathrm{ESS}_n(\theta_n) - \mathrm{ESS}_A^* = 0, \tag{5}$$

where ESS_A^* is a value chosen to compromise between efficiency and accuracy. Inspired by the Robbins-Monro recursion (see for example Kushner & Yin, 2003, page 3) for stochastic approximation, and aiming at the dynamical design of a sequence which keeps the ESS on average close to the threshold ESS_A^*, we define the updating scheme

$$\theta_n = \left[\theta_{n-1} + \left(\zeta_n \frac{\widetilde{\mathrm{ESS}}_n - \mathrm{ESS}_A^*}{M} \vee \Delta\theta_{\min}\right)\right] \wedge \theta_T, \tag{6}$$

where $\widetilde{\mathrm{ESS}}_n$ is the value observed for ESS at iteration n and the division by the number of particles M is only introduced for scaling purposes. Taking the maximum between the correction term and $\Delta\theta_{\min}$ ensures that the resulting sequence approaches the final target monotonically, while taking the minimum with θ_T ensures that the sequence ends at the desired target π_{θ_T}. Theoretically the ESS should ideally be equal to the total number of particles M of the SMC sampler, but to promote motion as a compromise between accuracy and efficiency, the threshold ESS_A^* can be fixed as a fraction $r \in (0, 1)$ of M, namely $\mathrm{ESS}_A^* = rM$. The number of iterations needed to reach the target π_T is reduced for smaller r. Similar adaptive ideas have also recently been applied to inference for stochastic volatility models by Jasra et al. (2011).

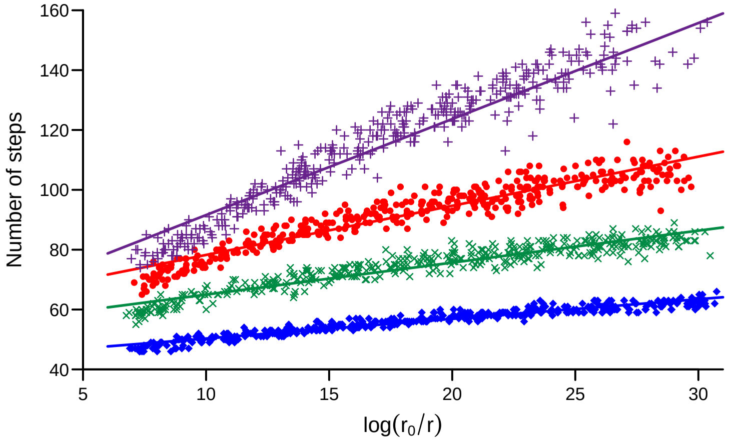

Fig. 1. The number of steps required for the SMC algorithm to reach a region of probability r for dimensions 2 (diamonds), 4 (crosses), 8 (dots) and 16 (pluses).

3·4. *Scaling behaviour*

The advantage of the SMC method, over alternatives which may be more efficient in sampling from truncated multivariate normals in low dimensions, is the scaling behaviour with the dimension p. Solving the adaptive equation (5) exactly means that we lose a fixed proportion of the probability mass at each iteration. The number of steps required to reach a target region of low probability r, then behaves like $\log(r)$, independently of p. This may not be true when using (6) as a numerical adaptive approximation to (5), especially as the number of steps for the adaption to settle grows linearly with p, so a weak dependence on the dimension could be expected.

A simulation study with targets of dimensions 2^n for $n = 1, \ldots, 4$ was performed. To limit the sources of variability, only one covariance structure was considered for the unconstrained distribution, with unit diagonals and a single non-zero off-diagonal element of 0·9. The SMC algorithm was initialized so that after an initial move the Student t target would be truncated to a region containing one quarter of the probability mass of an independent Gaussian, and we denote by r_0 the actual estimated probability. The cutoff for the final target, the same in all directions, was drawn so as to ensure that the log probability of a multivariate standard normal would be uniform on a given interval. The number of steps needed to reach the target are plotted against $\log(r_0/r)$ in Fig. 1, for 400 runs of a SMC sampler with 4000 particles for the different dimensions. A behaviour close to linear can be observed, though the offset increases by a factor of about 1·4 over the range of dimensions and the slope increases roughly linearly with p, which is likely due to any inexactness in the adaptation. The theoretical stability of these types of algorithms has recently been investigated in depth by Beskos et al. (2011).

3·5. *Sequential Monte Carlo EM*

After the initial sampling, which provides a particle approximation from the truncated target distribution corresponding to the initial parameter values, a sequential Monte Carlo approach can also be adopted to move between subsequent estimates $\psi^m = (\beta^m, \Sigma^m)$ without the need to perform the complete truncation again. Multiple sub-steps might be needed to update ψ^{m-1} to ψ^m, depending on how different the two corresponding targets are. For each observation j the local (to the EM iteration) initial and final distributions of the artificial sequence $\{\pi_n\}$ are defined as $\pi_0 = \mathrm{TMN}(A^j, X^j\beta^{m-1}, \Sigma^{m-1})$ and $\pi_T = \mathrm{TMN}(A^j, X^j\beta^m, \Sigma^m)$ respectively, while the parameter θ_n defining the intermediate targets moves from ψ^{m-1} to ψ^m, possibly in a single step. To avoid the situation where we would effectively need to move to a bigger region, which would prevent us from using a simplified version of the backward kernel L_n (see section 3.3.2.3 of Del Moral et al., 2006), we first rescale the previous sample to lie in the new truncation region, which can be done as long as the scaling factors are all positive, and then we apply the algorithm to update to the new covariance matrix.

4. Cycling conditional maximizations

4·1. *Two step maximization*

To overcome the difficulties associated with numerical maximization, Meng & Rubin (1993) suggested replacing the maximization over the full parameter space by a multi-step conditional maximization over several subspaces in turn. They treat the example of multivariate normal regression with incomplete data, where the parameters ψ^m at step m can again be split into Σ^m and β^m. This leads to a two-step conditional maximization which can be performed analytically. Keeping Σ fixed and maximizing equations (2) and (3) over β we obtain

$$\hat{\beta} = \left(\sum_{j=1}^{N} (X^j)^{\mathrm{T}} \Sigma^{-1} X^j \right)^{-1} \sum_{j=1}^{N} (X^j)^{\mathrm{T}} \Sigma^{-1} \sum_{k=1}^{M} \left(W^{j(k)} Z^{j(k)} \right), \tag{7}$$

so that by setting $\Sigma = \Sigma^m$ we can update the mean vector parameters for the next step as $\beta_{\mathrm{MR}}^{m+1} = \hat{\beta}$. Fixing β, the Σ which maximizes equation (2) is instead

$$\hat{\Sigma} = \frac{1}{N} \sum_{j=1}^{N} \sum_{k=1}^{M} W^{j(k)} (Z^{j(k)} - X^j \beta)(Z^{j(k)} - X^j \beta)^{\mathrm{T}}, \tag{8}$$

so that by setting $\beta = \beta_{\mathrm{MR}}^{m+1}$ we can the update the covariance matrix to $\Sigma_{\mathrm{MR}}^{m+1} = \hat{\Sigma}$ to give the new parameters ψ_{MR}^{m+1}. Though this two-step approach does not maximize ψ at each step, it removes the need for computationally intensive maximization and increases the likelihood at each step to ensure convergence of the (generalized) EM.

4·2. *Further maximization*

Since equation (8) maximizes $Q(\psi, \psi^m)$ over Σ for any value of β, we can substitute $\hat{\Sigma}$ into $Q(\psi, \psi^m)$ in (2) and obtain a function which only depends on β

$$\hat{Q}(\beta, \psi^m) = -\frac{N}{2} \log |\hat{\Sigma}| - \frac{Np}{2}, \tag{9}$$

Finding the value $\tilde{\beta}$ which maximizes (9) over β and setting $\tilde{\Sigma} = \hat{\Sigma}(\tilde{\beta})$ in (8) provides the new parameter $\tilde{\psi}$ which maximizes the likelihood. Performing the differential of (9) leads to the condition $\mathrm{tr}\{\hat{\Sigma}^{-1} \mathrm{d}\hat{\Sigma}\} = 0$. Though $\mathrm{d}\hat{\Sigma}$ is linear in the components of β, the inverse matrix $\hat{\Sigma}^{-1}$ leads to a system of coupled higher order polynomial equations. Solving these is impracticable, but one can proceed iteratively. As a starting point we can choose β^{m+1} from the conditional maximization of (7) so that $\hat{Q}$ has the value found from the two-step conditional maximization above (i.e. we set $\tilde{\beta}^{m+1,n} = \beta_{\mathrm{MR}}^{m+1}$ for $n = 0$). One option would be to perform Newton-Raphson iterations, but if the starting point is not too far from the maximum we can employ a simpler approximate maximization. Setting $\tilde{\Sigma}^{m+1,n} = \hat{\Sigma}(\tilde{\beta}^{m+1,n})$, we separate $\hat{\Sigma} = \tilde{\Sigma}^{m+1,n} + \Delta\tilde{\Sigma}$ and make the approximation $\log(1+x) \approx x$ to rewrite

$$\hat{Q}(\beta, \psi^m) \approx -\frac{N}{2} \mathrm{tr} \left\{ \log \tilde{\Sigma}^{m+1,n} \right\} - \frac{N}{2} \mathrm{tr} \left\{ (\tilde{\Sigma}^{m+1,n})^{-1} \Delta\tilde{\Sigma} \right\}.$$

Maximizing this is solving

$$-\frac{N}{2}\mathrm{tr}\left\{(\tilde{\Sigma}^{m+1,n})^{-1}\mathrm{d}\hat{\Sigma}\right\} = \mathrm{d}\beta^{\mathrm{T}}\sum_{j=1}^{N}\sum_{k=1}^{M} W^{j(k)}(X^j)^{\mathrm{T}}(\tilde{\Sigma}^{m+1,n})^{-1}(Z^{j(k)} - X^j\beta) = 0,$$

where we used the cyclicity of the trace to simplify. These are now linear equations in the components of β, which can easily be solved to find $\tilde{\beta}^{m+1,n+1}$. In fact the solutions are given precisely by (7) but now evaluated at the point $\tilde{\Sigma}^{m+1,n}$, so that $\tilde{\beta}^{m+1,n+1} = \hat{\beta}(\tilde{\Sigma}^{m+1,n})$ and $\tilde{\Sigma}^{m+1,n+1} = \hat{\Sigma}(\tilde{\beta}^{m+1,n+1})$ to give $\tilde{\psi}^{m+1,n+1}$.

4·3. *From generalized EM to EM*

Neatly, the logarithmic approximation and the two step conditional maximization of Meng & Rubin (1993) are equivalent when started at the same point, (ψ^m or ψ^{m+1}_{MR} for example). Because of the approximation, the values of β found in this way do not maximize $\hat{Q}$ but can be used as starting points for the next iteration to get closer to the maximum. In general with approximations the surety of convergence or even of not decreasing $\hat{Q}$ is lost, but, due to the equivalence above, each iteration does not decrease the likelihood and convergence follows from Meng & Rubin (1993). To complete the EM algorithm one can set $\psi^{m+1} = \lim_{n\to\infty}\tilde{\psi}^{m+1,n}$, and numerically stop the iterations when the Euclidean norm $||\tilde{\beta}^{m+1,n+1} - \tilde{\beta}^{m+1,n}||$ is small.

Though we have focused on multivariate normals, cycling through the conditional maximizations of Meng & Rubin (1993) until convergence can be applied more generally, turning the generalized EM of their single round procedure into an EM again. However, as they mention, it may be computationally advantageous to perform an E step between conditional maximizations when these are more demanding, and then the algorithm remains a generalized one.

5. IDENTIFIABILITY ISSUE

5·1. *Identifiability*

When the data is 'incomplete' maximization of the likelihood will not lead to uniquely identified parameters. Imposing constraints is a standard measure to ensure identifiability, but often with the effect of making the M step more involved (Bock & Gibbons, 1996; Chan & Kuk, 1997; Kuk & Chan, 2001). The issue is directly linked to symmetries of the likelihood, where it is invariant under some change of coordinates of the parameters. Focusing on *global* symmetries where the invariance of the likelihood $\mathcal{L}(\psi)$ does not depend on the particular value of $\psi \in \Psi$ we can decompose $\Psi = \Delta \times \Xi$ into an invariant space Δ and a reduced parameter space Ξ so that $\psi = (\delta, \xi)$ with $\delta \in \Delta$ and $\xi \in \Xi$. Due to the invariance of the likelihood over Δ

$$\mathcal{L}(\psi) = \mathcal{L}(\delta,\xi) = \hat{\mathcal{L}}(\xi) \quad \Rightarrow \quad \max_{\psi}\mathcal{L}(\psi) = \max_{\xi}\hat{\mathcal{L}}(\xi),$$

unconstrained maximization over the whole space Ψ is identical to performing it 'constrained' over the reduced space Ξ, with the difference that the parameters maximizing the likelihood in the larger space are $\psi^* = \Delta \times \xi^*$. Conversely, if the likelihood depended on some subspace of Δ then it would be identified during the maximization process. Therefore the dimension of Δ is the number of constraints needed to ensure identifiability.

In addition to any global symmetries, the likelihood function could also show a *local* symmetry so that $\hat{\mathcal{L}}(\xi)$ is maximized by a higher dimensional manifold rather than a single point (as discussed in Wu, 1983). In principle a local change of variables is possible (for example making the non-zero eigenvalues of the Hessian equal to -1 around the maximum) to decompose the space further, but in practice this presumes knowledge of the likelihood function. As above though, maximization over the subspace or the whole space are exactly equivalent because we still have (local) dimensions which do not affect the value of the likelihood.

Within the EM algorithm the identifiability issue becomes more subtle since the likelihood is not maximized directly, but by proxy through the function $Q(\psi, \psi^m)$. If this were to share the symmetries of the likelihood, then the simpler unconstrained maximization would be equivalent to the constrained version, as for the likelihood. If this is not the case, for example due to conditioning on the previous parameter value ψ^m, then any changes in Q arising from shifting ψ in the invariant space Δ of the likelihood must be exactly mimicked by changes in H. This spurious dependence can create differences between constrained and unconstrained maximization. The non decreasing behaviour of the likelihood remains preserved, since neither maximization decreases Q nor, because of Jensens's inequality, increases H. Hence either choice leads to the EM algorithm finding a maximum of the likelihood (though not necessarily the same one) and explains the conjecture of Bock & Gibbons (1996); Chan & Kuk (1997) and the agreement between constrained and unconstrained maximization found in Kuk & Chan (2001).

5·2. *Identifiability for the multivariate probit model*

In the multivariate probit model, the symmetries are related to the invariance of the likelihood under a rescaling of the coordinates of the normal variables. The full parameter space Ψ comprises $p(p+1)/2$ entries from the covariance matrix Σ and k regression coefficients from β. Scaling the coordinates $Z^j = DU^j$ by means of a diagonal matrix D with positive entries $(d_1, \ldots, d_p)$, transforms the covariance matrix to $\Omega = D^{-1}\Sigma D^{-1}$ and the vector β to $\lambda = (d_1^{-1}\beta_1^{\mathrm{T}}, \ldots, d_p^{-1}\beta_p^{\mathrm{T}})^{\mathrm{T}}$ but can easily be checked to leave the likelihood unchanged. Choosing the entries of D to be the square root of the diagonal elements of Σ reduces Ω to correlation form. The invariant space Δ can then be spanned by the p diagonal elements of Σ (i.e. $\delta_1 = 1/\sqrt{\sigma_{11}}$ etc.) while the reduced space Ξ includes the $p(p-1)/2$ rescaled upper triangular elements of Ω (i.e. $\omega_{ij} = \delta_i\delta_j\sigma_{ij}$) and the k elements of $\lambda = (\delta_1\beta_1^{\mathrm{T}}, \ldots, \delta_p\beta_p^{\mathrm{T}})^{\mathrm{T}}$.

The likelihood is not maximized directly, but through the function

$$Q(\psi, \psi^m) = \sum_{j=1}^{N} \int_{A^j} \log\left[\frac{1}{|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(z^{(j)} - X^j\beta)^{\mathrm{T}}\Sigma^{-1}(z^j - X^j\beta)\right)\right] \times \mathrm{TMN}(A^j, X^j\beta^m, \Sigma^m)\mathrm{d}z^j, \qquad (10)$$

which is only invariant under a change of integration variables $Z^j = DU^j$, for a diagonal matrix D, if we include a factor $|D|$ inside the log. Moreover, both ψ and ψ^m need to be scaled by same matrix so that essentially $\delta = \delta^m$. Although both ψ and ψ^m have independent invariant spaces for the likelihood, the Q function ties them together in this apparent constraint. Chib & Greenberg (1998) therefore maximized inside the constrained space Ξ, while keeping $\delta_i = 1$. Denote by ψ_{c} the parameter value found under such constraints, and by ψ_{u} the one obtained through unconstrained maximization of Q. Clearly $Q(\psi_{\mathrm{u}}, \psi^m) \geq Q(\psi_{\mathrm{c}}, \psi^m)$, but if we project ψ_{u} to a point ψ_{p} in the constrained space Ξ so that $\delta_i = 1$ then $Q(\psi_{\mathrm{p}}, \psi^m) \leq Q(\psi_{\mathrm{c}}, \psi^m)$. Since the likelihood is invariant under this projection

$$Q(\psi_{\mathrm{u}}, \psi^m) - Q(\psi_{\mathrm{p}}, \psi^m) = H(\psi_{\mathrm{u}}, \psi^m) - H(\psi_{\mathrm{p}}, \psi^m),$$

and without any information on $H(\psi_{\mathrm{u}}, \psi^m) - H(\psi_{\mathrm{c}}, \psi^m)$ it is impossible to say which maximization increases the likelihood most and is to be preferred in that respect.

5·3. *Reintroducing invariance*

To remove the above ambiguity, Q can be redefined to respect the invariance of the likelihood, for example by replacing (Σ, β) in (10) by their projection (Ω, λ). Such a replacement effectively enforces invariance of the resulting function $\tilde{Q}$ with respect to a rescaling of (Σ, β), making constrained and unconstrained maximization identical. However, this is no longer true if we perform a (cyclical) two-step conditional maximization. With the replacement $\tilde{Q}$ becomes

$$\tilde{Q}(\psi, \psi^m) = -\frac{N}{2}\left[\log\frac{|\Sigma|}{|D|^2} + \mathrm{tr}\Big\{D\Sigma^{-1}D\hat{S}\Big\}\right], \tag{11}$$

$$\hat{S} \simeq \frac{1}{N}\sum_{j=1}^{N}\sum_{k=1}^{M} W^{j(k)}(Z^{j(k)} - D^{-1}X^j\beta)(Z^{j(k)} - D^{-1}X^j\beta)^{\mathrm{T}}, \tag{12}$$

with D a diagonal matrix whose elements are the square roots of the diagonal elements of Σ (so that $\Omega = D^{-1}\Sigma D^{-1}$). Though $\tilde{Q}$ may appear to be limited to the constrained space, it depends on the full parameter space when one of Σ or β are given. Assume that for given ψ^m and $\beta^{m+1} = \lambda^{m+1}$ we wish to find Σ^{m+1}. Constrained maximization enforces $\delta_i = 1$ to find $\Omega_{\mathrm{c}}^{m+1}$ and hence ψ_{c}^{m+1}. An unconstrained maximization allows δ_i to vary, leading to $\Sigma_{\mathrm{u}}^{m+1}$ and correspondingly to ψ_{u}^{m+1}, such that $\tilde{Q}(\psi_{\mathrm{u}}^{m+1}, \psi^m) \geq \tilde{Q}(\psi_{\mathrm{c}}^{m+1}, \psi^m)$. Because of the invariance, the projection of ψ_{u}^{m+1} does not now change $\tilde{Q}$ resulting in a point in the constrained space with a higher value. In fact β is only defined up to a scale, which need not be preserved during each conditional maximization, nor given the stochastic nature of the estimation step.

Fixing Σ, the value of β maximizing equations (11) and (12) is as in (7), but with an extra factor D before the sum over k. Maximization with fixed β over Σ can in turn be done in two steps. The differential $\mathrm{d}\Sigma$ is split into a diagonal and an off-diagonal part. The condition for the latter to vanish is that $(\Omega^{-1} - \Omega^{-1}\hat{S}\Omega^{-1})$ be itself a diagonal matrix. As long as the diagonal elements of $\hat{S}$ are not too far from 1, a solution can be found by a simple iterative approach starting from an arbitrary Ω_0 and then solving for the diagonal matrix A the linear equations

$$\Omega_{k+1} = \hat{S} + \Omega_k A \Omega_k, \tag{13}$$

so that Ω_{k+1} is in correlation form. For fixed D the steps above allow us to perform constrained maximization for both (11) and (2). If D can vary, for the diagonal elements of $\mathrm{d}\Sigma$ to vanish

$$A - I + \Omega^{-1}\frac{1}{N}\sum_{j=1}^{N}\sum_{k=1}^{M} W^{j(k)}(Z^j - D^{-1}X^j\beta)(Z^j)^{\mathrm{T}}, \tag{14}$$

must have zero along the diagonal; a linear equation in the inverse elements of D. The solution depends on Ω, which in turn depends (through $\hat{S}$) on D so to perform the unconstrained maximization of (11) over Σ for a given β we would need to cycle through solving (14) and (13). As such, the difference between constrained and unconstrained maximization is made transparent.

5·4. *Model constraints*

In practice some constraints might already be imposed at the modelling stage. Typical for the multivariate probit model is to require that all the regression vectors are identical: $\beta_i = \beta_1$, replacing $X^j\beta$ by $X_{\mathrm{c}}^j\beta_1$, with X_{c}^j a matrix whose ith row is $(x_i^j)^{\mathrm{T}}$. The conditional maximization steps in Section 4 then allow one to maximize over the constrained space of (Σ, β_1).

However, the invariance of the likelihood needs to be reconsidered in light of the new constraints, which are broken when scaling the coordinate directions, and hence the β_i, by different positive factors. The likelihood is now left unchanged, independently of X^j, only when rescaling all the directions by the same amount, corresponding to a one dimensional invariant space. A reduced space can be defined by fixing the first diagonal element of the covariance matrix to 1, call (Ω', λ_1') the corresponding parameters. An invariant $\tilde{Q}$ is obtained by replacing X^j,Σ and β in (10) by X_c^j, Ω' and λ_1' respectively and by setting all the elements of D in (11) to be the square root of the first element of Σ. Constrained and unconstrained maximization follow from subsection 5·3 but with the slight changes that only the first element of the matrix A in (13) is non-zero and just the trace of (14) needs to be 0.

The effect on the invariance of assuming equal regression coefficients across components seems to have been overlooked by Chib & Greenberg (1998) as they required Ω' to be in correlation form. Maximizing over an overly constrained space leads in general to a lower likelihood than when only imposing the conditions needed to ensure identifiability. Nevertheless, were the correlation form desired for modelling reasons, one can perform the maximization by setting D to be the identity matrix and using X_c^j in the formulae in subsection 5·3.

6. Comparison to existing approaches

6·1. *The data and model used*

To assess the performance of our method, we treat the widely analysed data set from the Six Cities longitudinal study on the health effects of air pollution, for which a multivariate probit model was considered by Chib & Greenberg (1998), who conducted both Bayesian and non-Bayesian analysis. Later Song & Lee (2005) proposed a confirmatory factor analysis for the same model. More recently Craig (2008) used the example as a test case for his new method of evaluating multivariate orthant probabilities.

The study was meant to model a probabilistic relation over time between the wheezing status of children, the smoking habit of their mother during the first year of observation and their age. In particular the subset of data considered for analysis refers to the observation of 537 children from Stueberville, Ohio. The wheezing condition y_i^j of each child j at age $i \in \{7, 8, 9, 10\}$ and the smoking habit h^j of their mother are recorded as binary variables, with value 1 indicating the condition (wheezing/smoking) present. Three covariates are assumed for each component i, namely the age $x_{i1}^j = i - 9$ of child j centred at 9, the smoking habit $x_{i2}^j = h^j$ and an interaction term $x_{i3}^j = (i - 9)h^j$ between the two. A probit model can then be constructed

$$\mathrm{pr}\{y_i^j = 1\} = \mathrm{pr}(z_i^j > 0) = \Phi(\beta_0 + \beta_1 \cdot x_{i1}^j + \beta_2 \cdot x_{i2}^j + \beta_3 \cdot x_{i3}^j),$$

where z_i^j is the i-th component of a multivariate random variable $Z^j \sim \mathcal{N}(X_c^j\beta, \Sigma)$ and Φ is the cumulative distribution function of a standard normal random variable.

6·2. *Testing our algorithm*

To fit the model, a SMC sampler was implemented with the number of particles increasing from a starting value of 100 by 100 at each iteration up to 40, followed by 10 further steps of variance reduction (described below) with 4000 particles. Results for the constrained maximization are presented in Table 1 along with those of Chib & Greenberg (1998) and Craig (2008). Good agreement both for the estimates and the standard errors can be observed. Also given are average values of the corresponding log-likelihoods, easily obtained as a by-product of the SMC samplers, together with the standard deviation estimates over 40 runs. No real differences can be seen, with likelihoods comparable to, but slightly below, the estimate of -794·74 in Craig (2008).

Table 1. *Maximum likelihood estimates for the six cities dataset as obtained by using the constrained SMC algorithm with variance reduction and for a single run where the samples are recycled. Included for comparison are the results of Chib & Greenberg (1998) and Craig (2008).*

	Chib & Greenberg (1998)		Craig (2008)		variance reduction		recycled samples	
β_0	-1118	(65)	-1122	(62)	-1123	(62)	-1124	(62)
β_1	-79	(33)	-78	(31)	-78	(31)	-79	(32)
β_2	152	(102)	159	(101)	159	(101)	159	(101)
β_3	39	(52)	37	(51)	37	(51)	37	(51)
σ_{12}	584	(68)	585	(66)	582	(67)	582	(66)
σ_{13}	521	(76)	524	(72)	522	(72)	523	(71)
σ_{14}	586	(95)	579	(74)	575	(75)	572	(74)
σ_{23}	688	(51)	687	(56)	684	(57)	683	(56)
σ_{24}	562	(77)	559	(74)	557	(75)	554	(74)
σ_{34}	631	(77)	631	(67)	629	(68)	629	(68)
$l(\psi)$	-795·26	(0·75)	-795·21	(0·97)	-795·22	(0·82)	-795·30	(0·91)

The value in brackets next to each estimate is the estimated standard error. The values of the parameters (and their errors) have all been multiplied by 1000.

Table 2. *Example maximum likelihood estimates for the six cities dataset obtained using the unconstrained SMC algorithm for non-invariant Q, invariant $\tilde{Q}$ and by fixing $\sigma_{11} = 1$*

	β_0	β_1	β_2	β_3	σ_{12}	σ_{13}	σ_{14}	σ_{22}	σ_{23}	σ_{24}	σ_{33}	σ_{34}	σ_{44}	$l(\psi)$
Q	-1176	84	159	41	647	592	572	1208	855	619	1255	715	1001	-793·37
$\tilde{Q}$	-1235	-113	168	47	664	622	612	1275	921	683	1383	802	1146	-793·15
fixed σ_{11}	-1241	-116	169	48	666	626	615	1279	927	686	1395	809	1158	-793·07

The standard deviations of the log-likelihood estimates are 0·90, 0·75 and 0·70 respectively. The values of the parameters have all been multiplied by 1000.

Results from recycling the samples in a SMC EM algorithm as in subsection 3·5, with 4000 particles and 40 iterations are in the last column of Table 1. Since oscillations before the variance reduction step were around 0·001 between interations (with 4000 particles), parameter estimates when recycling the sample are essentially equivalent, at a much reduced computational cost.

An additional 20 iterations with 4000 particles are included before the variance reduction step for the unconstrained maximization, since it may take longer for the EM algorithm to explore a larger space. A fairly robust point is found with the non-invariant Q, while the invariant $\tilde{Q}$ seems to lead to a flatter likelihood neighbourhood, with the solution appearing more sensitive to the number of particles during earlier iterations or on imposing the constraint of fixing σ_{11} to 1. Results are given in Table 2, and again can be quite closely reproduced by recycling the samples in a sequential manner between parameter updates. Unfortunately, noise in the estimation of the observed information matrix overly influenced its numerical inversion, so that robust standard errors could not be obtained.

6·3. *Variance reduction*

To reduce the variance associated with the stochastic nature of the Monte Carlo E step, the parameter can be updated according to a stochastic approximation type rule

$$\psi^m = \psi^{m-1} + \zeta_m(\hat{\psi}^m - \psi^{m-1}) \equiv (1 - \zeta_m)\psi^{m-1} + \zeta_m\hat{\psi}^m,$$

where $\hat{\psi}_m$ is the actual estimate obtained from the M-step and $\zeta_m \in (0,1)$ a stepsize with the purpose of gradually shifting the relative importance from the innovation $(\hat{\psi}^m - \psi^{m-1})$ to the value of the parameter ψ_{m-1} learned through the previous iterations. The scheme is like taking a weighted average of the previous estimates, so we have referred to it as a 'variation reduction' step. This way the monotonicity property of the EM algorithm is not guarateed, but as long as the parameter remain within a neighbourhood of the maximum likelihood where it can be approximated quadratically, monotonicity follows so that in many practical cases this matter may not cause any issues.

REFERENCES

ANDRIEU, C. & THOMS, J. (2008). A tutorial on adaptive MCMC. *Statistics and Computing* **18**.

ASHFORD, J. R. & SOWDEN, R. R. (1970). Multi-variate probit analysis. *Biometrics* **26**, 535–546.

ATCHADÉ, Y. F. & ROSENTHAL, J. S. (2005). On adaptive Markov chain Monte Carlo algorithms. *Bernoulli* **11**, 815–828.

BESKOS, A., CRISAN, D. & JASRA, A. (2011). On the stability of sequential Monte Carlo methods in high dimensions. Preprint, arXiv:1103.3965.

BOCK, R. D. & GIBBONS, R. D. (1996). High-dimensional multivariate probit model. *Biometrics* **52**, 1183–1194.

CHAN, J. S. K. & KUK, A. Y. C. (1997). Maximum likelihood estimation for probit-linear mixed models with correlated random effects. *Biometrics* **53**, 86–97.

CHIB, S. & GREENBERG, E. (1998). Analysis of multivariate probit models. *Biometrika* **85**, 347–361.

CHOPIN, N. (2002). A sequential particle filter method for static models. *Biometrika* **89**, 539–551.

CRAIG, P. (2008). A new reconstruction of multivariate normal orthant probabilities. *Journal of the Royal Statistical Society Series B* **70**, 227–243.

DEL MORAL, P., DOUCET, A. & JASRA, A. (2006). Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society Series B* **68**, 411–436.

DEMPSTER, A. P., LAIRD, N. M. & RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* **39**, 1–38.

DOUC, R., CAPPE, O. & MOULINES, E. (2005). Comparison of resampling schemes for particle filtering. *Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis* , 64–69.

DOUCET, A., DE FREITAS, N. & GORDON, N. (2001). *Sequential Monte Carlo methods in practice*. Statistics for engineering and information science. Springer.

EMRICH, L. J. & PIEDMONTE, M. R. (1991). A method of generating high-dimensional multivariate binary variables. *The American Statistician* **45**, 302–304.

GELMAN, A. & MENG, X.-L. (1998). Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Statistical Science* **13**, 163–185.

GEWEKE, J. (1991). Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints. *Computing Science and Statistics: Proceedings of the Twenty-Third Symposium on the Interface* , 571–578.

GILKS, W. R., ROBERTS, G. O. & SAHU, S. K. (1998). Adaptive Markov chain Monte Carlo through regeneration. *Journal of the American Statistical Association* **93**, 1045 – 1054.

GORDON, N. J., SALMOND, D. J. & SMITH, A. F. M. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE-F* **140**, 107–113.

HAARIO, H., SAKSMAN, E. & TAMMINEN, J. (2001). An adaptive Metropolis algorithm. *Bernoulli* **7**, 223–242.

HAYES, J. F. & HILL, W. G. (1981). Modification of estimates of parameters in the construction of genetic selection indices ('bending'). *Biometrics* **37**, 483–493.

IMAI, K. & VAN DYK, D. A. (2005). A Bayesian analysis of the multinomial probit model using marginal data augmentation. *Journal of Econometrics* **124**, 311–334.

JASRA, A., STEPHENS, D. A., DOUCET, A. & TSAGARIS, T. (2011). Inference for Lévy driven stochastic volatility models via adaptive sequential Monte Carlo. *Scandinavian Journal of Statistics* **38**, 1–22.

KALMAN, R. E. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering* **82**, 35–45.

KITAGAWA, G. (1996). Monte Carlo filter and smoother for non-Gaussian non-linear state space model. *Journal of Computational and Graphical Statistics* **5**, 1–25.

KUK, A. Y. C. & CHAN, J. S. K. (2001). Three ways of implementing the EM algorithm when parameters are not identifiable. *Biometrical Journal* **43**, 207–218.

KUK, A. Y. C. & NOTT, D. J. (2000). A pairwise likelihood approach to analyzing correlated binary data. *Statistics and Probability Letters* **47**, 329–335.

KUSHNER, H. J. & YIN, G. G. (2003). *Stochastic approximation and recursive algorithms and applications*. Applications of mathematics. Springer, 2nd ed.

LIU, J. S. & CHEN, R. (1998). Sequential Monte Carlo methods for dynamic systems. *Journal of the American Statistical Association* **93**, 1032–44.

MCCULLOCH, C. E. (1994). Maximum likelihood variance components estimation for binary data. *Journal of the American Statistical Association* **89**, 330–335.

MCCULLOCH, R. & ROSSI, P. E. (1994). An exact likelihood analysis of the multinomial probit model. *Journal of Econometrics* **64**, 207–240.

MENG, X.-L. & RUBIN, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* **80**, 267–278.

NADARAJAH, S. & KOTZ, S. (2005). Sampling distributions associated with the multivariate t distribution. *Statistica Neerlandica* **59**, 214–234.

NATARAJAN, R., MCCULLOCH, C. E. & KIEFER, N. M. (2000). A Monte Carlo EM method for estimating multinomial probit models. *Computational Statistics & Data Analysis* **34**, 33–50.

RENARD, D., MOLENBERGHS, G. & GEYS, H. (2004). A pairwise likelihood approach to estimation in multilevel probit models. *Computational Statistics & Data Analysis* **44**, 649–667.

SONG, X.-Y. & LEE, S.-Y. (2005). A multivariate Probit latent variable model for analyzing dichotomous responses. *Statistica Sinica* **15**, 645–664.

VARIN, C. & CZADO, C. (2010). A mixed autoregressive probit model for ordinal longitudinal data. *Biostatistics* **11**, 127–138.

WEI, G. C. G. & TANNER, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association* **85**, 699–704.

WU, C. G. J. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics* **1**, 95–103.