J. R. Statist. Soc. B (2008)
70, *Part* 1, *pp.* 227–243

A new reconstruction of multivariate normal orthant probabilities

Peter Craig

University of Durham, UK

[Received September 2006. Final revision July 2007]

Summary. A new method is introduced for geometrically reconstructing orthant probabilities for non-singular multivariate normal distributions. Orthant probabilities are expressed in terms of those for auto-regressive sequences and an efficient method is developed for numerical approximation of the latter. The approach allows more efficient accurate evaluation of the multivariate normal cumulative distribution function than previously, for many situations where the original distribution arises from a graphical model. An implementation is available as a package for the statistical software R and an application is given to multivariate probit models.

Keywords: Cumulative distribution function; Fractional fast Fourier transform; Multivariate normal distribution; Multivariate probit model; Orthant probabilities; Orthoscheme; Polyhedral cones

1. Introduction

The pervasive role of the multivariate normal distribution in probability and statistics means that its cumulative distribution function is of intrinsic interest and there is value in any additional insight, be it algebraic, geometric or numerical. Moreover, from an applied viewpoint, the cumulative distribution function is a fundamental component of the multivariate probit models that were applied by Chib and Greenberg (1998), Gueorguieva and Agresti (2001), Papathomas and O'Hagan (2005) and others, where efficient computation is important.

We seek efficient accurate computation of orthant probabilities

$$P(X \geqslant 0) = P(X_i \geqslant 0, i = 1, \ldots, p) \qquad X = (X_1, \ldots, X_p)' \sim N_p(\mu, \Sigma) \tag{1}$$

for non-singular multivariate normal distributions. Note that, by changing the value of μ, we may obtain the multivariate cumulative distribution function at any point in R^p.

Many approaches have been suggested for numerical approximation of expression (1); for a review see the references in Gassmann *et al.* (2002) and Miwa *et al.* (2003). There is a line of geometric thinking running from Schläfli (1858) via Abrahamson (1964) and Miwa *et al.* (2003) to this paper although we exploit a somewhat different aspect of the geometry.

For many practical purposes, the efficient approach is via Monte Carlo (Genz, 1992, 1993) or random quasi-Monte Carlo (Genz and Bretz, 2002) sampling applied to a transformed version of the integral that is implied by expression (1), an implementation of which is available in R through the `mvtnorm` package (R Development Core Team, 2006; Genz *et al.*, 2005). However, the randomness that is inherent in such methods is sometimes more than a minor nuisance. For example, maximum likelihood estimation for multivariate probit models is much

Address for correspondence: Peter Craig, Department of Mathematical Sciences, University of Durham, South Road, Durham, DH1 3LE, UK.
E-mail: P.S.Craig@durham.ac.uk

1369–7412/08/70227

more straightforward when the multivariate normal distribution function may be calculated with sufficient precision to permit use of finite differences to approximate derivatives numerically with respect to parameters. The methods that are described here and in Miwa *et al.* (2003) make such computations possible.

A brute force approach would require the use of a p-dimensional numerical integration scheme. Efficient general purpose schemes are difficult to derive and implement even for moderately small p and are unsuitable for higher p as the computational storage and effort that are required grow exponentially in p. We shall describe an approach which uses one-dimensional convolutions and which may scale well in many contexts.

The basis of the approach here and in Miwa *et al.* (2003) is that, for any factorization $\Sigma = AA'$, we can write $X = \mu + AZ$ where $Z \sim N_p(0, I)$ and so $P(X \geqslant 0) = P(A^{-1}\mu + Z \in Q)$ where $Q = \{x : Ax \geqslant 0\}$ is a subset of R^p of a kind that is known as a polyhedral cone.

The interesting consequence is that we can exploit the geometry of cones to represent the original orthant probability in terms of others which are easier to approximate numerically. Miwa *et al.* (2003) showed that the orthant probability is easily computed when the cone is 'orthoscheme' and that any cone Q may be reconstructed in terms of at most $(p-1)!$ orthoscheme cones; the actual number required depends on the correlation structure of X. A cone is orthoscheme when there is a corresponding X which has a first-order moving average representation. We shall see in Section 3 that orthant probabilities are also fairly easily computed when the random vector may be represented as a first-order auto-regression and in Section 2 that there is a corresponding reconstruction for any orthant probability.

The intuition underlying the new reconstruction suggests that it should be particularly well suited to situations where X has a sparse graphical model. This frequently occurs when the original covariance matrix derives from hierarchical modelling or as a Bayesian network or from conditionally specified spatial modelling. Section 4 compares the efficiency of the two methods of reconstruction for some graphical models and shows that the new method is preferable in many situations.

The author has created a package called `orthants` (Craig, 2007) for the R statistical software, which implements the methods that are described here and in Miwa *et al.* (2003). The package may be used together with standard optimization functions in R to maximize the likelihood for multivariate probit models without having to turn to the hybrid Markov chain Monte Carlo expectation–maximization procedure that was introduced by Chib and Greenberg (1998). As an example, the 'six cities' data set of Chib and Greenberg (1998) is analysed in Section 5 by using this direct method.

2. Cone reconstructions

In this section we consider the two geometric approaches to reconstructing an arbitrary cone in terms of other cones which correspond respectively to orthant probabilities for moving average and auto-regressive sequences.

We are concerned with polyhedral cones. A polyhedral cone is a convex subset of R^p defined by $Ax \geqslant 0$ for some $n \times p$ matrix A, i.e. it is the intersection of n half-spaces. The rows $a'_1, \ldots, a'_n$ of A are known as the 'face vectors'. The ith face of the cone is the intersection of the cone with $a_i^{\perp}$ (the hyperplane through the origin orthogonal to a_i) and the boundary of the cone is the union of the faces. Faces are orthogonal if and only if their face vectors are. The face vectors may be scaled by positive constants without changing the cone.

A polyhedral cone can also be described in terms of a generating set. Vectors $v_1, \ldots, v_n$ in R^p generate a polyhedral cone by taking linear combinations of the vectors using only non-negative coefficients; if V is the matrix that is formed by taking the vectors as columns, the cone is $\{V\lambda : \lambda \geqslant 0\}$ and is unchanged if the vectors are scaled by positive constants.

In particular, a polyhedral cone that is defined by p linearly independent face vectors is also generated by its edge vectors $v_1, \ldots, v_p$. Geometrically, v_i lies in the one-dimensional intersection of all the faces other than the ith face and has positive inner product with a_i. We say that v_i is the edge opposite the ith face; $a_i^{\perp}$ is also the span of the edge vectors other than v_i. If we choose to scale each v_i so that $a_i' v_i = 1$, then V is the inverse of A.

The sum of cones, C_1 and C_2, is defined to be $C_1 + C_2 = \{x_1 + x_2 : x_i \in C_i\}$. If C_1 and C_2 are defined by generating sets, the union of those sets generates $C_1 + C_2$. If the subspaces of R^p that are spanned by C_1 and C_2, or equivalently by generating sets of the two cones, are linearly independent (the intersection contains only the zero vector), the values of x_1 and x_2 are unique for each element of $C_1 + C_2$.

Returning to multivariate normal distributions, the factorization $\Sigma = AA'$ is not unique but different choices correspond to rotations of R^p and the geometry of the cone is the same for any rotation. For a cone Q, the matrix A is not unique but the effect of different choices on Σ is to change only the order and standard deviations of the components of X; the correlation structure is invariant. In effect there is a one-to-one correspondence between cone geometry and correlation structure and so we can move back and forth between the two views. In the reconstructions that are described below, we first express an original orthant probability as a cone probability, then express the cone in terms of a number of other cones and finally express those cone probabilities as orthant probabilities for different multivariate normal distributions. Note that A (or equivalently $V = A^{-1}$) does determine Σ and Q.

2.1. Reconstruction using cones corresponding to moving averages (Miwa et al., 2003)

Miwa *et al.* (2003) expressed the original orthant probability in terms of what they called orthoscheme orthant probabilities. An orthoscheme probability is where Σ is tridiagonal (0 above and below the first off-diagonal band). The term orthoscheme was introduced by Schläfli (1858) and describes the geometry of the cone corresponding to the probability: faces i and j are orthogonal if $|i - j| > 1$. Stochastically, we can write $X_1 = \mu_1 + \alpha_1 Z_1$ and $X_i = \mu_i + \beta_{i-1} Z_{i-1} + \alpha_i Z_i$ for $i > 1$, i.e. X is a first-order moving average process. If some $\beta_i = 0$, Σ is block diagonal and the original orthant probability factors as the product of two orthoscheme orthant probabilities.

Further examination of the geometry reveals that a cone is orthoscheme and each corresponding β_i non-zero if, and only if, the edges of the cone (suitably scaled) form a sequence of orthogonal projections, i.e., for $i > 1$, v_i is the orthogonal projection of v_{i-1} onto the span of $v_i, \ldots, v_n$, or equivalently $v_{i-1} - v_i$ is orthogonal to v_j for $j \geqslant i$. For an orthoscheme cone, the moving average representation corresponds to factorizing $\Sigma = AA'$ so that entries in A are 0 apart from the diagonal and first subdiagonal: $A_{ii} = \alpha_i$ and $A_{i+1,i} = \beta_i$. Then, taking $\alpha_1^* = \alpha_1^{-1}$, $\gamma_i = (\alpha_i^* \alpha_i)^{-1}$ and $\alpha_i^* = -\beta_{i-1}\alpha_{i-1}^*/\alpha_i$, V is also lower triangular with $V_{ij} = \alpha_i^* \gamma_j$ for $i \geqslant j$ and, ignoring the scaling constants γ_j, column $i+1$ of V is obtained by zeroing the ith entry in column i. The reverse implication follows by rotating a sequence $v_1, \ldots, v_n$ of orthogonal projections so that the resulting matrix V is lower triangular.

Miwa *et al.* (2003) showed how an arbitrary cone may be reconstructed in terms of orthoscheme cones. The process sequentially increases the degree to which cones in the reconstruction are orthoscheme. Their language is that a cone is 'orthoscheme of order q' if faces i and j are orthogonal for $i \leqslant q$ and $j > i + 1$. The case $q = p - 2$ is the original definition of orthoscheme. Their presentation is quite algebraic and the following paragraphs emphasize the geometry. The important insight is that the reconstruction is based on constructing the sequence of orthogonal projections of an edge of the original cone.

Begin by picking one edge (v_1 say) to play a special role. Define a new edge v^* by projecting v_1 orthogonally onto the hyperplane $a_1^\perp$ that is spanned by the other edges; $a_1^\perp$ contains the face opposite v_1. Now define new polyhedral cones $Q_2, \ldots, Q_p$: Q_i has edges v_1, v^* and all the remaining original edges other than v_i. Each new cone is orthoscheme of order 1; it has two original face vectors (a_1 and a_i) and $p-2$ new face vectors that are orthogonal to a_1. The new cones divide into two groups; disregarding boundary sets of measure 0 which have no probability content, the difference between the union of one group and the other is the original cone and within each group the cones are disjoint. Hence it is easy to compute the probability content of the original cone from the probability contents of the new cones. If $a_1 \perp a_i$, the edges of Q_i are linearly dependent and so Q_i has dimension less than p; it has no probability content and may be ignored subsequently.

If $p > 3$, we can now reconstruct each Q_i in much the same way. Project v^* onto the span of the remaining original edges other than v_1 to obtain a new edge v_i^* and form $p-2$ new cones Q_{ij} which are orthoscheme of order 2. Q_{ij} has edges v_1, v^*, v_i^* and all the original edges other than v_i and v_j. It retains the original face vectors a_1 and a_i and one other face vector of Q_i (orthogonal to a_1) and has $p-3$ new face vectors which are orthogonal to both a_1 and a_i. As before, Q_i may be expressed in terms of the new cones and its probability content straightforwardly obtained from theirs.

If $p > 4$, continue by reconstructing each Q_{ij}; project v_i^* onto the span of all the remaining original edges other than v_1 to obtain a new edge v_{ij}^* and construct $p-3$ cones Q_{ijk} with edges v_1, v^*, v_i^*, v_{ij}^* and all the original edges other than v_i, v_j and v_k. In total we now have up to $(p-1)(p-2)(p-3)$ cones which are orthoscheme of order 3.

The original edge v_1 is being projected sequentially onto the span of smaller subsets of the other original edges. For any p, the process may be continued until the final projection is onto the span of two original edges, creating a total of at most $(p-1)!$ 'final stage' orthoscheme cones $Q_{i_1 i_2 \ldots i_{p-2}}$ having edges $v_1, v^*, v_{i_1}^*, \ldots, v_{i_1 i_2 \ldots i_{p-3}}^*$ and $v_{i_{p-1}}$ where i_{p-1} is the (unique) index distinct from $1, i_1, \ldots, i_{p-2}$.

Fig. 1 illustrates the general final projection, and also the initial projection when $p=3$. It shows a cross-section through the cone that is generated by the edges which are active in the final projection and how it is divided by the projection into two subcones, each having two

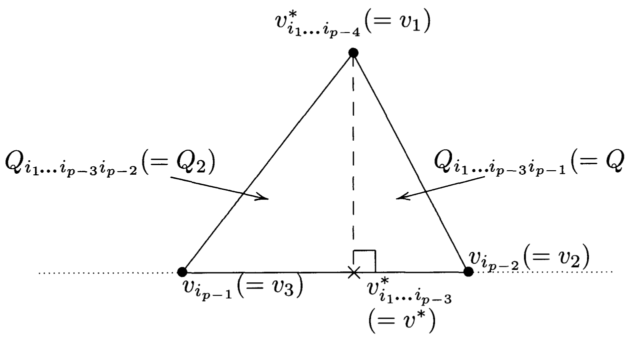
$v^*_{i_1 \ldots i_{p-4}} (= v_1)$
$Q_{i_1 \ldots i_{p-3} i_{p-2}} (= Q_2)$
$Q_{i_1 \ldots i_{p-3} i_{p-1}} (= Q_3)$
$v_{i_{p-2}} (= v_2)$
$v_{i_{p-1}} (= v_3)$
$v^*_{i_1 \ldots i_{p-3}}$
$(= v^*)$

Fig. 1. Illustration of the final stage of the Miwa *et al.* (2003) reconstruction, showing a cross-section of the cone that is generated by $v^*_{i_1 \ldots i_{p-4}}$, $v_{i_{p-2}}$ and $v_{i_{p-1}}$ in the subspace that is spanned by those vectors: cone $Q_{i_1 \ldots i_{p-3}}$ is represented by the full line triangle and is divided into two parts by the broken line, which shows the direction of orthogonal projection of $v^*_{i_1 \ldots i_{p-4}}$ onto the span of $v_{i_{p-2}}$ and $v_{i_{p-1}}$, the resulting edge being denoted $v^*_{i_1 \ldots i_{p-3}}$; the new cones $Q_{i_1 \ldots i_{p-3}, i_{p-2}}$ and $Q_{i_1 \ldots i_{p-3}, i_{p-1}}$ are represented respectively by the triangles falling to the left and right of the broken line; the parenthetic labelling illustrates the reconstruction for $p=3$

orthogonal faces. As shown, the original cone is the union of the two new cones but, if the new edge were to fall outside the original cone, the original would be expressed as the difference between the two new cones.

The description above assumes that $v^* \neq 0$; if $v^* = 0$, then $v_1 \perp v_j$ for $j > 1$, which implies that X_1 is independent of $X_2, \ldots, X_n$ and so $P(X \geqslant 0) = P(X_1 \geqslant 0)\, P(Y \geqslant 0)$ where $Y' = (X_2, \ldots, X_n)$ and the procedure may be applied to $P(Y \geqslant 0)$. Similarly if, at some later stage, $v^*_{i_1,\ldots,i_k} = 0$, the corresponding orthant probability factorizes into two orthant probabilities each of which may be computed by using the reconstruction described.

2.2. *Reconstruction using cones corresponding to auto-regressive sequences*

The new reconstruction that is described here expresses the original orthant probability in terms of orthant probabilities for a number of first-order auto-regressive (Markov) sequences. The original spark for doing so was the duality between moving average and auto-regressive processes: the covariance matrix of any moving average sequence is the precision matrix of some auto-regressive sequence and vice versa. An auto-regressive sequence is one for which $X_1 = \mu_1 + \alpha_1 Z_1$ and $X_i = \mu_i + \beta_i(X_{i-1} - \mu_{i-1}) + \alpha_i Z_i$ for $i > 1$. Thus, writing $Z = V(X - \mu)$, $V = A^{-1}$ is zero except on the diagonal and first subdiagonal and so the precision matrix $P = \Sigma^{-1} = V'V$ is tridiagonal. Geometrically, edges i and j of the corresponding cone are orthogonal if $|i - j| > 1$ and we say that the cone is edge orthoscheme.

Definition 1. A cone with edges $v_1, \ldots, v_p$ is said to be edge orthoscheme of order q if $v_i' v_j = 0$ for $i \leqslant q$ and $j > i + 1$. If $q = p - 2$, the cone is said to be edge orthoscheme.

Miwa *et al.* (2003) successively obtained cones of higher orthoscheme order. Here, we obtain cones of higher edge orthoscheme order. Whereas their reconstruction replaces one original edge by a new edge at each stage in the process, we shall replace an original face vector by a new face vector and possibly change the sign of others. Full details are given later in the theorems and corollaries but first we give a geometric description.

Start with an original edge v_1 and, for each pair of other edges, define $v_{ij} = v_{ji}$ to be a linear combination of v_i and v_j orthogonal to v_1. Except in degenerate cases, this exists and is unique apart from norm, which is unimportant, and orientation, which is defined in theorem 1. Form new cones $Q_2, \ldots, Q_p$ where Q_i has edges v_1, v_i and v_{ij} $(j \neq i)$. In Q_i, all edges except the second are orthogonal to the first, i.e. it is edge orthoscheme of order 1. The face vectors of Q_i are those of Q with two exceptions: the sign may change for some, and a^* replaces a_i; a^* is the projection of a_1 onto the span of $a_2, \ldots, a_p$, which is computable as a linear combination, orthogonal to v_1, of a_1 and v_1. The new cones divide into two groups; the union of one together with Q equals the union of the other, both unions being effectively disjoint. If $v_i \perp v_1$, Q_i is effectively empty and may subsequently be ignored.

Fig. 2 is primarily an illustration of theorem 1 but may also be seen as an illustration of this process for $p = 4$. It shows a cross-section through the cone that is generated by v_2, v_3 and v_4 in the space that is spanned by the same vectors and also the intersection of the cross-section with the hyperplane orthogonal to v_1. The original cone Q (which is represented by the triangle v_4, v_2, v_3) satisfies the equation $Q \cup Q_3 = Q_2 \cup Q_4$.

If $p > 3$, reconstruct each Q_i in a similar manner in terms of cones Q_{ij} where Q_{ij} has edges v_1, v_i and v_{ij} and edges orthogonal to v_i (and v_1) formed as linear combinations of v_{ij} and v_{ik} for k ranging from 2 to p omitting i and j. Apart from some sign changes, the face vectors of Q_{ij} are those of Q_i except that a_j has been replaced by a_i^*, which is the projection of a^* onto the span of the remaining original face vectors other than a_1. As before, Q_i may be expressed in terms of the new cones and its probability content obtained from theirs.

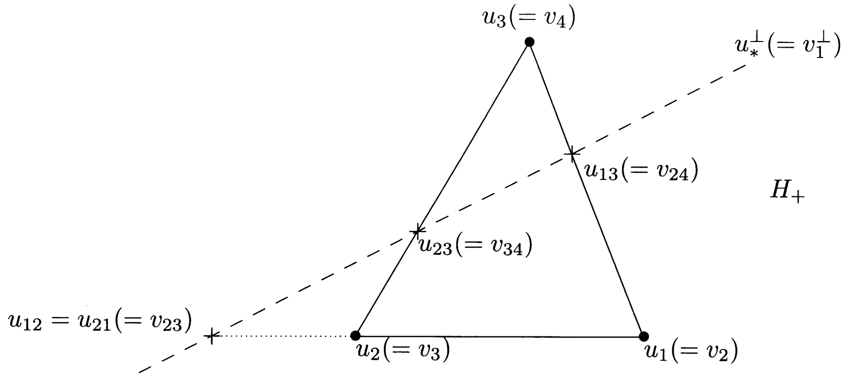

Fig. 2. Illustration of theorem 1 for $n=3$ and $m_-=m_+=2$, showing a cross-section of the cone that is generated by u_1, u_2 and u_3 in the subspace that is spanned by those vectors: the subset that is orthogonal to u_* is indicated by the broken line below which lies H_+; the original cone C is represented by the triangle with vertices u_1, u_2 and u_3; C_+ is represented by the quadrilateral with vertices u_1, u_2, u_{23} and u_{13} and together with the small triangle C_2 forms the larger triangle u_{12}, u_{13}, u_1 representing C_1; the parenthetic labelling illustrates the first reconstruction stage for $p=4$

If $p>4$, each Q_{ij} may be reconstructed in terms of cones Q_{ijk} with edges v_1, v_i, v_{ij} and v_{ijk} and edges orthogonal to v_{ij} (and v_1 and v_i) formed as linear combinations of v_{ijk} and v_{ijl} for l ranging from 2 to p omitting i,j and k.

The original face vector a_1 is being projected sequentially onto the span of smaller subsets of the other original face vectors and the precision matrices corresponding to the cones are becoming increasingly tridiagonal. For any $p\geqslant 3$, continue until at stage $p-2$ the projection is onto the span of two original face vectors and each new cone is edge orthoscheme. There are up to $(p-1)!$ such 'final stage' cones.

Theorem 1. Let $u_*,u_1,\ldots,u_n$ be linearly independent vectors in R^p $(n<p)$ such that $u_*'u_i\neq 0$ for some i, and let C be the cone that is generated by $u_1,\ldots,u_n$. Define $H_+=\{u:u_*'u\geqslant 0\}$ and $C_+=C\cap H_+$.

Order and scale $u_1,\ldots,u_n$ so that

$$u_*'u_j=\begin{cases}1 & j\leqslant m_+,\\ 0 & m_+<j\leqslant m_-,\\ -1 & j>m_-\end{cases}$$

and then, for each $i\leqslant m_+$, define u_{ij} orthogonal to u_* for $j\neq i$ by

$$u_{ij}=\begin{cases}u_i-u_j & \text{for } 1\leqslant j<i,\\ u_j-u_i & \text{for } i<j\leqslant m_+,\\ u_j & \text{for } m_+<j\leqslant m_-,\\ u_i+u_j & \text{for } m_-<j\leqslant n.\end{cases}$$

Finally, for $i\leqslant m_+$, define C_i to be the cone that is generated by u_i and u_{ij} $(j\neq i)$. Then, if $m_+>0$,

$$C_+\cup\bigcup_{\text{even } i}C_i=\bigcup_{\text{odd } i}C_i \tag{2}$$

where, ignoring sets of dimension less than n, the unions on both sides are disjoint. If $m_+=0$, C_+ has dimension less than n.

Note that changing the sign of u_* provides a similar reconstruction of C_- (the intersection of C with the complement of H_+), and so combining the two reconstructions provides a reconstruction for C.

Proof. Any $u \in C$ may be written as a linear combination of $u_1,\ldots,u_n$ with non-negative coefficients. When $m_+ = 0$, $u \in C_+$ implies that the coefficient of u_j is 0 for $j > m_-$ so C_+ has dimension m_- and the condition that $u_*' u_i \neq 0$ for some i implies that $m_- < n$.

Now assume that $m_+ > 0$. We first show that C_i is the intersection of H_+ and the set $\{\Sigma_k b_k u_k : b_k < 0 \text{ for } k < i, b_k > 0 \text{ for } k > i\}$. For $u \in C_i$ if and only if $u = c_i u_i + \Sigma_{j \neq i} c_j u_{ij}$ where the coefficients $c_1,\ldots,c_n$ are non-negative, expanding u_{ij} we obtain

$$u = -\sum_{k<i} c_k u_k + \sum_{k>i} c_k u_k + \Big(c_i + \sum_{k<i} c_k - \sum_{i<k\leqslant m_+} c_k + \sum_{k>m_-} c_k\Big) u_i.$$

Writing b_k for the coefficient of u_k, we see that constraining each $c_k \geqslant 0$ is equivalent to constraining $b_k \leqslant 0$ for $k < i$, $b_k \geqslant 0$ for $k > i$ and $b_i \geqslant -\Sigma_{k<i} b_k - \Sigma_{i<k\leqslant m_+} b_k + \Sigma_{k>m_-} b_k$; the last constraint may be expressed as $\Sigma_{k\leqslant m_+} b_k - \Sigma_{k>m_-} b_k \geqslant 0$, which is equivalent to $u_*' u \geqslant 0$.

For $1 \leqslant i \leqslant m_+ + 1$, define Δ_i to be the intersection of H_+ and the cone that is generated by $-u_1, -u_2, \ldots, -u_{i-1}, u_i, u_{i+1}, \ldots, u_n$. Then $C_+ = \Delta_1$ and $C_i = \Delta_i \cup \Delta_{i+1}$ for $1 \leqslant i \leqslant m_+$. Observe that $u_*' u < 0$ if $b_{m_+} > 0$ for u in the cone that is generated by $-u_1, \ldots, -u_{m_+}, u_{m_+ +1}, \ldots, u_n$ and so $b_{m_+} = 0$ in $\Delta_{m_+ +1}$, which implies that $\Delta_{m_+ +1} \subset \Delta_{m_+}$. Therefore $C_{m_+} = \Delta_{m_+}$. Hence, by grouping adjacent pairs of terms in the union $\cup_{i=1}^{m_+} \Delta_i$, we find that it may be expressed as either side of equation (2). However, ignoring sets of dimension less than n, the Δ_i are disjoint and so therefore are those unions.

Theorem 2. In addition to the conditions of theorem 1, suppose that there are $\tilde{u}_1,\ldots,\tilde{u}_q$ which, together with $u_*, u_1,\ldots,u_n$, form a basis. Let C^* be the cone that is generated by u_* and $\tilde{u}_1,\ldots,\tilde{u}_q$ and define $\tilde{C} = C^* + C$, $\tilde{C}_i = C^* + C_i$ and $\tilde{C}_+ = C^* + C_+$. Then either $m_+ = 0$, in which case $\tilde{C}_+$ has dimension less than p, or

$$\tilde{C}_+ \cup \bigcup_{\text{even } i} \tilde{C}_i = \bigcup_{\text{odd } i} \tilde{C}_i \tag{3}$$

where, ignoring sets of dimension less than p, the unions on both sides are disjoint. Changing the sign of u_* when applying theorem 1 leads to a reconstruction of $\tilde{C}_- = C^* + C_-$.

Proof. Since the vectors generating $\tilde{C}$ are a basis, the subspaces that are spanned by C^* and C are linearly independent. Moreover, the subspaces that are spanned by C_i and C_+ are subsets of the span of C and so are linearly independent of the span of C^*. Therefore, every vector lying in a cone appearing in equation (3) has a unique representation as the sum of a vector in C^* and a vector from the corresponding cone in equation (2). Consequently, equation (2) implies equation (3) and the disjointness property is inherited.

Corollary 1. Suppose that $\tilde{C}$ is edge orthoscheme of order q but not of order $q+1$. Then $\tilde{C}$ may be expressed as the difference of two disjoint unions of cones each of which is edge orthoscheme of order $q+1$.

Proof. Denote the edges of $\tilde{C}$ by $\tilde{u}_1,\ldots,\tilde{u}_q,u_*,u_1,\ldots,u_n$. Since $\tilde{C}$ is not orthoscheme of order $q+1$, $u_* u_i \neq 0$ for some i and so we may apply theorem 2. Since $\tilde{C}$ is edge orthoscheme of order q, each $\tilde{u}_j$ in theorem 2 is orthogonal to each u_k and hence to all the edges of C_i in theorem 1. Moreover, all edges of C_i other than u_i are orthogonal to u_* and so each $\tilde{C}_i$ is edge orthoscheme of order $q+1$. In effect u_* becomes $\tilde{u}_{q+1}$.

Corollary 2. Suppose that Q is edge orthoscheme of order q. Then the probability content of Q may be expressed as the difference of two sums of probabilities for cones which have edge orthoscheme order $q+1$.

3. Numerical computation

Recursive application of corollary 2 reconstructs an arbitrary cone in terms of at most $(p-1)!$ final stage edge orthoscheme cones. However, the rescaling of u_i in the statement of theorem 1 is there only to make the proof easier to read and if implemented naively can easily lead to floating-point overflow or underflow. It is wiser to stabilize the magnitudes of the vectors that are involved although there is no uniquely appropriate scaling. Our implementation scales all edge vectors to have unit norm. Thus, to find the edges of C_i, compute u_{ij} as a linear combination of u_i and u_j orthogonal to u_*, such as $(u_*'u_i)u_j - (u_*'u_j)u_i$, and then scale the result to have the correct sign and norm.

Theorem 1 does not determine the ordering of $u_1, \ldots, u_{m-}$; our implementation orders them to have decreasing inner product with u_*, which corresponds to increasing angle. We do not know this to be the best choice in general. However, if we exchange u_1 and u_2 in Fig. 2, we change the sign of u_{12}; it is difficult to visualize exactly what happens but it appears that cones C_1 and C_2 become much larger in relation to C_+ and the consequence is that the probability content of Q is expressed as the difference between two larger probabilities, which results in loss of precision.

The formal theory has been expressed entirely in terms of edges with no reference to faces; in effect it tells us how to compute V (and hence P) for each new cone. However, to compute the probability for an edge orthoscheme cone by using the methods in Section 3.2, we shall need to know the corresponding Σ and we also need to know the diagonals of Σ and P in Section 3.1. We could directly compute $\Sigma = AA'$ from $A = V^{-1}$ for each cone. However, it is not difficult to compute face vectors of each cone at the same time as the edges and, if we maintain the scaling where each has unit inner product with the opposite edge vector, we then know A and hence Σ. A further advantage is that, by computing A and V for each cone to some degree separately, we may detect numerical instability if the product of the resulting matrices A and V is significantly different from the identity.

Let the face vectors of $\tilde{C}$ in theorem 2 be $\tilde{a}_1, \ldots, \tilde{a}_q$, a_* and $a_1, \ldots, a_n$ labelled corresponding to the opposite edges and assume that edge and face vectors have been scaled as described in the previous paragraphs. Then, apart from rescaling to have unit inner product with opposite edge vectors, the face vectors of $\tilde{C}_i$ are the same except that

(a) for $j \neq i$, a_j is now opposite u_{ij} and needs to change sign when $j < i$ and
(b) the face vector that is opposite u_i, formerly a_i, is now proportional to $u_* - a_*$ when $q = 0$ and $u_* - a_* - (\tilde{u}_q' u_*)\tilde{a}_q$ for $q > 0$.

3.1. Testing and allowing for near orthogonality of edge vectors

In finite precision arithmetic, we cannot decide whether two edges are exactly orthogonal. Practically, we should treat edges as orthogonal if they are nearly so. Otherwise there are undesirable consequences:

(a) we must calculate orthant probabilities for cones which have nearly parallel edges and these are difficult to approximate numerically by using the methods that are described later,
(b) the decomposition becomes numerically unstable as the recursion progresses and
(c) we compute orthant probabilities for a larger than necessary number of final stage cones.

Thus, we must understand when it is reasonable to treat two edges as orthogonal when applying theorem 2. For multivariate normal X, $-P_{ij}/P_{jj}=-v_i'v_j/v_j'v_j$ is the coefficient of X_i in the conditional mean of X_j given all the other components of X. We now examine the sensitivity of the orthant probability to changes in the coefficient.

Without loss of generality (simple reordering), suppose that $j=p$. Write $W=X-\mu$ and $W_p|W_1,\ldots,W_{p-1}\sim N(\Sigma_{k<p}b_kW_k,\sigma^2)$. Then, denoting the standard normal probability density function by ϕ,

$$\left|\frac{\partial}{\partial b_i}P(X\geqslant 0)\right|\leqslant\int_{-\mu_1}^{\infty}\cdots\int_{-\mu_p}^{\infty}\left|\frac{\partial}{\partial b_i}p(w_1,\ldots,w_p)\right|\mathrm{d}w_1\ldots\mathrm{d}w_p$$

$$\leqslant\int_{R^p}\frac{|w_p-\sum_k b_kw_k|}{\sigma}\,\frac{|w_i|}{\sigma}\,\frac{1}{\sigma}\phi\left(\frac{w_p-\sum_{k<p}b_kw_k}{\sigma}\right)p(w_1,\ldots,w_{p-1})\,\mathrm{d}w_1\ldots\mathrm{d}w_p$$

$$=\frac{E[|Z_p|]\,E[|W_i|]}{\sigma}=\frac{2\Sigma_{ii}^{1/2}}{\pi\sigma}$$

and so the error that is induced by setting $b_i=0$ is less than $(2\Sigma_{ii}^{1/2}/\pi\sigma)|b_i|$.

For our choice of scaling, $\sigma^2=P_{jj}=v_j'v_j=1$ and $\Sigma_{ii}=a_i'a_i$. Therefore, we compare $v_i'v_j(a_i'a_i)^{1/2}$ with some threshold, the correct choice of which is hardware dependent. For standard double precision, 10^{-12} works well and permits the computation of orthant probabilities to approximately this accuracy. Larger thresholds restrict computational accuracy whereas smaller values lead to numerical instability.

3.2. Approximation of orthant probabilities for auto-regressive sequences

We wish to compute $P(X\geqslant 0)$ when $X_1,\ldots,X_p$ are an auto-regressive sequence. To make the calculation easier to structure, to simplify the subsequent presentation and to facilitate analysis of accuracy, we shall restrict to the case of unit variances and express the calculation in terms of $W=X-\mu$. Then, writing $\rho_i=\mathrm{corr}(X_i,X_{i+1})$ and $\sigma_i^2=1-\rho_i^2$, the sequence $W_1,\ldots,W_p$ is Markov and $W_{i+1}|W_i\sim N(\rho_iW_i,\sigma_i^2)$.

Taking $\psi_1(w_1)=p(w_1)=\phi(w_1)$, we sequentially compute approximations to the functions $\psi_2,\ldots,\psi_p$ where

$$\psi_{n+1}(w_{n+1})=\int_{-\mu_n}^{\infty}\cdots\int_{-\mu_1}^{\infty}p(w_1,\ldots,w_{n+1})\,\mathrm{d}w_1\ldots\mathrm{d}w_n$$

$$=\int_{-\mu_n}^{\infty}p(w_{n+1}|w_n)\int_{-\mu_{n-1}}^{\infty}\cdots\int_{-\mu_1}^{\infty}p(w_1,\ldots,w_n)\,\mathrm{d}w_1\ldots\,\mathrm{d}w_{n-1}\,\mathrm{d}w_n$$

$$=\frac{1}{\sigma_n}\int_{-\mu_n}^{\infty}\phi\left(\frac{w_{n+1}-\rho_nw_n}{\sigma_n}\right)\psi_n(w_n)\,\mathrm{d}w_n \tag{4}$$

$$=\frac{1}{|\rho_n|\sigma_n}\int_{-|\rho_n|\mu_n}^{\infty}\phi\left\{\frac{w_{n+1}-\mathrm{sgn}(\rho_n)u}{\sigma_n}\right\}\psi_n\left(\frac{u}{|\rho_n|}\right)\mathrm{d}u \tag{5}$$

so that $P(X\geqslant 0)$ is obtained by integrating ψ_p from $-\mu_p$ to ∞.

The Markov property has previously been exploited, as in equation (4), by Armitage *et al.* (1969), Williams (1971), Hirotsu *et al.* (1992) and others to simplify calculation of particular multivariate normal probabilities arising from statistical tests. Hayter (2006) provided a more general framework for exploiting factorization of joint densities for efficient computation of a variety of orthant and other probabilities. Armitage *et al.* (1969) also considered efficient computation for their special case which has additional smoothness properties.

In principle, calculation of equation (4) appears to require the numerical evaluation of a one-dimensional integral for each value of w_{n+1}. However, equation (5) is in the form of a convolution and may be amenable to calculation by using the fast Fourier transform (FFT). The FFT approach is efficient except when σ_n is small; moreover there is an efficient linear filtering algorithm to handle small σ_n. When ρ_n is 0 (or effectively so), equation (4) reduces to integrating ψ_n and restarting the process with ψ_{n+1} proportional to $\phi(w_{n+1})$.

For numerical calculations, we replace the lower limit of integration in equation (4) by $L_n = \max(-U, -\mu_n)$ and the upper limit by U. In doing so, we lose some probability from the final estimate of $P(X \geqslant 0)$ but the loss is less than $\Sigma_i P(|W_i| > U) = 2p\,\Phi(-U)$ where Φ is the standard normal distribution function. When linear filtering, we also truncate ϕ to 0 outside the interval $[-U, U]$ and again lose less than $2p\,\Phi(-U)$. In practice, it should be easy to choose U so that the total loss is negligible; for double precision, $U = 8$ works well.

The strategy is to approximate ψ_n on a grid (a sequence of equally spaced values) covering $[L_n, U)$ with a spacing Δ_n which we try to keep close to a fixed value Δ which is our basic control on the accuracy of approximation. The decision concerning when to use FFT or linear filtering relies on estimating the relative computational effort that is required, which depends on both hardware and the choice of FFT library. Details of this and other aspects of the algorithm may be found in a document distributed as part of the `orthants` (Craig, 2007) package for R.

3.2.1. *Using the fast Fourier transform*

The Fourier transform of ψ_{n+1} is

$$\Psi_{n+1}(t) = \int \exp(\mathrm{i}tx)\,\psi_{n+1}(x)\,\mathrm{d}x = \frac{1}{\sigma_n}\int \exp(\mathrm{i}tv)\,\phi\left(\frac{v}{\sigma_n}\right)\mathrm{d}v \int_{-\mu_n}^{\infty} \exp(\mathrm{i}t\rho_n y)\,\psi_n(y)\,\mathrm{d}y$$
$$= \exp\left(-\frac{\sigma_n^2 t^2}{2}\right)\Psi_n^+(\rho_n t)$$

where Ψ_n^+ is the Fourier transform of the truncation of ψ_n to 0 below $-\mu_n$. Thus we can compute ψ_{n+1} by finding the Fourier transform of the truncated version of ψ_n, stretching it by factor $1/\rho_n$, multiplying by $\exp(-\sigma_n^2 t^2/2)$ and finally inverting the Fourier transform.

Numerically, we can approximate this operation by using the fractional FFT of Bailey and Swarztrauber (1991) to compute Fourier transforms on suitable grids of values. The fractional FFT differs from the conventional FFT in that the output fundamental frequency (the spacing of the grid of values of t) may be freely chosen rather than being determined by the spacing in the x-grid and the number of grid points. As described in detail in section 13.9 of Press *et al.* (1993), we may increase the accuracy of the calculation of a Fourier integral for a fixed grid spacing by using Filon's (1928) method for Fourier integrals of continuous functions. The Filonized fractional FFT is most efficient when the grid length is a power of 2.

The main numerical difficulty is that the Fourier transform of a truncated smooth function decays to zero very slowly (asymptotically at rate $1/t$). Hence the decay to zero of $\exp(-\sigma_n^2 t^2/2)$ is the limiting control on the range over which we need to approximate Ψ_{n+1} to compute ψ_{n+1}. In effect, this also determines the range over which we must evaluate Ψ_n^+. For $\sigma_n = 1$, approximating Ψ_{n+1} on $[-U, U]$ should suffice but in general the range needs to be proportional to $1/\sigma_n$, which means that we must either use a larger grid or lose precision by increasing the spacing; we do the former as it is easy to use the Filonized fractional FFT to approximate Ψ_n^+ on a grid whose length is an integer multiple g_n of the length of the grid holding the truncation of ψ_n. In principle, g_n could become arbitrarily large. However, the difficulty only arises when

σ_n is near to 0. In such situations, the convolution may be computed directly by linear filtering of the function ψ_n evaluated on a grid.

3.2.2. Using linear filtering

Linear filtering is an efficient way to approximate equation (5) when σ_n is small since the integral will effectively only involve a small range of u for each x. Suppose that we have already calculated an approximation to $\psi_n(x)$ so that $\hat{\psi}_{n,k} \approx \psi_n(x_{n,k})$ for $0 \leqslant k < G$ where $x_{n,k} = L_n + k\Delta_n$. Linear filtering means that we plan (for most values of k) to compute $\hat{\psi}_{n+1,k} = \Sigma_{m=0}^{M_n} c_{n,m}\hat{\psi}_{n,k_0+k+m}$ where k_0 (depending on n) is an offset between the grids; for negative ρ_n, replace $\hat{\psi}_{n,k_0+k+m}$ by $\hat{\psi}_{n,k_0-k+m}$. For this to be possible, the grids that are used for x and u must have the same spacing and so $\Delta_{n+1} = |\rho_n|\Delta_n$.

We could easily obtain coefficients $c_{n,m}$ by using an elementary quadrature formula such as Simpson's rule to approximate equation (5). However, because σ_n is small, the integrand in equation (5) is not really sufficiently smooth for such a simple approach. The smoothness of the underlying multivariate normal probability density function means that ψ_n should be smooth apart from the step at $x = -\mu_n$, a view that is supported by some numerical experiments. This suggests the use of an approach to computing the convolution which takes advantage of the smoothness of ψ_n while allowing for the lack of smoothness in $\phi(\cdot/\sigma_n)$.

We replace ψ_n in equation (5) by a piecewise cubic interpolant. On each interval $(x_{n,k}, x_{n,k+1})$, approximate $\psi_n(x)$ by the cubic interpolant of $\hat{\psi}_n$ for the four nearest grid points. Hence, the contribution of the interval to $\hat{\psi}_{n+1,j}$ may be obtained as a linear combination of the values of $\hat{\psi}_n$ at those points. The coefficients may be calculated as fairly simple expressions in terms of the functions ϕ and Φ. To avoid excessive truncation error, those expressions need to be replaced by series approximations in certain situations.

4. Numerical examples

Table 1 compares accuracy and computation time for orthant probabilities computed by using the two reconstructions. It also shows the numbers N_{m} and N_{c} of final stage cones that are required respectively by the Miwa *et al.* (2003) method and the new method. The two examples considered are drawn from Miwa *et al.* (2003). In both cases, $p = 9$ and $\mu = 0$ and it can be shown that the true orthant probability is $1/10$. In the first example, the components of X are exchangeable (equicorrelated) with unit variance and correlation $\rho = 0.5$. In the second, Σ^{-1} is tridiagonal having unit diagonal and -0.5 on the bands.

In the exchangeable case, the two methods of reconstruction require the same number of final stage cones and the new method is typically 5–10 times as long for comparable accuracy, requiring a grid up to an order of magnitude larger. The second example is in fact an auto-regressive sequence. Consequently, the new method needs just one final stage cone which is the original cone whereas $N_{\mathrm{m}} = 323$. This disparity is enough to make the new method faster by an order of magnitude for comparable accuracy. Generally, doubling the grid size typically reduces error by an order of magnitude and roughly doubles the computation time although the computation time grows a little faster for the new method. It should be noted that, for the usual situation where the true answer is not known, additional computational effort would need to be expended to ascertain the accuracy of calculation for a particular grid size.

The time that is required for computing the reconstructions is negligible compared with the time that is spent computing final stage cone probabilities. Ratios of time per cone for the two methods are not entirely consistent owing to the variable grid sizing in our algorithm and the choice of FFT *versus* linear filtering but, across a range of examples including those in Table 1,

Table 1. Error and computation time depending on grid size by using the two methods of reconstruction to compute orthant probabilities for two nine-dimensional examples†

G	*Results for example (i)*				*Results for example (ii)*			
	Orthoscheme, $N_m = 40320$		*Edge orthoscheme, $N_c = 40320$*		*Orthoscheme, $N_m = 323$*		*Edge orthoscheme, $N_c = 1$*	
	Error	*Time (s)*	*Error*	*Time (s)*	*Error*	*Time (s)*	*Error*	*Time (s)*
32	-2.5×10^{-5}	8.2	5.0×10^{-3}	6.3	1.4×10^{-5}	0.07	-4.2×10^{-3}	0.00
64	-1.0×10^{-6}	16.2	1.9×10^{-3}	11.6	6.2×10^{-7}	0.12	-5.6×10^{-4}	0.00
128	-5.1×10^{-8}	32.3	1.4×10^{-4}	24.2	3.2×10^{-8}	0.25	-5.4×10^{-5}	0.00
256	-3.2×10^{-9}	64.4	4.7×10^{-6}	52.5	2.0×10^{-9}	0.50	-3.8×10^{-6}	0.00
512	-1.8×10^{-10}	128.7	2.9×10^{-7}	120.3	1.1×10^{-10}	0.99	-2.4×10^{-7}	0.01
1024	-1.1×10^{-11}	258.1	1.6×10^{-8}	279.4	6.9×10^{-12}	1.97	-1.7×10^{-8}	0.02
2048	-1.0×10^{-12}	518.0	1.0×10^{-9}	698.3	4.7×10^{-13}	3.96	-2.2×10^{-9}	0.04
4096	-9.3×10^{-14}	1037.9	-8.0×10^{-11}	2130.1	7.7×10^{-14}	7.94	-1.3×10^{-10}	0.08

†Example (i) is a unit variance exchangeable sequence with $\rho = 0.5$ and example (ii) is tridiagonal $P = \Sigma^{-1}$ having $P_{ii} = 1$ and $P_{i,i+1} = -0.5$. N_m and N_c are the numbers of final stage cones that are required by using respectively the reconstruction of Miwa *et al.* (2003) and the new reconstruction. G is the number of grid points that were used in computing final stage cones.

it appears that for the same grid size the new method takes typically twice as long per cone. Overall, it seems that $N_c/N_m = 0.1$ is approximately the point at which the two methods typically deliver the same accuracy for the same amount of effort.

Miwa *et al.* (2003) compared the performance of their method with the Monte Carlo method of Genz (1992) and found that for low accuracy the Monte Carlo method was much faster, whereas for high accuracy their method was usually faster unless the number of final stage cones became very great. We have compared both reconstruction approaches with the random quasi-Monte Carlo algorithm QRSVN of Genz and Bretz (2002) as implemented in the `mvtnorm` package for R. Table 1 shows that both reconstruction approaches typically obtain an order of magnitude reduction in error by doubling computational time. A small simulation study using the same two examples suggested that, for QRSVN, the typical magnitude of error for a particular probability is inversely proportional to the computation time raised to power $\frac{2}{3}$, which is better than ordinary Monte Carlo sampling but still implies that an order of magnitude reduction in error requires a 30-fold increase in effort. A direct comparison of typical error and the errors in Table 1 showed that, for example (i), QRSVN was slower than the Miwa *et al.* (2003) method for absolute error less than 10^{-7} and slower than the new method for error less than 10^{-8}; for example (ii), QRSVN was slower than either reconstruction method for error less than 10^{-4}.

4.1. Correlation structure and number of final stage cones

A key issue in comparing the efficiency of orthant probability calculations based on the two reconstructions is the number of final stage cones that are required by each method, which we lack the insight to compute theoretically. The following empirical results are subject to some numerical instability due to the threshold that was used to decide when to treat nearly orthogonal edges as orthogonal; the method of Miwa *et al.* (2003) requires a similar threshold.

In general, the number of cones that is required depends on which edge or face is chosen to be the special one to start the algorithm. For the new method, the ratio of worst to best number of cones varied by two orders of magnitude between examples but was generally greater than the typical value of 7 for the method of Miwa *et al.* (2003). The empirical evidence suggests that the optimal choice is a parentless or childless node for directed graphs and is a corner for the spatial lattice. The following results relate to the optimal choice.

Exchangeability leads to worst case behaviour for both methods, the number of cones being $N_e = (p-1)!$. For moving average and auto-regressive sequences respectively, the Miwa *et al.* (2003) and new methods require just a single cone. Duality of the two approaches to reconstruction indicates that the number of cones that was required by Miwa *et al.* (2003) for an auto-regressive sequence should be the same as for the new method applied to a moving average sequence. Empirically, some differences were observed and are believed to be due to different thresholding for near orthogonality in the two implementations. The original intuition underlying the development of the new reconstruction was that N_c would be substantially less than N_m for X having a sparse directed or undirected graphical model. This turns out to be true in many but not all cases.

Fig. 3 shows results for auto-regressive sequences, the directed net and binary tree graphs (Figs 4(a) and 4(b)) and lattice graphs (Fig. 4(c)) with undirected and directed edges. Fig. 3(a) compares the number of cones N_m that were required by Miwa *et al.* (2003) with N_e whereas Fig. 3(b) compares the number N_c that is required by the new method with N_m. We see that N_m is often very much lower than the worst case and N_c often much lower than N_m, both ratios generally decreasing as the dimension p increases for a particular graph. Overall, the ratios are lowest for the auto-regressive case, lower for the binary tree than for the directed net and generally lower for the directed net than for the spatial lattices where the ratios are nearly always slightly lower for the undirected than the directed version of the same graph. The shape of the lattice affects the outcome for lattices of the same dimension.

For 'loop' graphs (Fig. 4(d)), $p = m + n$. Investigating all loops with $p \leqslant 15$, we found that $N_c = 2^p/4$ whereas $N_m \approx 3.24^p/25$. Thus, for higher p, loop graphs would lie between auto-regressions and binary trees in Fig. 3(a) and near binary trees in Fig. 3(b).

All star graphs (Fig. 4(e)) having up to 13 nodes were investigated. Empirically, we found that N_c depends on the number of edges on all except the longest branch. Labelling the branches so that m_1 is the largest, N_c is the multinomial coefficient

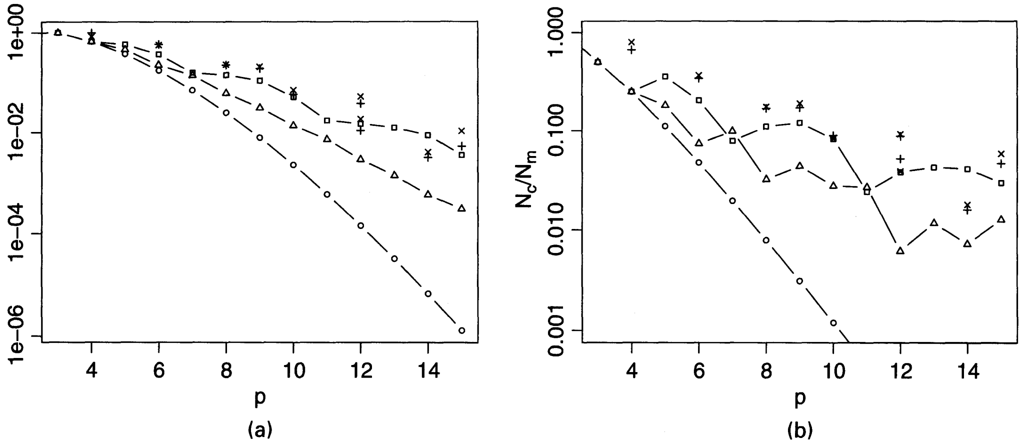

Fig. 3. Dependence of the optimal number of final stage cones on dimension p and various correlation structures (○, auto-regression; □, directed net; △, binary tree; +, lattice; ×, directed lattice): (a) number of cones, N_m, that are required by the method of Miwa *et al.* (2003) relative to the worst case number, $N_e = (p-1)!$; (b) number required by the new method, N_c, relative to N_m

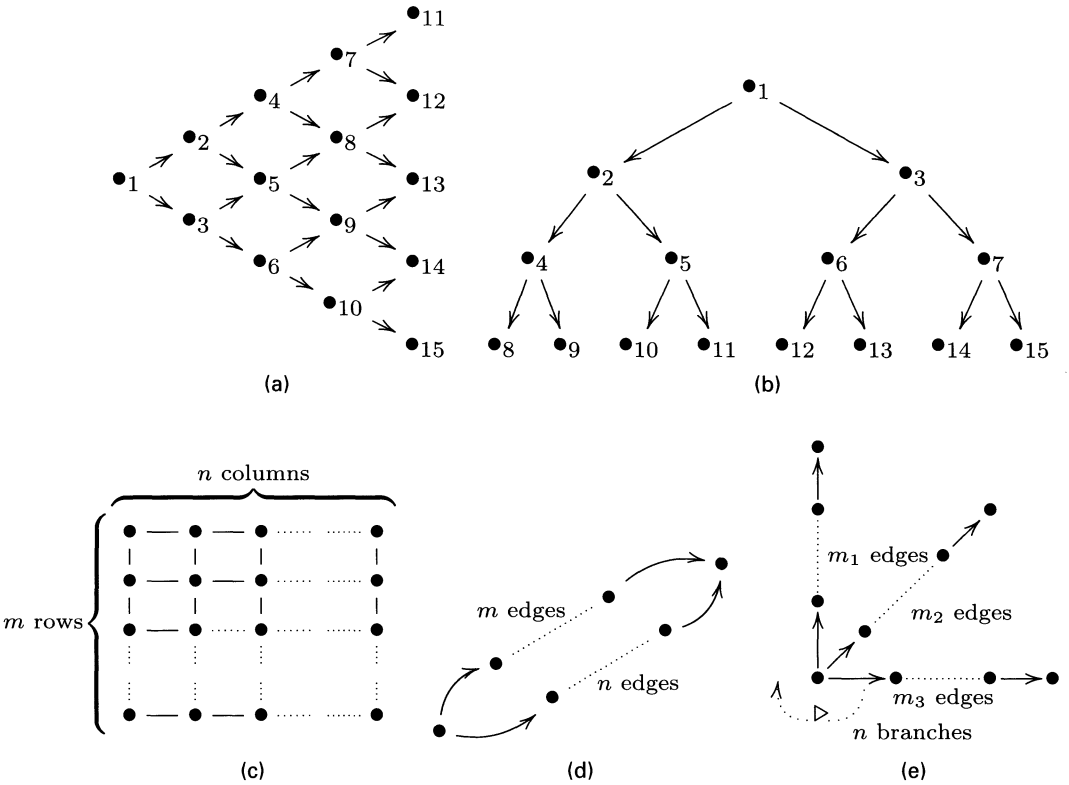

Fig. 4. Graphical models for which numbers of final stage cones are considered (for models (a) and (b), the graph that is used for $p < 15$ is obtained by deleting higher numbered nodes and the edges that are connected to them): (a) directed net ($p = 15$); (b) binary tree ($p = 15$); (c) lattice ($p = mn$); (d) loop ($p = m + n$); (e) star ($p = m_1 + \ldots + m_n + 1$)

$$(m_2 + \ldots + m_n)! \Big/ \prod_{i=2}^{n} m_i!.$$

For fixed p, the maximum of N_c is $(p-2)!$, which is attained by the graph having $p-1$ branches each of length 1; this is an example of a sparse graph which requires many cones. In contrast, an auto-regression is a star graph ($n = 1$) for which $N_c = 1$. For any star, merging branches i and j reduces N_c by a factor of at least $\binom{m_i+m_j}{m_i}$. In all cases with $p > 2$, $N_c < N_m$, and frequently much less. A purely empirical approximation to the number of Miwa cones is $N_m \approx 0.03\sqrt{(8^p N_c)}$, which was almost always accurate to within a factor of 2 although this is unlikely to extrapolate to substantially higher p.

The overall picture which emerges is that the new method is generally competitive for these graphs for $p \geqslant 7$ and nearly always preferable for high p although N_c may then be so large as to be impractical.

5. A statistical application

The frequently cited paper by Chib and Greenberg (1998) on multivariate probit modelling includes an example which we shall reconsider here. In probit modelling, binary outcomes are obtained by thresholding latent normally distributed continuous variables.

Chib and Greenberg (1998) considered a subset of the data from the 'six cities' longitudinal study on health effects of air pollution. Their data recorded whether or not each of 537 children suffered from wheeze at ages 7, 8, 9 and 10 years and whether or not the mother smoked during the first year of the study. Denoting the outcome (wheezing status) for child i at age j by y_{ij} $(j=7,8,9,10)$ and the mother's (binary) smoking status by s_i, the model is that $y_{ij}=1$ when $w_{ij}>0$ where the vector w_i follows a multivariate normal distribution with variance matrix Σ and the covariates are incorporated through $E[w_{ij}]=\beta_1+\beta_2(j-9)+\beta_3 s_i+\beta_4(j-9)s_i$. For identifiability, Σ is taken to be a correlation matrix and three different correlation structures were considered by Chib and Greenberg (1998): independence (no correlation parameters), exchangeability (one parameter) and the general correlation matrix (six parameters). Here, in addition, we consider auto-regressive structure (three parameters).

Chib and Greenberg (1998) considered maximum likelihood estimation, using a Monte Carlo expectation–maximization algorithm where the expectation step is with respect to the latent variables w_i and is approximated by simulation from the conditional distribution of w_i given y_i and the parameters. That conditional distribution is a truncated multivariate normal distribution; there is no direct sampling method and Gibbs sampling is used.

It is not trivial to implement the Chib and Greenberg (1998) method in a statistical package such as R or S-PLUS. However, the likelihood for a single observation may be calculated directly as an orthant probability and so computation of the likelihood function takes only a few lines of code in R provided that a suitable function is available for computing orthant probabilities. There is already the `mvtnorm` package, which is based on Monte Carlo integration (see Genz *et al.* (2005)). However, it is not suitable for use of standard optimization routines to maximize the likelihood, which require the objective function to be deterministic and which, for efficiency, rely on the use of finite differences to approximate partial derivatives of the likelihood function. The methods that are presented here and in Miwa *et al.* (2003) are suitable for that approach and the various models were easily estimated by using R.

Table 2 shows the results that were obtained for the four models fitted to the six cities data. The labelling $\mathcal{M}_i$ coincides with that used by Chib and Greenberg (1998) and the estimates here

Table 2. Maximum of likelihood, maximum likelihood estimates and standard errors for the four models that were fitted to the six cities wheeze data†

	Results for the following models:							
	$\mathcal{M}_1$		$\mathcal{M}_{\text{ar}}$		$\mathcal{M}_2$		$\mathcal{M}_3$	
β	−1.122	0.062	−1.130	0.062	−1.119	0.062	−1.126	0.047
	−0.078	0.031	−0.079	0.036	−0.078	0.030	−0.077	0.038
	0.159	0.101	0.155	0.100	0.161	0.100	0.171	0.076
	0.037	0.051	0.039	0.058	0.039	0.049	0.037	0.061
ρ	0.585	0.066	0.671	0.060	0.599	0.041		
	0.524	0.072	0.728	0.049				
	0.579	0.074	0.623	0.060				
	0.687	0.056						
	0.559	0.074						
	0.631	0.067						
$L(\hat{\theta})$	−794.74		−802.70		−797.67		−909.72	

†$\mathcal{M}_1$ is the unrestricted model, $\mathcal{M}_{\text{ar}}$ is the auto-regressive model, $\mathcal{M}_2$ is the exchangeability model and $\mathcal{M}_3$ is the independence model.

differ only by at most 0.001. The standard errors also agree except for $\mathcal{M}_2$ where it appears that their values are erroneous as they differ substantially from the Bayesian posterior standard deviations that they quoted for that model. Their method does not provide a value for the log-likelihood at the maximum although they could, for example, have used the methods that are available in the `mvtnorm` package to compute them with very good accuracy.

We note also that the availability of derivatives of the log-likelihood function makes it straightforward to implement for this problem a Bayesian analysis using a Laplace approximation to the posterior distribution as the basis for a Metropolis–Hasting independence sampler written in R; hence it was easy to verify the Bayes estimates that were provided in Chib and Greenberg (1998), which they obtained by random-walk Metropolis–Hastings steps with latent variables. The availability of the likelihood function also makes it possible to use the method of Chib and Jeliazkov (2001) to compute marginal likelihoods without having to use kernel density estimation of the posterior. The values that were obtained agreed with those of Chib and Greenberg (1998) up to the first decimal place for M_1 and M_3 but differed for M_4 where the value that was obtained here was −926.8 compared with their value of −931.2.

Finally, we note that Chib and Jeliazkov (2006) used auto-regressive normally distributed latent variables as the basis for a semiparametric analysis of longitudinal binary data and that such calculations should also benefit from the availability of efficient methods, such as proposed here, for computing orthant probabilities when the precision matrix is sparse. A detailed account is beyond the scope of this paper.

6. Possible developments

Geometrically, there are some interesting possible areas of further development. First, exchangeable sequences correspond to cones where the angle between each pair of edges is the same as is the angle between each pair of faces. It is conceivable that there might be another method of reconstruction which uses such cones. More generally, first-order moving averages, first-order auto-regressions and exchangeable sequences have in common the fact that the sequence of conditional means $E_k = E[X_k|X_1, \ldots, X_{k-1}]$ is Markov. In the case of first-order moving averages, $E_2, \ldots, E_p$ are independent, which is why orthant probability computation is fundamentally more efficient for moving average sequences than for auto-regressive sequences. The general situation in which the sequence $E_2, \ldots, E_p$ is Markov is for a (non-stationary) auto-regressive moving average ARMA(1,1) process. In principle, it should be possible to find a more efficient reconstruction of an arbitrary cone in terms of such cones. However, the geometry is more complex and it is not obvious how to exploit it.

Alternatively, we could seek ways to augment the quantities considered to increase computational efficiency. For example, the orthant probability for an exchangeable sequence may be computed efficiently by using Hayter's (2006) method if we augment the sequence by the 'population mean' so that the directed acyclic graph becomes a tree. The idea may be extended to provide efficient computation for the structures in Papathomas and O'Hagan (2005).

Acknowledgements

Thanks go to the Joint Editor, an Associate Editor and two referees for their helpful suggestions for improving the manuscript and to Dr T. Miwa for his computer programs.

References

Abrahamson, I. G. (1964) Orthant probabilities for the quadrivariate normal distribution. *Ann. Math. Statist.*, **35**, 1685–1703.

Armitage, P., McPherson, C. K. and Rowe, B. C. (1969) Repeated significance tests on accumulating data. *J. R. Statist. Soc.* A, **132**, 235–244.

Bailey, D. H. and Swarztrauber, P. N. (1991) The fractional Fourier transform and applications. *SIAM Rev.*, **33**, 389–404.

Chib, S. and Greenberg, E. (1998) Analysis of multivariate probit models. *Biometrika*, **85**, 347–361.

Chib, S. and Jeliazkov, I. (2001) Marginal likelihood from the Metropolis-Hastings output. *J. Am. Statist. Ass.*, **96**, 270–281.

Chib, S. and Jeliazkov, I. (2006) Inference in semiparametric dynamic models for binary longitudinal data. *J. Am. Statist. Ass.*, **101**, 685–700.

Craig, P. (2007) Orthants: multivariate normal orthant probabilities. University of Durham, Durham. (Available from `http://www.maths.dur.ac.uk/~dma0psc/orthants/`.)

Filon, L. (1928) On a quadrature formula for trigonometric integrals. *Proc. R. Soc. Edinb.*, **49**, 38–47.

Gassmann, H. I., Deák, I. and Szántai, T. (2002) Computing multivariate normal probabilities: a new look. *J. Computnl Graph. Statist.*, **11**, 920–949.

Genz, A. (1992) Numerical computation of multivariate normal probabilities. *J. Computnl Graph. Statist.*, **1**, 141–150.

Genz, A. (1993) Comparison of methods for the computation of multivariate normal probabilities. *Comput. Sci. Statist.*, **25**, 400–405.

Genz, A. and Bretz, F. (2002) Comparison of methods for the computation of multivariate t probabilities. *J. Computnl Graph. Statist.*, **11**, 950–971.

Genz, A., Bretz, F. and Torsten Hothorn, R. (2005) mvtnorm: multivariate normal and t distribution. Friedrich-Alexander-Universität, Erlangen-Nürnberg. (Available from `http://cran.r-project.org/src/contrib/Descriptions/mvtnorm.html`.)

Gueorguieva, R. V. and Agresti, A. (2001) A correlated probit model for joint modeling of clustered binary and continuous responses. *J. Am. Statist. Ass.*, **96**, 1102–1112.

Hayter, A. J. (2006) Recursive integration methodologies with statistical applications. *J. Statist. Planng Inf.*, **136**, 2284–2296.

Hirotsu, C., Kuriki, S. and Hayter, A. J. (1992) Multiple comparison procedures based on the maximal component of the cumulative chi-squared statistic. *Biometrika*, **79**, 381–392.

Miwa, T., Hayter, A. J. and Kuriki, S. (2003) The evaluation of general non-centred orthant probabilities. *J. R. Statist. Soc.* B, **65**, 223–234.

Papathomas, M. and O'Hagan, A. (2005) Updating beliefs for binary variables. *J. Statist. Planng Inf.*, **135**, 324–338.

Press, W. H., Flannery, B. P., Teukolsky, S. A. and Vetterling, W. T. (1993) *Numerical Recipes in C: the Art of Scientific Computing*, 2nd edn. Cambridge: Cambridge University Press.

R Development Core Team (2006) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Schläfli, L. (1858) On the multiple integral $\int^n dxdy\ldots dz$, whose limits are $p_1 = a_1x + b_1y + \cdots + h_1z > 0, p_2 > 0, \ldots, p_n > 0$ and $x^2 + y^2 + \cdots + z^2 < 1$. *Q. J. Pure Appl. Math.*, **2**, 269–301.

Williams, D. (1971) A test for differences between treatment means when several dose levels are compared with a zero dose control. *Biometrics*, **27**, 103–117.