Graphical models for skew-normal variates

A. CAPITANIO

University of Bologna

A. AZZALINI

University of Padua

E. STANGHELLINI

University of Perugia

ABSTRACT. This paper explores the usefulness of the multivariate skew-normal distribution in the context of graphical models. A slight extension of the family recently discussed by Azzalini & Dalla Valle (1996) and Azzalini & Capitanio (1999) is described, the main motivation being the additional property of closure under conditioning. After considerations of the main probabilistic features, the focus of the paper is on the construction of conditional independence graphs for skew-normal variables. Necessary and sufficient conditions for conditional independence are stated, and the admissible structures of a graph under restriction on univariate marginal distribution are studied. Finally, parameter estimation is considered. It is shown how the factorization of the likelihood function according to a graph can be rearranged in order to obtain a parameter based factorization.

Key words: conditional independence graph, graphical models, parameter based factorization, skew-normal distribution, skewness

1. Introduction

Graphical models currently represent one of the most active areas of statistical research, with an increasing impact in applications. In spite of substantial advances, however, the distributional assumption on the continuous components of these models is invariably the Gaussian one, as far as we are aware. This constraint is closely related to the strong convenience of working with a parametric family of distributions which allows simple formal manipulation of the variables under consideration. In particular, closure of the parametric class under marginalization and conditioning are very convenient features to rely on, and these properties seldom hold outside the class of multivariate normal distributions.

The present paper explores the potential usefulness in this context of the class of skew-normal distributions, recently studied by Azzalini & Dalla Valle (1996) and Azzalini & Capitanio (1999). This class extends the Gaussian one by introducing a vector parameter α which regulates the shape; when $\alpha = 0$ we are back to the normal distribution. Specifically, the density function of a skew-normal variate is

$$2\phi_k(y-\xi;\Omega)\Phi\big(\alpha^{\mathrm{T}}\omega^{-1}(y-\xi)\big), \quad y \in \mathbb{R}^k, \tag{1}$$

where $\phi_k(y;\Omega)$ is the density function of a k-dimensional $N_k(0,\Omega)$ variable, $\Omega = (\Omega_{ij})$ is a full rank covariance matrix, ξ is a k-vector of location parameters, Φ is the distribution function of $N(0,1)$, and

$$\omega = \mathrm{diag}(\Omega_{11}, \Omega_{22}, \ldots, \Omega_{kk})^{1/2}.$$

The class of densities (1) shares a number of properties with the normal one, which justify the use of the name skew-normal. Among those properties, we mention in particular closure under affine transformations, χ^2 distribution of certain quadratic forms, closure under marginalization and an approximate form of closure under conditioning; see Azzalini & Capitanio (1999, sect. 3 and 4) for details.

The above list of appealing formal properties makes this class a good candidate for work in graphical models. However, in this context, exact closure under conditioning is a very convenient feature to have available. This can be achieved by considering a slight extension of the class (1), namely

$$f(y) = \phi_k(y-\xi;\Omega)\Phi\big(\alpha_0 + \alpha^{\mathrm{T}}\omega^{-1}(y-\xi)\big)\big/\Phi(\tau), \quad y \in \mathbb{R}^k, \tag{2}$$

where τ is an additional real parameter and α_0 is a function of (Ω, α, τ) to be specified later. When $\tau = 0$, also $\alpha_0 = 0$ and (2) reduces to (1). The cost to be paid for gaining closure under conditioning is the loss of the χ^2 distribution of certain quadratic forms, which holds for (1).

The density form (2) arose in Azzalini & Capitanio (1999) from a conditioning operation on (1); see specifically their (13). Arnold & Beaver (2000) have examined (2), along with other extensions to (1), and noticed the property of closure under conditioning. Their work, however, is focused on a different direction, and has almost no overlap with this paper.

The next section presents a derivation of (2) and its basic probabilistic properties. This study is intended as a preliminary step to the central theme of the paper which is dealt with in section 3 and 4. In section 3, we focus on the use of density (2) in graphical models examining in particular the construction of conditional independence graphs and the derivation of their basic properties. This plan implies dealing, in section 4.1, with estimation of parameters, which is also of independent interest. In section 4.2, likelihood based factorizations are considered, in the sense examined by Cox & Wermuth (1999). The type of likelihood functions considered here have a clear similarity with those of their paper. The key distinction is that in our case there is a single Gaussian variable being dichotomized and it is unobserved; hence we deal only with continuous components, which are observed only conditionally on a given event.

The end conclusion of the paper is that the skew-normal distribution is a viable extension of the Gaussian one. It provides increased flexibility of the distributional assumption with limited additional complexity and computational burden. The present paper provides a set of basic results for the construction of graphical models. From this point, further progress is possible; potential directions of development include directed graphs and models for mixed distributions, skew-normal and discrete.

2. An extension of the skew-normal distribution

2.1. Definition and simple properties

Consider a $(k+1)$-dimensional normal random vector

$$W^* = (W_0, W_1, \ldots, W_k)^{\mathrm{T}} = \begin{pmatrix} W_0 \\ W \end{pmatrix} \sim N_{k+1}(0, \bar{\Omega}^*) \tag{3}$$

where

$$\bar{\Omega}^* = \begin{pmatrix} 1 & \delta^{\mathrm{T}} \\ \delta & \bar{\Omega} \end{pmatrix} \tag{4}$$

is a full-rank correlation matrix. The probability density function of $Z = (W|W_0 + \tau > 0)$ is

$$\phi_k(z;\bar{\Omega})\Phi(\alpha_0 + \alpha^{\mathrm{T}}z)/\Phi(\tau) \tag{5}$$

where

$$\alpha_0 = \tau\big(1 - \delta^{\mathrm{T}}\bar{\Omega}^{-1}\delta\big)^{-1/2}, \quad \alpha = \big(1 - \delta^{\mathrm{T}}\bar{\Omega}^{-1}\delta\big)^{-1/2}\bar{\Omega}^{-1}\delta. \tag{6}$$

For later use, it is also useful to write, after some algebraic manipulation,

$$\alpha_0 = \tau\big(1 + \alpha^{\mathrm{T}}\bar{\Omega}\alpha\big)^{1/2}, \quad \delta = \frac{1}{\big(1 + \alpha^{\mathrm{T}}\bar{\Omega}\alpha\big)^{1/2}}\bar{\Omega}\alpha.$$

The corresponding cumulant generating function is essentially as given by Azzalini & Capitanio (1999, sect. 4.2); by adapting their expression to the present notation, this becomes

$$K(t) = \frac{1}{2}t^{\mathrm{T}}\bar{\Omega}t + \zeta_0(\tau + \delta^{\mathrm{T}}t) - \zeta_0(\tau)$$

where $\zeta_0(t) = \log\{2\Phi(t)\}$. Simple differentiation gives immediately

$$\mathbb{E}\{Z\} = K'(0) = \zeta_1(\tau)\delta, \quad \mathrm{var}\{Z\} = K''(0) = \bar{\Omega} + \zeta_2(\tau)\delta\delta^{\mathrm{T}}.$$

where $\zeta_m(\cdot)$ is the mth derivative of $\zeta_0(\cdot)$.

For use in statistics, we also clearly need to introduce a location and a scale parameter. Hence define

$$Y = \xi + \omega Z$$

where $\xi \in \mathbb{R}^k$ and ω is a $k \times k$ diagonal matrix with positive diagonal elements. The density function of Y is then (2) where $\Omega = \omega\bar{\Omega}\omega$, and the cumulants generating function is

$$K_Y(t) = \xi^{\mathrm{T}}t + \tfrac{1}{2}t^{\mathrm{T}}\Omega t + \zeta_0(\tau + \delta^{\mathrm{T}}\omega t) - \zeta_0(\tau). \tag{7}$$

If a random variable Y has density function (2), we shall say that it has a (extended) skew-normal distribution with parameters $(\xi, \Omega, \alpha, \tau)$, and write

$$Y \sim \mathrm{SN}_k(\xi, \Omega, \alpha, \tau). \tag{8}$$

It can be shown that each of the four component parameters $(\xi, \Omega, \alpha, \tau)$ can be chosen independently of the others, which explains why the notation (8) has been adopted. In a number of aspects, it would be simpler to regard δ as the shape parameter, rather than α; for instance, the components of δ identify the marginal indices of skewness of the corresponding components of Y. Unfortunately, δ cannot be chosen independently of Ω, since we must ensure that $\bar{\Omega}^* > 0$, and this makes δ poorly suited as a parametrization component.

Some additional formal properties of the family (2) are given in the appendix. These results are not of direct use for the sequel of the paper but they are of independent interest.

2.2. Marginal and conditional distributions

Consider the following partition of Y and its parameters

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}, \quad \xi = \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}, \quad \Omega = \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{pmatrix}, \quad \alpha = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} \tag{9}$$

where Y_1 is of size h. The marginal distribution of Y_1 still belongs to the family (2); this statement follows both from the stochastic construction of Y itself and also from consideration of (7) evaluated at the point $(t_1, 0)$. Specifically, on partitioning $\bar{\Omega} = \omega^{-1}\Omega\omega^{-1}$ similarly to Ω, and writing

$$\bar{\Omega}_{11}^{-1} = (\bar{\Omega}_{11})^{-1}, \quad \bar{\Omega}_{22\cdot 1} = \bar{\Omega}_{22} - \bar{\Omega}_{21}\bar{\Omega}_{11}^{-1}\bar{\Omega}_{12},$$

the marginal distribution turns out to be

$$Y_1 \sim \mathrm{SN}_h(\xi_1, \Omega_{11}, \alpha_{1(2)}, \tau)$$

where

$$\alpha_{1(2)} = \frac{\alpha_1 + \bar{\Omega}_{11}^{-1}\bar{\Omega}_{12}\alpha_2}{\left(1 + \alpha_2^{\mathrm{T}}\bar{\Omega}_{22\cdot1}\alpha_2\right)^{1/2}}. \tag{10}$$

These formulae can also be obtained as a special case of the more general result on affine transformations of Y given in the appendix.

The density of the conditional distribution $(Y_2|Y_1 = y_1)$ is

$$\begin{aligned} f_{2|1}(y_2) &= \phi_{k-h}(y_2 - \xi_{2\cdot1}; \Omega_{22\cdot1}) \\ &\quad \times \Phi\{\alpha_0 + (\alpha_1^{\mathrm{T}} + \alpha_2^{\mathrm{T}}\bar{\Omega}_{21}\bar{\Omega}_{11}^{-1})\omega_1^{-1}(y_1 - \xi_1) + \alpha_2^{\mathrm{T}}\omega_2^{-1}(y_2 - \xi_{2\cdot1})\}/\Phi(\tau_{2\cdot1}) \\ &= \phi_{k-h}(y_2 - \xi_{2\cdot1}; \Omega_{22\cdot1})\Phi\{\alpha_0' + \alpha_2^{\mathrm{T}}\omega_2^{-1}(y_2 - \xi_{2\cdot1})\}/\Phi(\tau_{2\cdot1}) \end{aligned} \tag{11}$$

where

$$\xi_{2\cdot1} = \xi_2 + \Omega_{21}\Omega_{11}^{-1}(y_1 - \xi_1) \tag{12}$$

and

$$\begin{aligned} \Omega_{22\cdot1} &= \Omega_{22} - \Omega_{21}\Omega_{11}^{-1}\Omega_{12}, \\ \tau_{2\cdot1} &= \tau\left(1 + \alpha_{1(2)}^{\mathrm{T}}\bar{\Omega}_{11}\alpha_{1(2)}\right)^{1/2} + \alpha_{1(2)}^{\mathrm{T}}\omega_1^{-1}(y_1 - \xi_1), \\ \alpha_0' &= \tau_{2\cdot1}\left(1 + \alpha_{2\cdot1}^{\mathrm{T}}\bar{\Omega}_{22\cdot1}\alpha_{2\cdot1}\right)^{1/2}. \end{aligned}$$

From the pattern of (11), one notices that this is still of type (2), namely

$$(Y_2|Y_1 = y_1) \sim \mathrm{SN}_{k-h}(\xi_{2\cdot1}, \Omega_{22\cdot1}, \alpha_{2\cdot1}, \tau_{2\cdot1})$$

where

$$\alpha_{2\cdot1} = \omega_{22\cdot1}\omega_2^{-1}\alpha_2, \quad \omega_{22\cdot1} = (\Omega_{22\cdot1} \odot I)^{1/2}, \tag{13}$$

and $\odot$ denotes the Hadamard product, that is the component-wise product.

These expressions for the marginal and the conditional distributions are essentially those given by Azzalini & Capitanio (1999, sect. 4), except for some algebraic simplifications and the introduction of τ, of course.

As mentioned in the introduction, the addition of the new parameter τ allows closure of the class (2) under the operation of conditioning, instead of an approximate form of closure as discussed by Azzalini & Capitanio (1999, sect. 4.2). This has some costs, however, in particular the loss of chi-square properties of certain quadratic forms.

2.3. Independence and local dependence function

We now establish some results on independence and conditional independence among components of Y which will play a central role in the development of section 3.

First of all, independence of Y_1 and Y_2 introduced in section 2.2 is ensured when at least one of α_1 and α_2 is the null vector and Ω_{12} is the null matrix, hence also the corresponding block, Ω^{12}, in the inverse matrix is 0. As a corollary it follows that at least one of Y_1 and Y_2 must be Gaussian. More generally, independence of two affine transformations A_1Y and A_2Y holds under the conditions of prop. 6 of Azzalini & Capitanio (1999).

The similar situation holds for the case of conditional independence, except that one works with the components of the conditional distribution, $\alpha_{2\cdot1}$ and $\Omega_{22\cdot1}$. Since

$$\Omega^{22} = (\Omega^{-1})_{22} = (\Omega_{22\cdot 1})^{-1},$$

and the components of $\alpha_{2\cdot 1}$ are positive multiples of those of α_2, then conditional independence of two components of Y given the others can be examined by simple inspection of α and Ω^{-1}. In connection with the problem of graphical models it is useful to re-phrase this result in the following form.

Proposition 1
Consider the three block partition $Y^T = (Y_A^T, Y_B^T, Y_C^T)$ where A, B and C are sets of indices of Y. Then Y_A and Y_B are conditionally independent given Y_C, that is

$$Y_A \perp\!\!\!\perp Y_B | Y_C,$$

if and only if the two following conditions hold simultaneously:

(i) $\Omega^{AB} = 0$,
(ii) at least one of α_A *and* α_B tis the null vector,

where α_A and α_B denote the blocks of α associated to A and B, respectively, and Ω^{AB} denotes the block of Ω^{-1} with subscripts (A, B).

The proofs of all the above statements are similar to those for the case $\tau = 0$; see Azzalini & Capitanio (1999, sect. 3, 4, 6.3).

It is interesting to consider the local dependence function studied by Holland & Wang (1987) and developed further by Jones (1996, 1998), to measure the dependence among the components of a bivariate variable; this is defined as

$$\gamma(x, y) = \frac{\partial^2}{\partial x \partial y} \log f(x, y).$$

In the case of a 'standard' bivariate variable $\mathrm{SN}_2(0, \bar{\Omega}, \alpha, \tau)$, we obtain

$$\gamma(x, y) = \frac{\rho}{1 - \rho^2} + \alpha_1 \alpha_2 \zeta_2(\alpha_0 + \alpha_1 x + \alpha_2 y)$$

where ρ is the off-diagonal term of $\bar{\Omega}$. This is identically 0 if $\rho = 0$ and $\alpha_1 \alpha_2 = 0$, in agreement with the statements above. Notice that, if $\alpha_1 \alpha_2 = 0$, then $\gamma(x, y) =$ constant, a case connected to the problem studied by Jones (1998). In section 3.4 a consequence of this fact will be given.

It is worth remarking that $\gamma(x, y)$ has been used in graphical models as a measure of interaction. See Whittaker (1990, ch. 2) for related results, especially those concerning parametric collapsibility.

3. Graphical models

In this section, we shall consider a random variable $Y \sim \mathrm{SN}_k(\xi, \Omega, \alpha, \tau)$, with the aim of examining the properties of the corresponding conditional independence graph $\mathcal{G}(V, E)$. To avoid trivialities, we assume that Y is a "proper" SN variate, in the sense that $\alpha \neq 0$. In particular, we shall deal with the following aspects:

(i) how to build $\mathcal{G}(V, E)$;
(ii) the relationships between $\mathcal{G}$ and the similar graph, say $\mathcal{G}_{W^*}$, of the generating Gaussian variate W^*, as defined in (3);
(iii) the admissible conditional independence relationships between the components of Y, when some marginal distributions are Gaussian;
(iv) some properties related to decomposable graphs.

Some standard definitions and concepts will be used. For background material, the reader is referred to the first two chapters of Lauritzen (1996). Among these standard results, recall the factorization of the density according to a graph and the equivalence of the three Markov properties; see Lauritzen (1996, pp. 29–36). These facts can be immediately employed here taking into account that an SN variate has a continuous and strictly positive density function.

3.1. Some preliminary results

In order to build the conditional independence graph of an SN variate, conditions for pairwise conditional independence are required. They follow as a special case of proposition 1.

Proposition 2 (Pairwise conditional independence)
If $Y \sim \mathrm{SN}_k(\xi, \Omega, \alpha, \tau)$, *then*

$$Y_i \perp\!\!\!\perp Y_j \mid \textit{all other variables}$$

if and only if the following conditions simultaneously hold:

(a) $\Omega^{ij} = 0$,
(b) $\alpha_i \alpha_j = 0$

where Ω^{ij} *denotes the* (i, j)*-th entry of* Ω^{-1}.

Hence, given the parameters of the distribution of Y, the above result leads immediately to the construction of $\mathcal{G}(V, E)$, since

$$(i, j) \in E \iff \Omega^{ij} \neq 0 \quad \text{or} \quad \alpha_i \alpha_j \neq 0. \tag{14}$$

However, things are very different working in the reverse direction. In fact, if we take the topology of $\mathcal{G}(V, E)$ as given, and examine the set of compatible null entries of α and Ω^{-1}, then from proposition 2 it is easy to note that, given a conditional independence graph, there exists more than one configuration of consistent non-null entries in α and Ω^{-1}. The next proposition states how this set can be identified.

Proposition 3
Denote by $\{g_h(V_h, E_h), h = 1, \ldots, q\}$ *the set of all the complete subgraphs of* $\mathcal{G}$, *and consider the two sets* $I_\alpha = \{u : \alpha_u \neq 0\}$, $I_\Omega = \{(u, v) : \Omega^{uv} \neq 0\}$. *Then a pair* (Ω^{-1}, α) *is consistent with* $\mathcal{G}(V, E)$ *if and only if there exists* $j \in \{1, 2, \ldots, q\}$ *such that the two following conditions hold simultaneously:*

(i) $V_j = I_\alpha$,
(ii) $E \backslash E_j \subseteq I_\Omega \subseteq E$.

Proof. We prove sufficiency first. Suppose that (Ω^{-1}, α) is consistent with $\mathcal{G}(V, E)$; then either $\alpha_u \alpha_v \neq 0$ or $\Omega^{uv} \neq 0$ if and only if $(u, v) \in E$, such that $I_\Omega \subseteq E$. Since for all $u, v \in I_\alpha$ we have $\alpha_u \alpha_v \neq 0$, it follows that I_α induces a complete subgraph g_j, say, of $\mathcal{G}$. Furthermore, if $(u, v) \in E \backslash E_j$, then $\alpha_u \alpha_v = 0$, so that the relationship $E \backslash E_j \subseteq I_\Omega$ follows from (14).

To prove necessity suppose that for some j the relationships $V_j = I_\alpha$ and $E \setminus E_j \subseteq I_\Omega \subseteq E$ both hold true. Then $\alpha_u \alpha_v \neq 0$ if and only if $(u, v) \in E_j$, and since $(u, v) \in E \backslash E_j$ implies $\Omega^{uv} \neq 0$ it follows that relation (14) holds, so that (Ω^{-1}, α) is consistent with $\mathcal{G}(V, E)$.

It is well-known that, when the graph associated to a conditional independence structure is decomposable, the estimation procedure can be simplified. Such simplifications are strictly related to zero constraints on the parameters, and the results contained in proposition 3 reveal that a number of different situations must be taken into account. Specifically, it follows from the above proposition that the number of admissible pairs (Ω^{-1}, α), counting the possible choices of null and non-null terms, is

$$\sum_{h=1}^{q} 2^{\binom{q_h}{2}}, \tag{15}$$

where q_h is the number of elements of V_h.

3.2. Relationships between $\mathcal{G}_Z$ and $\mathcal{G}_{W^*}$

Recall from section 2.1 that, given $W^* = (W_0^{\mathrm{T}}, W^{\mathrm{T}})$, $Z = (W|W_0 + \tau > 0) \sim \mathrm{SN}_k(0, \bar{\Omega}, \alpha, \tau)$. It is easy to see that, replacing W^* with $(0, \xi^{\mathrm{T}})^{\mathrm{T}} + \mathrm{diag}(1, \Omega_{11}, \Omega_{22}, \ldots, \Omega_{kk})^{1/2} W^*$, the variate $Y \sim \mathrm{SN}_k(\xi, \Omega, \alpha, \tau)$ can be generated in the same manner. However, since the conditional independence graph $\mathcal{G}_Z(V, E)$ of Z coincides with $\mathcal{G}(V, E)$, the first one will be considered. Therefore, if we observe Gaussian variables conditionally on a given event, then the model for the observed variables is the SN distribution. In this section some relationships between the "generating" Gaussian variate and the resulting skew-normal one are studied.

The relationships between the parameters of Z and W^* are one to one, such that, if the parameters of the SN are known, then the ones of the generating Gaussian variable are uniquely identified, and vice-versa. The same is not true if our knowledge is restricted to the conditional independence graph. More precisely, if $\mathcal{G}_{W^*}$ is given then $\mathcal{G}_Z$ is uniquely defined, while a set of consistent $\mathcal{G}_{W^*}$ corresponds to $\mathcal{G}_Z$. The situation is summarized by the following propositions.

Proposition 4
If $\mathcal{G}_{W^*}$ is known, then $\mathcal{G}_Z$ is uniquely identified. Furthermore, given $\mathcal{G}_{W^*}$ the graph $\mathcal{G}_Z$ can be obtained by adding those edges needed to make $bd(0)$ complete and by deleting 0 and corresponding edges.

Proof. From (6) we have

$$(\bar{\Omega}^*)^{-1} = \begin{pmatrix} c^2 & -\alpha^{\mathrm{T}} c \\ -\alpha c & (\bar{\Omega}^{-1} + \alpha\alpha^{\mathrm{T}}) \end{pmatrix} \tag{16}$$

where $c = (1 - \delta^{\mathrm{T}} \bar{\Omega}^{-1} \delta)^{-1/2} > 0$. Taking into account proposition 2, the result follows.

Proposition 5
If $\mathcal{G}_Z$ is known, then $\mathcal{G}_{W^*}$ is not uniquely identified. Specifically, the number of conditional independence graphs for W^* consistent with $\mathcal{G}_Z$ is given by (15).

Proof. It is immediate taking into account expression (16) and considering the number of admissible pairs (Ω^{-1}, α).

3.3. Some restrictions on marginal distribution

As stated in section 2.2, the SN family is closed under marginalization, and contains the set of normal distributions. Hence a k-dimensional variable possessing joint SN distribution can have some components with Gaussian marginal distribution. It will be shown that, if some marginals

are Gaussian, then some configurations of the conditional independence graph must be excluded. This result can be relevant in the model selection context, since the potentially very large number $2^{\binom{k}{2}}$ of admissible conditional independence graphs can be reduced.

In the following, the two sets of vertices corresponding to marginally Gaussian and marginally skew-normal univariate components of $Y \sim \mathrm{SN}_k(\xi, \Omega, \alpha, \tau)$ will be denoted by Γ and Σ, respectively. The graph will then be a marked graph, according as to whether a node belongs to Γ or to Σ. Examples of marked graphs are given in Fig. 1.

Proposition 6
Consider the three block partition $Y^{\mathrm{T}} = (Y_A^{\mathrm{T}}, Y_B^{\mathrm{T}}, Y_C^{\mathrm{T}})$ where A, B and C are disjoint subsets of indices. If C separates A from B, i.e. $Y_A \perp\!\!\!\perp Y_B | Y_C$, then one among the three following conditions must hold:

(i) $A \cup C \subseteq \Gamma$
(ii) $B \cup C \subseteq \Gamma$
(iii) $C \not\subseteq \Gamma$.

Proof. Note that by the conditional independence assumption $\bar{\Omega}^{AB} = 0$ and at least one of α_A and α_B is the null vector. Therefore, from the second equality in (6), at least one of the two following equalities must hold:

$$\bar{\Omega}^{AA}\delta_A + \bar{\Omega}^{AC}\delta_C = 0, \quad \bar{\Omega}^{BB}\delta_B + \bar{\Omega}^{BC}\delta_C = 0.$$

The result then follows from the fact that $\bar{\Omega}$ is strictly positive definite.

Corollary 1
Let (A, B, C) be a partition of V such that $A \cup C \subseteq \Gamma$. If C separates A from B, then $\alpha_A = 0$.

Proposition 7
If $i \in \Gamma$ and $bd(i) \cap \Sigma = \{h\}$, i.e $bd(i)$ has only one vertex in Σ, then $\alpha_i \neq 0$.

Proof. Let h be the unique SN vertex in $bd(i)$. Then, from the second equality of (6), we have $\alpha_i \propto \Omega^{ih}\delta_h$. Since $\delta_h \neq 0$, it follows that $\alpha_i = 0$ if and only if $\Omega^{ih} = 0$, implying $(i, h) \notin E$.

Corollary 2
If $i, j \subseteq \Gamma$ and both $bd(i)$ and $bd(j)$ have exactly one vertex in Σ, then $(i, j) \in E$.

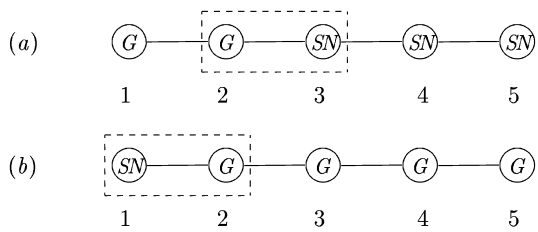

Fig. 1. Two examples of decomposable marked graphs. Here *G* and *SN* denote Gaussian and skew-normal nodes, respectively; the dashed boxes indicate the cliques with non-null αs in the joint 5-dimensional distribution.

Proof. Immediate from propositions 3 and 7.

The algebraic conditions for existence of a pair (Ω^{-1}, α) consistent with a specified structure for a marked graph can be obtained from (6), taking into account that Ω and δ are not variation independent. The above two propositions provide two necessary conditions which can be easily checked for the admissibility of a marked graph in a practical situation.

(i) In any three-set partition of a marked graph, a subset of Gaussian vertices cannot separate two subsets each containing some skew-normal vertices.
(ii) In a marked graph, there cannot exist two not connected Gaussian vertices having on their boundaries exactly one skew-normal vertex.

As an example consider the marked graph in Fig. 1(a). From proposition 7, $\alpha_2 \neq 0$ and, since $\{2\}$ is a separator, $\alpha_1 = 0$ from corollary 1. Then using proposition 3 the set $\{i : \alpha_i \neq 0\} \subseteq \{2, 3\}$. If node $\{5\}$ were in Γ, then the graph would not be admissible, since also $\alpha_5 \neq 0$, but $\{2, 5\}$ is not a complete subgraph.

3.4. Decomposable graphs

A primary role in graphical models context is played by decomposable graphs. If a graphical model is decomposable, then simplification occurs in both the interpretation of data and the estimation procedure. In fact, models can be specified in terms of conditional and marginal probability distributions, leading to a simplified analysis based on lower dimensional components.

Specifically, if a graph $\mathcal{G}$ is decomposable, then the joint density of the associated variables can be factorized according to a perfect sequence of cliques $C_1, C_2, \ldots, C_m$; see Lauritzen (1996, sect. 5.3.1) for details. Hence, on defining

$$H_j = \bigcup_{h=1}^{j} C_h, \quad R_j = C_j \setminus H_{j-1}, \quad \text{and} \quad S_j = H_{j-1} \cap C_j,$$

the triplet (H_{j-1}, R_j, S_j), for all $j \in \{1, 2, \ldots, m\}$, decomposes the subgraph induced by H_j, and the joint density admits the factorization

$$f = \prod_{j=1}^{m} f_{C_j} \Big/ \prod_{j=2}^{m} f_{S_j} \tag{17}$$

where f_A denotes the joint marginal density of the A components.

Starting from this general result, some specific properties can be stated. It will be shown that (17) can be rewritten as the product of a skew-normal and $m - 1$ Gaussian densities. In order to achieve this result the following lemma is needed.

Lemma 1
Let $Y \sim \mathrm{SN}_k(\xi, \Omega, \alpha, \tau)$ and consider $S \subseteq V \setminus I_\alpha$. Then the conditional distribution of Y_S given the remaining variables is Gaussian.

Proof. The result follows taking into account the expression of the shape parameter of a conditional SN distribution given in (13), and observing that $\alpha_i = 0$, for all $i \in V \setminus I_\alpha$.

Lemma 1 implies that the measure of local dependence defined in section 2.3, when applied to any pair of variables in S conditionally on the remaining, is constant. As a

consequence it becomes a partial measure of interaction, in the terminology of Whittaker (1990, sect. 2.3), since it does not depend on the values taken on by the conditioning variables.

Proposition 8
Suppose that $\mathcal{G}$ is decomposable and let $C_1, C_2, \ldots, C_m$ be a perfect sequence of cliques associated to $\mathcal{G}$. Let $j \in \{1, 2, \ldots, m\}$ *such that* $I_\alpha \subseteq C_j$. *Then*

$$f = f_{C_j} \prod_{h \neq j} \phi^{(h)} \tag{18}$$

where f_{C_j} is the density of a proper skew-normal variate, and the $\phi^{(h)}$s are suitable Gaussian densities. Moreover, the shape parameter of f_{C_j} is equal to the block α_{C_j} of α.

Proof. Notice that, from proposition 3, it follows that an index $j \in \{1, \ldots, m\}$ such that $I_\alpha \subseteq C_j$ must exist. Furthermore, from lemma 1, the conditional distribution of any subset of variables corresponding to vertices in $S \subseteq V \setminus C_j$, given the remaining ones, is Gaussian. Then (17) can be rearranged into

$$f = f_{C_j} \prod_{h \neq j} f_{C_h} \Big/ \prod_{h=2}^{m} f_{S_h} = f_{C_j} \prod_{h<j} \frac{f_{C_h}}{f_{S_{h+1}}} \prod_{h>j} \frac{f_{C_h}}{f_{S_h}} \tag{19}$$

where, considering the partitions of $\mathcal{G}$ induced by each separator S_h and taking into account the global Markov property, $(Y_{C_h \setminus S_{h+1}} | Y_{S_{h+1}})$ has a Gaussian distribution when $h < j$; a similar fact holds for $(Y_{C_h \setminus S_h} | Y_{S_h})$ when $h > j$. Finally, the fact that the shape parameter is α_{C_j} follows by consideration of (10). The terms of the two products in (19) determine the specific form of the $\phi^{(h)}$s. The explicit expression of their parameters will be given later.

These results will be useful for parameter estimation. In fact, it will be shown in section 4.2 that (19) identifies a parameter based factorization of the likelihood, such that estimation of the parameters can be performed separately for each clique. Furthermore, since the conditional distributions $\phi^{(h)}$ are Gaussian, known results concerning parameter estimation can be applied, reducing the computational complexity of the procedure.

The discussion of this section has made no assumption on the type of vertices, Gaussian or skew-normal. In case we are dealing with a marked graph, then this information can be incorporated, in the sense that, in general, some of the m factorizations (18) can be discarded, possibly down to only one admissible factorization.

4. Parameter estimation

4.1. Computational aspects

Consider the case where $Y_i \sim \mathrm{SN}_k(\xi_i, \Omega, \alpha, \tau)$ for $i = 1, \ldots, n$, and the components are independent. Moreover assume that the regression model

$$(\xi_1, \ldots, \xi_n)^{\mathrm{T}} = X\beta$$

holds for a $n \times p$ design matrix X of full rank p. To estimate the parameters $(\beta, \Omega, \alpha, \tau)$, the corresponding log-likelihood is

$$\begin{aligned} \log L &= \sum_{i=1}^{n} \Big\{ -\tfrac{1}{2} \log |\Omega| - \tfrac{1}{2} u_i^{\mathrm{T}} \Omega^{-1} u_i + \zeta_0\Big(\tau(1 + \eta^{\mathrm{T}} \Omega \eta)^{1/2} + \eta^{\mathrm{T}} u_i\Big) - \zeta_0(\tau) \Big\} \\ &= \tfrac{1}{2} n \log |\Omega^{-1}| - \tfrac{1}{2} \mathrm{tr}(\Omega^{-1} U^{\mathrm{T}} U) + \textstyle\sum_{i=1}^{n} \zeta_0(v_i) - n\zeta_0(\tau) \end{aligned}$$

where

$$u_i = y_i - x_i^{\mathrm{T}}\beta, \quad U = (u_1, \ldots, u_n)^{\mathrm{T}}$$
$$v_i = \tau(1 + \eta^{\mathrm{T}}\Omega\eta)^{1/2} + \eta^{\mathrm{T}}u_i, \quad \eta = \omega^{-1}\alpha.$$

The above function cannot be maximized in closed form and we have to resort on numerical methods. Details on the computational aspects, including expressions for the derivatives of the log-likelihood, are given in the appendix.

There is, however, a difficulty to bear in mind. As demonstrated by Azzalini & Capitanio (1999, sect. 4.2), the class (2) is quite closely approximated by the subclass given by the restriction $\tau = 0$. More explicitly, for each member of the four-parameter class (2), there is a member of the three-parameter class (1) which is close to it as for numerical values of the density. When this fact is translated into the context of parameter estimation, the implication is that it can be difficult to locate the parameters, since there can be more parameter combinations which have about the same likelihood.

Some numerical work using the scheme described in the appendix has confirmed that the procedure works but, in some cases, it can have difficulties in converging or equivalently it might require a very large n, which typically means a few hundred cases even in the case of equally distributed observations. If these sorts of problems occur, we have found it useful to construct the profile log-likelihood as a function of τ. The source of the above problem is related to the presence of the parameter τ in conjunction with the other parameters, and τ is effectively removed when the log-likelihood is evaluated at any given value of its range. Therefore the "near unidentifiability" problem is also removed, leading to a much more stable behaviour of the optimization algorithms.

An example of the outcome is shown in Fig. 2, which refers to the data (Ht, Wt) of the Australian Institute of Sport, already used for illustration in related problems by Azzalini & Dalla Valle (1996) and Arnold & Beaver (2000). Direct global maximization of the log-likelihood function with respect to all four parameters simultaneously appeared troublesome, while the construction of the profile log-likelihood was much more stable and numerically satisfactory, as illustrated by Fig. 2. Only for very large negative values of τ was there some erratic behaviour, but this can presumably be due to numerical instability, especially with $\Phi(\tau)$ when τ is less than, say, -7.

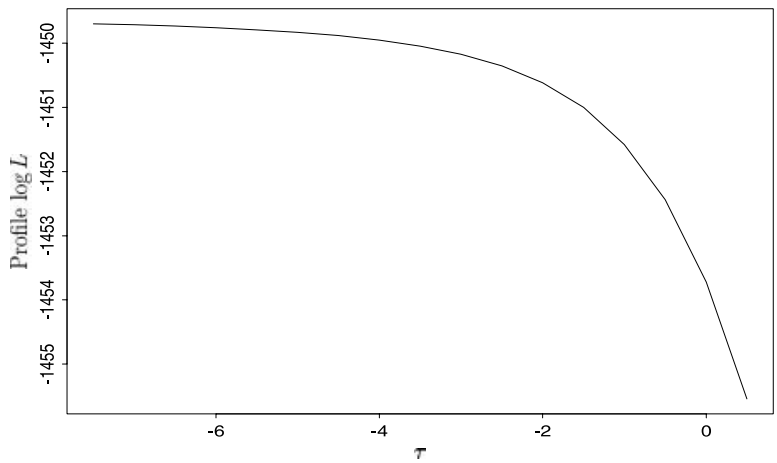

Fig. 2. Profile log-likelihood for τ for the data (Ht, Wt).

The shape of the profile log-likelihood in the specific case of Fig. 2 appears peculiar, in that there seems to exist no finite maximum likelihood estimate. This sort of behaviour has been discussed by Azzalini & Capitanio (1999, sect. 5.3), and it must be regarded as a failure of the maximum likelihood method. In practical terms, there is no actual difficulty, in the sense that each value of τ smaller than, say, -4 has associated estimates of the other parameters which produce about the same density function; this latter sort of plot has not been reported here.

Notice that τ is effectively removed from the expression of (2) when $\alpha = 0$. Hence the above discussion applies to the case when α is known to be different from 0.

4.2. Parameter based factorizations

In this section we look at the conditions under which factorizations of the likelihood function according to a conditional independence graph are parameter based. Parameter based factorizations of the likelihood lead to simplification of the inference on the parameters of a model. The idea of exploiting the conditional independencies in a graph to derive parameter based factorizations has been widely used in the statistical literature on graphical models, and it is formalized in Cox & Wermuth (1999).

For a family of models specified by a parameter θ taking values in a parameter space Θ, the likelihood of an observed vector x admits a parameter based factorization if

$$L(\theta; x) = L_1(\theta_1; x)L_2(\theta_2; x)$$

where $\theta_1 \in \Theta_1$ and $\theta_2 \in \Theta_2$ and θ_1 and θ_2 are variation independent, that is $\Theta = \Theta_1 \times \Theta_2$.

A parameter based factorization of the likelihood leads to a simplification in the maximum likelihood estimation, as the maximum likelihood estimate is obtained by separate maximization of the factors. Moreover, inference about, say, θ_1 may be performed from the factor L_1 solely. The definition extends directly to more than two factors.

We shall show that some factorizations that exploit the conditional independencies in a graph are parameter based. As stated in the next proposition, some hypotheses on the structure of zeros in α are required to establish which factorizations are parameter based.

In the following we denote by $L(\cdot)$ the likelihood based on y, by $L_{C_j}(\cdot)$ the likelihood based on y_{C_j} and by $L_{\bar{C}_j|C_j}(\cdot)$ the likelihood based on the variables belonging to the complement set $\bar{C}_j = V \setminus C_j$, conditionally on y_{C_j}.

Proposition 9

Suppose that $\mathcal{G}$ is decomposable, and let $C_1, C_2, \ldots, C_m$ be a perfect sequence of cliques associated to $\mathcal{G}$. Assume that the admissible zero constraints on α and Ω^{-1} are as follows:

1. $\alpha_u = 0$ when $u \notin C_j$,
2. $\Omega^{uv} = 0$ when $(u, v) \notin E$.

Then the factorization

$$L = L_{C_j} L_{\bar{C}_j|C_j} = L_{C_j} \prod_{h<j} L_{C_h \setminus S_{h+1}|S_{h+1}} \prod_{h>j} L_{C_h \setminus S_h|S_h}$$

is parameter based for every choice of $j \in \{1, 2, \ldots, m\}$.

Proof. Consider the block partition of Y, ξ, Ω and α as defined in (9), where $Y_1 = Y_{C_j}$ and $Y_2 = Y_{\bar{C}_j}$. Furthermore Y_1 is partitioned into $Y_{1R} = Y_{R_j}$ and $Y_{1S} = Y_{S_j}$. Under the above hypothesis, we get $\alpha_2 = 0$. Hence, applying proposition 8, the factor $L_{2|1}$ corresponds to a

Gaussian likelihood, and $\alpha_{1(2)} = \alpha_1$. Taking into account that Y_1 contains the separator Y_{1S}, the factorization reduces to

$$L(\xi, \Omega, \alpha, \tau) = L_1(\xi_1, \Omega_{11}, \alpha_1, \tau) L_{2|1S}(\xi_{2\cdot 1S}, \Omega_{22\cdot 1S}).$$

According to (18), the term $L_{2|1S}$ factorizes into a product of conditional Gaussian likelihood functions of the form $L_{R_h^*|S_{h+1}}(\xi_{R_h^*\cdot S_{h+1}}, \Omega_{R_h^* R_h^*\cdot S_{h+1}})$ and $L_{R_h|S_h}(\xi_{R_h\cdot S_h}, \Omega_{R_h R_h\cdot S_h})$, where $R_h^* = C_h \backslash S_{h+1}$. Since ξ, α, Ω and τ are variation independent, and using known results on Gaussian models, the m parameter spaces

$$\begin{aligned} &(\xi_1, \Omega_{11}, \alpha_1, \tau), \\ &(\xi_{R_h^*\cdot S_{h+1}}, \Omega_{R_h^* S_{h+1}} \Omega_{S_{h+1} S_{h+1}}^{-1}, \Omega_{R_h^* R_h^*\cdot S_{h+1}}, h < j), \\ &(\xi_{R_h\cdot S_h}, \Omega_{R_h S_h} \Omega_{S_h S_h}^{-1}, \Omega_{R_h R_h\cdot S_h}, h > j)) \end{aligned} \tag{20}$$

are variation independent and together span the full space of the original specification under the constraints defined by the graph $\mathcal{G}$. This concludes the proof.

Notice that (i) the factorization given by this proposition is of type (19); (ii) the iterative computational procedure described in section 4 needs to be applied only to the SN component (20), while the others are Gaussians.

When information on the type of each node is given, propositions 6 and 7, and corollary 1 can be used to reduce the number of admissible cliques, possibly down to a single admissible clique corresponding to non-null αs. For instance, the marked graph in Fig. 1(a) can be decomposed by the perfect sequence of cliques $C_1 = \{1, 2\}$, $C_2 = \{2, 3\}$, $C_3 = \{3, 4\}$, $C_4 = \{4, 5\}$, and the separators are $S_2 = \{2\}$, $S_3 = \{3\}$, $S_4 = \{4\}$. As already seen in section 3.3, the set $\{i : \alpha_i \neq 0\}$ is a subset of C_2, so that the factorization

$$L_{\{2,3\}}(\cdot) L_{\{1|2\}}(\cdot) L_{\{4|3\}}(\cdot) L_{\{5|4\}}(\cdot)$$

is parameter based.

Another particular situation is of special interest. From proposition 9, it follows that the existence of a parameter based factorization strictly depends on the presence of only one non-Gaussian density among the factors. As a matter of fact, if all but one marginal densities over the cliques in a marked graph are Gaussian, then another parameter based factorization does exist. An example of such a graph is given in Fig. 1(b); here both factorizations

$$L_{\{2,3\}}(\cdot) L_{\{1|2\}}(\cdot) L_{\{4|3\}}(\cdot) L_{\{5|4\}}(\cdot), \quad L_{\{4,5\}}(\cdot) L_{\{1|2\}}(\cdot) L_{\{2|3\}}(\cdot) L_{\{3|4\}}(\cdot)$$

are parameter based.

In general situations when information about the clique containing the vertices corresponding to non-null αs is not available, an iterative procedure of estimation must be defined. Proposition 3 defines the whole set of admissible pairs (Ω^{-1}, α) which are compatible with a given conditional independence graph, but clearly some of these are negligible in estimation. Suppose that a conditional independence structure is defined, and that the parameters of the model are to be estimated under suitable zero constraint. A good rule is to impose at first the minimum number of constraints, i.e. those strictly necessary to guarantee coherence with the given independence structure. From proposition 3 they correspond to those defined on the basis of the pairs (Ω^{-1}, α) such that $I_\alpha = C_j$ and $I_\Omega = E$, for all cliques C_j, i.e. all maximally complete subgraphs of $\mathcal{G}$. Then

(a) for each clique C_j, consider a model as defined in proposition 9;
(b) for each model, compute the maximum likelihood estimates using the simplified procedure based on separate maximization of each factor;
(c) select the model and the parameter estimates with highest value of the likelihood function.

Notice that, while this procedure does lead to the maximum likelihood estimate, the search is performed over a parameter space which is not an open set of a Euclidean space. Hence standard results on the asymptotic distribution of the estimator are not automatically applicable.

Acknowledgements

We would like to thank two anonymous referees for useful comments which led to improved presentation of the material. The first author is grateful to Pier Luigi Conti and Paolo Paruolo for helpful discussions. Part of this work was completed while the third author was visiting Nuffield College, Oxford, supported by the Jemolo Fellowship scheme. The generous hospitality of the College is gratefully acknowledged. This research was supported partly by the Consiglio Nazionale delle Ricerche (grant No. 98.01532.CT10), partly by MURST (grant PRIN 2000), Italy.

References

Arnold, B. C. & Beaver, R. J. (2000). Hidden truncation models. *Sankhyā Ser. A* **62**, 22–35.
Azzalini, A. & Capitanio, A. (1999). Statistical applications of the multivariate skew-normal distribution. *J. Roy. Statist. Soc. Ser. B* **61**, 579–602.
Azzalini, A. & Dalla Valle, A. (1996). The multivariate skew-normal distribution. *Biometrika* **83**, 715–726.
Cox, D. R. & Wermuth, N. (1999). Likelihood factorizations for mixed discrete and continuous variables. *Scand. J. Statist.* **26**, 209–220.
Genz, A. (1993). Comparison of methods for the computation of multivariate normal probabilities. In: *Computing science and statistics. Proceedings of the 25th symposium on the interface*, 400–405. Interface Foundation of North America, Fairfax Station, VA.
Holland, P. W. & Wang, Y. J. (1987). Dependence function for continuous bivariate densities. *Comm. Statist. Theory Methods* **16**, 863–876.
Jones, M. C. (1996). The local dependence function. *Biometrika* **83**, 899–904.
Jones, M. C. (1998). Constant local dependence. *J. Multivariate Anal.* **64**, 148–155.
Lauritzen, S.L. (1996). *Graphical models*. Oxford University Press, Oxford.
Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika* **57**, 519–530.
Paulsen, V. I., Power, S. C. & Smith, R. R. (1989). Schur products and matrix completions. *J. Funct. Anal.* **85**, 151–78.
Roverato, A. (2000). Cholesky decomposition of a hyper inverse Wishart matrix. *Biometrika* **87**, 99–112.
Whittaker, J. (1990). *Graphical models in applied multivariate statistics*. Wiley, Chichester.

Received February 2001, in final form December 2001

Antonella Capitanio, Dipartimento di Scienze Statistiche, Università di Bologna, via delle Belle Arti 41, I-40126 Bologna, Italy.
E-mail: capitani@stat.unibo.it

Appendix

Some properties of Z

The higher derivatives of the cumulant generating function of Z are

$$\frac{d^m K(t)}{dt_i dt_j \cdots dt_r} = \zeta_m(\tau + \delta^{\mathrm{T}} t)\delta_i \delta_j \cdots \delta_r, \quad (m > 2),$$

which allow computation of the corresponding cumulants; these in turn lead to

$$\gamma_{1,k} = \zeta_3(\tau)^2(\delta^{\mathrm{T}}\Upsilon\delta)^3, \quad \gamma_{2,k} = \zeta_4(\tau)(\delta^{\mathrm{T}}\Upsilon\delta)^2$$

i.e. the indices of multivariate skewness and kurtosis of Mardia (1970), where $\Upsilon = \mathrm{var}\{Z\}^{-1}$.

To compute the distribution function of Z, write

$$\begin{aligned}\mathbb{P}\{Z \le z\} &= \mathbb{P}\{W \le z | W_0 + \tau > 0\} \\ &= \mathbb{P}\{-W_0 < \tau \cap W \le z\}/\mathbb{P}\{-W_0 < \tau\} \\ &= \Phi_{k+1}((\tau, z^{\mathrm{T}})^{\mathrm{T}}; \tilde{\Omega})/\Phi(\tau)\end{aligned}$$

where $\tilde{\Omega}$ is a matrix similar to $\bar{\Omega}^*$, but with δ replaced by $-\delta$, and $\Phi_m(x; A)$ denotes the integral of $\phi_m(x; A)$. Therefore, the distribution function of Z can be obtained from an algorithm which produces the distribution function of a $(k+1)$-dimensional normal variate. For a discussion of the latter problem, see Genz (1993).

For the generation of random numbers, it is natural to exploit the "definition via conditioning" itself, namely $Z = (W | W_0 + \tau > 0)$. This defines a procedure which is conceptually simple and easy to simulate on a computer. The only drawback is that it leads to the rejection of a fraction $\Phi(-\tau)$ of the simulated W vectors, and this fraction becomes large if $\tau \to -\infty$. To decrease the rejection rate in the case $\tau \le 0$, notice that one can generate data with distribution (5) by setting $Z = (-W | -W_0 + \tau > 0)$, when the condition $W_0 + \tau > 0$ fails; this device doubles the acceptance rate when $\tau < 0$. The overall fraction p of accepted samples with this combination of rules is now

$$p = \begin{cases} \Phi(\tau) & \text{if } \tau > 0, \\ 2\Phi(\tau) & \text{if } \tau \le 0. \end{cases}$$

Affine transformations

The distribution of the affine transformation

$$T = AY + b,$$

where A is a $h \times k$ matrix of rank h, can be obtained from its cumulant generating function,

$$K_T(t) = K_Y(A^{\mathrm{T}}t) + b^{\mathrm{T}}t,$$

which is still of type (7). Again, the expressions for the parameters of T are similar to those of the case $\tau = 0$; see Azzalini & Capitanio (1999, sect. 3.2, and 4.1). Specifically, we have

$$T \sim \mathrm{SN}_h(\xi_T, \Omega_T, \alpha_T, \tau),$$

where

$$\begin{aligned}\xi_T &= A\xi + b, \\ \Omega_T &= A\Omega A^{\mathrm{T}}, \\ \alpha_T &= \frac{1}{\left(1 + \alpha^{\mathrm{T}}(\bar{\Omega} - B\Omega_T^{-1}B^{\mathrm{T}})\alpha\right)^{1/2}} \omega_T \Omega_T^{-1} B^{\mathrm{T}} \alpha, \\ B &= \omega^{-1}\Omega A^{\mathrm{T}}.\end{aligned}$$

Aspects of numerical maximization of the log-likelihood

It is convenient to reparametrize the problem, partly because the space of positive definite matrices Ω or equivalently Ω^{-1} is difficult to handle directly. Therefore we write

$$\Omega^{-1} = A^{\mathrm{T}}DA = A^{\mathrm{T}}\ \mathrm{diag}(\exp(\psi))A \tag{21}$$

where A is an upper triangular matrix with diagonal elements all equal to 1, and D is the diagonal matrix of positive values. For numerical optimization, it is convenient to

reparametrize the diagonal elements of D to $\exp(\psi)$, to deal with unbounded parameters. Hence, the new parametrization is then $(\beta, A, \psi, \eta, \tau)$ and the log-likelihood is now written as

$$\log L = \frac{1}{2} n \log |A^{\mathrm{T}} DA| - \frac{1}{2} \mathrm{tr}(A^{\mathrm{T}} DAQ) + 1_n^{\mathrm{T}} \zeta_0(v) - n\zeta_0(\tau) \tag{22}$$

where $Q = UU^{\mathrm{T}}$ and

$$v = (v_1, \ldots, v_n)^{\mathrm{T}}, \quad \zeta_0(v) = (\zeta_0(v_1), \ldots, \zeta_0(v_n))^{\mathrm{T}}, \quad 1_n = (1, \ldots, 1)^{\mathrm{T}}.$$

Note that, as remarked in Roverato (2000), reparametrization (21) is particularly suitable in graphical model context. In fact, using th. 2.4 in Paulsen *et al.* (1989), it is easy to show that, if the graph $\mathcal{G}(V, E)$ is decomposable and the set I_Ω is equal to E, then the rows (and columns as well) of Ω^{-1} can be ordered according to a perfect vertex elimination scheme such that $\Omega^{ij} = 0$ implies $A^{ij} = 0$. As a consequence, zero constraints imposed to some entries of Ω^{-1} can be directly applied to the corresponding entries of A.

To increase the efficiency of numerical optimization, it is useful to supply the algorithm with the derivatives of (22) with respect to $\beta, A, \psi, \eta, \tau$; these are as follows:

$$\begin{aligned}
\frac{\partial \log L}{\partial \beta} &= X^{\mathrm{T}} U \Omega^{-1} - X^{\mathrm{T}} \zeta_1(v) \eta^{\mathrm{T}}, \\
\frac{\partial \log |A^{\mathrm{T}} DA|}{\partial A} &= \frac{\partial \log |A|^2}{\partial A} = 0, \\
\frac{\partial \log |A^{\mathrm{T}} AD|}{\partial D} &= D^{-1}, \\
\frac{\partial \mathrm{tr}(A^{\mathrm{T}} DAQ)}{\partial A} &= \text{upper triangle of } (2DAQ), \\
\frac{\partial \mathrm{tr}(A^{\mathrm{T}} DAQ)}{\partial D} &= I \odot (AQA^{\mathrm{T}}), \\
\frac{\partial \zeta_0(v)}{\partial A} &= \zeta_1(v) \frac{\tau}{2(1 + \eta^{\mathrm{T}} \Omega \eta)^{1/2}} \text{upper triangle of } (-2(A^{-1})^{\mathrm{T}} \eta \eta^{\mathrm{T}} \Omega), \\
\frac{\partial \zeta_0(v)}{\partial D} &= \zeta_1(v) \frac{\tau}{2(1 + \eta^{\mathrm{T}} \Omega \eta)^{1/2}} I \odot (-D^{-1}(A^{-1})^{\mathrm{T}} \eta \eta^{\mathrm{T}} A^{-1} D^{-1}), \\
\frac{\partial \log L}{\partial \eta} &= 1_n^{\mathrm{T}} \zeta_1(v) \frac{\tau}{(1 + \eta^{\mathrm{T}} \Omega \eta)^{1/2}} \Omega \eta + U^{\mathrm{T}} \zeta_1(v), \\
\frac{\partial \log L}{\partial \tau} &= 1_n^{\mathrm{T}} \zeta_1(v)(1 + \eta^{\mathrm{T}} \Omega \eta)^{1/2} - n\zeta_1(\tau)
\end{aligned}$$

where $\odot$ denotes the Hadamard product. On writing $d = \exp(\psi)$ and using the chain rule

$$\frac{\partial \log L}{\partial \psi} = \frac{\partial \log L}{\partial d} d,$$

one converts the derivatives is with respect to d into those for ψ.