On generalizing Gaussian graphical models

Daniel Vogel

Technische Universität Dortmund
Vogelpothsweg 87, 44221 Dortmund, Germany

Abstract

We explore elliptical graphical models as a generalization of Gaussian graphical models, that is, we allow the population distribution to be elliptical instead of normal. Towards a statistical theory for such graphical models, consisting of estimation, testing and model selection, we consider the problem of estimating partial correlations. We derive the asymptotic distribution of a class of partial correlation matrix estimators based on affine equivariant scatter estimators.

Keywords: elliptical distribution, partial correlation, Tyler scatter matrix, robustness.

AMS: 62G35, 62H12

1. Introduction: partial correlations and graphical models

Let $p \geq 3$ and $\boldsymbol{X} = (\boldsymbol{Z}, \boldsymbol{Y})$ with $\boldsymbol{Z} = (Z_1, Z_2)$, $\boldsymbol{Y} = (Y_1, ..., Y_{p-2})$, be a p-dimensional random vector having distribution F and a non-singular covariance matrix Σ. Let furthermore $\hat{Z}_i(\boldsymbol{Y})$, $i = 1, 2$, be the projection of Z_i onto the space of all affine linear functions of $\boldsymbol{Y}$. Then the *partial correlation of Z_1 and Z_2 given $\boldsymbol{Y}$* is defined as

$$\varrho_{1,2\bullet Y} = \frac{\mathrm{cov}\left(Z_1 - \hat{Z}_1(\boldsymbol{Y}), \, Z_2 - \hat{Z}_2(\boldsymbol{Y})\right)}{\sqrt{\mathrm{var}\left(Z_1 - \hat{Z}_1(\boldsymbol{Y})\right)\mathrm{var}\left(Z_2 - \hat{Z}_2(\boldsymbol{Y})\right)}},$$

i.e. it is the correlation between the residuals $Z_1 - \hat{Z}_1(\boldsymbol{Y})$ and $Z_2 - \hat{Z}_2(\boldsymbol{Y})$. One can extend the definition of partial correlation (and thus partial uncorrelatedness) to vector-valued random variables in a straightforward manner. The partial correlation $\varrho_{1,2\bullet Y}$ can be computed from the covariance matrix Σ of $\boldsymbol{X}$:

$$\varrho_{1,2\bullet Y} = -\frac{k_{1,2}}{\sqrt{k_{1,1}k_{2,2}}},$$

where $k_{i,j}$, $i, j = 1, ..., p$, are the elements of $K = \Sigma^{-1}$, see e.g. [8], p. 143. K is called the *concentration matrix* (or *precision matrix*) of $\boldsymbol{X}$. Let

$$P = (p_{i,j})_{i,j=1,...,k} = K_D^{-\frac{1}{2}} K K_D^{-\frac{1}{2}},$$

where K_D denotes the diagonal matrix having the same diagonal as K and $K_D^{-\frac{1}{2}}$ is to be read as $(K_D)^{-\frac{1}{2}}$. The matrix P equals 1 on the diagonal and contains the negative partial correlations as its off-diagonal elements, i.e. $\varrho_{1,2\bullet Y} = -p_{1,2}$. We will also refer to P as partial

correlation matrix even though it contains negative partial correlations. In this paper we consider the task of estimating P in the elliptical model, which is a popular generalization of the multivariate normal model. Its first and second order characteristics provide an intuitive description of the geometry of the distribution, and it is mathematically tractable. In addition it allows to model different tail behaviours, and is often chosen to model data with heavy tails.

Our interest in partial correlation is originated in its application in graphical models. A thorough introduction of the latter would go beyond the scope of this exposition, we refer to standard volumes, e.g. [4] or [8]. If the population distribution is jointly normal, due to the particular properties of the normal family (most notably that it is closed under conditioning, and that correlation zero implies independence) partial uncorrelatedness implies conditional independence. A spherical distribution, however, has independent margins if and only if it is a normal distribution. This is also known as the Maxwell-Hershell-theorem cf. e.g. [2], p. 51. Consequently, in the elliptical model partial uncorrelatedness (i.e. an off-diagonal zero entry in the precision matrix K) does not imply conditional independence. It does, however, imply *conditional uncorrelatedness*, cf. [1], i.e. the conditional distribution of (Z_1, Z_2) given $\boldsymbol{Y} = \boldsymbol{y}$ (which is a bivariate distribution depending on $\boldsymbol{y}$) is for almost all values $\boldsymbol{y}$ such that it has correlation zero. Thus, in the elliptical model partial correlation is a measure of conditional correlation.

2. Elliptical distributions and shape matrices

In this introduction to elliptical distributions we mainly follow the notation of chapter 13 of [2]. A continuous distribution F in $\mathbb{R}^p$ is said to be *elliptical* if it has a Lebesgue-density f of the form

$$f(\boldsymbol{x}) = \det(S)^{-\frac{1}{2}} g\big((\boldsymbol{x} - \boldsymbol{\mu})^T S^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\big). \tag{1}$$

for some $\boldsymbol{\mu} \in \mathbb{R}^p$ and symmetric, positive definite $p \times p$ matrix S. We call $\boldsymbol{\mu}$ the symmetry center and S the shape matrix of F, and denote the class of all continuous elliptical distributions on $\mathbb{R}^p$ having these parameters by $E_p(\boldsymbol{\mu}, S)$. If second-order moments of $\boldsymbol{X} \sim F$ exist, then $\mathbb{E}(\boldsymbol{X}) = \boldsymbol{\mu}$, and $\text{Var}(\boldsymbol{X}) = \Sigma(F)$ is proportional to S. In the parametrization $(\boldsymbol{\mu}, S)$, the symmetry center $\boldsymbol{\mu}$ is uniquely defined whereas the matrix S is unique only up to scale, that is, $E_p(\boldsymbol{\mu}, S) = E_p(\boldsymbol{\mu}, cS)$ for any $c > 0$. One is tempted to impose some form of general standardization on S (several have been suggested in the literature, e.g., setting the trace to p or the determinant or a specific element of S to 1) and thus uniquely defining *the* shape matrix of an elliptical distribution. However, we refrain from such a standardization and call any matrix S satisfying (1) for a suitable function g a shape matrix of F. This allows, for example, to work always with the "simplest" function g. We want to mention two examples of elliptical distributions, the normal distribution $N_p(\boldsymbol{\mu}, S)$, which corresponds to $g_{N_p}(y) = (2\pi)^{-\frac{p}{2}} \exp\big(-\frac{1}{2}y\big)$, and the multivariate $t_{\nu,p}$-family with

$$g_{t_{\nu,p}}(y) = \frac{\Gamma(\frac{\nu+p}{2})}{(\nu\pi)^{\frac{p}{2}}\Gamma(\frac{\nu}{2})}(1 - \frac{y}{\nu})^{-\frac{\nu+p}{2}}.$$

Here the first subscript ν denotes the degrees of freedom. The $t_{\nu,p}(\boldsymbol{\mu}, S)$ distribution converges to $N_p(\boldsymbol{\mu}, S)$ as $\nu \to \infty$ and is, for small ν, a popular example of a heavy-tailed

distribution. Its moments are finite only up to order $(\nu - 1)$. For $\nu \geq 3$ its covariance is $\Sigma(t_{\nu,p}(\boldsymbol{\mu}, S)) = \frac{\nu}{\nu-2}S$.

We now turn to our statistical problem of interest: to estimate P in the elliptical model. Let $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ be i.i.d. random variables with elliptical distribution $F \in E_p(\boldsymbol{\mu}, S)$ and co-variance matrix Σ. Let furthermore $\mathbb{X}_n = (\boldsymbol{X}_1^T, ..., \boldsymbol{X}_n^T)^T$ be the $n \times p$ data matrix containing the data points as rows and $\hat{S}_n = \hat{S}_n(\mathbb{X}_n)$ a scatter estimator. Here we use the term *scatter estimator* in a very informal way for any symmetric matrix-valued estimator that gives some form of information about the spread of the data. In a narrower sense scatter estimators aim at estimating the covariance matrix. Hence it is a desirable property of such estimators to transform in the same way as the covariance matrix under affine linear transformations, that is, they satisfy $\hat{S}_n(\mathbb{X}_n A^T + \mathbf{1}\boldsymbol{b}^T) = A\hat{S}_n(\mathbb{X}_n)A^T$ for any full rank matrix $A \in \mathbb{R}^{p \times p}$ and vector $\boldsymbol{b} \in \mathbb{R}^p$. This property of a scatter estimator is called *affine equivariance*. However, there are estimators that do not satisfy affine equivariance, but a slightly weaker condition which we want to call *affine pseudo-equivariance* or *proportional affine equivariance*.

Condition C1 $\hat{S}_n(\mathbb{X}_n A^T + \mathbf{1}\boldsymbol{b}^T) = h(A)A\hat{S}_n(\mathbb{X}_n)A^T$ *for* $\boldsymbol{b} \in \mathbb{R}^p$, $A \in \mathbb{R}^{p \times p}$ *with full rank, and* $h : \mathbb{R}^{p \times p} \to \mathbb{R}$ *satisfying* $h(H) = 1$ *for any orthogonal matrix* H.

Estimators satisfying C1 shall also be called shape estimators: they give information about the shape (orientation and relative length of the axes of the contour-ellipses of F), but not the overall scale. Since the overall scale is irrelevant for (partial) correlations, i.e.

$$P = V_D^{-\frac{1}{2}}VV_D^{-\frac{1}{2}}, \quad \text{where} \quad V = S^{-1}, \tag{2}$$

for any shape matrix S of F, shape estimators are useful for estimating partial correlations, and we will turn our attention to this class of estimators in the following. A variety of shape estimators have been proposed and extensively studied, primarily in the robustness literature, see e.g. [9] for a review, but also the MLE of the covariance matrix at an elliptical distribution possesses this property. Affine (pseudo-)equivariance is indeed a very handy property, and such estimators are particularly suited for the elliptical model. Their variance (which then appears as asymptotic variance if the estimator is asymptotically normal) can be shown to have a rather simple general form under elliptical population distributions, which is given below in condition C2, and is basically due to Tyler [5]. We need to introduce some matrix notation.

For matrices $A, B \in \mathbb{R}^{p \times p}$, the Kronecker product $A \otimes B$ is the $p^2 \times p^2$ matrix with entry $a_{i,j}b_{k,l}$ at position $(i(p-1) + k, j(p-1) + l)$. Let $\boldsymbol{e}_1, ..., \boldsymbol{e}_p$ be the unit vectors in $\mathbb{R}^p$ and define the following matrices: $J_p = \sum_{i=1}^p \boldsymbol{e}_i\boldsymbol{e}_i^T \otimes \boldsymbol{e}_i\boldsymbol{e}_i^T$, $K_p = \sum_{i=1}^p \sum_{j=1}^p \boldsymbol{e}_i\boldsymbol{e}_j^T \otimes \boldsymbol{e}_j\boldsymbol{e}_i^T$ (the *commutation matrix*), I_{p^2} the $p^2 \times p^2$ identity matrix and $N_p = \frac{1}{2}(I_{p^2} + K_p)$. Finally $\text{vec}(A)$ is the p^2 vector obtained by stacking the columns of $A \in \mathbb{R}^{p \times p}$ from left to right underneath each other. Many shape estimators have been shown to satisfy the following condition in the elliptical model (possibly under additional assumptions on the population distribution F).

Condition C2 *There exist constants* $\eta \geq 0$, $\sigma_1 \geq 0$ *and* $\sigma_2 \geq -2\sigma_1/p$ *such that*

$$\hat{S}_n \xrightarrow{p} \eta S \qquad and \qquad \sqrt{n}\,\text{vec}(\hat{S}_n - \eta S) \xrightarrow{\mathscr{L}} N_{p^2}(\mathbf{0}, W),$$

where
$$W = 2\sigma_1\eta^2 N_p(S \otimes S) + \sigma_2\eta^2 \text{vec}(S)\big(\text{vec}(S)\big)^T,$$

and the constants σ_1 and σ_2 do not depend on S.

By means of the CMT and the multivariate delta method one can derive the general form of the asymptotic variance of any partial correlation estimator derived from a scatter estimator satisfying C2.

Proposition 1 *If $\hat{S}_n$ fullfills C2, the corresponding partial correlation estimator*
$$\hat{P}_n = (\hat{S}_n^{-1})_D^{-\frac{1}{2}} \hat{S}_n^{-1} (\hat{S}_n^{-1})_D^{-\frac{1}{2}}$$

satisfies
$$\hat{P}_n \xrightarrow{p} P \quad and \quad \sqrt{n}\,\text{vec}(\hat{P}_n - P) \xrightarrow{\mathscr{L}} N_{p^2}(\mathbf{0}, 2\sigma_1 \Gamma N_p(V \otimes V)\Gamma^T) \tag{3}$$

with P and V as in (2) and $\Gamma = (V_D^{-\frac{1}{2}} \otimes V_D^{-\frac{1}{2}}) - N_p(P \otimes V_D^{-1})J_p$.

Remark. In the expression for the asymptotic variance of $\hat{P}_n$ the constant η obviously has to cancel out. But also the constant σ_2 does not appear. Thus the comparison of the asymptotic efficiencies of partial correlation matrix estimators based on affine (pseudo-) equivariant scatter estimators reduces to the comparison of the respective values of the scalar σ_1. This is generally true for "scale-free" functions of $\hat{S}_n$ and has already been noted by Tyler [6].

3. Example: Tyler's M-estimator of scatter

Strictly speaking, two examples are given: Tyler's estimator mentioned in the title and, for comparison, the empirical covariance matrix $\hat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^n (\mathbf{X}_i - \overline{\mathbf{X}}_n)(\mathbf{X}_i - \overline{\mathbf{X}}_n)^T$, which is the maximum likelihood estimator for Σ at the multivariate normal distribution. $\hat{\Sigma}_n$ fulfills condition C1 with $h \equiv 1$, and we have the following asymptotic result.

Proposition 2 *If $\mathbf{X}_1, ..., \mathbf{X}_n$ are i.i.d. with distribution $F \in E_p(\boldsymbol{\mu}, \alpha\Sigma)$, $\alpha > 0$, and $\mathbb{E}||\mathbf{X} - \boldsymbol{\mu}||^4 < \infty$, then $\hat{\Sigma}_n$ fulfills C2 with $\eta = \alpha^{-1}$, $\sigma_1 = 1 + \kappa/3$ and $\sigma_2 = \kappa/3$, where κ is the kurtosis excess of the first (or any other) component of $\mathbf{X}_1$.*

The Tyler scatter estimator $\hat{T}_n = \hat{T}_n(\mathbb{X}_n)$ is defined as the solution of
$$\frac{p}{n}\sum_{i=1}^n \frac{(\mathbf{X}_i - \overline{\mathbf{X}}_n)(\mathbf{X}_i - \overline{\mathbf{X}}_n)^T}{(\mathbf{X}_i - \overline{\mathbf{X}}_n)^T \hat{T}_n^{-1}(\mathbf{X}_i - \overline{\mathbf{X}}_n)} = \hat{T}_n \tag{4}$$

which satisfies $\text{tr}(\hat{T}_n) = p$. It is regarded as the most robust M-estimator. Existence, uniqueness and asymptotic properties are treated in [7]. Apparently $\hat{T}_n$ satisfies
$$\hat{T}_n(\mathbb{X}_n A^T + \mathbf{1}\boldsymbol{b}^T) = \frac{p}{\text{tr}(A\hat{T}_n(\mathbb{X}_n)A^T)} A\hat{T}_n(\mathbb{X}_n)A^T$$

for $\boldsymbol{b} \in \mathbb{R}^p$ and any full rank $A \in \mathbb{R}^{p\times p}$, but not condition C1. As a consequence the asymptotic variance of $\hat{T}_n$ has a slightly different form than W in condition C2. Nonetheless the corresponding partial correlation estimator $\hat{P}_n^{(T)} = (\hat{T}_n^{-1})_D^{-\frac{1}{2}}\hat{T}_n^{-1}(\hat{T}_n^{-1})_D^{-\frac{1}{2}}$ satisfies (3). This is simply because, by choosing a suitable alternative normalization instead of setting the trace to p, one can obtain an estimator satisfying C1, which leads to the same partial correlation estimator $\hat{P}_n^{(T)}$. Precisely, we have the following result.

Proposition 3 *If $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are i.i.d. with distribution $F \in E_p(\boldsymbol{\mu}, \alpha\Sigma)$, $\alpha > 0$, and $\mathbb{E}||\boldsymbol{X} - \boldsymbol{\mu}||^2 < \infty$ and $\mathbb{E}||\boldsymbol{X} - \boldsymbol{\mu}||^{-\frac{3}{2}} < \infty$, then $\hat{P}_n^{(T)}$ fulfills (3) with $\sigma_1 = 1 + \frac{2}{p}$.*

Thus the scalar σ_1 is constant for the Tyler matrix, irrespective of the function g, i.e. the Tyler matrix (and hence the resulting partial correlation estimator) is distribution-free within the elliptical model. Moreover, it is more efficient than $\hat{\Sigma}_n$ at distributions with large (positive) kurtosis, i.e. heavy-tailed distributions. For instance, this holds true for the $t_{\nu,p}$-distribution if $\nu < p + 4$.

Final remark. Both moment conditions in Proposition 3 are only due to the location estimation in (4). Location estimators other than the mean are also possible and, in view of robustness, might be more appropriate, most notably the Hettmansperger-Randles median, cf. [3]. However, the inverse moment condition $\mathbb{E}||\boldsymbol{X} - \boldsymbol{\mu}||^{-\frac{3}{2}} < \infty$ can generally not be avoided by choosing a different location estimator, cf. [7]. But this is a fairly mild condition: for $p \geq 2$ it is fulfilled if g has no singularity at 0, thus including normal and $t_{\nu,p}$-distributions.

REFERENCES

[1] Baba, K., Shibata, R. and Sibuya, M. (2004) Partial correlation and conditional correlation as measures of conditional independence, *Aust. N. Z. J. Stat.* 46, 657-664.

[2] Bilodeau, M. and Brenner, D. (1999) *Theory of multivariate statistics*, Springer.

[3] Hettmansperger, T. and Randles, R. (2002) A practical affine equivariant multivariate median, *Biometrika* 89, 851-860.

[4] Lauritzen, S. (1996) *Graphical Models*, Oxford University Press.

[5] Tyler D. (1982), Radial estimates and the test for sphericity, *Biometrika* 69, 429-436.

[6] Tyler D. (1983), Robustness and efficiency properties of scatter matrices, *Biometrika* 70, 411-420.

[7] Tyler D. (1987), A distribution-free M-estimator of multivariate scatter, *Annals of Statistics* 15, 234-251.

[8] Whittaker, J. (1990) *Graphical models in applied multivariate statistics*, John Wiley & Sons.

[9] Zuo, Y. (2006) Robust location and scatter estimators in multivariate analysis, in Fan, J. and Koul, H. (ed.), *Frontiers in statistics. Dedicated to Peter John Bickel on honor of his 65th birthday.* Imperial College Press, 467-490.