# Inequalities on partial correlations in Gaussian graphical models containing star shapes

**Edmund Jones and Vanessa Didelez, School of Mathematics, University of Bristol**

## Abstract

This short paper proves inequalities that restrict the magnitudes of the partial correlations in star-shaped structures in Gaussian graphical models. These inequalities have to be satisfied by distributions that are used for generating simulated data to test structure-learning algorithms, but methods that have been used to create such distributions do not always ensure that they are. The inequalities are also noteworthy because stars are common and meaningful in real-world networks.

## 1. Introduction and definitions

Networks that model real-world phenomena often have few edges but several hubs, which are vertices that are connected to many others. Hubs are often important. For example, when Gaussian graphical models (GGMs) are used to model gene regulation networks, as for example in Friedman et al. (2000), Castelo and Roverato (2006, 2009), and Edwards et al. (2010), hubs are likely to correspond to genes that code for transcription factors that regulate other genes. If a GGM is used to model currency values, as in Carvalho et al. (2007), then a hub might correspond to a country that has large-scale trade with many others.

In a GGM the strength of the direct association between two vertices is measured by the magnitude of their partial correlation. The partial correlation is the correlation between the two vertices given the values of all the other vertices, and if there is no edge between the vertices then the partial correlation is zero (Lauritzen, 1996, section 5.1). This paper shows that the magnitudes of the partial correlations are always small, in a certain sense, in the case of a star, which is a structure that consists of a hub and a set of vertices that have edges to the hub but not to each other. (We use the term “star” because the results do not apply to a hub where two or more of the vertices that have edges to the hub also have edges to each other.)

Several definitions and relations will be needed. Let $G = (V, E)$ be an undirected graph, $V = \{1, \dots, n\}$, and $X \sim N_n(\mu, \Sigma)$, and suppose that $G$ is the graph for a graphical model that includes the distribution of $X$; this means that if $(i, j) \notin E$ then $X_i \perp\!\!\!\perp X_j \mid X_{V \setminus \{i,j\}}$ ($X_i$ is conditionally independent of $X_j$ given $X_{V \setminus \{i,j\}}$). For the multivariate Gaussian distribution, $X_i \perp\!\!\!\perp X_j \mid X_{V \setminus \{i,j\}} \Leftrightarrow \Omega_{ij} = 0$, where $\Omega = \Sigma^{-1}$ is the precision matrix. Let $M$ be the standardized precision matrix, which means that $M = D^{-1/2} \Omega D^{-1/2}$, where $d_{ij} = 0$ for $i \neq j$ and $d_{ii} = \omega_{ii}$. It follows that $m_{ij} = \omega_{ij} / \sqrt{\omega_{ii} \omega_{jj}}$ and $-1 \leq m_{ij} \leq 1$. The partial correlation between $X_i$ and $X_j$ is then $p_{ij} = -m_{ij}$, for $i \neq j$.

## 2. Sylvester’s criterion

Sylvester’s criterion states that a matrix is positive-definite if and only if the determinants of all its square upper-left submatrices are positive—these determinants are called the leading principal minors of the matrix. The origins of this result are obscure but some light is shed on them by Smith (2008). A proof is given in Gilbert (1991).

For GGMs it is common to assume that $\Sigma$ is positive-definite, which is equivalent to the support of $X$ being $\mathbb{R}^n$. This assumption is made in the propositions below. It is also equivalent to $\Omega$ or $M$ being positive-definite.

If $\Sigma$ is not assumed to be positive-definite, it must at least be positive-semidefinite. It might be conjectured that Sylvester’s criterion could be adapted to this case, to state that a matrix is positive-semidefinite if and only if all its leading principal minors are non-negative. But this is not true, and a counterexample is given by Swamy (1973).

## 3. The inequalities

This section presents four inequalities. The first two are the main results and the other two are corollaries. Proposition 1 is about graphs that consist entirely of a single star-structure.

**Proposition 1.** Suppose $E = \{\{1,2\},\{1,3\},\dots,\{1,n\}\}$, so that $G$ is a star-shaped graph centered at vertex 1. Then a necessary and sufficient condition for $M$ to be positive-definite is that $\sum_{i=2}^{n} p_{1\mathrm{i}}^2 < 1$.

**Proof.** One of the square upper-left submatrices of $M$ is the whole matrix $M$ itself.

$$M = \begin{pmatrix} 1 & m_{12} & m_{13} & \cdots & m_{1n} \\ m_{12} & 1 & 0 & \cdots & 0 \\ m_{13} & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ m_{1n} & 0 & 0 & \cdots & 1 \end{pmatrix}.$$

The determinant of $M$ is $1 - \sum_{i=2}^{n} m_{1i}^2$, and this being positive is equivalent to $\sum_{i=2}^{n} m_{1i}^2 < 1$. If this inequality holds then the other determinants in Sylvester's criterion are also positive. So this inequality on its own is a necessary and sufficient condition for $M$ being positive-definite, and it is obviously equivalent to $\sum_{i=2}^{n} p_{1i}^2 < 1$.

Proposition 2 is about graphs that contain star-shaped subgraphs. It states that if $G$ contains a star as an induced subgraph (Lauritzen, 1996, section 2.1.1), then a similar inequality to Proposition 1 holds in that subgraph as a necessary condition.

**Proposition 2.** Suppose $\{i,j_1\},\dots,\{i,j_s\} \in E$ and $\{j_a,j_b\} \notin E$ for all $a,b \in \{1,\dots,s\}$. Then $\sum_{a=1}^{s} p_{ij_a}^2 < 1$.

**Proof.** Relabel the vertices as follows: $i \to 1, j_1 \to 2, \dots, j_s \to s+1$. Sylvester's criterion implies that the determinant of the upper-left $(s+1) \times (s+1)$ matrix must be positive. This implies that $\sum_{i=2}^{s+1} m_{1i}^2 < 1$ and $\sum_{i=2}^{s+1} p_{1i}^2 < 1$, which is equivalent to the inequality in the proposition.

Proposition 3 is a corollary of Proposition 2 in which the inequality may be easier to interpret.

**Proposition 3.** If the graph is as in Proposition 2, then the mean magnitude of the partial correlations $p_{ij_1}, \dots, p_{ij_s}$ must be less than $1/\sqrt{s}$.

**Proof.** Suppose that $q_1, \dots, q_s \geq 0$ and $\sum_{a=1}^{s} q_a^2 = 1$. The method of Lagrange multipliers can be used to show that the maximum value of $(\sum_{a=1}^{s} q_a)/s$ is $1/\sqrt{s}$. Proposition 2 implies that $|p_{ij_1}|, \dots, |p_{ij_s}|$ satisfy the same conditions as $q_1, \dots, q_s$, except that the equals sign is replaced by a less-than sign. It follows that $\left(\sum_{a=1}^{s} |p_{ij_a}|\right)/s < 1/\sqrt{s}$.

For example, in a star with $s$ edges, at least one of these edges must have the magnitude of the corresponding partial correlation being less than $\sqrt{1/s}$. In any graph that contains a V-shape (three vertices with two edges), which means any graph that does not consist entirely of disjoint cliques, there must be at least one partial correlation on an edge that has magnitude less than $\sqrt{1/2} \approx 0.707$.

Two classes of graphs that contain many stars are forests and trees. Forests can be defined as graphs that have no cycles, and trees are connected forests. These are very restricted classes of graphs, but forest and tree graphical models have several advantages and have been widely studied and used (Willsky et al. 2002; Meilă and Jaakkola 2006; Eaton and Murphy 2007; Edwards et al. 2010; Anandkumar et al. 2012). If $G$ is a forest or tree, then Proposition 2 holds with $i$ as any of the vertices. In Proposition 4 this fact is used to make an inequality on the partial correlations throughout the graph.

**Proposition 4.** Suppose $G$ is a forest, for $i \in V$ let $\deg(i) = |\{j \in V : (i,j) \in E\}|$ be the degree of $i$, and let $L = \{i \in V : \deg(i) = 1\}$ be the set of leaves in $G$. Then

$$2 \sum_{\substack{(i,j)\in E,\\ i\notin L \text{ and } j\notin L}} p_{ij}^2 + \sum_{\substack{(i,j)\in E,\\ i\in L \text{ or } j\in L\\ \text{but not both}}} p_{ij}^2 < n - |L|.$$

**Proof.** Apply Proposition 2 to all the vertices in $V \setminus L$ in turn, and sum all these $n - |L|$ inequalities. For each $(i, j) \in E$, if $i \notin L$ and $j \notin L$ then $p_{ij}^2$ will appear twice on the left-hand side, if $i \in L$ or $j \in L$ but not both then it will appear once, and if both $i \in L$ and $j \in L$ then it will not appear at all.

It might be conjectured that if the inequality in Proposition 2 is satisfied for all stars that appear as induced subgraphs of $G$, then $M$ is positive-definite. But this does not hold. For example, if $n = 4$, $E = \{\{1,2\},\{2,3\},\{3,4\},\{4,1\}\}$, and $p_{12} = p_{23} = p_{34} = p_{41} = 0.7$, then this condition is satisfied but $M$ is not positive-definite. It might then be conjectured that if you add the further condition that $G$ is decomposable then $M$ will be positive-definite, but this does not hold either—for example, let $E = \{\{1,2\},\{2,3\},\{3,4\},\{4,1\},\{1,3\}\}$, and $p_{12} = p_{23} = p_{34} = p_{41} = p_{13} = 0.7$.

For shapes other than stars, there do not seem to be any inequalities that are as notable as Propositions 1 and 2. It is possible to write down the inequalities that result from Sylvester's criterion, but it is generally not easy to rearrange them into a meaningful form.

The conditional independence relations shown by graphs also imply conditions on the marginal correlations (which are usually just called correlations). These can be found by inverting $M$ and then standardizing to find the correlation matrix $C$. For example, if the graph is as in Proposition 1 then $c_{jk} = c_{1j}c_{1k}$ for all $j, k \in \{2, \dots, n\}$. This is a special case of the fact that the correlation between two vertices in a tree is the product of the correlations along the edges that connect them (Pearl, 1988, section 8.3.4; Tan et al., 2010).

## 4. Relevance to experiments on structure-learning algorithms

Proposition 2 arises when doing a certain type of experiment on algorithms for learning the structure of GGMs from data. In these experiments, simulated data is generated from a multivariate Gaussian distribution that corresponds to a known graph, then the structure-learning algorithm is used on the data, and finally the output of the algorithm is compared with the original graph. The first step in making the simulated data is to create a covariance matrix $\Sigma$ that corresponds to the original graph, and naturally this $\Sigma$ has to fulfil the inequality in Proposition 2.

Numerous publications describe experiments of this type, for example Friedman et al. (2007), Moghaddam et al. (2009), Albieri (2010), Wang and Li (2012), Wang (2012), and Green and Thomas (2013). Most of these do not mention the issue of ensuring that $\Sigma$ is positive-definite, suggesting that it was not a problem. One experiment that does mention the issue appears in Meinshausen and Bühlmann (2006). They used large graphs whose vertices have maximum degree 4, and chose all the partial correlations to be 0.245. They state without proof that absolute values less than 0.25 guarantee that $\Omega$ is positive-definite—this condition is stronger than Proposition 2, which implies only that the mean absolute value has to be less than 0.5.

One detailed procedure for creating $\Sigma$ is described for the first example in section 4.1 of Guo et al. (2011). This procedure presumably gave positive-definite matrices when it was used for this example, with $n = 100$ and small numbers of extra edges, but it does not always do so. The procedure starts with $V = \{1, \ldots, n\}$, $E = \big\{\{1,2\}, \{2,3\}, \ldots, \{n-1, n\}\big\}$, and $\Sigma$ defined by $\sigma_{ij} = \exp\big(-\big|s_i - s_j\big|/2\big)$, where $s_i - s_{i-1} \sim Unif(0.5,1)$. (This formula for $\sigma_{ij}$ guarantees that $\Omega$ is tridiagonal, as required by the graph.) Extra edges are then added at random, presumably from a uniform distribution, and for each extra edge the two corresponding elements of $\Omega$ are set to be a value from $Unif([-1,-0.5] \cup [0.5,1])$.

To show how this procedure sometimes fails, it is convenient to start by considering specific values from the uniform distributions that are used, though obviously the probability that these exact values would be drawn is zero. Suppose that $n \geq 4$, $s_i - s_{i-1} = 0.9$ for $i = 2, \ldots, n$, the extra edges include {1,3} and {1,4} but no other edges between any of the first five vertices (the first four if $n = 4$), and the corresponding four new elements of $\Omega$ are all 0.95. The exact value of $\Omega$ can now be calculated by using equation 3.2 in Barrett (1979) to calculate the tridiagonal $\Omega$ and then adding the new elements. Let

$$\alpha = \frac{1}{1 - \mathrm{e}^{-0.9}}\,, \quad \beta = \frac{1 - \mathrm{e}^{-1.8}}{(1 - \mathrm{e}^{-0.9})^2}\,, \quad \text{and} \quad \gamma = -\frac{\mathrm{e}^{-0.45}}{1 - \mathrm{e}^{-0.9}}\,.$$

Then

$$\Omega = \begin{pmatrix} \alpha & \gamma & 0.95 & 0.95 & 0 & \cdots \\ \gamma & \beta & \gamma & 0 & 0 & \cdots \\ 0.95 & \gamma & \beta & \gamma & 0 & \cdots \\ 0.95 & 0 & \gamma & \beta & \gamma & \cdots \\ 0 & 0 & 0 & \gamma & \beta & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

For all $n \geq 6$ the upper-left $5 \times 5$ submatrix of $\Omega$ is the same, and its determinant is negative, which means that $\Omega$ is not positive-definite; the cases $n = 4$ and $n = 5$ can be checked separately. This argument still holds if all the instances of 0.9 and 0.95 are replaced by slightly different values, because if the determinant of the upper-left $5 \times 5$ submatrix is written as a function of $s_1, \ldots, s_n, \omega_{13}$, and $\omega_{14}$, then this function is continuous. It follows that the procedure fails with positive probability for all $n \geq 4$, assuming that it is possible for the new edges between the first five vertices to be as in this counterexample.

## 5. Discussion

Proposition 1 is a necessary and sufficient condition for the covariance matrix to be positive-definite, but it only applies to graphs that consist of a single star-structure. Proposition 2 applies much more widely, to graphs that contain star-structures, but it is only necessary, not sufficient. Nevertheless, Proposition 2 is useful and important in practice. When creating a covariance matrix it is natural to want to choose specific values for the partial correlations, and Proposition 2 places strong restrictions on what these can be.

Proposition 2 has several other interesting consequences or interpretations. If $X_1$ has sufficiently strong direct associations (partial correlations) with $X_2$ and $X_3$ then there must also be a direct association between $X_2$ and $X_3$. On the other hand, if $X_2$ and $X_3$ are both almost deterministic functions of $X_1$ (and not of each other), then the marginal correlations $c_{12}$ and $c_{13}$ will be close to 1 or $-1$, but at least one of the partial correlations $p_{12}$ and $p_{13}$ must have magnitude less than $1/\sqrt{2}$. Obviously both of these consequences generalize to larger $n$.

The proofs of Propositions 1 and 2 are straightforward applications of Sylvester's criterion. This criterion is well known in some fields but does not seem to have been previously used or even mentioned in connection with partial correlation matrices for GGMs.

Since Proposition 2 is not a sufficient condition, the question arises of how to create a possible covariance matrix for an arbitrary given graph. There are several methods that are guaranteed to work, though these do not easily allow specific values to be chosen for the partial correlations or elements of the covariance matrix. One method is described in the appendix of Roverato (2002). This uses the Cholesky decomposition $\Omega = \Phi^T\Phi$, where $\Phi$ is an upper-triangular matrix. The diagonal elements of $\Phi$ and the elements that correspond to edges in the graph can be chosen freely, and the other elements have to be calculated according to Roverato's equation (10). For decomposable graphs, the calculations for this second set of elements can be avoided—if the vertices are ordered according to a perfect vertex elimination scheme (Lauritzen, 1996, section 2.1.3), then these elements are all zero.

An alternative method to create a covariance matrix for any graph is as follows. Start with any $n \times n$ symmetric matrix in which the diagonal elements are positive and the elements corresponding to absent edges are zero, find its eigenvalues, and if any of these are negative then let $-\lambda$ be the lowest one and add $(\lambda + \epsilon)I_n$ to the matrix, for some $\epsilon > 0$. The resulting matrix's eigenvalues are all positive, which means that it is positive-definite, and it still has the symmetry and the zeroes in the same places.

Propositions 1–4 also apply to directed acyclic graphical models (also known as Bayesian networks), because stars in undirected graphical models are equivalent to stars in directed acyclic graphical models, if the edges are all directed from the hub to the other vertices. The edges are oriented like this if the hub corresponds to a gene that codes for a transcription factor, for example.

Inequalities that are essentially the same as Propositions 1–4 also apply to covariance graphical models (Wermuth & Cox 2001), in which an edge that is absent from the graph means that the two variables are marginally independent (rather than conditionally independent as in GGMs) and corresponds to zeroes in the covariance and correlation matrices (rather than the precision and partial correlation matrices). The four propositions hold for these models if $M$ is just replaced by the correlation matrix and $p_{ij}$ is replaced by the correlation between $X_i$ and $X_j$. Covariance graphical models are an active topic of research (Chaudhuri et al. 2008; Drton and Richardson 2008; El Karoui 2008; Bien and Tibshirani 2011; Wang 2014a,b) and can be used to analyze gene expression data, protein networks, and financial data.

## Acknowledgements

This work was supported by the Engineering and Physical Sciences Research Council (grant number EP/P50483X/1). Financial support from the Leverhulme Trust (Research Fellowship RF-2011-320) is gratefully acknowledged. The funding organizations had no direct involvement in the research work or the decision to submit the paper for publication.

# References

Albieri, V. (2010). A comparison of procedures for structural learning of biological networks. PhD thesis, Università degli Studi di Padova.

Anandkumar, A., Tan, V.Y.F., Huang, F., Willsky, A.S. (2012). High-dimensional structure learning of Ising models: local separation criterion. *The Annals of Statistics* 40 (3): 1346–1375.

Barrett, W.W. (1979). A theorem on inverses of tridiagonal matrices. *Linear Algebra and its Applications* 27:211–217.

Bien, J., Tibshirani, R. J. (2011). Sparse estimation of a covariance matrix. *Biometrika* 98(4):807–820.

Carvalho, C.M., Massam, H., West, M. (2007). Simulation of hyper-inverse Wishart distributions in graphical models. *Biometrika* 94 (3):647–659.

Castelo, R., Roverato, A. (2006). A robust procedure for Gaussian graphical model search from microarray data with *p* larger than *n*. *Journal of Machine Learning Research* 7:2621–2650.

Castelo, R., Roverato, A. (2009). Reverse engineering molecular regulatory networks from microarray data with qp-graphs. *Journal of Computational Biology* 16 (2):213–227.

Chaudhuri, S., Drton, M., Richardson, T.S. (2007). Estimation of a covariance matrix with zeros. *Biometrika* 94 (1): 199–216.

Drton. M., Richardson, T.S. (2008). Graphical methods for efficient likelihood inference in Gaussian covariance models. *Journal of Machine Learning Research* 9: 893–914.

Eaton, D., Murphy, K. (2007). Bayesian structure learning using dynamic programming and MCMC. *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence* 101–108.

Edwards, D., De Abreu, G.C.G., Labouriau, R. (2010). Selecting high-dimensional mixed graphical models using minimal AIC or BIC forests. *BMC Bioinformatics* 11 (18).

El Karoui, N. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *The Annals of Statistics* 36 (6): 2717–2756.

Friedman, N., Linial, M., Nachman, I., Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *Journal of Computational Biology* 7 (3/4):601–210.

Friedman, J., Hastie, T., Tibshirani, R. (2007). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 9 (3):432–441.

Gilbert, G.T. (1991). Positive definite matrices and Sylvester's criterion. *The American Mathematical Monthly* 98 (1):44–46.

Green, P.J., Thomas, A. (2013). Sampling decomposable graphs using a Markov chain on junction trees. *Biometrika* 1–20.

Guo, J., Levina, E., Michailidis, G., Zhu, J. (2011). Joint estimation of multiple graphical models. *Biometrika* 98 (1):1–15.

Lauritzen, S.L. (1996). *Graphical Models*. Oxford: Oxford University Press.

Meilă, M., Jaakkola, T. (2006). Tractable Bayesian learning of tree belief networks. *Statistics and Computing* 16:77–92.

Meinshausen, N., Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics* 34 (3):1436–1462.

Moghaddam, B., Marlin, B.M., Khan, M.E., Murphy, K.P. (2009). Accelerating Bayesian structural inference for non-decomposable Gaussian graphical models. *Advances in Neural Information Processing Systems 22* 1285–1293.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco: Morgan Kaufmann.

Roverato, A. (2002). Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models. *Scandinavian Journal of Statistics* 29 (3):391–411.

Smith, J.T. (2008). Epilog: Sylvester's criterion. http://math.sfsu.edu/smith/Math880/General/Epilog.pdf. Accessed on 14th July 2014.

Swamy, K.N. (1973). On Sylvester's criterion for positive-semidefinite matrices. *IEEE Transactions on Automatic Control* 18 (3):306.

Tan, V.Y.F., Anandkumar, A., Willsky, A.S. (2010). Learning Gaussian tree models: analysis of error exponents and extremal structures. *IEEE Transactions on Signal Processing* 58 (5):2701–2714.

Wang, H. (2012). Bayesian graphical lasso models and efficient posterior computation. *Bayesian Analysis* 7 (4):867–886.

Wang, H. Li, S.Z. (2012). Efficient Gaussian graphical model determination under G-Wishart prior distributions. *Electronic Journal of Statistics* 6:168–198.

Wang, H. (2014a). Coordinate descent algorithm for covariance graphical lasso. *Statistics and Computing* 24:521-529.

Wang, H. (2014b). Scaling it up: Stochastic graphical model determination under spike and slab prior distributions. (Working paper.)

Wermuth, N., Cox, D.R. (2001). Graphical models: overview. In Baltes, P.B. & Smelser, N.J., eds. *International Encyclopedia of the Social and Behavioral Sciences*. Elsevier: Amsterdam.

Willsky, A.S. (2002). Multiresolution Markov models for signal and image processing. *Proceedings of the IEEE* 90 (8):1396–1458.