Parameterising Gaussian Graphical Models

Jack Storror Carter JACK.CARTER@UPF.EDU
Department of Business and Economics
Universitat Pompeu Fabra
Barcelona, Spain

Editors: Gustau Camps-Valls, Manuele Leonelli and Gherardo Varando

Abstract

Gaussian graphical models (GGMs) describe the dependence structure among jointly Gaussian random variables. However, the most common parameterisation of GGMs, the precision matrix, describes both the dependence and scale of the variables. This has been shown to lead to model selection methods that depend on the scale of the variables, despite graphical models being scale invariant. Even after standardising data to have unit sample variances, entries of the precision matrix can be on different scales leading to poor model selection. This paper decomposes the precision matrix into marginal variances and interpretable scale-invariant parameters - the partial correlations and variance inflation factors. This decomposition gives new insights into the precision matrix entries and the dynamics of model selection methods such as penalised likelihoods. In particular, it explains the observed phenomenon that penalties on the precision matrix perform poorly at selecting hub variables and motivates the necessity of data standardisation. It also shows why methods based on partial correlations have better hub detection properties. The effect of penalisation of different quantities on the estimation of marginal variances is then investigated and an interesting simplification of the log-likelihood is shown when using maximum likelihood estimation of the marginal variances.

Keywords: Gaussian graphical model, precision matrix, correlation matrix, partial correlation, model selection, penalised likelihood.

1. Introduction

The distribution of a Gaussian random vector $X = (X_1, \ldots, X_p)$ is fully specified by its first and second order moments. The univariate moments correspond to the marginal properties of location, through the first order moments $\mathbb{E}[X_i]$, and scale, through the second order moments $\mathbb{E}[X_i^2]$. The bivariate second order moments $\mathbb{E}[X_i X_j]$ describe the dependence structure between variables - in the case of Gaussian random vectors, dependence is completely determined by these pairwise relationships. However, the standard Gaussian parameterisation of the mean vector $\mu = 0$, which is assumed equal to 0 for simplicity, and the positive definite covariance matrix Σ does not cleanly separate the concepts of scale and dependence. This is because the off-diagonals of Σ are a function of both the marginal variances and the pairwise correlations. For this reason a popular alternative is to parameterise directly by the marginal variances and the correlation matrix. The correlation matrix is scale invariant and therefore gives a clean description of the dependence structure.

Gaussian graphical models (GGMs) combine the Gaussian distributional assumption with the conditional independence assumptions of an undirected graphical model (see Uhler (2019) for an overview). A graphical model is a model of only the dependence structure of

© 2026 J.S. Carter.

X and is invariant to scale. As such, a graphical model only makes assumptions about the correlation matrix. However, the connection between marginal correlations and conditional independence is non-trivial which lead to alternative parameterisations.

The most popular parameterisation for GGMs is the precision matrix $\Theta := \Sigma^{-1}$ due to its connection to conditional independence - $\Theta_{ij} = 0$ if and only if $X_i \perp\!\!\!\perp X_j \mid X_{-(i,j)}$. The sparsity structure of Θ therefore directly implies a graphical model. However, while the zeros in Θ indicate conditional independence, the non-zeros are not a good measure for conditional dependence leading to two main issues. First, being a one to one transformation of Σ rather than the correlation matrix, Θ is not scale invariant. The issue is more pronounced here because, unlike Σ where only the off-diagonals conflate scale and dependence, every entry of Θ is a function of both scale and dependence. Second, as will be demonstrated in Section 3, even when data has been standardised, the non-zero off-diagonals of Θ are not easy to interpret and are not a clear measure of strength of conditional dependence.

These issues were partly addressed by Carter et al. (2024) who pointed out the lack of scale invariance of Θ and instead advocated for encoding conditional independence by partial correlations. Partial correlations are inherently scale invariant and are a clear measure of conditional dependence. However, this only addressed part of the issue with the remaining diagonal entries of Θ also being functions of both scale and dependence. Therefore, a parameterisation combining partial correlations with the diagonal of Θ does not fully separate the two concepts. This paper will further decompose the diagonal of Θ into marginal variances and a scale invariant component - the variance inflation factors (VIFs).

It will be shown that the VIFs are functions of the partial correlations, therefore allowing Θ to be reparameterised by the pair of marginal variances and partial correlations. This fully separates scale and dependence while also encoding conditional independence. The VIFs also have a clear interpretation as a measure of overall dependence of X_i on X_{-i}. Decomposing the entries of Θ into partial correlations, marginal variances and VIFs gives an interpretable explanation for the precision matrix entries and new understanding on how methods based on the precision matrix work.

The main focus of this paper will be on model selection through the lens of penalised likelihoods. For the precision matrix Θ, this is an estimator of the form

$$\hat{\Theta}_{\text{Pen}}(S) := \underset{\Theta \succ 0}{\operatorname{argmax}}\, l(\Theta \mid S) - \sum_{i=1}^{p} \text{pen}_D(\theta_{ii}) - \sum_{i<j} \text{pen}(\theta_{ij}),$$

where S is the sample covariance matrix, l is the log-likelihood function, pen_D is a penalty function on the diagonals and pen is a penalty function on the off-diagonals. Penalised likelihood estimates can analogously be defined for other parameterisations. The penalty function pen is generally chosen to be decreasing on $(-\infty, 0)$ and increasing on $(0, \infty)$ so that estimates are shrunk towards 0. When the penalty is non-differentiable at 0, this can result in sparse estimation of Θ.

Many popular methods for GGM selection fit into this framework, for example l_1 penalisation via graphical LASSO (GLASSO) (Yuan and Lin, 2007; Banerjee et al., 2008; Friedman et al., 2008), non-convex penalties (Fan and Li, 2001; Fan et al., 2009; Zhang, 2010) and penalties that approximate the l_0 (Dicker et al., 2013; Wang and Zhu, 2016; Williams and Rast, 2020). They are popular due to their simple construction and generally fast computation.

However, any such penalised likelihood method on Θ, other than l_0 penalisation, is not scale invariant (Carter et al., 2024, Proposition 1), with the resulting estimate and selected graphical model being sensitive to the scaling of the data or standardisation choices. As an alternative, Carter et al. (2024) proposed direct penalisation of the partial correlations, for which estimation is scale invariant. An l_1 penalisation, called partial correlation graphical LASSO (PCGLASSO) was implemented and shown in simulated experiments to outperform GLASSO, even when data was standardised. The advantage of PCGLASSO was most pronounced in hub settings where the data generating graphical model has high degree vertices. This experimental observation was then backed up theoretically by Bogdan et al. (2025) who proved model selection consistency of PCGLASSO under a weaker condition than GLASSO, with the greatest gap in these conditions being for hub graphs. Further real data and simulated experiments backed up this theoretical result.

Previously, the poor performance of GLASSO at selecting hubs has been attributed, using a Bayesian argument, to it apriori assigning equal and independent probability to each edge being present, resulting in a small prior probability of high degree nodes (Tan et al., 2014). This motivated an adaptation of the GLASSO penalty to actively encourage hubs (Wang et al., 2025). However, the improved performance of PCGLASSO in detecting hubs goes against this reasoning. The reparameterisation of Θ in this paper presents an alternative explanation, showing that GLASSO systematically places higher penalty on hubs, even when data has been standardised. This explanation extends beyond GLASSO to any strictly increasing penalty on Θ, showing that this is an inherent problem of penalised likelihood methods on Θ. On the other hand, penalties on the partial correlations do not have this property of penalising hubs.

Many other methods have been proposed for GGM selection. Early methods did focus on partial correlations through hypothesis testing of conditional independence, see for example Smith and Whittaker (1998); Drton and Perlman (2007). However, hypothesis testing procedures are typically computationally complex which lead to the popularity of shrinkage methods which allow for model selection in large dimensions. At first this was done by relating a GGM to a set of linear regressions for which shrinkage methods were well established (Meinshausen and Bühlmann, 2006; Liu and Luo, 2015; Peng et al., 2009). At a similar time GLASSO was proposed and this sparked the popularity of shrinkage methods on the precision matrix. For example, some methods optimise a function other than the log-likelihood, usually combined with an l_1 penalty - SPLICE uses a pseudo-likelihood (Rocha et al., 2008), CLIME minimises the L_1 norm of Θ under a constraint that Θ^{-1} remain close to S (Cai et al., 2011) and a D-trace loss function has also been proposed (Zhang and Zou, 2014). A fully Bayesian approach to model selection would usually use the conjugate G-Wishart prior (Wang and Li, 2012). Other Bayesian methods are more similar to the penalised likelihood approach and use a shrinkage prior to encourage sparsity - these include the Bayesian GLASSO (Wang, 2012), graphical horseshoe (Li et al., 2019) and spike and slab priors (Banerjee and Ghosal, 2015; Gan et al., 2019; Jewson et al., 2024; Sulem et al., 2025). Although the focus of this paper is on penalised likelihood methods, the reparameterisation of Θ in this paper can be used to give new insights and interpretations in any of these methods.

A number of different parameterisations will be introduced through the paper. For reference, these are summarised in Table 1.

Parameter	Description
Σ	Covariance matrix
$\sigma^2 = \text{diag}(\Sigma)$	Marginal variances
$R = \text{cov2cor}(\Sigma)$	Correlation matrix
$\Theta = \Sigma^{-1}$	Precision matrix
$\theta = \text{diag}(\Theta)$	Precision matrix diagonal (inverse partial variances)
$\Delta = \text{cov2cor}(\Theta)$	Negative partial correlation matrix
$\psi = \text{diag}(\Delta^{-1})$	Variance inflation factors
$K = R^{-1}$	Inverse correlation matrix

Table 1: Summary of different parameters.

2. Covariance and Correlation

The most common parameterisation of a Gaussian distribution is the covariance matrix $\Sigma = (\Sigma_{ij})$. The diagonals are marginal variances, $\Sigma_{ii} = \text{Var}(X_i)$, and the off-diagonals are pairwise covariances, $\Sigma_{ij} = \text{Cov}(X_i, X_j)$. For an i.i.d. sample $x_1, \ldots, x_n \sim N(0, \Sigma)$ and defining the sample covariance matrix as $S := \frac{1}{n}\sum_{i=1}^n x_i x_i^\top$, the log-likelihood for Σ, after removing additive and multiplicative constants (as they will be for subsequent log-likelihoods), is

$$l(\,\Sigma \mid S\,) := -\log\det\Sigma - \text{tr}(S\Sigma^{-1}),$$

and the maximum likelihood estimate (MLE) is $\hat{\Sigma} = S$. The MLE follows a rescaled Wishart distribution $nS \sim \mathcal{W}_p(\Sigma, n)$ and is an unbiased estimate of Σ. Marginally, the diagonal entry S_{ii} follows a rescaled χ^2_n distribution with mean Σ_{ii} and variance $2\Sigma_{ii}^2/n$ (Muirhead, 2009, Section 3.2). Notice that the variance does not depend on the dimension p and so, for moderate sample sizes, one can be reasonably assured that $S_{ii} \approx \Sigma_{ii}$ regardless of dimension.

While the diagonals Σ_{ii} describe the scale of the variables and are invariant to dependence between variables, seen by $\text{Var}(X_i)$ being a marginal property, the off-diagonals Σ_{ij} are a function of both the scale of and dependence between the variables. That they are not scale invariant follows because scalar multiples of the variables generally have different Σ_{ij} - letting $Y_i = a_i + b_i X_i$ with $b_i \neq 0$, $\text{Cov}(Y_i, Y_j) = b_i b_j \text{Cov}(X_i, X_j)$ which are only equal for all b_i, b_j when $\text{Cov}(X_i, X_j) = 0$.

To separate the concepts of scale and dependence, one can instead combine the marginal variances, now denoted by the diagonal matrix $\sigma^2 = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$ with $\sigma_i^2 := \Sigma_{ii}$, with the correlation matrix, the matrix $R = (R_{ij})$ with unit diagonal satisfying $\Sigma = \sigma R \sigma$. Using the name of the `R` function, denote by $R = \text{cov2cor}(\Sigma)$ the function that scales a general positive definite matrix into the corresponding correlation matrix with unit diagonal. Since Σ is a one to one function of (σ^2, R), this is a valid parameterisation and fully specifies the Gaussian distribution.

The correlations $R_{ij} = \text{Corr}(X_i, X_j)$ are scale invariant because

$$\text{Corr}(Y_i, Y_j) = \frac{\text{Cov}(Y_i, Y_j)}{\sqrt{\text{Var}(Y_i)\text{Var}(Y_j)}} = \frac{b_i b_j \text{Cov}(X_i, X_j)}{\sqrt{b_i^2 \text{Var}(X_i) b_j^2 \text{Var}(X_j)}} = \text{Corr}(X_i, X_j).$$

Hence the (σ^2, R) parameterisation cleanly separates the concepts of scale, through σ^2, and dependence, through R. Therefore a graphical model, a model of the conditional independence relationships between variables, only places restrictions on R with no restrictions or assumptions made on σ^2.

However, because the relationship between correlations and conditional independence is non-trivial, different parameterisations are generally used for GGMs that directly encode conditional independence via sparsity.

3. Precision and partial correlation

The most popular parameterisation of a GGM is the precision matrix $\Theta := \Sigma^{-1}$. The precision is the canonical parameter of a Gaussian distribution with a particularly simple form for the log-likelihood (Uhler, 2019)

$$l(\Theta \mid S) := \log\det\Theta - \operatorname{tr}(S\Theta). \tag{1}$$

Its diagonal entries are equal to the inverse partial variances $\Theta_{ii} = 1/\operatorname{Var}(X_i \mid X_{-i})$. The off-diagonals do not have such an obvious interpretation, but the reparameterisations in subsequent sections will provide one.

The popularity of Θ for parameterising GGMs comes from its clear encoding of conditional independence by its zero entries with $\Theta_{ij} = 0$ if and only if $X_i \perp\!\!\!\perp X_j \mid X_{-(i,j)}$ (Uhler, 2019). However, this comes at the cost of harder inference. The MLE of Θ only exists when S is positive definite, which happens with probability 1 when $n \geq p$ but probability 0 when $n < p$ (Mathai et al., 2022, Section 8.3).[1] When it exists, the MLE is $\hat{\Theta} = S^{-1}$ which follows a scaled inverse-Wishart distribution $S^{-1}/n \sim \mathcal{W}_p^{-1}(\Theta, n)$. This is a biased estimate of Θ with mean $\frac{n}{n-p-1}\Theta$ and the diagonals of S^{-1} follow a scaled inverse-Gamma distribution $(S^{-1})_{ii}/n \sim IG((n-p+1)/2, \Theta_{ii}/2)$ so that $(S^{-1})_{ii}$ has mean $\frac{n}{n-p-1}\Theta_{ii}$ and variance $\frac{2n^2}{(n-p-1)^2(n-p-3)}\Theta_{ii}^2$ (Gupta and Nagar, 1999, Section 3.3). Note that both the bias and variance are increasing with p, and the variance is generally much larger than that of S_{ii}. Hence one would need a much larger sample size in order to have any reasonable assurance that $(S^{-1})_{ii} \approx \Theta_{ii}$, with the required n increasing with p.

A second issue with the precision matrix is that, being a function of Σ, it is a measure of both scale and dependence, and the conflation of the two concepts is worse here because every entry of Θ depends on both scale and dependence. For example, the diagonal of Θ being the inverse partial variances shows that this is a multivariate property and therefore a function of the dependence between variables. That all entries of Θ are not scale invariant follows from $\Theta(Y) = b^{-1}\Theta b^{-1}$ where $\Theta(Y)$ is the precision matrix of $Y = a + bX$ and $b = \operatorname{diag}(b_1, \ldots, b_p)$.

A first step to decoupling scale and dependence in Θ is to consider the negative partial correlation matrix, $\Delta := \operatorname{cov2cor}(\Theta)$. The entries of Δ are $\Delta_{ij} = \Theta_{ij}/\sqrt{\Theta_{ii}\Theta_{jj}} = -\operatorname{Corr}(X_i, X_j \mid X_{-(i,j)})$. That Δ is scale invariant follows from $\Delta(Y) = \operatorname{cov2cor}(b^{-1}\Theta b^{-1}) = \operatorname{cov2cor}(\Theta) = \Delta$. Writing the diagonal of Θ as $\theta = \operatorname{diag}(\theta_1, \ldots, \theta_p)$ with $\theta_i := \Theta_{ii}$,

1. This is the case when the mean μ is known. When μ is unknown, the sample covariance matrix $S = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^\top$ is positive definite only when $n > p$.

$\Theta = \theta^{1/2}\Delta\theta^{1/2}$. Hence (θ, Δ) is a one to one transformation of Θ and a valid parameterisation. The log-likelihood is

$$l(\theta, \Delta \mid S) = \sum_{i=1}^{p} \log \theta_i + \log\det \Delta - \mathrm{tr}(S\theta^{1/2}\Delta\theta^{1/2}).$$

Since the covariance matrix is related to the correlation matrix via $\Sigma = \sigma R \sigma$ and can also be written in terms of the partial correlation matrix as $\Sigma = (\theta^{1/2}\Delta\theta^{1/2})^{-1} = \theta^{-1/2}\Delta^{-1}\theta^{-1/2}$ it follows that $R = \sigma^{-1}\theta^{-1/2}\Delta^{-1}\theta^{-1/2}\sigma^{-1} = \mathrm{cov2cor}(\Delta^{-1})$, or conversely, $\Delta = \mathrm{cov2cor}(R^{-1})$. This is a one to one transformation meaning that Δ contains the same information as the correlation matrix and fully determines the dependence structure.

However, under this parameterisation θ is still a function of both scale and dependence and, as seen above, its estimation is challenging, requiring a large sample size relative to dimension for an accurate MLE.

4. Variance inflation factors

As noted above, the diagonals of the precision matrix are equal to the inverse partial variances $\theta_i = 1/\mathrm{Var}(X_i \mid X_{-i})$. It was seen that this quantity is not scale invariant, and it is also a function of dependence - for fixed marginal variances, $\mathrm{Var}(X_i \mid X_{-i})$ gets smaller as dependence between X_i and X_{-i} increases. To separate scale and dependence, θ_i can be decomposed into the marginal variance σ_i^2 and a scale invariant parameter

$$\psi_i := \mathrm{Var}(X_i)/\mathrm{Var}(X_i \mid X_{-i}) = \sigma_i^2\theta_i,$$

which will be referred to as the variance inflation factor (VIF) of X_i. This is the factor by which the variance of X_i increases when conditioning on X_{-i} is removed. This is related to the coefficient of determination R^2 in linear regression with $\psi_i = 1/(1 - R_i^2)$. It can be shown that $\mathrm{Var}(X_i \mid X_{-i}) \leq \mathrm{Var}(X_i)$ (intuitively, conditioning cannot increase the variance) and therefore $\psi_i \in [1, \infty)$. When $\psi_i = 1$, X_i is independent of all of X_{-i}, while $\psi_i \to \infty$ means that X_i is a linear function of X_{-i} resulting in a degenerate distribution. A large value of ψ_i implies that the variance of $X_i \mid X_{-i}$ is much smaller than that of X_i, meaning that X_{-i} are good predictors of X_i. The value of ψ_i therefore gives a measure of the total dependence between X_i and X_{-i}. In the graphical model setting, a large ψ_i typically occurs when X_i is of high degree (has many edges) with dependence on many of the X_{-i}. However, large ψ_i can also occur when X_i has few edges with particularly strong dependencies.

It can easily be shown that ψ_i is scale invariant

$$\psi_i(Y) = \frac{\mathrm{Var}(Y_i)}{\mathrm{Var}(Y_i \mid Y_{-i})} = \frac{b_i^2\mathrm{Var}(X_i)}{b_i^2\mathrm{Var}(X_i \mid X_{-i})} = \psi_i.$$

This implies that ψ_i is only a measure of dependence and is therefore a function of Δ, which was shown to fully determine the dependence structure. Recalling the above connection that $\Delta^{-1} = \theta^{1/2}\sigma R \sigma\theta^{1/2} = \psi^{1/2}R\psi^{1/2}$ where $\psi = \mathrm{diag}(\psi_1, \ldots, \psi_p)$, the diagonal of Δ^{-1} is equal to ψ. Similarly, the inverse correlation matrix is $R^{-1} = \sigma\theta^{1/2}\Delta\theta^{1/2}\sigma = \psi^{1/2}\Delta\psi^{1/2}$. So ψ can be calculated directly by the diagonal of either Δ^{-1} or R^{-1}. This also implies that the

pair (σ^2, Δ) of marginal variances and partial correlation matrix is a one to one function of Σ and so fully determines the Gaussian distribution.

The log-likelihood for (σ^2, Δ) is

$$l(\sigma^2, \Delta \mid S) = -\sum_{i=1}^{p} \log \sigma_i^2 + \sum_{i=1}^{p} \log \psi_i + \log\det \Delta - \operatorname{tr}(S\sigma^{-1}\psi^{1/2}\Delta\psi^{1/2}\sigma^{-1}).$$

This can be simplified by considering the inverse correlation matrix $K := R^{-1} = \psi^{1/2}\Delta\psi^{1/2}$,

$$l(\sigma^2, K \mid S) = -\sum_{i=1}^{p} \log \sigma_i^2 + \log\det K - \operatorname{tr}(S\sigma^{-1}K\sigma^{-1}). \tag{2}$$

The off-diagonals of K are $\Delta_{ij}\sqrt{\psi_i\psi_j}$ so sparsity still indicates conditional independence.

5. Effects on model selection

The previous sections decomposed the precision matrix into three interpretable parameters: the partial correlations Δ_{ij} which measure the dependence between X_i and X_j as well as indicating conditional independence; the VIFs ψ_i which measure the total dependence of X_i on X_{-i}; and the marginal variances σ_i^2. The precision matrix off-diagonals can be written in terms of these interpretable parameters as

$$\Theta_{ij} = \frac{\Delta_{ij}\sqrt{\psi_i\psi_j}}{\sigma_i\sigma_j}. \tag{3}$$

A penalty function on Θ therefore has the effect of penalising the partial correlations Δ_{ij} (the desired effect) as well as penalising the VIFs and inflating the marginal variances (see Section 6). In other words, partial correlations associated to larger VIFs or smaller marginal variances receive larger penalty. In the case of marginal variances, this is bad because the magnitude of penalisation depends on the arbitrary scaling of the variables, however this can be mitigated by data standardisation. In the case of VIFs, this means that variables with the highest overall dependence will receive the most penalisation. When a large ψ_i is due to a small number of strong dependencies, the likelihood is typically strong enough to overpower this high penalty. However, when a large ψ_i is due to a large number of weaker dependencies, i.e. when X_i is a hub, the likelihood function may not have enough evidence for $\Delta_{ij} \neq 0$ and remove these hub edges due to the high penalty associated to them. These phenomena will be demonstrated by an example.

Consider the most extreme hub graph with partial correlations $\Delta_{1j} = -1/\sqrt{p}$, $j = 2, \ldots, p$ and $\Delta_{ij} = 0$ otherwise. For dimension $p = 10$, this has $\psi_1 = 10$ associated to the hub variable and $\psi_i = 2$ for all other variables. First consider an idealised setting where $\Theta = \Delta$. Generate S from a sample of size $n = 50$ and show the regularisation path for the standard GLASSO (with no diagonal penalty) across different penalty parameters (Figure 1, top left). It is able to recover the true model across quite a large range of penalty parameters. However, if X_2 is multiplied by 10, so that now $\theta_2 = 0.01$, then the true model is not recovered for any penalty parameter (Figure 1, top right). The edges associated to X_1 are shrunk to 0 before those of X_2. Due to X_2 having large variance, it receives less

penalisation and is selected as a hub variable. To account for variables being on different scales, one should standardise the data such that S has unit diagonal. As discussed in Section 2, for moderate sample sizes there is some assurance that $\sigma_i^2 \approx 1$ for all i after standardisation. However, now the effect of variables having different ψ_i becomes evident. Again, the true model is not recovered for any penalty parameter (Figure 1, bottom left). In contrast, PCGLASSO is able to select the true model for a large range of penalty parameter values (Figure 1, bottom right) and, being scale invariant, this does not depend on data scales or standardisation.

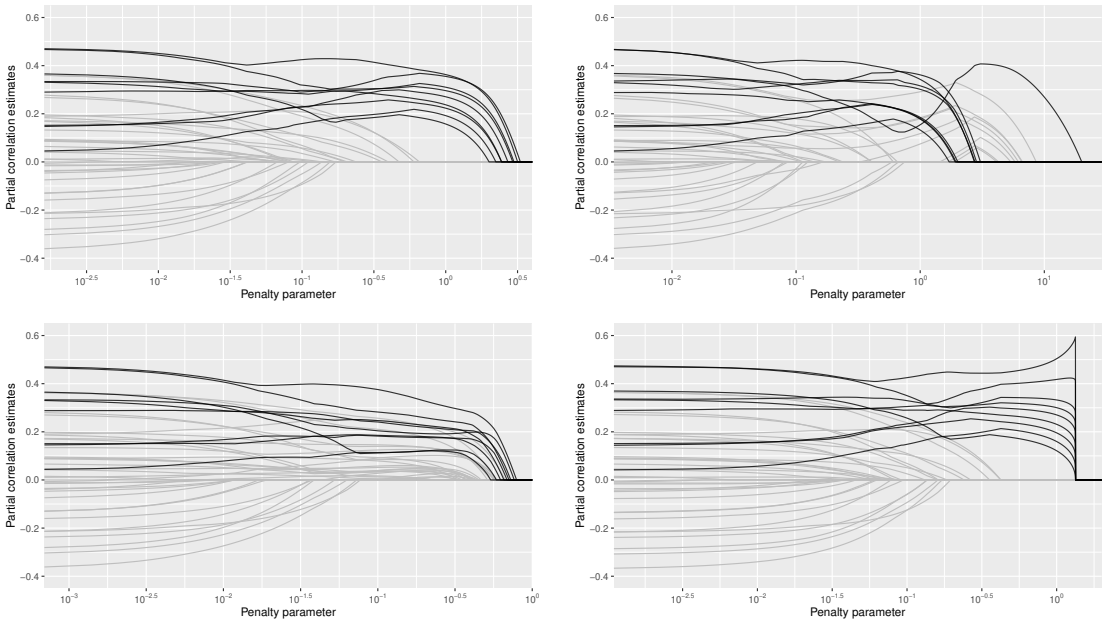


Figure 1: Regularisation paths for hub example. Truly non-zero Δ_{ij} estimates in black. Top left: GLASSO with $\Theta = \Delta$. Top right: GLASSO with high variance variable. Bottom left: GLASSO on standardised data. Bottom right: PCGLASSO.

It is evident from this example that data standardisation is vital when using GLASSO or other similar methods. Without standardisation, high variance variables receive less penalisation and are often selected as hubs, regardless of the evidence for dependence in the likelihood. However, even after standardisation such that S has unit diagonal and $\sigma_i^2 \approx 1$ (in which case $\Theta \approx K$), model selection can suffer when there are large differences in the ψ_i values. These issues have largely gone unnoticed in simulation studies because an idealised setting with $\Theta = \Delta$ is often chosen with no data standardisation. This assumes that $\psi_i = \sigma_i^2$ for all i, an unreasonable assumption in any real data.

One might argue that, instead of standardising so that $\sigma_i^2 \approx 1$, one should standardise so that $\theta_i \approx 1$. However, estimation of θ_i is a fundamentally more difficult task than estimating σ_i^2 and so such a standardisation is more prone to error. For example, standardising such that S^{-1} has unit diagonal is not possible when $n < p$ and, as pointed out in Section 3, a large sample size relative to p is required to have any assurance that $(S^{-1})_{ii} \approx \theta_i$. One

alternative is an adaptive approach where different penalty parameters are used for each Θ_{ij} with the penalty scaled based on the currently estimated θ.

A further aspect of penalised likelihood methods is their treatment of the diagonal θ with shrinkage often also being applied to these parameters. Although not directly indicating model selection through sparsity, the decomposition $\theta_i = \psi_i/\sigma_i^2$ shows that its estimation can still have a great effect on model selection. For fixed σ_i^2, large θ_i corresponds to large ψ_i and therefore high dependence between X_i and X_{-i}, while small $\theta_i \approx 1/\sigma_i^2$ indicates near independence between X_i and X_{-i}. Shrinkage of θ_i therefore encourages shrinkage of ψ_i and models with smaller degrees.

6. Estimation of σ^2

A graphical model is a model of dependence and therefore places no assumptions on the scale parameters σ^2. One might therefore consider σ^2 a nuisance parameter in GGM selection. However, estimation of σ^2 can still play a role in practice since tuning parameters are often learned from the data. If estimation of σ^2 is poor for certain tuning parameter values, then these are less likely to be chosen regardless of how well they estimate Δ or the graphical model.

This section proposes two strategies for estimating σ^2 when treated as a nuisance parameter. It is then shown how penalised likelihood estimates of the various parameterisations actually estimate σ^2 and whether they correspond to either of these two strategies.

6.1. Marginal estimation

An obvious strategy for estimating σ^2 is to simply fix them equal to the global MLE $\hat{\sigma}_i^2 = S_{ii}$. This is how the other nuisance parameter, the mean μ, is often treated in practice in GGMs, with its estimate fixed equal to the sample mean and analysis continuing contingent on this estimate. The log-likelihoods for Δ and K with the fixed $\hat{\sigma}_i^2 = S_{ii}$ estimate are

$$
\begin{aligned}
l(\Delta \mid S, \sigma^2 = \operatorname{diag}(S)) &= \sum_{i=1}^{p} \log \psi_i + \log\det \Delta - \operatorname{tr}(C\psi^{1/2}\Delta\psi^{1/2}), \\
l(K \mid S, \sigma^2 = \operatorname{diag}(S)) &= \log\det K - \operatorname{tr}(CK),
\end{aligned}
$$

where $C = \operatorname{cov2cor}(S)$ is the sample correlation matrix. Note that the log-likelihood for K has the same form for the log-likelihood of Θ in (1), but with S replaced by C. It therefore seems like estimation of K is analogous to estimation of Θ with standardised data. However, the difference is that K is an inverse correlation matrix, i.e. K^{-1} has unit diagonal, and so optimisation must be performed on this subspace. This restriction directly implies that the estimated σ^2 satisfy $\sigma^2 = \operatorname{diag}(S)$, while, as will be seen later, penalised likelihood estimation of Θ on standardised data can result in vastly different estimates for σ^2 when the diagonals of Θ are penalised.

6.2. Joint estimation

The previous approach of fixing the estimated σ^2 equal to $\operatorname{diag}(S)$ might seem like the correct approach since it is the same way as the nuisance parameter μ is dealt with. However, these

two things are fundamentally different: the MLE for μ is always the sample mean, no matter the estimated Σ, while the MLE for σ^2 depends on the estimated correlation matrix $R = K^{-1}$. To see this, differentiate the penalised likelihood (2) for (σ^2, K) with respect to σ_i to arrive at the first order conditions

$$\sigma_i = \sum_{j=1}^{p} \frac{S_{ij}K_{ij}}{\sigma_j}, \; i = 1, \ldots, p, \tag{4}$$

This is only satisfied at $\sigma_i^2 = S_{ii}$ when $K = C^{-1}$ (the global MLE), when $K = I_p$ or some linear combination of the two. By writing the optimisation problem in terms of the Hadamard product $S \odot K$, which is guaranteed to be positive definite when S is positive semidefinite with positive diagonal and K is positive definite (Horn and Johnson, 2012, Theorem 7.5.3), it can be shown that the solution to these equations exists and is unique. However, the solution is not generally available in closed form. Denote the solution for a given K or Δ as $\hat{\sigma}^2(K)$ and $\hat{\sigma}^2(\Delta)$.

Since S_{ij} and the global MLE $\hat{K}_{ij} = (C^{-1})_{ij}$ usually have opposite signs (having the same signs corresponds to a Simpson's paradox setting where conditioning changes the direction of dependence), shrinking K_{ij} away from its MLE $\hat{K}_{ij}$ towards 0 will generally result in smaller $\hat{\sigma}^2(K)$. Hence, for penalised likelihood estimates, usually $\hat{\sigma}^2(K)_i \leq S_{ii}$.

Interestingly, letting $\sigma^2 = \hat{\sigma}^2(K)$, the trace term in the log-likelihood is constant in K

$$\operatorname{tr}(S\sigma^{-1}K\sigma^{-1}) = \sum_{i=1}^{p}\sum_{j=1}^{p} \frac{S_{ij}K_{ij}}{\sigma_i\sigma_j} = p,$$

with the second equality following from the first order condition (4). The log-likelihood in the K and Δ parameterisations are then

$$l(K \mid S, \sigma^2 = \hat{\sigma}^2(K)) = -\sum_{i=1}^{p} \log \hat{\sigma}^2(K)_i + \log\det K,$$
$$l(\Delta \mid S, \sigma^2 = \hat{\sigma}^2(\Delta)) = -\sum_{i=1}^{p} \log \hat{\sigma}^2(\Delta)_i + \sum_{i=1}^{p} \log \psi_i + \log\det \Delta.$$

6.3. How shrinkage of different parameterisations affects σ^2

Since the decomposition of Θ_{ij} in (3) includes $\sigma_i\sigma_j$ in the denominator, penalisation of Δ might be expected to inflate the estimates for σ^2. This is most pronounced when the diagonal of Θ is penalised. When GLASSO includes the penalty on the diagonal with parameter λ, the estimated σ_i^2 is equal to $S_{ii} + \lambda$ (Friedman et al., 2008). The diagonal estimates are similarly inflated when any penalisation is directly applied the diagonal, although the exact form of the inflation depends on the choice of penalty function. These estimates are inflated above the global MLE and are likely even further from the conditional MLE. Hence, for large λ the estimated σ_i^2 do not describe the data well. This means that, when treated as a tuning parameter, large λ values will perform relatively poorly resulting in smaller λ values, and more dense models, being selected.

When the diagonal penalty is excluded from GLASSO, the estimated marginal variances are exactly equal to S_{ii}, the global MLE values. The same can be shown to occur for any penalised likelihood on Θ which excludes the diagonal penalty (Lauritzen and Zwiernik, 2022).

When the penalty is instead applied only to the partial correlation matrix Δ (or to the inverse correlation matrix K) the penalty is not a function of σ^2 and so the resulting estimated marginal variances satisfy (4) and are equal to $\hat{\sigma}^2(\Delta)$.

Penalising the off-diagonal of Θ therefore results in the global MLE of σ^2, whereas penalising Δ results in the MLE conditional on the estimated Δ. Since the likelihood for σ^2 can vary greatly for different fixed Δ, the second option of penalising Δ seems preferable.

7. Singular S

The previous two sections show the advantages of using penalised likelihoods on Δ

$$\underset{\theta \succ 0, \Delta \succ 0}{\operatorname{argmax}} \; l(\theta, \Delta \mid S) - \sum_{i<j} \operatorname{pen}(\Delta_{ij}) \tag{5}$$

On top of being scale invariant, they also better estimate graphs with high order nodes and estimate σ^2 by the MLE given the estimated Δ.

However, without further penalisation of other parameters, the solution to (5) does not exist when S is not positive definite, because the likelihood diverges along a path with $\|\theta\| \to \infty$ and Δ converging to a singular matrix. The penalty on Δ cannot counteract this, even with $\operatorname{pen}(\Delta_{ij}) \to \infty$ as $|\Delta_{ij}| \to 1$, because there are singular Δ with all $\Delta_{ij} \in (-1, 1)$. See Carter (2026) for a thorough analysis of the existence of penalised likelihood estimates for GGMs. Instead, one must either use a non-separable penalty on Δ which actively penalises matrices with small eigenvalues, or the direction $\|\theta\| \to \infty$ must be penalised.

It was shown in Carter et al. (2024) that a logarithmic penalty on θ maintains scale invariance of the selected GGM and in Carter and Molinari (2025) that a sufficiently strong logarithmic penalty is enough to ensure existence of the penalised likelihood estimate for singular S. However, as with diagonal penalisation for the precision matrix, penalisation of θ results in inflated estimates of the marginal variances.

Instead existence of the penalised likelihood estimate may be achieved through penalisation of ψ. The resulting penalised likelihood estimate is

$$\underset{\sigma^2 \succ 0, \Delta \succ 0}{\operatorname{argmax}} \; l(\sigma^2, \Delta \mid S) - \sum_{i=1}^{p} \operatorname{pen}_D(\psi_i) - \sum_{i<j} \operatorname{pen}(\Delta_{ij}).$$

Because there is no penalisation of σ^2, it will be estimated via the conditional MLE $\hat{\sigma}^2(\Delta)$.

Since ψ is the diagonal of Δ^{-1} and Δ has minimum eigenvalue tending to 0 if and only if a diagonal entry of Δ^{-1} goes to ∞, the direction $\psi_i \to \infty$ exactly characterises how $l(\sigma^2, \Delta \mid S) \to \infty$. Hence, sufficient penalisation of ψ is enough to ensure existence of the penalised likelihood estimate. Letting m_0 be the number of eigenvalues equal to 0 in S, as a corollary to Carter and Molinari (2025) Theorem 1.1, I conjecture that the solution exists as long as

$$\liminf_{\psi_i \to \infty} \frac{\operatorname{pen}_D(\psi_i)}{\log \psi_i} > m_0/p.$$

Although penalisation of ψ results in penalisation of the total dependence of X_i on X_{-i}, and therefore larger penalisation of high degree nodes, this is exactly what is required when S is singular. This is because in the small sample size setting $n < p$, within the observed data X_i is a linear function of X_{-i}, which is characterised by $\psi_i \to \infty$. If one wishes to keep the possibility of high degree nodes in the estimated graph, it therefore seems prudent to set the penalty on ψ_i to be as small as possible while ensuring existence, i.e. a logarithmic penalty. However, penalising ψ instead of θ does give the possibility of moving away from logarithmic penalties while maintaining scale invariance of the estimator. This might be for computational reasons or because an estimated graph with low node degrees is desired, in which case a stronger penalty on ψ can be used.

8. Discussion

The decomposition of Θ into partial correlations, VIFs and marginal variances, demonstrates some fundamental issues with penalised likelihood methods on Θ. I believe this is enough evidence to redirect attention towards directly penalising partial correlations.

The main challenge for penalising Δ is in computation. This is because this parameterisation introduces non-convexity in the log-likelihood through the trace term. Progress has already been made on this front with two algorithms proposed for optimisation of PC-GLASSO (Carter and Molinari, 2025; Bogdan et al., 2025). Further research is required for implementation of non-L_1 penalties on Δ and for applying a penalty directly to ψ as recommended in Section 7. The observation in Section 6.2 that the trace term of the log-likelihood is constant when the marginal variances are estimated via $\hat{\sigma}^2(\Delta)$ is also an interesting avenue that could lead to computational advances.

Although this paper has focused on penalised likelihood estimates, the decomposition of Θ can be used to gain insight into any shrinkage method. For example, a number of papers reframe estimation of Θ through linear regression $X_i = \sum_{j \neq i} \beta_j^{(i)} X_j + \epsilon_i$. The regression coefficients and error variance can be written as (Meinshausen and Bühlmann, 2006)

$$\beta_j^{(i)} = -\frac{\Theta_{ij}}{\Theta_{ii}} = -\Delta_{ij}\sqrt{\frac{\sigma_i^2 \psi_j}{\sigma_j^2 \psi_i}}, \qquad \text{Var}(\epsilon_i) = \frac{1}{\Theta_{ii}} = \frac{\sigma_i^2}{\psi_i}.$$

Data can been standardised to remove the effect of σ_i^2, σ_j^2 and ψ_i is common to all coefficients in this regression. So the important aspect is that Δ_{ij} is multiplied by $\sqrt{\psi_j}$. Penalised likelihood methods will therefore place higher penalty on $\beta_j^{(i)}$ associated to large ψ_j, i.e. when X_j has high dependence on the other predictors. An interesting recent method of this style that aims to mitigate the effects of scaling is Nguyen and Basu (2026).

Another important application of this decomposition of Θ is in the Bayesian setting. The current standard is to use a marginal exponential prior on the diagonals θ, but an exponential has a heavy lower tail and so will tend to encourage smaller values of θ. As was shown, this can lead to smaller node degrees and poor estimation of σ^2. Additionally, an important aspect of the Bayesian framework is the ability to incorporate prior information. By separating the concepts of scale and dependence by (σ^2, Δ), this task is made easier.

Acknowledgments

This work was supported by the EUTOPIA Science and Innovation Fellowship Programme and funded by the European Union Horizon 2020 programme under the Marie Skłodowska-Curie grant agreement No 945380.

References

Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *The Journal of Machine learning Research*, 9:485–516, 2008.

Sayantan Banerjee and Subhashis Ghosal. Bayesian structure learning in graphical models. *Journal of Multivariate Analysis*, 136:147–162, 2015. ISSN 10957243. doi: 10.1016/j.jmva.2015.01.015.

Małgorzata Bogdan, Adam Chojecki, Ivan Hejnỳ, Bartosz Kołodziejek, and Jonas Wallin. Identifying network hubs with the partial correlation graphical lasso. *arXiv preprint arXiv:2508.12258*, 2025.

Tony Cai, Weidong Liu, and Xi Luo. A constrained l1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494): 594–607, 2011.

Jack Storror Carter. Existence of penalised likelihood estimates and posterior propriety of separable prior distributions for gaussian precision matrices. *arXiv preprint arXiv:2505.20005v2*, 2026.

Jack Storror Carter and Cesare Molinari. Existence and optimisation of the partial correlation graphical lasso. *arXiv preprint arXiv:2510.25712*, 2025.

Jack Storror Carter, David Rossell, and Jim Q Smith. Partial correlation graphical lasso. *Scandinavian Journal of Statistics*, 51(1):32–63, 2024.

Lee Dicker, Baosheng Huang, and Xihong Lin. Variable selection and estimation with the seamless-l0 penalty. *Statistica Sinica*, pages 929–962, 2013.

Mathias Drton and Michael D. Perlman. Multiple testing and error control in gaussian graphical model selection. *Statistical Science*, 22(3):430 – 449, 2007. doi: 10.1214/088342307000000113. URL https://doi.org/10.1214/088342307000000113.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

Jianqing Fan, Yang Feng, and Yichao Wu. Network exploration via the adaptive lasso and scad penalties. *The annals of applied statistics*, 3(2):521, 2009.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Lingrui Gan, Naveen N Narisetty, and Feng Liang. Bayesian regularization for graphical models with unequal shrinkage. *J. Amer. Statist. Assoc.*, 114(527):1218–1231, 2019.

Arjun K Gupta and Daya K Nagar. *Matrix variate distributions*, volume 104. CRC Press, 1999.

Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

Jack Jewson, Li Li, Laura Battaglia, Stephen Hansen, David Rossell, and Piotr Zwiernik. Graphical model inference with external network data. *Biometrics*, 80(4):ujae151, 2024.

Steffen Lauritzen and Piotr Zwiernik. Locally associated graphical models and mixed convex exponential families. *The Annals of Statistics*, 50(5):3009–3038, 2022.

Yunfan Li, Bruce A Craig, and Anindya Bhadra. The graphical horseshoe estimator for inverse covariance matrices. *Journal of Computational and Graphical Statistics*, 28(3): 747–757, 2019.

Weidong Liu and Xi Luo. Fast and adaptive sparse precision matrix estimation in high dimensions. *Journal of multivariate analysis*, 135:153–162, 2015.

Arak M Mathai, Serge B Provost, and Hans J Haubold. *Multivariate statistical analysis in the real and complex domains*. Springer Nature, 2022.

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

Robb J Muirhead. *Aspects of multivariate statistical theory*. John Wiley & Sons, 2009.

Ha Nguyen and Sumanta Basu. Large: A locally adaptive regularization approach for estimating gaussian graphical models. *arXiv preprint arXiv:2601.09686*, 2026.

Jie Peng, Pei Wang, Nengfeng Zhou, and Ji Zhu. Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 104(486):735–746, 2009.

Guilherme V Rocha, Peng Zhao, and Bin Yu. A path following algorithm for sparse pseudo-likelihood inverse covariance estimation (splice). *arXiv preprint arXiv:0807.3734*, 2008.

Peter WF Smith and Joe Whittaker. Edge exclusion tests for graphical gaussian models. In *Learning in Graphical Models*, pages 555–574. Springer, 1998.

Deborah Sulem, Jack Jewson, and David Rossell. Bayesian computation for high-dimensional gaussian graphical models with spike-and-slab priors. *arXiv preprint arXiv:2511.01875*, 2025.

Kean Ming Tan, Palma London, Karthik Mohan, Su-In Lee, Maryam Fazel, and Daniela Witten. Learning graphical models with hubs. *Journal of Machine Learning Research*, 15:3297–3331, 2014.

Caroline Uhler. Gaussian graphical models. In Marloes Maathuis, Mathias Drton, Steffen Lauritzen, and Martin Wainwright, editors, *Handbook of Graphical Models*, pages 217–238. CRC Press, 2019.

Chengjing Wang, Peipei Tang, Wenling He, and Meixia Lin. Learning the hub graphical lasso model with the structured sparsity via an efficient algorithm. *Journal of Scientific Computing*, 103(3):98, 2025.

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

Hao Wang and Sophia Zhengzi Li. Efficient gaussian graphical model determination under g-wishart prior distributions. *Electronic Journal of Statistics*, 6:168–198, 2012.

Yanxin Wang and Li Zhu. Variable selection and parameter estimation with the atan regularization method. *Journal of Probability and Statistics*, 2016(1):6495417, 2016.

Donald R Williams and Philippe Rast. Back to the basics: Rethinking partial correlation network methodology. *British Journal of Mathematical and Statistical Psychology*, 73(2): 187–212, 2020.

Ming Yuan and Yi Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894, 2010.

Teng Zhang and Hui Zou. Sparse precision matrix estimation via lasso penalized d-trace loss. *Biometrika*, 101(1):103–120, 2014.