The Wrong Tool for Inference
A Critical View of Gaussian Graphical Models

Kevin R. Keane and Jason J. Corso

Computer Science and Engineering, University at Buffalo, SUNY, Buffalo, New York, U.S.A.

Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, Michigan, U.S.A.

Keywords: Multivariate Normal Distributions, Gaussian Graphical Models, Degenerate Priors.

Abstract: Myopic reliance on a misleading first sentence in the abstract of *Covariance Selection*[a] Dempster (1972) spawned the computationally and mathematically dysfunctional Gaussian graphical model (GGM). In stark contrast to the GGM approach, the actual (Dempster, 1972, § 3) *algorithm* facilitated elegant and powerful applications, including a "texture model" developed two decades ago involving arbitrary distributions of 1000+ dimensions Zhu (1996). The "Covariance Selection" *algorithm* proposes a greedy sequence of increasingly constrained maximum entropy hypotheses Good (1963), terminating when the observed data "fails to reject" the last proposed probability distribution. We are mathematically critical of GGM methods that address a continuous convex domain with a discrete domain "golden hammer". Computationally, selection of the wrong tool morphs polynomial-time algorithms into exponential-time algorithms. GGMs concepts are at odds with the fundamental concept of the invariant spherical multivariate Gaussian distribution. We are critical of the Bayesian GGM approach because the model selection process derails at the start when virtually all prior mass is attributed to comically precise multi-dimensional geometric "configurations" (Dempster, 1969, Ch. 13). We propose two Bayesian alternatives. The first alternative is based upon (Dempster, 1969, Ch. 15.3) and (Hoff, 2009, Ch. 7). The second alternative is based upon Bretthorst (2012), a recent paper placing maximum entropy methods such as the "Covariance Selection" *algorithm* in a Bayesian framework.

1 INTRODUCTION

Gaussian graphical models (GGMs) have a nice interpretation: the absence of an edge implies conditional independence between the corresponding pair of variables (Whittaker, 1990, Ch. 6). Both the search based GGM approach, for example Jones et al. (2005); Moghaddam et al. (2009); Wang et al. (2011) and, the l_1 regularization based GGM approach, for example Dahl et al. (2005); Meinshausen and Bühlmann (2006); Banerjee et al. (2006); Yuan and Lin (2007); Friedman et al. (2008) focus on interpretation and exploitation of the *pairwise* Markov property. Given an undirected dependency graph $G = (V, E)$ with node set V and edge set E for a set of random variables X, two variables x_j and x_k are independent given all other variables $X_{V\setminus\{j,k\}}$ if the edge $\{j,k\}$ is not in the edge set E,

$$X_j \perp\!\!\!\perp X_k \mid X_{V\setminus\{j,k\}} \quad \text{if } \{j,k\} \notin E \quad . \tag{1}$$

A zero in the precision matrix elements (j,k) and (k,j) corresponds to $\{j,k\} \notin E$. We are concerned that certain fundamental Gaussian and Bayesian concepts fade from consciousness with myopic focus on these graph representations of the multivariate Gaussian distribution.

2 GAUSSIAN GRAPHICAL MODELS

The concept of a *finite* enumeration of graphs (Whittaker, 1990, Ch. 6) clouds the natural characterization of the multivariate Gaussian as an invariant spherical distribution (Dempster, 1969, Ch. 12). A graph's structure corresponds to strict constraints on the angles among the random variables. Adhesion to the original coordinates of a data set is at odds with a typical approach for multivariate Gaussian analysis where measured data $x \sim \mathrm{N}(\mu, \Sigma)$ is translated, rotated and scaled $y = \Sigma^{-\frac{1}{2}}(x - \mu)$ to equivalent linear combina-

[a] "The covariance structure of a multivariate normal population can be simplified by setting elements of the inverse of the covariance matrix to zero."

Keane, K. and Corso, J.
The Wrong Tool for Inference - A Critical View of Gaussian Graphical Models.
DOI: 10.5220/0006644604700477
In *Proceedings of the 7th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2018)*, pages 470-477
ISBN: 978-989-758-276-9

tions which are independent and normally distributed $y \sim \mathrm{N}(0,\mathrm{I})$. The concept of search over a discrete space is at odds with geometric exploitation of a continuous convex distribution.

2.1 Imposition of Graph Structure

Conditional independence corresponds to a precise alignment of the measured variables. Even in a man-made setting – sensors in a building – the simple logic and attractiveness of a GGM may not prevail Gonzalez and Hong (2008):

> We can see that adding the graphical interpretation gave slightly worse predictions than using just the kernel function. One explanation may be that the graph does not accurately reflect the conditional independence structure of the room. For example, all sensors near windows were linked by the outside temperature and therefore not conditionally independent even though the floor plan does not suggest strong spatial linkage between them.

We are somewhat sympathetic to the attractiveness of specifying a GGM in scenarios with comparable exogenous structural information. But, we will make two points. First, the graph in Gonzalez and Hong (2008) was not obtained by *search* over $2^{(p-1)p/2}$ candidate graphs, but from architectural plans. Second, constraining inference to the graph did not yield superior performance. We greatly appreciate access to this experimental result as it effectively illustrates our concern with GGMs: the focus on pairwise interaction and desperate desire to specify models that "make sense" risks misspecification for subtle factors.

2.1.1 Relative Alignment of Variables

The off-diagonal elements of the variance matrix specify the relative alignment of a pair of random variables. Consider the case of two zero mean, unit variance Gaussian variables. The variance σ_{12} implies an angle γ_{12} between the two variables since $\sigma_{12} = \mathrm{E}(x_1 \cdot x_2) = \mathrm{E}(|x_1||x_2|\cos(\gamma_{12})) = \sigma_{x_1}\sigma_{x_2}\cos(\gamma_{12}) = \cos(\gamma_{12})$. When $x_1 \perp\!\!\!\perp x_2$, $\sigma_{12} = \cos(\gamma_{12}) = 0$, and the variables are independent. When a GGM's graph omits one or more edges from the complete graph, a rigid alignment of the variables is imposed. Point estimates for continuous parameters such as γ_{12} should raise a large red flag for Bayesians; but, we will delay that discussion to Section 3.

2.2 Adhesion to the Initial Basis

Multivariate Gaussian inference is fundamentally based upon the concept of a *spherical* distribution that is *invariant* under all linear transformations which carry an origin-centered sphere into itself (Dempster, 1969, Ch. 12.2). A concept of *special* coordinates, including the original coordinates of the data set, is problematic. GGMs appear to be stuck in the original coordinates whereas a change of basis is a fundamental technique in analysis of Gaussian data.

2.2.1 Univariate Change of Basis

The concept of the standard normal distribution is widely understood. To display a histogram of $x \sim \mathrm{N}(\mu, \sigma^2)$ observations, the mean μ is subtracted, and the data is scaled by its standard deviation σ to obtain $y \sim \mathrm{N}(0,1)$, $y = \dfrac{x-\mu}{\sigma}$. To sample the distribution of x, standard normal variate y is obtained, scaled, and translated to yield $x = \sigma y + \mu$. This fluid change of basis — well known for univariate data — applies equally to multivariate data.

2.2.2 (Dempster, 1969, Thrm. 12.4.1)

> Suppose that $\mathbf{X}$ has the $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution where $\mathbf{X}$ and $\boldsymbol{\mu}$ have dimensions $1 \times p$ and $\boldsymbol{\Sigma}$ is a $p \times p$ positive definite, or semi-definite, symmetric matrix of rank $q \leq p$. Suppose that $\boldsymbol{\Delta}$ is any $p \times q$ matrix such that $\boldsymbol{\Sigma} = \boldsymbol{\Delta}\boldsymbol{\Delta}^\mathsf{T}$ and suppose that $\boldsymbol{\Gamma}$ is a pseudoinverse of $\boldsymbol{\Delta}$. Then $\mathbf{Y} = (\mathbf{X} - \boldsymbol{\mu})\boldsymbol{\Gamma}^\mathsf{T}$ has the $\mathrm{N}(\mathbf{0}, \mathbf{I})$ distribution where $\mathbf{Y}$, $\mathbf{0}$, and $\mathbf{I}$ have dimensions $1 \times q$, $1 \times q$, and $q \times q$, respectively. Furthermore, $\mathbf{X}$ may be recovered from $\mathbf{Y}$ with probability 1 using $\mathbf{X} = \boldsymbol{\mu} + \mathbf{Y}\boldsymbol{\Delta}^\mathsf{T}$.

The GGM community appears opposed to (Dempster, 1969, Thrm. 12.4.1) and stuck in arbitrary measurement bases. This makes no sense for the multivariate Gaussian distribution with readily accessible, analytically attractive coordinates. All the GGM discussions of decomposable and non-decomposable graphs are a red herring. The conventional and simple mathematical approach to analyzing multivariate Gaussian data is to translate, rotate, and scale the data to a multivariate standard normal distribution which is trivial to manipulate and interpret. Inference computations for graphs with no edges, the $\mathrm{N}(\mathbf{0}, \mathbf{I})$ graphs, are trivial.

2.2.3 Discarding Information

In an experimental setting, more likely than not the subtle factors are unknown. The problem with incomplete graphs in measurement coordinates is that the sample statistics corresponding to missing edges on the graph are discarded – an obstructive approach to inference. Maximum entropy and Bayesian methods begin with a simple distribution typically characterized by a diagonal precision matrix and incorporate structure as justified by the data. It is an entirely different approach to discard sample statistics that do not conform to an arbitrary graph.

GGM methods that set elements of the precision matrix to zero is in direct opposition to the spirit of the "Covariance Selection" maximum entropy *algorithm* where constraints are introduced when the data demands doing so as determined by a statistical test. Setting precision matrix elements to zero risks destruction of subtle (and not so subtle) structure in data sets. (Dempster, 1972, Introduction, second paragraph) warns "errors of misspecification are introduced because the null values are incorrect." (Tibshirani, 1996, § 11(c)) identifies a similar problem for subset selection in the presence of a "large number of small effects". (West and Harrison, 1997, Ch. 16.3.1) warns (emphasis theirs) "These factors, that dominate variations at the macro level, often have relatively little *apparent* effect at the disaggregate level and so are ignored." Our fear is that the pairwise removal of structure corresponds to a scenario where one "can't see the forest for the trees." Starting with a diagonal precision matrix and adding structure demonstrably necessary seems more prudent.

2.3 Computational Considerations

A final complaint we will raise for the search based GGM approach is the acceptance of exponential-time discrete search algorithms when a distribution defined by a log quadratic density function should clearly exploit more efficient polynomial-time algorithms. This appears to be an example of a discrete "golden hammer" inappropriately applied to a continuous convex domain.

3 BAYESIAN GAUSSIAN GRAPHICAL MODELS

Bayesians typically prefer minimally informative priors and produce posterior *distributions*, not point estimates or points with probability mass. For all GGM graphs *except* the complete graph, one or more natural parameters are constrained to a point or set of points which would be expected to reflect true continuous parameter values *with probability zero*. In high dimension, the concept of a uniform prior over the graphs (Giudici and Green, 1999, § 1.2) results in the allocation of virtually all prior mass, $\left(2^{(p-1)p/2}-1\right)/2^{(p-1)p/2}$, to point estimates for the continuous natural parameters.

3.1 Bayesian Model Selection

Giudici and Green (1999) utilize a model selection framework described by MacKay (1992). There are $2^{(p-1)p/2}$ potential graphs for p-variables. Giudici and Green (1999) limit consideration to d decomposable graphs, therefore the uniform prior for the graph g is:

$$P(g) = d^{-1} \quad . \tag{2}$$

In Figure 1, Giudici and Green (1999) would assign priors for the graphs:

$$p(\mathbf{G}_0) = p(\mathbf{G}_1) = \frac{1}{2} \tag{3}$$

The problem from a Bayesian perspective is that assigning probability mass to a graph assigns probability mass to a *point* in the natural parameters. The priors for continuous model parameter θ given the graph g, illustrated in Figure 2, are:

$$p(\theta|\mathbf{G}_0) = \begin{cases} \frac{1}{2} & \text{if } \theta = \frac{1}{2}\pi \quad , \\ \frac{1}{2} & \text{if } \theta = \frac{3}{2}\pi \quad , \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

$$p(\theta|\mathbf{G}_1) = \frac{1}{2\pi} d\theta \quad . \tag{5}$$

We find the model parameter prior $p(\theta|\mathbf{G}_0)$ objectionable. Trading technical precision for intuition, we consider $p(\theta|\mathbf{G}_0)$ to be a *degenerate prior*[1]. To the extent $p(\theta|\mathbf{G}_0)$ is justifiable, we would propose consideration of an equally "justifiable" infinite class of *Sure Thing* hypotheses (attributed to E.T. Jaynes in MacKay, 1992, p. 12) with unit mass at $\theta = \frac{1}{2}\pi + \phi, \quad \phi \in [0, 2\pi]$.

[1] A degenerate distribution places all probability mass on one point; we mean to describe a broader concept inclusive of mixtures of degenerate and non-degenerate distributions characterized by probability mass greater than zero occurring at a finite set of points.

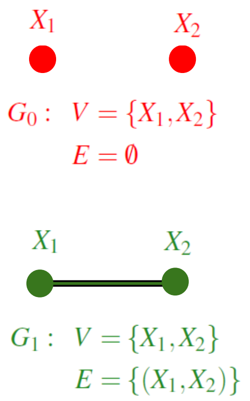

Figure 1: An enumeration of the Gaussian graphical models for the bivariate normal distribution. Graph $\mathbf{G}_0$ corresponds to independent normal variables x_1 and x_2. Graph $\mathbf{G}_1$ corresponds to the general case where covariance structure between normal variables x_1 and x_2 is unrestricted. Equation 3 defines the "uniform prior" for these two graphs (Giudici and Green, 1999, § 1.2).

The unconditional prior $p(\theta)$ illustrated in Figure 3 is:

$$p(\theta) = p(\theta|\mathbf{G}_0)\, p(\mathbf{G}_0) + p(\theta|\mathbf{G}_1)\, p(\mathbf{G}_1) \qquad (6)$$

$$= \begin{cases} \frac{1}{4} & \text{if } \theta = \frac{1}{2}\pi \quad , \\ \frac{1}{4} & \text{if } \theta = \frac{3}{2}\pi \quad , \\ \frac{1}{4\pi} d\theta & \text{otherwise.} \end{cases} \qquad (7)$$

We find priors assigning point mass to continuous parameters objectionable. With that caveat, the remainder of the Bayesian GGM model comparison framework proceeds as follows. The *evidence* $P(X|g)$ for structure g is:

$$P(X|g) = \int P(X|\theta, g)\, P(\theta|g)\, d\theta \quad , \qquad (8)$$

and the probability of graph g given the data X is:

$$P(g|X) \propto P(X|g)P(g) \quad . \qquad (9)$$

We like the Bayesian model selection approach in MacKay (1992) and the more recent computational advances in Skilling (2004). However, Bayesian GGM approach never gives the process a fair shake. The number of graphs is $2^{(p-1)p/2}$. Assuming a uniform prior over the graphs as proposed in Giudici and Green (1999), $\left(2^{(p-1)p/2} - 1\right) / 2^{(p-1)p/2} \approx 1$ of the prior mass is assigned to specified points for the continuous natural parameters. Any finite set of points should collectively have zero probability with a "reasonable prior" (for continuous parameters). We view the Bayesian GGM priors as so inequitable, only an unrealistic number of observations $n \to \infty$ will mitigate its effect.

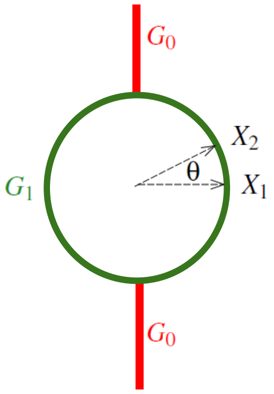

Figure 2: Geometric interpretation of the relative alignment parameter θ for a bivariate standard normal distribution. $\sigma_{1\,2} = \mathrm{E}(x_1 \cdot x_2) = \mathrm{E}(|x_1||x_2|\cos(\theta)) = \cos(\theta)$.
Equation 4 permits $\theta = \frac{1}{2}\pi$ and $\theta = \frac{3}{2}\pi$ for $\mathbf{G}_0$;
Equation 5 permits $0 \leq \theta \leq 2\pi$ for $\mathbf{G}_1$.

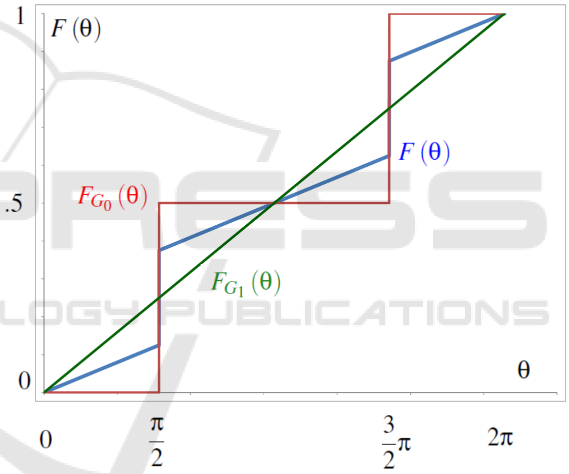

Figure 3: A "uniform prior" on the graphs in Figure 1 results in a "degenerate prior" for θ in Figure 2. $p(\theta|\mathbf{G}_0)$ defined in Equation 4, $p(\theta|\mathbf{G}_1)$ defined in Equation 5, and $p(\theta)$ defined in Equation 7. Cumulative probability $F(\phi) = \int_0^\phi p(\theta)d\theta$.

4 RELATED ARGUMENTS

The beauty of Bayesian methods is the ability to generate reasonable inference from "complex" models with limited data. Andrew Gelman's blog provides many insightful comments and references relevant to the issues we wrestle with in this paper. A number of lively, good natured debates on the blog encouraged the use of "complex"[2] models. We view sparse GGMs as a misguided attempt to maintain parsimony and simplicity. The following comments encouraged us to question the wisdom of pursuing simplicity or parsimony with GGMs.

Gelman (2004) identifies (Neal, 1996, pp. 103-104) as a favorite quote:

> Sometimes a simple model may outperform a more complex model, at least when the training data is limited. Nevertheless, I believe that deliberately limiting the complexity of the model is not fruitful when the problem is evidently complex. Instead, if a simple model is found that outperforms some particular complex model, the appropriate response is to define a different complex model that captures whatever aspect of the problem led to the simple model performing well.

A comment that appears specifically related to our discomfort with uniform priors over the graphs and point mass distributions for continuous model parameters appears in Gelman (2011):

> The Occam applications I don't like are the discrete versions such as advocated by Adrian Raftery and others, in which some version of Bayesian calculation is used to get results saying that the posterior probability is 60%, say, that a certain coefficient in a model is exactly zero. I'd rather keep the term in the model and just shrink it continuously toward zero.

Gelman (2013) nicely clarified that over-fitting is not attributable to flexibility alone (i.e. the complete graph in GGMs):

> Overfitting comes from a model being flexible and unregularized. Making a model inflexible is a very crude form of regularization. Often we can do better.

5 PREFERABLE METHODS

5.1 Option One

For high dimensional Gaussian inference we first suggest a full Bayesian implementation as outlined in (Dempster, 1969, Ch. 15.3) and its equivalent (Hoff, 2009, Ch. 7). Starting with a prior for the mean $p(\mu) \sim \mathrm{N}(\mu_0, \Lambda_0)$ and the variance $p(\Sigma) \sim$ inverse-Wishart$\left(\nu_0, \mathbf{S}_0^{-1}\right)$, the conditional posterior distributions are:

$$p(\mu|x_1,\ldots,x_n,\Sigma) \sim \mathrm{N}(\mu_n, \Lambda_n) \tag{10}$$

$$p(\Sigma|x_1,\ldots,x_n,\mu) \sim \text{inverse-Wishart}\left(\nu_n, \mathbf{S}_n^{-1}\right) \tag{11}$$

where

$$\Lambda_n = \left(\Lambda_0^{-1} + n\Sigma^{-1}\right)^{-1} \tag{12}$$

$$\mu_n = \Lambda_n\left(\Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\bar{\mathbf{x}}\right) \tag{13}$$

$$\nu_n = \nu_0 + n \tag{14}$$

$$\mathbf{S}_n = \mathbf{S}_0 + \mathbf{S}_\mu \tag{15}$$

$$\mathbf{S}_\mu = \sum_{i=1}^{n} (x_i - \mu)(x_i - \mu)^\intercal \tag{16}$$

The joint posterior $p(\mu, \Sigma|x_1,\ldots,x_n)$ is available from a Gibbs sampler using these conditional distributions Equation 10 and Equation 11.

5.1.1 Implementation Considerations

Transforming the sampling problem to a set of independent variables, (Dempster, 1969, Thrm. 12.4.1) quoted in subsubsection 2.2.2 facilitates straight forward parallel implementation of Equation 10 and Equation 11 in a Gibbs sampler. Sherman-Morrison-WoodburyBindel (2009) will be helpful in computing $\mathbf{S}_n^{-1}$ in Equation 11 and Λ_n in Equation 12, treating the $n \ll p$ samples as low n rank updates to the $p \times p$ diagonal matrices $\mathbf{S}_0^{-1}$ and Λ_0 respectively.

5.2 Option Two

The second alternative where both p and n are very large would be to use a maximum entropy algorithm. Assuming streaming data, one would define a set of domain specific marginals of interest – for example, the filters in Zhu (1996) and the gene regulatory network modules in Celik et al. (2014). We would then implement a maximum entropy algorithm beginning with the identity matrix and use the framework of Bretthorst (2012) to determine a posterior distribution for both the number of constraints and the range of Lagrange multiplier values defining the synthesized distribution. Bretthorst (2012) nicely demonstrates Bayesian inference of the appropriate number of marginal constraints and inference as to the distribution of Lagrange multipliers enforcing a particular constraint. A final consideration in a dynamic environment would be a method to gracefully forget past observations – perhaps randomly removing one observation at each iteration to keep a recent weighted constant size sample; or perhaps weighting the observations vectors directly for a finite horizon.

[2] We put "complex" in quotes because its not clear that high dimensionality alone equates to complexity; and, a log quadratic density certainly is not that "complex."

6 CONCLUSION

The dominant discrete theme of GGM obscures the continuous convex properties of the multivariate Gaussian distribution. Restricting inference to a particular graphical model obstructs accumulation of information describing the underlying distribution. For Bayesian GGMs, uniform priors over the graphs results in extremely concentrated probability mass in the natural parameters.

We support the use of GGMs for *interpretation* and *communication* of approximate inference results from multivariate Gaussian distributions. We strongly discourage the use of GGMs directly for multivariate Gaussian *inference*.

REFERENCES

Banerjee, O., Ghaoui, L. E., d'Aspremont, A., and Natsoulis, G. (2006). Convex optimization techniques for fitting sparse gaussian graphical models. In *Proceedings of the 23rd international conference on Machine learning*, pages 89–96. ACM.

Bindel, D. (2009). Sherman-Morrison-Woodbury. Matrix Computations (CS 6210), Cornell University lecture.

Bretthorst, G. (2012). The maximum entropy method of moments and Bayesian probability theory. In *32nd International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Carching, Germany*, pages 3–15.

Celik, S., Logsdon, B., and Lee, S. (2014). Efficient dimensionality reduction for high-dimensional network estimation. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1953–1961.

Dahl, J., Roychowdhury, V., and Vandenberghe, L. (2005). Maximum likelihood estimation of Gaussian graphical models: numerical implementation and topology selection. Technical report, Department of Electrical Engineering, University of California, Los Angeles.

Dempster, A. (1969). *Elements of continuous multivariate analysis*. Addison-Wesley.

Dempster, A. (1972). Covariance selection. *Biometrics*, pages 157–175.

Dobra, A. and West, M. (2004). Bayesian covariance selection. *Duke Statistics Discussion Papers*, 23.

Fan, J., Feng, Y., and Wu, Y. (2009). Network exploration via the adaptive lasso and scad penalties. *The Annals of Applied Statistics*, 3(2):521.

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

Gelman, A. (2004). Against parsimony. [Online; accessed 15-April-2015].

Gelman, A. (2011). David MacKay and Occam's Razor. [Online; accessed 15-April-2015].

Gelman, A. (2013). Flexibility is good. [Online; accessed 20-May-2015].

Giudici, P. and Green, P. (1999). Decomposable graphical Gaussian model determination. *Biometrika*, 86(4):785–801.

Gonzalez, J. and Hong, S. (2008). Linear-time inverse covariance matrix estimation in Gaussian processes. Technical report, Computer Science Department, Carnegie Mellon University.

Good, I. J. (1963). Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables. *The Annals of Mathematical Statistics*, 34(3):911–934.

Hoff, P. (2009). *A first course in Bayesian statistical methods*. Springer Science & Business Media.

Jalobeanu, A. and Gutiérrez, J. (2007). Inverse covariance simplification for efficient uncertainty management. In *27th MaxEnt workshop, AIP Conference Proceedings, Saratoga Springs, NY*.

Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C., and West, M. (2005). Experiments in stochastic computation for high-dimensional graphical models. *Statistical Science*, 20(4):388–400.

Knuiman, M. (1978). Covariance selection. *Advances in Applied Probability*, pages 123–130.

Lian, H. (2011). Shrinkage tuning parameter selection in precision matrices estimation. *Journal of Statistical Planning and Inference*, 141(8):2839–2848.

MacKay, D. (1992). *Bayesian methods for adaptive models*. PhD thesis, California Institute of Technology.

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

Moghaddam, B., Marlin, B., Khan, M., and Murphy, K. (2009). Accelerating Bayesian structural inference for non-decomposable Gaussian graphical models. In *NIPS*.

Neal, R. (1996). *Bayesian Learning for Neural Networks*. Springer.

Skilling, J. (2004). Nested sampling. *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, 735:395–405.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Wang, H., Reeson, C., and Carvalho, C. (2011). Dynamic financial index models: Modeling conditional dependencies via graphs. *Bayesian Analysis*, 6(4):639–664.

West, M. and Harrison, J. (1997). *Bayesian forecasting and dynamic models*. Springer Verlag.

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

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35.

Zhu, S. (1996). *Statistical and computational theories for image segmentation, texture modeling and object recognition*. PhD thesis, Harvard University.

APPENDIX

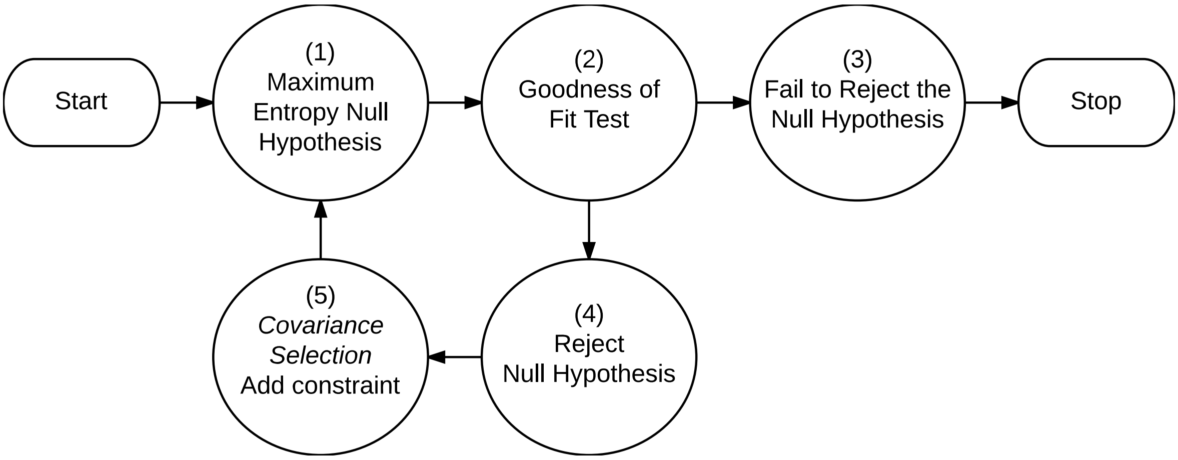
Start
(1) Maximum Entropy Null Hypothesis
(2) Goodness of Fit Test
(3) Fail to Reject the Null Hypothesis
Stop
(4) Reject Null Hypothesis
(5) *Covariance Selection* Add constraint

Figure 4: "Covariance Selection" algorithm. "The principle of maximum entropy generates much of statistical mechanics as a null hypothesis, to be tested by experiment" (Good, 1963, p. 912). The above diagram is the algorithm *demonstrated* in Dempster (1972). The diagram accurately describes the algorithm appearing in Zhu (1996).

The Covariance Selection Algorithm

The first sentence of the abstract (Dempster, 1972, Summary) is misleading:

> The covariance structure of a multivariate normal population can be simplified by setting elements of the inverse of the covariance matrix to zero.

With respect to the *demonstrated algorithm* in (Dempster, 1972, § 3), the widely repeated assertion that "covariance selection" inserts zeros in a precision matrix[3] is *false*. Non-zero entries are placed in a precision matrix as covariance constraints are added to a maximum entropy distribution. The matrices generated are maximally sparse, with non-zeros corresponding to statistically significant structure in the observed data.

The technique demonstrated in (Dempster, 1972, § 3) is *not* about *setting elements* of the precision matrix (inverse of the covariance matrix) *to zero*. As shown in Figure 4, the technique is as follows: 1) propose a maximum entropy distribution for the "null hypothesis"; 2) test the "null hypothesis" using observed data; 3) if you "fail to reject the null hypothesis," STOP; otherwise, 4) "reject the null hypothesis;" 5) "Covariance Selection" – add a covariance constraint requiring the proposed distribution match the observed distribution for the marginal with the worst discrepancy, this augmented proposal is a new "null hypothesis," loop to step 1.

Sparsity in the Precision Matrix

Sparsity is a pervasive topic in papers citing Dempster (1972). It is important to observe that the algorithm *directly constructs sparse precision matrices.* The maximum number of zeros in the precision matrix occurs at initialization, when the precision matrix and the variance matrix for the proposed distribution are both diagonal. Under duress, as a sequence of proposed models are rejected by the observed data, "Covariance Selection" *adds non-zeros to the precision matrix.* In Table 1, we provide a sequence of correlation matrices that match each stage of (Dempster, 1972, § 3 Tbl. 1 and system output) exactly and we provide the corresponding sequence of inverse correlation matrices to clarify the non-zero fill pattern to show that the maximum entropy algorithm in "Covariance Selection" defines sparse precision matrices *by construction.*

Replicating the Covariance Selection Example

In Table 1, we are able to fully replicate (Dempster, 1972, § 3 Tbl. 1 and system output) using the algorithm defined in Figure 4 by selecting for inclusion the pair $(i,j) = \underset{(i,j)}{\arg\max} |\mathbf{S}_{i,j} - \mathbf{\Sigma}_{i,j}|$ in algorithm step five "Covariance Selection".

[3] For example, ("setting concentrations (elements of the inverse covariance matrix) to zero" Knuiman, 1978); ("specifies that certain elements in the inverse of the variance matrix are zero" Whittaker, 1990, p. 11); ("by setting to zero selected elements of the precision matrix" Dobra and West, 2004); ("setting to zero some of the elements of the inverse covariance matrix" Jalobeanu and Gutiérrez, 2007); ("setting some elements of the precision matrix to zero" Fan et al., 2009) ("simplified the matrix structure by setting some entries to zero." Lian, 2011)

Table 1: Using the algorithm in Figure 4 and specifying at each iteration a covariance constraint for the variable pair with the maximum absolute discrepancy between the observed covariance s_{ij} and the synthesized covariance σ_{ij}, the column $\mathbf{\Sigma}$ below exactly replicates the output of (Dempster, 1972, § 3). Although more iterations are shown in Dempster (1972) and below, Dempster suggests stopping the algorithm after stage 5 based upon a statistical significance test. We provide for review the precision matrices at each stage. Note in the "Covariance Selection" algorithm only one symmetric pair of non-zeros (in **bold**) enters the precision matrix $\mathbf{\Sigma}^{-1}$ at each iteration. The algorithm of Dempster (1972) is widely misrepresented as "setting elements of the precision matrix to zero." Clearly, zeros reside in the precision matrix $\mathbf{\Sigma}^{-1}$ from initialization, dropping out as constraints are imposed.

	$\mathbf{\Sigma}^{-1}$	$\mathbf{\Sigma}$
0	$\begin{matrix} 1.000000 & & & & & \\ & 1.000000 & & & & \\ & & 1.000000 & & & \\ & & & 1.000000 & & \\ & & & & 1.000000 & \\ & & & & & 1.000000 \end{matrix}$	$\begin{matrix} 1.000000 & & & & & \\ & 1.000000 & & & & \\ & & 1.000000 & & & \\ & & & 1.000000 & & \\ & & & & 1.000000 & \\ & & & & & 1.000000 \end{matrix}$
1	$\begin{matrix} 1.000000 & & & & & \\ & 1.000000 & & & & \\ & & 1.000000 & & & \\ & & & 1.279033 & \mathbf{0.597405} & \\ & & & \mathbf{0.597405} & 1.279033 & \\ & & & & & 1.000000 \end{matrix}$	$\begin{matrix} 1.000000 & & & & & \\ & 1.000000 & & & & \\ & & 1.000000 & & & \\ & & & 1.000000 & \mathbf{-0.467075} & \\ & & & \mathbf{-0.467075} & 1.000000 & \\ & & & & & 1.000000 \end{matrix}$
2	$\begin{matrix} 1.273150 & & & & \mathbf{0.589712} & \\ & 1.000000 & & & & \\ & & 1.000000 & & & \\ & & & 1.279033 & 0.597405 & \\ \mathbf{0.589712} & & & 0.597405 & 1.552183 & \\ & & & & & 1.000000 \end{matrix}$	$\begin{matrix} 1.000000 & & & 0.216345 & \mathbf{-0.463192} & \\ & 1.000000 & & & & \\ & & 1.000000 & & & \\ 0.216345 & & & 1.000000 & -0.467075 & \\ \mathbf{-0.463192} & & & -0.467075 & 1.000000 & \\ & & & & & 1.000000 \end{matrix}$
3	$\begin{matrix} 1.459781 & \mathbf{-0.470598} & & & 0.589712 & \\ \mathbf{-0.470598} & 1.186631 & & & & \\ & & 1.000000 & & & \\ & & & 1.279033 & 0.597405 & \\ 0.589712 & & & 0.597405 & 1.552183 & \\ & & & & & 1.000000 \end{matrix}$	$\begin{matrix} 1.000000 & \mathbf{0.396583} & & 0.216345 & -0.463192 & \\ \mathbf{0.396583} & 1.000000 & & 0.085799 & -0.183694 & \\ & & 1.000000 & & & \\ 0.216345 & 0.085799 & & 1.000000 & -0.467075 & \\ -0.463192 & -0.183694 & & -0.467075 & 1.000000 & \\ & & & & & 1.000000 \end{matrix}$
4	$\begin{matrix} 1.617232 & -0.470598 & \mathbf{-0.426898} & & 0.589712 & \\ -0.470598 & 1.186631 & & & & \\ \mathbf{-0.426898} & & 1.157451 & & & \\ & & & 1.279033 & 0.597405 & \\ 0.589712 & & & 0.597405 & 1.552183 & \\ & & & & & 1.000000 \end{matrix}$	$\begin{matrix} 1.000000 & 0.396583 & \mathbf{0.368826} & 0.216345 & -0.463192 & \\ 0.396583 & 1.000000 & 0.146270 & 0.085799 & -0.183694 & \\ \mathbf{0.368826} & 0.146270 & 1.000000 & 0.079794 & -0.170837 & \\ 0.216345 & 0.085799 & 0.079794 & 1.000000 & -0.467075 & \\ -0.463192 & -0.183694 & -0.170837 & -0.467075 & 1.000000 & \\ & & & & & 1.000000 \end{matrix}$
5	$\begin{matrix} 1.617232 & -0.470598 & -0.426898 & & 0.589712 & \\ -0.470598 & 1.186631 & & & & \\ -0.426898 & & 1.157451 & & & \\ & & & 1.279033 & 0.597405 & \\ 0.589712 & & & 0.597405 & 1.706470 & \mathbf{0.422009} \\ & & & & \mathbf{0.422009} & 1.154287 \end{matrix}$	$\begin{matrix} 1.000000 & 0.396583 & 0.368826 & 0.216345 & -0.463192 & 0.169344 \\ 0.396583 & 1.000000 & 0.146270 & 0.085799 & -0.183694 & 0.067159 \\ 0.368826 & 0.146270 & 1.000000 & 0.079794 & -0.170837 & 0.062458 \\ 0.216345 & 0.085799 & 0.079794 & 1.000000 & -0.467075 & 0.170763 \\ -0.463192 & -0.183694 & -0.170837 & -0.467075 & 1.000000 & \mathbf{-0.365602} \\ 0.169344 & 0.067159 & 0.062458 & 0.170763 & \mathbf{-0.365602} & 1.000000 \end{matrix}$