university of groningen / faculty of science and engineering / mathematics and applied mathematics

Gaussian graphical models.

Bachelor's Project Mathematics

July 2017

Student: D. Besteman

First supervisor: Prof. dr. E.C. Wit

Second assessor: Dr. W.P. Krijnen

Abstract

In this thesis we study Gaussian graphical models and how to use these to model the relationships between for example stock prices of different stocks. We will show and derive how these Gaussian Graphical models are defined and how to construct them. We will use two different methods, Hypotheses testing with respect to the partial correlation and the penalized maximum likelihood solved by the graphical Lasso to construct these models. Furthermore if we can construct different models, we will discus how to select the model that is a good fit to the data. These methods are then used to fit a model to data of a financial stock market.

Contents

1 Introduction **4**

2 graphical models **4**
2.1 Graph 5
2.1.1 Independence graphs 6
2.1.2 Markov properties 7
2.1.3 Factorization 9

3 Gaussian graphical models **10**

4 Estimation of a Gaussian graphical model **12**
4.1 The likelihood 12
4.2 The maximum likelihood estimators 13
4.2.1 Quadratic form and the trace 13
4.2.2 Determinant 14
4.2.3 MLE for μ 15
4.2.4 MLE for precision matrix K 15

5 Fit a model **16**

6 Test the goodness of fit **18**
6.1 Deviance 18
6.2 AIC 19
6.3 BIC 21

7 Hypotheses testing **22**

8 Graphical lasso **25**
8.1 Penalized maximum likelihood estimator 25
8.2 Calculate the PMLE 25

9 Financial network data analysis **31**
9.1 Correlation hypotheses testing 31
9.2 Graphical Lasso 32

10 Concluding remarks **35**

1 Introduction

Consider a financial stock market, we know that the price of the stocks are some how related to each other. In this thesis we will look at a way to model these relations between, for example the stocks prices, with use of Gaussian graphical models. first there will be an introduction about what a graphical model is and some important properties will be introduced. Next we will discuss the Gaussian graphical models. How this is defined and how to estimate a Gaussian graphical model for a specific data set. We will derive that if we assume that the data is from a multivariate normal distribution (Gaussian). The precision matrix, which is the inverse of the covariance matrix, will specify the relations between the stocks. Such that there is a relation between two stocks if the corresponding entry in the precision matrix is not zero. So if we want to model the relations, we want to estimate this precision matrix. We will discuss the maximum likelihood estimator and show that this is the inverse of the sample covariance matrix. Here we will find a problem since the sample covariance matrix is not sparse, which means it has no zero entries. Which then again means that all stock prices are always related. We will show how to work around this problem in different ways. In this thesis we will discus for example the correlation hypotheses testing and the penalized maximum likelihood estimator which we solve with the graphical Lasso algorithm. Now if we can construct different models we also need to discuss how to select a model that is a good fit to the data. —Therefor we will discuss the deviance, the Bayesian information criteria and the Akaike information criteria. In the end of the thesis we will use the discussed methods to specify models using data from different stocks from big companies and select the model that is a good fit to the data determined by model selection tools similar to the ones discussed.

2 graphical models

If we have a dataset which is sampled from random variables with a specific probability distribution. A graphical model is a way to express the conditional dependencies between the random variables, according to the data, in a graph. First we will now formally define some properties used here and needed to further define graphical models.

2.1 Graph

A graph $G = (V, E)$ is a set of vertices $V = \{1, \ldots, p\}$ combined with a set of edges $E \subset V \times V$ which consists of pairs of elements of V. Such that if the ordered pair $(i, j) \in E$ then there is a directed edge, which is drawn as an arrow, from the vertices i to j in the graph. When the pairs (i, j) and (j, i) are both elements of E. Then there is a undirected edge, drawn as a line, between vertices i and j in the graph. A graph is called undirected if all edges are undirected, so

$$\forall i, j \in V: \quad (i, j) \in E \implies (j, i) \in E. \tag{1}$$

For now we will only consider undirected graphs and so if we note that $(i, j) \in E$ we state automatically that $(j, i) \in E$ and we don't have to note that. Furthermore we will also consider simple graphs which means we also won't allow loops and multiple edges between two vertices. A loop is a edges from a point to itself i.e. $(i, i) \in E$. Now will follow an example of a graph:

Example 1.

$$\begin{aligned} G &= (V, E) \\ V &= \{1, 2, 3, 4, 5\} \\ E &= \{(1, 2), (1, 3), (2, 3), (3, 4), (3, 5), (4, 5)\} \end{aligned}$$

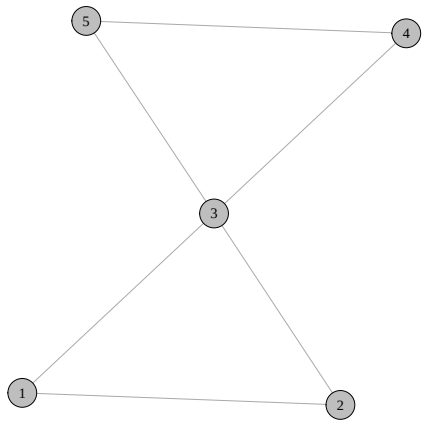

Figure 1: The graph G from example 1

2.1.1 Independence graphs

Lets start with the definition of independence for random variables. Such that we can then introduce conditional independence, and then continue with the principle of a independence graph.

Definition 1. The random variables X and Y are independent, denoted by $X \perp\!\!\!\perp Y$, when

$$\begin{aligned} P(X \leq x, Y \leq y) &= P(X \leq x)P(Y \leq y) \quad \forall x, y \\ &\Updownarrow \\ f_{XY}(x, y) &= f_X(x) f_Y(y) \quad \forall x, y. \end{aligned} \tag{2}$$

The independence relation is symmetric and so is the conditional independence relation. The definition of conditional independence is similar to the definition of unconditional independence. Only now there will be a conditional probability and conditional density functions.

Definition 2. The random variables X and Y are conditional independent given Z, denoted by $X \perp\!\!\!\perp Y|Z$, when

$$\begin{aligned} P(X \leq x, Y \leq y|Z) &= P(X \leq x|Z)P(Y \leq y|Z) \quad \forall x, y, z \quad P(Z \leq z) > 0 \\ &\Updownarrow \\ f_{XY|Z}(x, y; z) &= f_{X|Z}(x; z) f_{Y|Z}(y; z) \quad \forall x, y, z \quad f_Z(z) > 0. \end{aligned} \tag{3}$$

let $X = (X_1, X_2, \ldots, X_n)$ be a vector of random variables, with the corresponding set of vertices $V = \{1, 2, \ldots, n\}$. The graph $G = (V, E)$ is a conditional independence graph, also called independence graph, of X, when there is no edge between the vertices if the random variables corresponding to the vertices are conditional independent given the rest of the random variables. The graph G is an undirected graph, this because the conditional independence relation is symmetric.

Definition 3. The *conditional independent graph* of $X = (X_1, X_2, \ldots, X_n)$ is the undirected graph $G = (V, E)$ with $V = \{1, \ldots, n\}$ such that,

$$(i, j) \notin E \iff X_i \perp\!\!\!\perp X_j | X_{V \setminus \{i,j\}} \tag{4}$$

Example 2. Let $G = (V, E)$ be the graph of example 1 and $X = (x_1, x_2, x_3, x_4, x_5)$ be a vector of random variables, such that V corresponds with X. Now let G be the conditional independence graph of X, this means that: $X_1 \perp\!\!\!\perp X_4 | X_{V \setminus \{1,4\}}$, $X_1 \perp\!\!\!\perp X_5 | X_{V \setminus \{1,5\}}$, $X_2 \perp\!\!\!\perp X_4 | X_{V \setminus \{2,4\}}$ and $X_2 \perp\!\!\!\perp X_5 | X_{V \setminus \{2,5\}}$. Since these edges are missing in the graph.

In fact we use here a special conditional independence structure, namely the pairwise Markov property, to define the independence graph. The pairwise Markov property, as well as the two other Markov properties, will be explained in the next section.

2.1.2 Markov properties

The Markov properties are a way to interpret a graph from a graphical model. There are three Markov properties for a graph $G = (V, E)$, which are not always equivalent but there is a relation between the three of them.

1. pairwise Markov:

$$(i, j) \notin E \iff X_i \perp\!\!\!\perp X_j | X_{V \setminus \{i,j\}} \tag{5}$$

2. local Markov:

$$(i, j) \notin E \iff X_i \perp\!\!\!\perp X_j | X_{N(i)} \tag{6}$$

3. global Markov:

$$A, B, S \subset V \quad S \text{ seperates } A, B \iff X_A \perp\!\!\!\perp X_B | X_S \tag{7}$$

Pairwise Markov This is what we used to define the conditional independence graph. It is only a pairwise relation and therefor the weakest of the three properties.

Local Markov The local Markov property looks a lot like the pairwise Markov property. The difference is that X_i and X_j are independent conditional on the random variables X associated with the vertices $N(i)$, which means the neighborhood of $i \in V$, instead of the random variables X associated with all the vertices except i and j, i.e. $\{1, \ldots, n\} \setminus \{i, j\}$.

Definition 4. Let $G = (V, E)$ be a graph with vertices $V = \{1, \ldots, n\}$. Then the neighborhood of vertex i, denoted by $N(i)$, is the set of vertices j for which $(i, j) \in E$.

Global Markov The global Markov property is the strongest of the three properties. A graph is global Markov if for any two subsets A and B of V, X_A and X_B are independent conditional on X_S. Where S is a subset of V separating A and B in the graph G.

Definition 5. Let A, B, S be three distinct subsets of V. Then S *separates* A and B if and only if every path from a vertex in A to a vertex in B goes trough at least one vertex in S

Theorem 2.1 (Relation between Markov properties)**.** *Global Markov* $\implies$ *local Markov* $\implies$ *pairwise Markov*

Proof. Let $G = (V, E)$ be a graph and $X = (X_1, \ldots, X_n)$ a vector of random variables associated with the vertices $V = \{1, \ldots, n\}$.

- First lets look at why global Markov $\implies$ local Markov. We need to proof that if X with graph G has the global Markov property then,

$$(i, j) \notin E \iff X_i \perp\!\!\!\perp X_j | X_{N(i)}. \tag{8}$$

  So if $(i, j) \notin E$ then $j \notin N(i)$ and so $N(i)$ is separating i and j. Now since G has the global Markov property $X_i \perp\!\!\!\perp X_j | X_{N(i)}$

- Now lets look at local Markov $\implies$ pairwise Markov, if X with graph G has the local Markov property this means that $(i, j) \notin E \iff X_i \perp\!\!\!\perp X_j | X_{N(i)}$. Then this means $X_i \perp\!\!\!\perp X_b | X_{N(i)}$, where $b = V \backslash (\{i\} \cup N(i))$. If we now pick a vertex j such that $(i, j) \notin E$, then $j \in b$. Which means we can write $X_i \perp\!\!\!\perp (X_j, X_c) | X_{N(i)}$, with $c = V \backslash (i, j \cup N(i))$. This then can be written like, $X_i \perp\!\!\!\perp X_j | (X_{N(i)}, X_c)$. Since $a \cup c = V \backslash \{i, j\}$ this is the same as $X_i \perp\!\!\!\perp X_j | X_{V \backslash \{i,j\}}$. This means X with graph G has the pairwise Markov property.

- Note: If the joint density function $f_X(x) > 0$ then pairwise Markov $\implies$ global Markov, and so all Markov properties are equivalent.

□

Example 3. Let $X = (X_1, \ldots, X_6)$ be a vector of random variables with graph $G = (V, E)$ in figure 2. Now we will see what the different Markov properties mean for this graph.

- Assume the pairwise Markov property, then we can conclude, for example, that $X_3 \perp\!\!\!\perp X_6 | X_1, X_2, X_4, X_5$

- If we now assume the local Markov property. Then we can conclude, for example, that $X_3 \perp\!\!\!\perp X_6 | X_1, X_2$ since $N(3) = 1, 2$, Or even that $X_3 \perp\!\!\!\perp X_6 | X_5$ since $N(6) = 5$ and the independence relation is here symmetric.

- For the last example assume the global Markov property. Then we can conclude that, for example, $X_1, X_2, X_3 \perp\!\!\!\perp X_6, X_4 | X_5$ and that $X_1, X_2 \perp\!\!\!\perp X_6, X_4, X_5 | X_3$

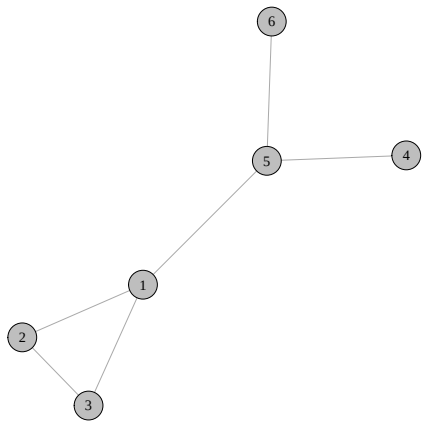

Figure 2: The graph G from example 3

2.1.3 Factorization

In this section we are going to introduce the concept of Factorization, which is used for defining a graphical model. First we will talk about *the factorization criteria* for conditional independence, such that we then can discuss the factorization of the joint density of the random variables associated with the graph G.

Proposition 2.2 (The factorization criteria for conditional independence)**.** *The random factors X and Y are independent given Z, $X \perp\!\!\!\perp Y|Z$, if and only if there exist two functions g and h such that,*

$$f_{XYZ}(x, y, z) = g(z, y)h(z, x). \tag{9}$$

Now before we go any further we will define the notion of a clique of a graph.

Definition 6. A clique of a graph $G = (V, E)$ is a *complete* subgraph of the graph G, not contained in any other complete subgraph of G. Where *complete* means that every distinct pair of vertices in the subgraph in connected with a unique edge.

So this means that a clique is the largest subgraph, where all vertices are connected to each other. There can be more cliques in a graph, and they don't have to be disjoint.

Theorem 2.3 (HammersleyClifford theorem)**.** *Let $G = (V, E)$ be a graph and $X = (X_1, \ldots, X_n)$ a vector of random variables associated with the vertices $V = \{1, \ldots, n\}$. And let $f_X(x)$ be the joint density function of X, which is equal to the likelihood function L_X. Now if $f_X > 0$ then $f_X = \prod_c h_c(x_c)$, where c are the cliques of the graph G and $h_c(x_c)$ are functions defined on the cliques.*

Corollary 2.4. *Let $G = (V, E)$ be a graph and $X = (X_1, \ldots, X_n)$ a vector of random variables associated with the vertices $V = \{1, \ldots, n\}$. And let $f_X(x)$ be the joint density function of X. If $f_X(x) = \prod_c h_c(x_c)$ then at least the pairwise Markov properties holds.*

From this we can conclude that the joint density function of the random variables and the way we can factorize this, is some how connected to the structure of the graph from the graphical model. If for the random variables X associated with graph G, $f_X > 0$ then f_X can be factorized over the cliques of the graph, $f_X(x) = \prod_c h_c(x_c)$, and then also all the Markov properties hold for the graph. Such that we can use these to interpreted the graph.

3 Gaussian graphical models

In this section we will take a closer look at a special type of graphical models, the Gaussian graphical models. We assume here that the random variables X_i for $i \in 1 \ldots n$ are normal distributed, $X_i \sim N(\mu_i, \sigma_i^2)$, and so the vector X of the n random variables X_i has a multivariate normal distribution. So let $X \sim \mathcal{N}(\mu, \Sigma)$. Now Since Σ is positive-definite and symmetric, Σ is invertible. Then we can say $\Sigma = K^{-1}$, where K is called the precision matrix. Then the probability density function of X is:

$$
\begin{aligned}
f_X(x) &= (2\pi)^{-n/2} |\Sigma|^{-1/2} e^{-1/2(x-\mu)^T \Sigma^{-1}(x-\mu)} && (10)\\
&= (2\pi)^{-n/2} |K|^{1/2} e^{-1/2(x-\mu)^T K(x-\mu)} && (11)\\
&= (2\pi)^{-n/2} |K|^{1/2} e^{-1/2 \sum_{i,j}(x_i-\mu_i)^T K_{ij}(x_j-\mu_j)} && (12)\\
&= (2\pi)^{-n/2} |K|^{1/2} \prod_{i,j} e^{-1/2(x_i-\mu_i)^T K_{ij}(x_j-\mu_j)}. && (13)
\end{aligned}
$$

Since $(2\pi)^{-n/2}|K|^{1/2}$ is a constant, we see in equation 13 a factorization of the joint probability density function, such that we have $f_X(x) = \prod_{i,j} h_{ij}(x_i, x_j)$ for all i and j with $K_{ij} \neq 0$. This because when $K_{ij} = 0$ we will only have the constant terms for this x_i and x_j since $e^0 = 1$. So we can factorize the probability density faction over all i, j for which $K_{i,j} \neq 0$. This doesn't mean that there exists only cliques of two vertices because we can say that $h_{ij}(i, j)h_{ik}(i, k)h_{kj}(k, j) = h_{ijk}(i, j, k)$. But this is a special property of the Multi normal distribution that shows that there are only pairwise interactions between the random variables. Now what we can see is that since for the multivariate distribution $f_X(x) > 0$ we can use theorem 2.3, from which we can conclude that for every $h_{ij}(x_i, x_j)$, there is a edge between the vertices of the

graph associated with X_i and X_j. So when $K_{ij} \neq 0$ there is a edge between vertices i and j in the graph associated with X.

Example 4. Let $X \sim \mathcal{N}(\mu, \Sigma)$, with

$$\Sigma = \begin{pmatrix} 1.5 & 0 & 0.5 & -1 \\ 0 & 1 & 0 & 0 \\ 0.5 & 0 & 1.5 & -1 \\ -1 & 0 & -1 & 2 \end{pmatrix}$$

Then

$$K = \Sigma^{-1} = \begin{pmatrix} 1 & 0 & 0 & 0.5 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0.5 \\ 0.5 & 0 & 0.5 & 1 \end{pmatrix}$$

Now since we know K, we know the structure of the graph $G = (V, E)$ associated with X. Namely, $V = (1, 2, 3, 4)$ and $E = \{(1, 3), (1, 2), (2, 3), (2, 4)\}$, such that G is the following graph.

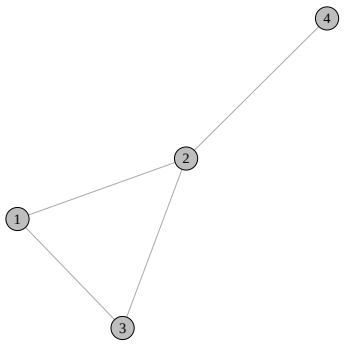

Figure 3: The graph G from example 4

4 Estimation of a Gaussian graphical model

We have seen in the previous section that for the Gaussian graphical model, the precision matrix determines the structure of the graph. So when we have a dataset sampled from a multivariate normal distribution we want to estimate this precision matrix K, such that we can estimate the graph G associated with X. We will do this by maximum likelihood estimation (MLE). The estimators $\hat{\mu}$ and $\hat{K}$ are called maximum likelihood estimators when these estimators are the ones that maximize the likelihood.

4.1 The likelihood

Now that we have seen the density function of a n-dimensional multivariate normal distributed random variable X, let $X_1, \ldots, X_p$ be independent and identically distributed samples from the random variable X, such that $p > n$ because then the sample covariance matrix S is positive definite (with probability 1), and so the inverse exists. Why we need to assume this will become clear later on. Now the likelihood function is then:

$$L(\mu, \Sigma) = (2\pi)^{-np/2}|\Sigma|^{-p/2} \prod_{i=1}^{p} e^{-1/2(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)} \tag{14}$$

$$= (2\pi)^{-np/2}|\Sigma|^{-p/2} e^{-\sum_{i=1}^{p} 1/2(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)}. \tag{15}$$

We are interested in the precision matrix $K = \Sigma^{-1}$, since this matrix determines the structure of the graph, as seen in section 3. So we can substitute this in the likelihood.

$$L(\mu, K) = (2\pi)^{-np/2}|K|^{p/2} e^{-\sum_{i=1}^{p} 1/2(x_i-\mu)^T K(x_i-\mu)} \tag{16}$$

To find the maximum likelihood estimators $\hat{\mu}$ and $\hat{K}$ we want to find the parameters K and μ that maximize the likelihood function. To do that we use the logarithm of the likelihood function, this can be done since the logarithmic function is a continuous strictly increasing function, and so the maximum will appear by the same value of $\hat{\mu}$ and $\hat{K}$. The log-likelihood function is then:

$$l(\mu, K) = -\frac{np}{2} log(2\pi) + \frac{p}{2} log(|K|) - \frac{1}{2} \sum_{i=1}^{p} (x_i - \mu)^T K(x_i - \mu). \tag{17}$$

Now we want to know for which μ and K the derivative, with respect to μ and K is zero, since that will be a maximum of the log-likelihood function.

4.2 The maximum likelihood estimators

Before we can determine the derivative of the log-likelihood function we need to introduce some matrix calculus, and some other properties.

4.2.1 Quadratic form and the trace

A quadratic form had the following form, $x^T Ax$, where x is a n-dimensional vector and A is a $n \times n$-dimensional matrix. This can also be written as, $x^T Ax = \sum_{i=i}^{n} \sum_{j=1}^{n} x_i A_{ij} x_j$. So $x^T Ax$ is a 1×1-dimensional matrix. The derivative of a quadratic form with respect to x is defined the following way.

Lemma 4.1 (The derivative of a quadratic form.)**.** *let x be a n dimensional vector and A be a $n \times n$ dimensional matrix, such that $B = x^T Ax$. Then*

$$\frac{dB}{dx} = x^T (A + A^T). \tag{18}$$

The trace of a matrix A is defined as the sum of the diagonal elements.

Definition 7. let A be a $m \times n$-dimensional matrix, with elements a_{ij} then,

$$\mathrm{tr}(A) = \sum_{i=1}^{\min(m,n)} a_{ii}. \tag{19}$$

A property of the trace of a matrix that we will use here is the invariance under cyclic permutations.

Lemma 4.2 (invariance under cyclic permutations)**.** *The trace of a matrix is invariant under cyclic permutations which means,*

$$tr(ABCD) = tr(DABC) = tr(CDAB) \neq tr(ABDC). \tag{20}$$

Where A, B, C and D are matrices.

Lemma 4.3 (The derivative of the trace)**.** *Let X be a matrix, then the derivative of the trace of X is:*

$$\frac{dtr(X)}{dX} = I \tag{21}$$

$$\frac{dtr(AX)}{X} = A \tag{22}$$

Now since a quadratic form is just a 1×1-dimensional matrix it is equal to its trace, i.e. $x^T Ax = \mathrm{tr}(x^T Ax)$. So then we can use lemma 4.2 and 4.3 to see that,

$$\frac{d}{dA} x^T Ax = \frac{d}{dA} \mathrm{tr}(x^T Ax) = \frac{d}{dA} \mathrm{tr}(xx^T A) = xx^T \tag{23}$$

4.2.2 Determinant

Definition 8. A way to define and calculate the determinant of a $n \times m$-dimensional matrix A with elements a_{ij} is:

$$|A| = \sum_{i=1}^{n} (-1)^{i+j} a_{ij} M_{ij} \quad \text{(for fixed } j\text{)} = \sum_{j=1}^{m} (-1)^{i+j} a_{ij} M_{ij} \quad \text{(for fixed } i\text{)} \tag{24}$$

Where M_{ij} is the determinant of the $(n-1) \times (n-1)$-dimensional matrix that follows from eliminating row i and column j from A. The elements $(-1)^{i+j} M_{ij}$ are also known as the cofactors of A.

Now we will use this definition of the determinant to define the derivative of the determinant.

Lemma 4.4 (The derivative of the determinant.)**.** *Let A be a matrix with elements a_{ij}, then*

$$\frac{d|A|}{da_{ij}} = \frac{d}{da_{ij}} \sum_{k=1}^{n} (-1)^{k+j} a_{kj} M_{kj} = (-1)^{i+j} M_{ij} = cofactor(A)_{ij}. \tag{25}$$

From this follows:

$$\frac{d|A|}{dA} = cofactor\ (A) = adj\ (A)^T \tag{26}$$

Where adj (A) is the adjugate matrix of A, which is defined as the transpose of the cofactor matrix.

Lemma 4.5. *The adjugate matrix is defined such that*

$$A\ adj\ (A) = |A| I \tag{27}$$

holds. From this follows that if A is invertible then,

$$adj\ (A) = |A| A^{-1} \tag{28}$$

Now we can use these two lemmas to calculate the following derivative.

$$\frac{d}{dX} \log(|X|) = \frac{1}{|X|} \frac{d|X|}{dX} = \frac{1}{|X|} \text{adj}(X)^T = \frac{1}{|X|} |X| {X^{-1}}^T = {X^{-1}}^T \tag{29}$$

4.2.3 MLE for μ

We will use these two lemma's to determine the derivative of the log-likelihood with respect to μ and K. Lets first start with the derivative with respect to μ.

$$\frac{dl(\mu,K)}{d\mu} = -\frac{1}{2}\sum_{i=1}^{p}\frac{d}{d\mu}(x_i-\mu)^T K(x_i-\mu) \tag{30}$$

$$= -\frac{1}{2}\sum_{i=1}^{p}\frac{d}{du_i}u_i^T K u_i \frac{d(x_i-\mu)}{d\mu} \tag{31}$$

$$= -\frac{1}{2}\sum_{i=1}^{p}u_i^T(K+K^T)\frac{d\mu}{d\mu} \tag{32}$$

$$= -\frac{1}{2}\sum_{i=1}^{p}(x_i-\mu)^T 2KI \tag{33}$$

$$= -\sum_{i=1}^{p}(x_i-\mu)^T K. \tag{34}$$

In this derivation we use lemma 4.1. Now if $\frac{dl(\mu,K)}{d\mu} = 0$ then this implies that $\sum_{i=1}^{p}(x_i-\mu) = 0$ and so we can conclude that $\hat{\mu} = \bar{x} = \frac{1}{p}\sum_{i=1}^{p}x_i$. Where $\bar{x}$ is called the sample mean.

4.2.4 MLE for precision matrix K

Now we will consider the derivative of the likelihood function with respect to K. Since we now are interested in K and we know the maximum likelihood estimator for μ we can substitute this in the likelihood function. Furthermore since $(x_i-\bar{x})^T K(x_i-\bar{x})$ is a quadratic form, we can use the results of equation 23.

$$\frac{dl(\mu,K)}{dK} = \frac{p}{2}\frac{d}{dK}log(|K|) - \frac{1}{2}\sum_{i=1}^{p}\frac{d}{dK}(x_i-\bar{x})^T K(x_i-\bar{x}) \tag{35}$$

$$= \frac{p}{2}\frac{d}{dK}log(|K|) - \frac{1}{2}\sum_{i=1}^{p}\frac{d}{dK}\mathrm{tr}[(x_i-\bar{x})^T K(x_i-\bar{x})] \tag{36}$$

$$= \frac{p}{2}\frac{d}{dK}log(|K|) - \frac{1}{2}\sum_{i=1}^{p}(x_i-\bar{x})(x_i-\bar{x})^T. \tag{37}$$

Now to differentiate the first part of the equation we can use the results in equation 29 such that

$$\frac{dl(\mu,K)}{dK} = \frac{p}{2}K^{-1^T} - \frac{1}{2}\sum_{i=1}^{p}(x_i-\bar{x})(x_i-\bar{x})^T \tag{38}$$

$$= \frac{p}{2}K^{-1} - \frac{1}{2}\sum_{i=1}^{p}(x_i-\bar{x})(x_i-\bar{x})^T. \tag{39}$$

Since K is symmetric K^{-1} is also symmetric. Now we want $\frac{dl(\mu,K)}{dK} = 0$, this implies the following,

$$\frac{p}{2}K^{-1} - \frac{1}{2}\sum_{i=1}^{p}(x_i - \bar{x})(x_i - \bar{x})^T = 0 \tag{40}$$

$$\frac{1}{2}\sum_{i=1}^{p}(x_i - \bar{x})(x_i - \bar{x})^T = \frac{p}{2}K^{-1} \tag{41}$$

$$\frac{1}{p}\sum_{i=1}^{p}(x_i - \bar{x})(x_i - \bar{x})^T = S = K^{-1}. \tag{42}$$

So the maximum likelihood estimator of the precision matrix K is $\hat{K} = S^{-1}$, where $S = \frac{1}{p}\sum_{i=1}^{p}(x_i - \bar{x})(x_i - \bar{x})^T$ is the sample variance matrix.

5 Fit a model

In this section we will show how to fit a specific model to the data. Lets start with an example.

Example 5. Let X_1, X_2, X_3 be samples from the multivariate normal distributed variable X with sample variance S. Now we want to fit the graphical model with the following graph to the data.

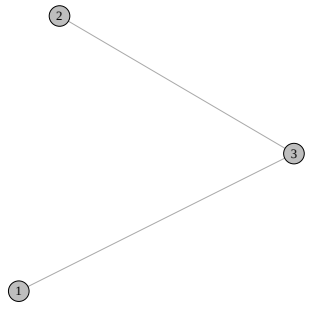

Figure 4: Graph example 5

So we know that elements K_{12} and K_{21} of the precision matrix K are zero.

$$K = \begin{pmatrix} k_{11} & 0 & k_{13} \\ 0 & k_{22} & k_{23} \\ k_{31} & k_{32} & k_{33} \end{pmatrix}$$

We now use maximum likelihood estimation to estimate K under this model. Substitute this K in the log likelihood function, then we get

$$l(\mu, K) = -\frac{np}{2}log(2\pi) + \frac{p}{2}log(|K|) - \frac{p}{2}\text{tr}(SK). \tag{43}$$

Since we want to maximize this again over K we can forget about the constant $-\frac{np}{2}log(2\pi)$ and the $\frac{p}{2}$ term in front of the log and the trace. So we focus on

$$\begin{aligned} l(K) &\propto \log(|K|) - \text{tr}(SK) && (44)\\ &= -s_{11}k_{11} - s_{22}k_{22} - s_{33}k_{33} - 2s_{23}k_{23} - 2s_{13}k_{13} \\ &\quad + \log(k_{11}k_{22}k_{33} - k_{23}^2k_{11} - k_{13}^2k_{22}). && (45)\end{aligned}$$

If we now differentiate this by K we get,

$$\begin{pmatrix} k_{22}k_{33} - k_{23}^2/|K| - s_{11} & * & -2k_{22}k_{13}/|K| - 2s_{13} \\ * & (k_{11}k_{33} - k_{13}^2)/|K| - s_{22} & -2k_{11}k_{23}/|K| - 2s_{23} \\ -2k_{22}k_{13}/|K| - 2s_{13} & -2k_{11}k_{23}/|K| - 2s_{23} & k_{11}k_{22}/|K| - s_{33} \end{pmatrix} \tag{46}$$

If we then set this equal to zero we get,

$$\begin{pmatrix} s_{11} & * & s_{13} \\ * & s_{22} & s_{23} \\ s_{13} & s_{23} & s_{33} \end{pmatrix} = \frac{1}{|K|}\begin{pmatrix} k_{22}k_{33} - k_{23}^2 & * & -k_{13}k_{22} \\ * & k_{11}k_{33} - k_{13}^2 & -k_{11}k_{23} \\ -k_{13}k_{22} & -k_{11}k_{23} & k_{11}k_{22} \end{pmatrix} \tag{47}$$

Now since we know for this model that the partial covariance of X_1 and X_2 given X_3 is zero, the following holds

$$cov(X_1, X_2) = \frac{cov(X_2, X_3)cov(X_1, X_3)}{var(X_3)}. \tag{48}$$

So for the left hand $*$ we can implement $s_{13}s_{23}/s_{33}$, from this follows then

$$\frac{s_{13}s_{23}}{s_{33}} = \frac{k_{13}k_{22}k_{23}k_{11}}{k_{11}k_{33}|K|} = \frac{k_{13}k_{23}}{|K|}. \tag{49}$$

If we then implement this in the right hand $*$, we see that the right hand side in equal to K^{-1} and so the MLE is

$$\hat{K} = \begin{pmatrix} s_{11} & \frac{s_{13}s_{23}}{s_{33}} & s_{13} \\ \frac{s_{13}s_{23}}{s_{33}} & s_{22} & s_{23} \\ s_{13} & s_{23} & s_{33} \end{pmatrix}^{-1} \tag{50}$$

In the example above we see that if we fit a specific Gaussian graphical model the entries of the maximum likelihood estimator $\hat{K}$ are the same as the entries of S, except for the entries $(i.j)$ for which there is no edge between vertices i and j in the graph. This is in general true and is proven in Whittaker (2009) [13].

6 Test the goodness of fit

In this section we will discuss the goodness of fit of a model. The maximum likelihood estimator $\hat{K} = S^{-1}$ is not sparse with probability 1, which means it has no 0 entries. So this implies that we would have a model associated with a full graph, a graph where all the points are connected with each other. This is called the *saturated model*. So this will be the model that fits the data best, but we are looking for a model that balances between goodness of fit and complexity. Since this will tell us a lot more about the relations between the variables. First we will look to the deviance, the deviance is a measure for how good a model fits the data. Later in this section we will also discuss the Akaikes information criterion (AIC) and the Bayesian information criterion (BIC) which next to goodness of fit also consider the complexity of the model, by the number of parameters.

6.1 Deviance

The deviance is a measure for the goodness of fit of a model, and we can use it to compare the goodness of fit of different models. Lets start with the definition of the deviance.

Definition 9. the deviance D of a model M is,

$$D = -2(l_M - l_s). \tag{51}$$

Where l_m is the log-likelihood of the model and l_s is the log-likelihood of the saturated model.

In our case this means,

$$l_s(\bar{x}, S) = -\frac{np}{2}\log(2\pi) - \frac{p}{2}\log(|S|) - \frac{1}{2}\sum_{i=1}^{p}(x_i - \bar{x})^T S^{-1}(x_i - \bar{x}) \tag{52}$$

$$= C - \frac{p}{2}\log(|S|) - \frac{1}{2}\mathrm{tr}[\sum_{i=1}^{p}(x_i - \bar{x})(x_i - \bar{x})^T S^{-1}] \tag{53}$$

$$= C - \frac{p}{2}\log(|S|) - \frac{1}{2}\mathrm{tr}[pI_n] = C - \frac{p}{2}\log(|S|) - \frac{np}{2}. \tag{54}$$

Where $C = -\frac{np}{2}\log(2\pi)$ is just a constant.

Before we are going to look at l_m remind that $\hat{\Sigma}$ and S only differ in the entries $\hat{\Sigma_{ij}}$ for which $K_{ij} = 0$, i.e. $\hat{\Sigma}_{ij} = s_{ij} \iff k_{ij} \neq 0$. Now since $\text{tr}(S\hat{k}) = \sum_{i,j} S_{ij}\hat{K}_{ij}$, we can write $\hat{\Sigma}$ instead of S, such that $\text{tr}(S\hat{k}) = \text{tr}(\hat{\sigma}\hat{K})$. But we already know by the way $\hat{K}$ is defined that $\hat{K} = \hat{\Sigma}^{-}1$, so $\text{tr}(\hat{\sigma}\hat{K}) = \text{tr}(I_n) = n$. Now we will use this to calculate l_m.

$$l_m(\bar{x}, \hat{K}) = C + \frac{p}{2}\log(|\hat{K}|) - \frac{1}{2}\text{tr}[\sum_{i=1}^{p}(x_i - \bar{x})(x_i - \bar{x})^T\hat{K}] \tag{55}$$

$$= C - \frac{p}{2}\log(|\hat{\Sigma}|) - \frac{p}{2}\text{tr}[S\hat{K}] \tag{56}$$

$$= C - \frac{p}{2}\log(|\hat{\Sigma}|) - \frac{np}{2}. \tag{57}$$

Now it is easy to see that the deviance D for a model M with estimated $\hat{K}$ is,

$$D = -p\log(|\hat{K}||S|) = p\log(|\hat{\Sigma}|/|S|). \tag{58}$$

It is also not hard to see is that the deviance difference for two models $M_0 \subseteq M_1$ with $\hat{K_0}$ and $\hat{K_1}$ is defined as,

$$\Delta D = p\log(|\hat{K_1}|/|\hat{K_0}|) = p\log(|\hat{\Sigma_0}|/|\hat{\Sigma_1}|) \tag{59}$$

This ΔD has a chi-square distribution (χ^2) with degrees of freedom equal to the difference in edges. Furthermore by Whittaker (2009) [13] is proven that the difference in deviance ΔD associated with removing an edge between vertices i and j is defined by,

$$\Delta D = -p\log(1 - \hat{\rho_{ij}}^2) \tag{60}$$

where ρ_{ij} is the partial correlation coefficient defined as

$$\rho_{ij} = \frac{-k_{ij}}{\sqrt{k_{11}k22}}, \tag{61}$$

and $\hat{\rho_{ij}}$ is the sample partial correlation coefficient. From this follows the relation $\rho_{ij} = 0 \iff k_{ij} = 0$, and so we can also use the partial correlation matrix to see the structure of the graph. We can use the facts that ΔD for removing an edge has a $\chi^2(1)$ distribution and that $\rho_{ij} = 0$ implies that there is no edge between vertices i and j to construct a model that fits the data. This will be shown in section 7

6.2 AIC

The Akaikes information criterion (AIC) is used to select the model that approximates the unknown true data generating process best, by the KullbackLeibler (KL) divergence. Which means that if we use the data $x_1, \ldots, x_p$ from the true unknown density function g to estimate the true model by different models $\hat{M}_i$, the model that minimizes the AIC value will be the model "closest" to the true model.

Definition 10 (AIC)**.** the AIC of a model $\hat{M}_i$ is defined as,

$$AIC(\hat{M}_i) = -2l(\hat{\theta}_i) + p_i \tag{62}$$

Where $\hat{\theta}_i$ is the MLE under the model $\hat{M}_i$ and p_i is the number of estimated parameters.

First let us consider this KL divergence.

Definition 11 (KullbackLeibler divergence)**.** Let M_1 and M_2 be continuous random variables with density functions f_1 and f_2. Then the KullbackLeibler divergence, which is a measure for the difference in information contained in a model, is defined as:

$$D_{KL}(M_1||M_2) = \int f_1(x)\log(f_1(x))dx - \int f_1(x)\log(f_2(x))dx \tag{63}$$

Now let in our case M be the true model with unknown true density function g and $\hat{M}_i$ be the estimated model from the data with density function f from the multivariate normal distribution. The KullbackLeibler divergence then is,

$$D_{KL}(M||\hat{M}) = \int g(x)\log(g(x))dx - \int g(x)\log(f(x;\hat{K}_i)). \tag{64}$$

Where $\hat{K}_i$ is the MLE of the precision matrix under the model $\hat{M}_i$, since the MLE $\hat{K}_i$ minimizes the KL divergence under model M_i. Now since $D_{KL}(M||\hat{M})$ is ar random variable, we will try to minimize the expected KL divergence, $E_g(D_{KL}(M||\hat{M}))$. But this value is unobservable since we don't know the true density g and so we need to estimate this value.

$$E_g(D_{KL}(M||\hat{M})) = E_g(\int g(x)\log(g(x))dx - \int g(x)\log(f(x;\hat{K}_i))) \tag{65}$$

$$= C - E_g(\int g(x)\log(f(x;\hat{K}_i))) \tag{66}$$

$C = \int g(x)\log(g(x))dx$ is just a constant and does not depend on the estimated model. This leaves us with estimating $l_i = E_g(\int g(x)\log(f(x;\hat{K}_i)))$. Akaike (1973) noted that $\hat{l}_i = l(\hat{K}_i)$ is a biased estimator of l_i and that the bias adjustment can be asymptotically approximated by p_i^*,

$$E(\hat{l}_i - l_i) \approx p_i^*. \tag{67}$$

Where $p_i^* = \text{tr}(J^{-1}F)$ and J and F are here forms of the expected observed Fisher information, such that if the density function f and the true density g are equal then $J = F$ and so $p_i^* = p_i$. Where p_i is the number of parameters estimated in the model $\hat{M}_i$. the matrices J and F can be estimated but this is very unstable, so it is better to use $p_i^* = p_i$. In our case p_i is equal to

the number of non zero elements in the upper diagonal part of the estimated precision matrix $\hat{K}_i$. So for the estimated KL divergence we get,

$$\hat{D}_{KL} = C - l(\hat{K}_i) + p_i. \tag{68}$$

Where we can take out C since this is an irrelevant constant, if we then multiple by 2 we get the AIC as defined in definition 10. So if we want to find the model that approximates the true data generating process best (according to the AIC), we select the model that minimizes the AIC value.

6.3 BIC

Were the AIC aims at selecting the model that approximates the unknown true data generating process best. The Bayesian information criteria (BIC) aims at selecting the model that is most likely to the data, $X = x_1, \ldots, x_p$. It does this by maximizing the posterior probability $P(\hat{M}_i|X)$ over the different models $\hat{M}_i$, with parameter set $\Theta_{\hat{M}_i}$, that estimate the true model M.

Definition 12 (BIC)**.** The Bayesian information criteria (BIC) of a model $\hat{M}_i$ is defined as,

$$BIC(\hat{M}_i) = -2l(\hat{\theta}_i) + p_i \log(p). \tag{69}$$

Where $\hat{\theta}_i$ is the MLE under the model $\hat{M}_i$, p_i is the number of estimated parameters and p is the number of data points in X.

Now from Bayes theorem we know that,

$$P(\hat{M}_i|X) = \frac{P(X|\hat{M}_i)P(\hat{M}_i)}{P(X)}. \tag{70}$$

Since $P(y)$ is a constant and $P(\hat{M}_i)$ is the prior information chosen irespective of the data, we will focus on the $P(X|\hat{M}_i)$ term. Where this can be written as,

$$P(X|\hat{M}_i) = \int_{\Theta_{\hat{M}_i}} P(X|\hat{M}_i, \theta)P(\theta|\hat{M}_i)d\theta \tag{71}$$

$$= \int_{\Theta_{\hat{M}_i}} L(\theta)d\theta \tag{72}$$

$$= \int_{\Theta_{\hat{M}_i}} \exp(n\bar{l}(\theta))d\theta. \tag{73}$$

Where $\bar{l}(\theta)$ is the mean log-likelyhood, $\bar{l}(\theta) = \frac{l(\theta)}{n}$. This integral can now be approximated by the Laplace approximation for integrals. Which states that for a function f with global maximum x_0 and a large C,

$$\int_{-\infty}^{\infty} e^{Cf(x)} \approx \sqrt{2\pi}C|f''(x_0)|e^{Cf(x_0)} \tag{74}$$

Such that in our case

$$P(X|\hat{M}_i) \approx e^{l(\hat{\theta})}(2\pi n)^{-p/2}|\frac{d^2}{d\theta^2}\bar{l}(\hat{\theta})|, \tag{75}$$

for a flat prior on $\Theta_{\hat{M}_i}$. Where p is the dimension of $\Theta_{\hat{M}_i}$. If we now ignore the terms that do not depend on n, since we consider a large n. Then we get

$$P(X|\hat{M}_i) \approx e^{l(\hat{\theta})}n^{-p/2} \tag{76}$$

Since we want to maximize this we can just take the logarithm, such that this becomes

$$l(\hat{\theta}) - \frac{p}{2}\log(n). \tag{77}$$

If we now multiply by -2 we should minimize the value instead of maximizing, and then we have the formula for the BIC as in equation 12. So to select the best model according to the BIC we want to minimize the value

$$-2l(\hat{\theta}) + p\log(n). \tag{78}$$

7 Hypotheses testing

We will now use hypotheses testing to construct a model that fits the data. let $X_1, \ldots, X_p$ be a random sample of a multivariate normal distributed random variable X, with sample mean $\bar{x}$ and sample variance S. Now $S^{-1} = \hat{K}$, where $\hat{K}$ is the estimated precision matrix of the random variable X and is the MLE for the saturated model. We can use this $\hat{K}$ to calculate the estimated partial correlation matrix $\hat{\rho}$, where we know that $\hat{\rho}$ has no zero entries with probability 1 since $\hat{K}$ has no zero entries with probability 1. We can now use hypotheses testing to see if we can exclude an edge from the saturated model. We will use the following hypotheses, $H_0 : \rho_{ij} = 0$ which means there is no edge between vertices i and j and $H_1 : \rho_{ij} \neq 0$ which means there is an edge between vertices i and j. Since $-p\log(1 - \hat{\rho_{ij}}^2)$ has a $\chi^2(1)$ distribution, we can use this to test H_0. Let the significance level be denoted by α then H_0 is rejected if

$$-p\log(1 - \hat{\rho_{ij}}^2) > \chi^2_\alpha(1), \tag{79}$$

which can be rewritten to

$$|\hat{\rho_{ij}}| > \sqrt{1 - \exp(-\frac{\chi^2_\alpha(1)}{p})}. \tag{80}$$

Now we can use this for every entry of the matrix $\hat{\rho}$, such that we accept H_0 for every $|\hat{\rho_{ij}}| < \sqrt{1 - \exp(-\frac{\chi^2_\alpha(1)}{p})}$ such that then $\hat{\rho_{ij}} = 0$. Now a example will follow to show the theory in practice.

Example 6 (World championships allround speedskating)**.** In this example we will analyze the relation between the times skated on the different distances. For this example we will use the following data.

Table 1: Data World championship allround speedskating

rank	**Name**	**500 time (s)**	**1500 time (s)**	**5000 time (s)**	**10000 time (s)**
1	Sven Kramer	36.41	105.78	372.33	790.45
2	Patrick Roest	36.27	105.88	375.10	801.11
3	Jan Blokhuijsen	36.34	108.27	375.99	793.39
4	Sverre Lunde Pedersen	36.54	106.23	383.61	804.57
5	Bart Swings	36.87	106.34	382.44	800.61
6	Andrea Giovannini	36.80	108.48	381.60	806.17
7	Patrick Beckert	37.74	107.88	379.88	800.84
8	Shota Nakamura	36.00	106.41	392.34	837.88

The sample variance matrix S of this data is,

$$S = \begin{pmatrix} 0.28 & 0.26 & -0.39 & -2.60 \\ 0.26 & 1.23 & 0.09 & -1.60 \\ -0.39 & 0.09 & 38.77 & 82.16 \\ -2.60 & -1.60 & 82.16 & 211.11 \end{pmatrix}. \tag{81}$$

Now we can calculate the inverse of the sample variance matrix $S^{-1} = \hat{K}$, which is the MLE of the saturated model and the sample partial correlation matrix $\hat{\rho}$.

$$\hat{K} = \begin{pmatrix} 5.94 & -0.90 & -0.45 & 0.24 \\ -0.90 & 1.01 & -0.02 & 0.01 \\ -0.45 & -0.02 & 0.19 & -0.08 \\ 0.24 & 0.01 & -0.08 & 0.04 \end{pmatrix} \tag{82}$$

$$\hat{\rho} = \begin{pmatrix} 1.00 & 0.37 & 0.42 & -0.50 \\ 0.37 & 1.00 & 0.05 & -0.03 \\ 0.42 & 0.05 & 1.00 & 0.93 \\ -0.50 & -0.03 & 0.93 & 1.00 \end{pmatrix} \tag{83}$$

As we can see both matrices have no entries equal to zero. And so for this model all vertices are connected with each other. But now we can use the hypothesis testing to exclude edges from the graph. Let for all $i, j \in \{1, 2, 3, 4\}$ the $H_0 : \hat{\rho}_{ij} = 0$ and the $H_1 : \hat{\rho}_{ij} \neq 0$ and let our significance level be $\alpha = 0.05$, then $\chi^2_{0.05}(1) = 3.481$. Now for every $|\hat{\rho}_{ij}| < \sqrt{1 - \exp(-\frac{3.841}{8})} = 0.617$ we accept H_0 so the new $\hat{\rho}$ is,

$$\hat{\rho} = \begin{pmatrix} 1.00 & 0 & 0 & 0 \\ 0 & 1.00 & 0 & 0 \\ 0 & 0 & 1.00 & 0.93 \\ 0 & 0 & 0.93 & 1.00 \end{pmatrix}. \tag{84}$$

Associated with this partial correlation matrix is the following graph,

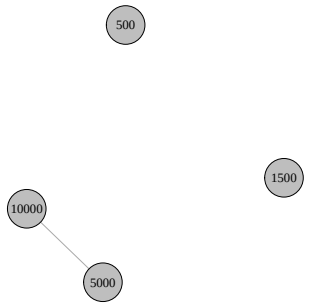

Figure 5: Graph example 6 with $\alpha = 0.05$

We can for example use an other significance level $\alpha = 0.25$, then $\chi^2_{0.25}(1) = 1.323$. We now accept H_0 for all $|\hat{\rho}_{ij}| < 0.39$ and so our new $\hat{\rho}$ is,

$$\hat{\rho} = \begin{pmatrix} 1.00 & 0 & 0.42 & -0.50 \\ 0 & 1.00 & 0 & 0 \\ 0.42 & 0 & 1.00 & 0.93 \\ -0.50 & 0 & 0.93 & 1.00 \end{pmatrix}. \tag{85}$$

And the graph associated is,

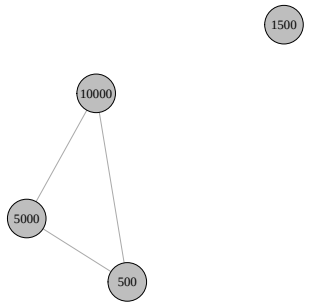

Figure 6: Graph example 6 with $\alpha = 0.25$

8 Graphical lasso

In section 7 we explained how to use hypothesis testing to find a model that fits the data. However we use here repeated hypothesis testing, in particular we use a hypothesis test for each entry of $\hat{K}$, which is not consistent. Furthermore we can only use this if the MLE $\hat{K}$ exists, this is only the case when the sample size p is larger then the number of parameters n, since then the inverse of the sample covariance matrix S exists, which in practice is not always true. So therefor in this section we will discuss another way to find a model that fits the data, using the penalized maximum likelihood solved by the graphical lasso.

8.1 Penalized maximum likelihood estimator

Since the MLE $\hat{K}$ doesn't always exist and if it exists is not sparse. We can look at the penalized maximum likelihood estimator (PMLE), $\hat{K}_p$. This PMLE $\hat{K}_p$ is defined as,

$$\hat{K}_p = \arg\max_K \quad l(K) - \text{pen}(K). \tag{86}$$

Where $l(K)$ is the log likelihood function and $\text{pen}(K)$ intended to penalize the model. In our case of the Gaussian graphical models $l(K) \propto \log(|K|) - \text{tr}(SK)$ and we will use $\text{pen}(K) = \lambda||K||_1$, where $\|K\|_1$ is the l_1-norm.

$$\hat{K}_p = \arg\max_K \quad l(K) - \lambda\|K\|_1. \tag{87}$$

We will use this to ensure that the estimate exist and to aim at a more sparse estimate. In our case we can discuss if we want to penalize the diagonal elements too, since we are interested in the conditional dependence between the variables and this concerns only the of-diagonal elements.

8.2 Calculate the PMLE

Compared to calculating the MLE calculating the PMLE by solving equation 87 is more complicate. Fortunate *Friedman et all (2008)*[5] derived a iterative algorithm to calculate the PMLE, and called this algorithm the *graphical lasso*. We want to maximize $l(K) - \lambda\|K\|_1$ over K, so lets take the derivative of this and set this equal to zero.

$$\frac{d}{dK}\log(|K|) - \text{tr}(SK) - \lambda\|K\|_1 = 0 \tag{88}$$

$$K^{-1} - S - \lambda\Gamma = 0 \tag{89}$$

Where Γ is a matrix, such that the entries $\Gamma_{ij} = \text{sign}(K_{ij})$ if $K_{ij} \neq 0$ and $\Gamma_{ij} \in [-1, 1]$ if $K_{ij} = 0$, since $\|K\|_1 = \sum_{i,j} |K_{ij}|$ the derivative with respect to K_{ij} of the norm is $\frac{d}{dK_{ij}}\|K\|_1 = \text{sign}(K_{ij})$. Now let W be the current estimate of $K^{-1} = \Sigma$, such that

$$W - S - \lambda\Gamma = 0. \tag{90}$$

In order to solve this equation we will now use block coordinate descent. Which means we partition each of these matrices in to one column versus the rest, such that

$$\begin{aligned} K = \begin{pmatrix} K_{11} & K_j \\ K_j^T & K_{jj} \end{pmatrix}, \quad W = \begin{pmatrix} W_{11} & W_j \\ W_j^T & W_{jj} \end{pmatrix} \\ \Gamma = \begin{pmatrix} \Gamma_{11} & \Gamma_j \\ \Gamma_j^T & \Gamma_{jj} \end{pmatrix}, \quad S = \begin{pmatrix} S_{11} & S_j \\ S_j^T & S_{jj} \end{pmatrix}. \end{aligned} \tag{91}$$

Where for example K_{11} is a $(n-1) \times (n-1)$ dimensional matrix, K_j is a $(n-1)$ dimensional vector and K_{jj} is the j^{th} diagonal element. Now we can write equation 90 in block form, such that

$$W_j - S_j - \lambda \Gamma_j = 0. \tag{92}$$

Note that $WK = I$ since by definition $W = K^{-1}$, in block form we then get

$$\begin{pmatrix} W_{11} & W_j \\ W_j^T & W_{jj} \end{pmatrix} \begin{pmatrix} K_{11} & K_j \\ K_j^T & K_{jj} \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & 1 \end{pmatrix}. \tag{93}$$

From this we can see that $W_{11}K_j + W_jK_{jj} = 0$, and hence $W_j = -W_{11}\frac{K_j}{K_{jj}}$. Now let $-\frac{K_j}{K_{jj}} = \beta$ such that $\text{sign}(K_j) = -\text{sign}(\beta)$ since $K_{jj} > 0$. Then we can write

$$W_{11}\beta - S_j + \lambda \text{ sign}(\beta) = 0. \tag{94}$$

We now note that this looks a lot like a Lasso optimization problem. Where the Lasso optimization problem want to minimize

$$\frac{1}{2N}(y - X\beta)^T(y - X\beta) + \lambda \| \beta\|_1 \tag{95}$$

with respect to β, where y are the outcome variables and X is the predictor matrix. Which is equivalent with solving

$$\frac{1}{N}X^TX\beta - \frac{1}{N}X^Ty + \lambda \text{ sign}(\beta) = 0 \tag{96}$$

for β.

In *Friedmann et all. (2007)* [4] an coordinate descent algorithm is discussed to solve such a Lasso problem, which means that . Furthermore as in *Friedmann et all. (2008)* [5] described we can use a slightly modified form of this algorithm to solve equation 94. The standard lasso estimates for the p^{th} variable takes as input S_{11} and S_p, instead we use W_{11} and S_p. Where W_{11} is our current estimate of W_{11}. Then update $W_j = W_{11}\hat{\beta}$ where $\hat{\beta}$ is the solution acquired from solving equation 94 for β. Note that if we penalize the diagonal elements then since $K_{ii} > 0$ from equation 90 follows that the solutions for the diagonal element of W are $W_{ii} = S_{ii} + \lambda$, if we chose to not penalize the diagonal elements we get that the solutions for the diagonal element of W are $W_{ii} = S_{ii}$. Now

we have estimated W but we want to know K. Remind that from equation 93 followed $W_{11}K_j + W_jK_{jj} = 0$, so $K_j = -K_{jj}W_{11}^{-1}W_j$. Since $W_j = W_{11}\hat{\beta}$ it follows that $W_{11}^{-1}W_j = \hat{\beta}$, and so $K_j = -K_{jj}\hat{\beta}$. In the same way follows from equation 93 that $W_j^TK_j + W_{jj}K_{jj} = 1$ and so $-W_j^TK_{jj}\hat{\beta} + W_{jj}K_{jj} = 1$. Which leads to $K_{jj} = {}^1/_{(W_{jj} - W_j^T\hat{\beta})}$. Now the algorithm described above is called the graphical lasso and works in the following way:

Algorithm (Graphical lasso)**.** *1. Start with $W = S + \lambda I$ such that the diagonal elements are correct since they don't change in what comes next. (we can chose to leave out the penalty term λI here, if we dont want to penalize the diagonal elements)*

2. repeat for $j = 1, 2, \ldots, n, 1, \ldots, n, \ldots,$

(a) partition W and S in all but the j^{th} row and column, W_{11} and S_{11}, the j^{th} diagonal element, W_{jj} and S_{jj}, and the j^{th} row,W_j^T and S_j^T, and the j^{th} column, W_j and S_j.

(b) Solve the lasso problem $W_{11}\beta - S_j + \lambda\ sign(\beta) = 0$ using the coordinate decent algorithm with input W_{11} and S_j. Which gives a $n-1$ dimensional vector solution $\hat{\beta}$.

(c) Update the corresponding row and and column vector $W_j = W_{11}\hat{\beta}$

(d) Calculate $K_{jj} = {}^1/_{(W_{jj} - W_j^T\hat{\beta})}$ and use this to update $K_j = -K_{jj}\hat{\beta}$. (You can also store for each j the vector $\hat{\beta}$ in a matrix and calculate K_j for each $j \in \{1, \ldots, n\}$ after convergence)

3. Continue until convergence.

Now we will follow with an example of this method used on the speedskating data from example 6.

Example 7. Let the data and so the sample variance matrix S be as in example 6. Now let us for example take $\lambda = 0.5$, then we find the following solution of the graphical lasso algorithm.

$$\hat{K}_{0.5} = \begin{pmatrix} 1.31117 & -0.00000 & -0.00000 & 0.01303 \\ -0.00000 & 0.57897 & -0.00118 & 0.00347 \\ -0.00000 & -0.00118 & 0.12887 & -0.04974 \\ 0.01303 & 0.00346 & -0.04974 & 0.02407 \end{pmatrix} \tag{97}$$

This suggests the following graph:

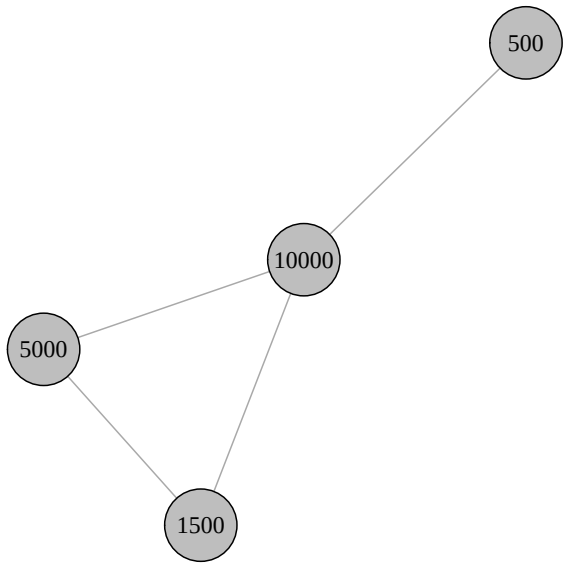

Figure 7: Graph example 7 with $\lambda = 0.5$

If we now take a bigger penalty term λ, say $\lambda = 1.5$. We have the following estimate of the precision matrix $\hat{K}_{1.5}$ and the associated graph,

$$\hat{K}_{1.5} = \begin{pmatrix} 0.56247 & -0.00000 & -0.00000 & 0.00292 \\ -0.00000 & 0.36591 & -0.00000 & 0.00017 \\ -0.00000 & -0.00000 & 0.10340 & -0.03923 \\ 0.00292 & 0.00017 & -0.03923 & 0.01960 \end{pmatrix} \tag{98}$$

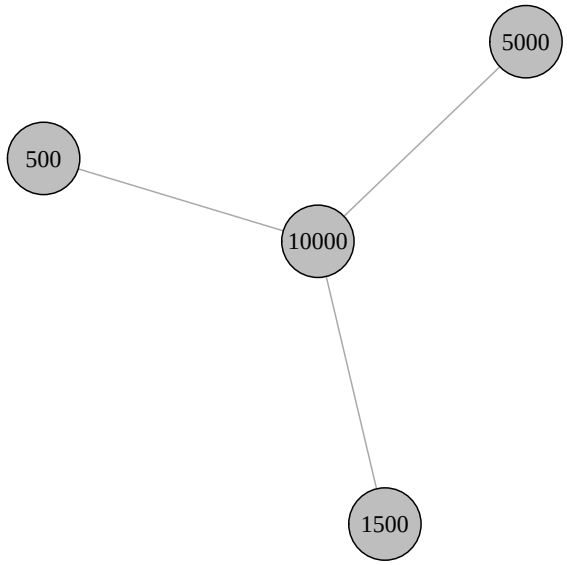

Figure 8: Graph example 7 with $\lambda = 1.5$

We see that all the entries of $\hat{K}$ are closer to 0 for this bigger λ. We even see that $\hat{K}_{23}$ has become 0 and so the edge between vertices 1500 and 5000 is excluded in this model. Furthermore when λ gets bigger the entries of $\hat{K}$ will converge to 0 and so there will be more edges excluded from the graph. This can be seen in the following plot, where we plotted the entries of $\hat{K}$ against λ.

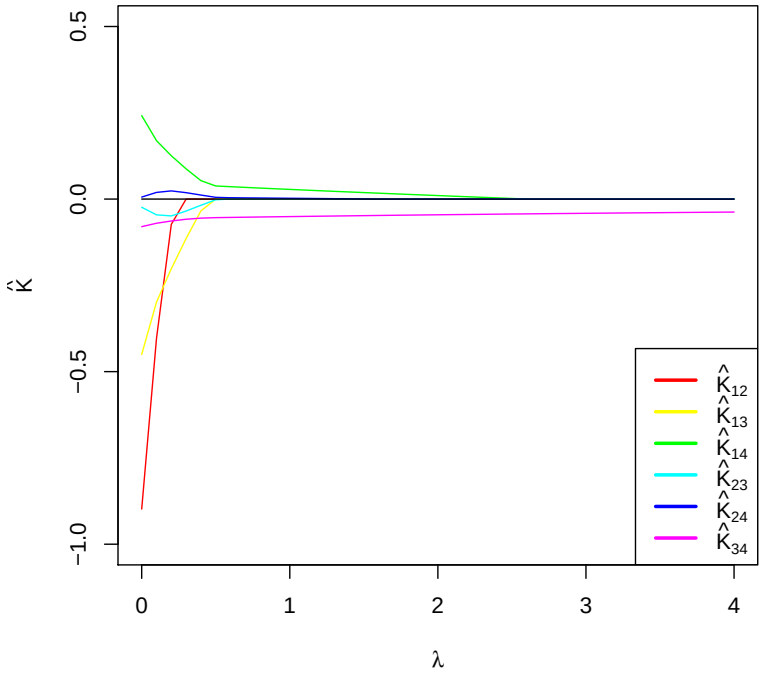

Figure 9: Plot example 7 showing the relation between the entries of $\hat{K}$ and λ

So we see that the algorithm has as input the sample covariance matrix S and the parameter λ. A larger λ means a more sparse estimate of K, but how to select a proper λ? We can do this for example with the model selection tools like BIC and AIC discussed in section 6. For selecting λ in paper we will use an extended form of the BIC, eBIC.

9 Financial network data analysis

In this section we will use the methods we discussed to model a financial network. We want to model a network of a financial market where stocks are sold. Of course the stock prices are correlated, so now we want to model these relations between the stock prices. We will look at 21 stocks from well known companies, and try to model the relations between these stocks. Our sample consists of the adjusted closing price for each stock each day during the period from 01-06-2015 till 26-05-2017, which are 504 data points per stock. The adjusted closing price is the closing price of the stock at the end of the day adjusted due to corporate actions like splits, dividends and so on.

List of stocks.

1. Apple (AAPL)
2. Amazon (AMZN)
3. Alibaba (BABA)
4. Facebook (FB)
5. GOOGLE (GOOG)
6. GOOGLE (GOOGL)
7. Intel (INTC)
8. Coca-Cola (KO)
9. Mastercard (MA)
10. McDonald's (MCD)
11. Microsoft (MSFT)
12. Netflix (NFLX)
13. Pepsi (PEP)
14. Royal Dutch Shell (RDS-B)
15. Twitter (TWTR)
16. Visa (V)
17. Starbux (SBUX)
18. Nike (NKE)
19. Total (TOT)
20. BP (BP)
21. GoPro (GPRO)

9.1 Correlation hypotheses testing

The first method we will use is the correlation hypotheses testing, explained in section 7. So from the data we can calculate the sample partial correlation matrix $\hat{\rho}$, which is a 21×21-dimensional symmetric matrix. Now we will test the following hypotheses $H_0 : \hat{\rho}_{ij} = 0$ and $H_1 : \hat{\rho}_{ij} \neq 0$ such that if

$$|\hat{\rho_{ij}}| > \sqrt{1 - \exp(-\frac{\chi^2_\alpha(1)}{504})}, \tag{99}$$

we reject H_0 and otherwise we accept H_0 such that $\hat{\rho_{ij}} = 0$. Now we set the significance level $\alpha = 0.05$, then $\chi^2_{0.05}(1) = 3.841$ and then

$$\sqrt{1 - \exp(-\frac{\chi^2_{0.05}(1)}{504})} = 0.08713 \tag{100}$$

So if we now for every $|\hat{\rho_{ij}}| < 0.08713$ accept H_0 such that these $\hat{\rho_{ij}} = 0$. We now have a model for this financial stock market and we can visualize in the following graph.

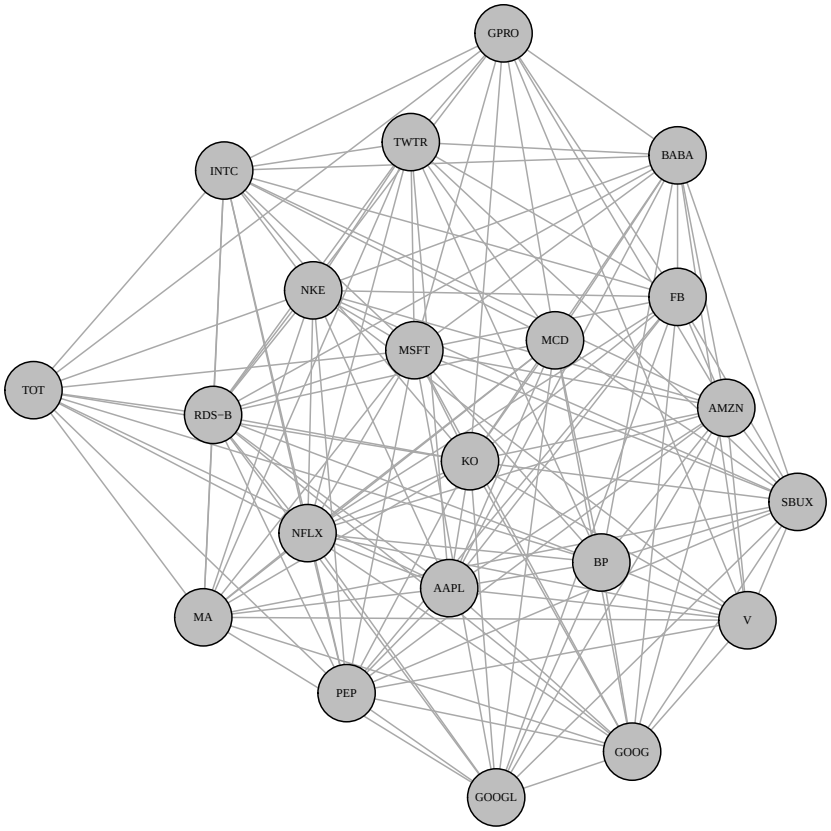

Figure 10: Graph analyzing stocks using correlation hypothesis testing

9.2 Graphical Lasso

We will now model the financial stock market using the Graphical Lasso method described in section 8. For this method we will use the "huge"-package in R. We will construct a model for different values of λ and use model selection tools to select the best model. Now first we will construct models for 100

different values of λ between 0 and 1. Then we will use the model selection tool, extended Bayesian information criteria (eBIC), which is implemented in the "huge"-package and is an extended form of the in section 6.3 explained BIC. The eBIC selects $\lambda = 0.09981$, which corresponds to the following graphical model.

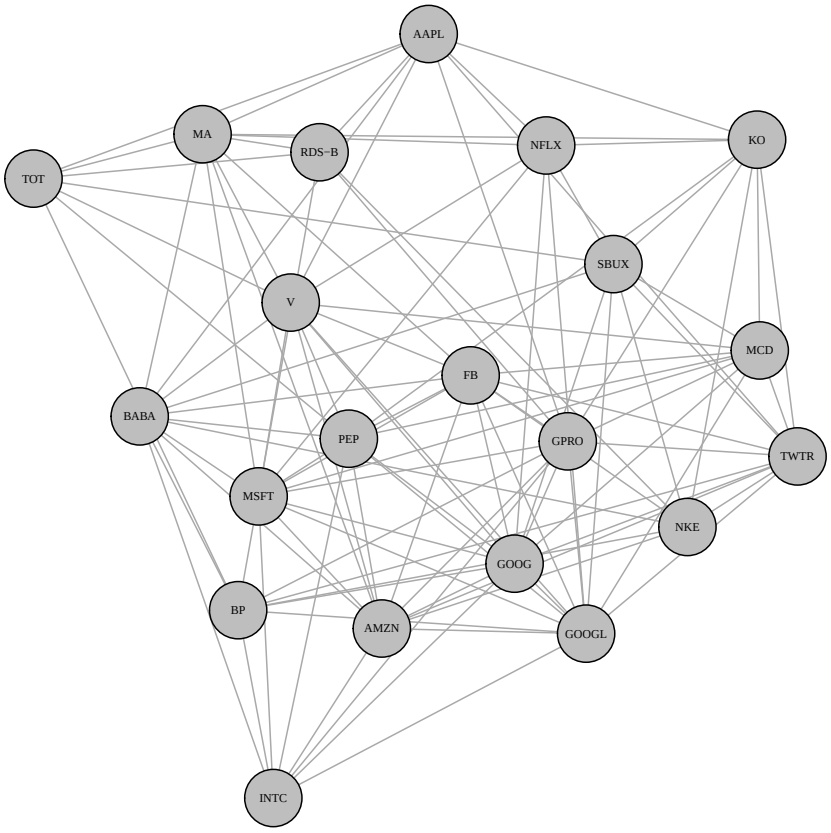

Figure 11: Graph analyzing stocks using graphical Lasso with model selection using eBIC

Furthermore another model selection tool is implemented in the "huge"-package, that is called stability approach to regularization selection (stars). The stars method selects $\lambda = 0.2472$, which then corresponds to the following graphical model.

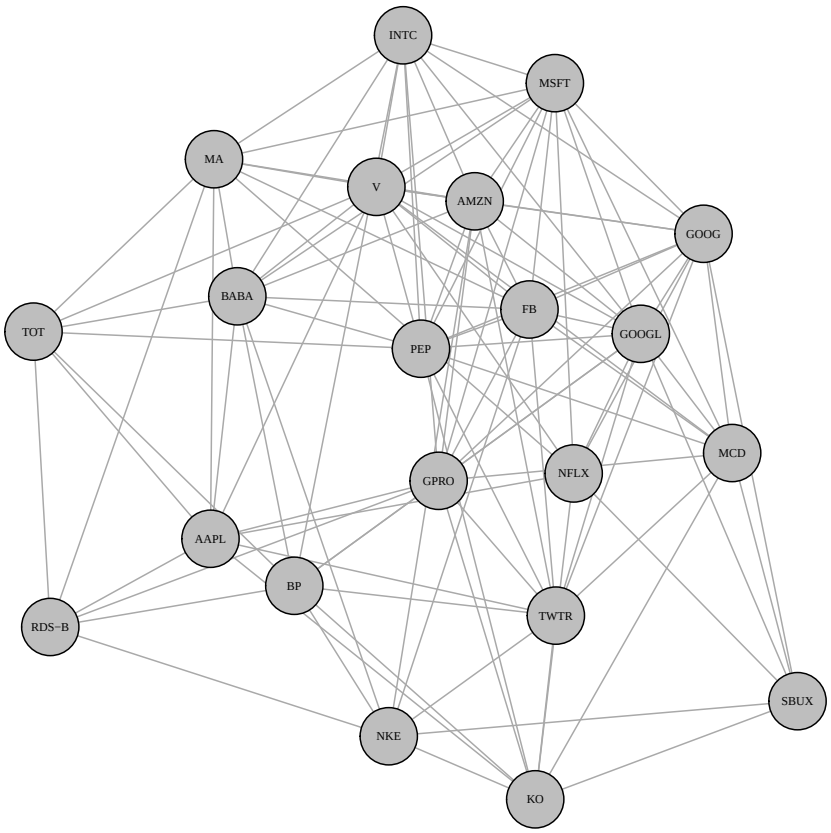

Figure 12: Graph analyzing stocks using graphical Lasso with model selection using "stars"

10 Concluding remarks

In this thesis we have explained and derived the notion of Gaussian graphical models and its most important properties. We have shown different methods of how to construct Gaussian graphical models and we have used the discussed methods to construct such models for a financial stock market. We then used model selection tools similar to the ones we described to select a model that fits the data. We have seen that the methods work and we have constructed three models using different methods visualized in different graphs. We can see the difference in density of the graphs for different methods. The method of the penalized maximum likelihood solved by the graphical Lasso should be preferred above the partial correlation hypotheses testing since repeated hypotheses testing has been proved to be not consistent. However we can discus for instance also the assumption of normality of the data since we sample the data over time and don't consider this in the model. Further more I think there should be more research done on the subject of model selection for these kind of models. The model selection tools we used worked, but I think that with more research there should be a more precise way of selecting these kind of graphical models.

References

[1] Onureena Banerjee, Laurent El Ghaoui, and Alexandre dAspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine learning research*, 9(Mar):485–516, 2008.

[2] Joseph E Cavanaugh. Unifying the derivations for the akaike and corrected akaike information criteria. *Statistics & Probability Letters*, 33(2):201–208, 1997.

[3] David Edwards. *Introduction to graphical modelling*. Springer Science & Business Media, 2012.

[4] Jerome Friedman, Trevor Hastie, Holger Höfling, Robert Tibshirani, et al. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332, 2007.

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

[6] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity*. CRC press, 2015.

[7] Søren Højsgaard, David Edwards, and Steffen Lauritzen. *Graphical models with R*. Springer Science & Business Media, 2012.

[8] Shuhua Hu. Akaike information criterion. *Center for Research in Scientific Computation*, 93, 2007.

[9] Michael Irwin Jordan. *Learning in graphical models*, volume 89. Springer Science & Business Media, 1998.

[10] Steffen L Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.

[11] Andrew A Neath and Joseph E Cavanaugh. The bayesian information criterion: background, derivation, and applications. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(2):199–203, 2012.

[12] Pradeep Ravikumar, Martin J Wainwright, Garvesh Raskutti, Bin Yu, et al. High-dimensional covariance estimation by minimizing 1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

[13] Joe Whittaker. *Graphical models in applied multivariate statistics*. Wiley Publishing, 2009.

[14] Ernst Wit, Edwin van den Heuvel, and Jan-Willem Romeijn. all models are wrong...: an introduction to model uncertainty. *Statistica Neerlandica*, 66(3):217–236, 2012.