Determining conditional Gaussian distributions for decomposable graphical models

-An new method-

Master thesis

Peter Enemark Lund

Aalborg University
Department of Mathematics

AALBORG UNIVERSITY
STUDENT REPORT

Department of Mathematics
Fredrik Bajers Vej 7G
Phone 99 40 99 40
www.math.aau.dk

Title:
Determining conditional Gaussian distributions for decomposable graphical models
- A new method

Project Period:
Fall 2015

Author:
Peter Enemark Lund

Supervisor:
Poul Svante Eriksen

Copies: 4

Page Numbers: 37

Date of Completion:
January 11, 2016

Abstract:

This report introduces a new method for calculating conditional Gaussian distributions for decomposable models.
Basic graph theory is explained followed by multivariate normal distribution and maximum likelihood estimation. The new method is then presented along with the general case.
Lastly the implementation in R is explained, and comparison with a simple method is made.

Resumé

Ved beregning af betingede Gaussiske fordelinger, stiger beregningstiden når antallet af betingende variable går op. Det er derfor nødvendigt med nye metoder for at reducere beregningstiden, da den datamængde vi kan indsamle konstant vokser og beregningstiden dermed øges.

I dag findes metoder til at beregne betingede Gaussiske fordelinger, hvor variablene kan repræsenteres via en dekomposabel model, der er hurtigere end den simple direkte metode, som kræver invertering af kovariance matricen for de betingende variable. I rapporten præsenteres en ny metode som ikke benytter matrix inversion.

Den grundlæggende grafteori bliver gennemgået, efterfulgt af den dekomposable flerdimensionale normalfordeling og maximum likelihood estimationen af denne.

Den nye metode bliver derefter præsenteret, opdelt i tre dele. En initialisering, som omskriver simultan fordelingen til et produkt af potentialer. Herefter en frem og tilbage propagering hvor de enkelte variable gennemløbes. Dernæst præsenteres en general udgave hvor hele klynger opdateres på samme tid.

Slutteligt beskrives R kode implementeringen af algoritmen og R koden brugt til at teste hastigheden ved den nye metode. Den bliver derefter holdt op mod den direkte metode.

Contents

Chapter 1

Introduction

When calculating conditional Gaussian distributions, the computational time goes up with the number of conditioning variables. This creates a need to find faster ways of obtaining the same results, as the amount of data we can collect constantly increases thereby increasing the computational time.

The simple method of calculating conditional Gaussian Distributions involves inverting a large matrix. Today there are, in the case of decomposable models, a faster approaches.

We will present a new method that does not rely on matrix inversion, except for an initialization. The original draft for the algorithm in this method was provided by Poul Svante Eriksen.

Basic graph theory is first presented, which defines the terminology used later. The decomposable multivariate normal distribution and its maximum likelihood estimation are then presented as they form the basis for the new method. The new method is laid out along with a more general case, which also covers an existing method. Lastly, the implemented code for the new method and testing is presented followed by a comparison of this and a more simple method.

Chapter 2

Basic Graph theory

[Cowell et al., 1999][Chapter 4]

A graph $\mathcal{G} = (V, E)$ is a finite set V (vertices) and a set E of ordered (a, b) and unordered $\{a, b\}$ pairs (edge) of elements from V. In this report we will restrict $\mathcal{E}$ so $a \neq b$ and it only can have one pair containing both a and b. An unordered edge $\{a, b\}$ is called undirected. A graph is called undirected if all edges in E are unordered and we say that a and b are neighbors which we write as $a \sim b$. All neighbors of a vertex a is denoted $ne(a)$. If an edge (a, b) in E is ordered, we call the edge directed $a \rightarrow b$, say that a is a parent of b and that b is a child of a. A graph is called directed if all edges in E are directed and all parents of a vertex a is denoted $pa(a)$.

Let A be a subset of V and let $\mathcal{G}^A$ denote the graph restricted to A, then $\mathcal{G}^A$ is called a *subgraph* of $\mathcal{G}$.

A graph $\mathcal{G} = (V, E)$ is called *complete* if all vertices are connected, i.e. for all $a, b \in V | a \neq b$, E contains a pair with a and b. A subgraph $\mathcal{G}^A = (A, E_A)$ and the subset A are called complete if E_A contains a pair with a and b for all $a, b \in A | a \neq b$. Given a complete subgraph A, if there exists no subgraph B such that $A \subset B$ and B is complete, then A is called a *clique*.

The cliques $(C_1, \ldots, C_q)$ of a graph is said to have a *running intersection property* if they can be ordered such that for $i = 1, \ldots, q-1$ there is a $j \in \{C_{i+1}, \ldots, C_q\}$ such that

$$C_i \cap (C_{i+1} \cup \ldots \cup C_q) \subset C_j$$

A *Path* from vertex c to d in a graph is a sequence of vertices $a_0, a_1, \ldots, a_n$ such that $c = a_0$, $d = a_n$ and $\{a_{i-1}, a_i\} \in E$ or $(a_{i-1}, a_i) \in E$ for $i = 1, \ldots, n$. If at least one of these edges is directed $a_{i-1} \rightarrow a_i$ it is a *directed path.* A *n-cycle* is a path where $a_0 = a_n$ and is called a *directed cycle* if it is a directed path. We call a graph *acyclic* if it contains no cycles. If a graph is directed and acyclic it is called a *Directed Acyclic Graph* (DAG).

If $pa(a_i)$ for all a_i in a DAG form a complete subset, the DAG is called *perfect.* The numbering/sequence of the vertices $a_1, \ldots, a_n$ in a DAG $\mathcal{G}$ is *perfect* if $a_j \in pa(a_i) \Rightarrow i < j$ for all $a_i \in \mathcal{G}$. The sequence of the vertices in an undirected graph is *perfect* if $ne(a_i) \cap \{a_{i+1}, \ldots a_n\}$ for all $a_i \in \mathcal{G}$ forms a complete sub-graph. Given a perfect sequence of an undirected graph, a directed version can be obtained by directing all edges so they go from higher to lower numbered vertices. Note that other litterateur defines a perfect sequence/numbering as the reverse.

A subset $C \subseteq V$ is said to be a (a, b)-*separator*, or *separate* A and B, if every path from any $a \in A$ to any $b \in B$ intersects C in the undirected version of V. If no proper subset of an (a, b)-separator C exist that also is an (a, b)-separator, then C is a *minimal* (a, b)-separator.

Given two non-consecutive vertices a_i a_j in a n-cycle in a graph $\mathcal{G}$, then if $\mathcal{G}$ contains the edge (a_i, a_j), that edge is called a *chord.* We call an undirected graph *chordal* or *triangulated* if all its n-cycles with $n \geq 4$ contains a chord.

Definition 2.1 (Decomposition)
Given an undirected graph $\mathcal{G} = (V, E)$ and let A, B and C be three disjoint subsets of V, then we say (A, B, C) form a decomposition of $\mathcal{G}$ or decompose $\mathcal{G}$, if

- *$V = A \cup B \cup C$;*
- *A from B are separated by C;*
- *C is complete.*

Note that any of the three subsets can be empty, and if A and B are non-empty subsets the decomposition is said to be proper.

Definition 2.2 (Decomposable graph)
An undirected graph $\mathcal{G}$ is said to be decomposable either $\mathcal{G}$ is complete or there exist a proper decomposition of $\mathcal{G}$ such that the subgraphs $\mathcal{G}_{A \cup C}$ and $\mathcal{G}_{B \cup C}$ are decomposable.

With this we can now show the connection between decomposable and chordal graphs

Theorem 2.3
Given an undirected graph $\mathcal{G}$ the following is equivalent:

1. *$\mathcal{G}$ is decomposable;*
2. *$\mathcal{G}$ is chordal;*
3. *Every minimal (a, b)-separator is complete.*

Proof.
The conditions are always true for three or fewer vertices. The proof now follows by induction. Assume that it is true if the number of vertices $|V| \geq n$ and consider a graph where $|V| = n + 1$.

$1 \Rightarrow 2$:
If $\mathcal{G}$ is decomposable it can either be complete, in which case $\mathcal{G}$ is also chordal, or properly decomposed into $\mathcal{G}_{A \cup C}$ and $\mathcal{G}_{B \cup C}$. These both contain less than $n + 1$ vertices and are therefore chordal, so for $\mathcal{G}$ not to be chordal there must be a chordless cycle intersecting A and B. Since C is the separator, this cycle must intersect C at least twice. However, then the cycle contains a chord since C is assumed to be complete.

$2 \Rightarrow 3$:
Now assume that $\mathcal{G}$ is chordal and we have a minimal $(a, b) - separator\ C$. If C is a single vertex, it is complete. Assume C is not complete and contain more than one vertex. Then select two non-adjacent vertices s_1 and s_2. There is a path going from a through s_1 to b and back to a through s_2 since C is minimal. This forms a cycle if we allow it to have repeated vertices. These repeated vertices and chords in the cycle are now used to shorten the cycle, still leaving one vertex in A and one in B. This must lead to a cycle with a length of at least 4. This must contain a chord, and as there can be no chord from A to B, it must connect s_1 and s_2. Since it is not possible not have two non-adjacent vertices in C, it is complete.

$3 \Rightarrow 1$:
We now suppose that every minimal (a, b)-separator is complete. If $\mathcal{G}$ is complete it is decomposable, so assume there are two non-adjacent vertices a and b. We also suppose that this result holds for every proper sub-graph of $\mathcal{G}$. Let C be a minimal (a, b)-separator that divides the vertices in $\mathcal{G}$ into A, B, C and D (where D is the

remaining vertices and might be empty). Since C is complete, $(A \cup D, B, C)$ form a proper decomposition of $\mathcal{G}$. For $\mathcal{G}$ to be decomposable, the sub-graphs $\mathcal{G}_{A \cup D \cup C}$ and $\mathcal{G}_{B \cup C}$ must also be decomposable. Let C_1 be a minimal (a_1, b_1)-separator in $\mathcal{G}_{A \cup D \cup C}$, then it is also a minimal separator in $\mathcal{G}$ and by assumption complete. The same goes for $\mathcal{G}_{B \cup D}$ and since these are proper sub-graphs it holds for them by assumption. We can therefore decompose $\mathcal{G}$ into decomposable sub-graphs.

□

The following is from [Lauritzen, 2004, Proposition 2.17]

Theorem 2.4
A undirected graph $\mathcal{G}$ is decomposable if and only if the vertices in $\mathcal{G}$ admits a perfect numbering.

To obtain a perfect directed graph, use the perfect numbering of the undirected version and direct all edges to go from higher to lower numbered vertices.

From this we see that $\mathcal{G}$ just needs to be chordal to admit a perfect numbering. To determine whether an undirected graph is chordal a *Maximum Cardinality Search* can be preformed:

Algorithm 2.5 (Maximum Cardinality Search)

```
• Set Output:="$\mathcal{G}$ is chordal"
• Set counter $i := 1$
• Set $L = \emptyset$
• For all $v \in V$, set $c(v) := 0$
• While $L \neq V$:
    – Set $U := V \setminus L$
    – Select any vertex $v \in U$ such that $c(v) \geq c(w) \forall w \in U$ and label it $i$.
    – If $\prod_{v_i} := ne(v_i) \cap L$ is not complete in $(G)$ {
          Set Output:="$\mathcal{G}$ is not chordal" break}
      else Set $c(w) := c(w) + 1 \forall w \in ne(v_i) \cap U$
    – Set $L := L \cup \{v_i\}$
    − Set $i := i + 1$
• return Output
```

Not only does this help to figure out if $\mathcal{G}$ is chordal, but it also determine a perfect numbering of $\mathcal{G}$ which is contained in L.

Theorem 2.6

[Tarjan and Yannakakis, 1984] If $\mathcal{G}$ is chordal, then the reverse numbering provided by the Maximum Cardinality Search is a perfect numbering of $\mathcal{G}$

Chapter 3

Multivariate Normal Distribution

[Seber, 1984]

Let $\boldsymbol{x} = (x_1, \ldots, x_d)^T$ be a vector of random variables. Then $\boldsymbol{x}$ is said to have a multivariate normal distribution with mean $\boldsymbol{\mu}$ ($d \times 1$ vector) and covariance matrix $\boldsymbol{\Sigma}$ ($d \times d$ matrix) if its density function is

$$f(\boldsymbol{x}) = (2\pi)^{-d/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \right), \quad \boldsymbol{x} \in \mathbb{R}^d$$

We write that $\boldsymbol{x} \sim \mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\mu_i = \mathbb{E}[X_i]$, $\Sigma_{ij} = \mathrm{Cov}[X_i, X_j]$ and $\Sigma_{ii} = \mathrm{Var}[X_i] > 0$.

Let $\boldsymbol{x} \sim \mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and dividing $\boldsymbol{x}$ into two parts $\boldsymbol{x} = (\boldsymbol{x}_1^T, \boldsymbol{x}_2^T)^T$, then we can split $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ accordingly so

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \text{ and } \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$$

The conditional properties of the normal distribution of $\boldsymbol{X}_1$ given $\boldsymbol{X}_2 = \boldsymbol{x}_2$ is $\mathcal{N}(\boldsymbol{\mu}_{1|2}, \boldsymbol{\Sigma}_{1|2})$ where

$$\begin{aligned} \boldsymbol{\mu}_{1|2} &= \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} (\boldsymbol{x}_2 - \boldsymbol{\mu}_2) \\ \boldsymbol{\Sigma}_{1|2} &= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{21} \end{aligned}$$

Proof.
[Lauritzen, 2004][C.1]
The following matrix equality is used later. Let

$$E = A - BD^{-1}C \qquad F = D^{-1}C \qquad G = BD^{-1}$$

then

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} E^{-1} & -E^{-1}G \\ -FE^{-1} & D^{-1} + FE^{-1}G \end{pmatrix} \tag{3.1}$$

Which can be seen by finding the product of the matrices

$$\begin{aligned}
&\begin{pmatrix} A & B \\ C & D \end{pmatrix} \times \begin{pmatrix} E^{-1} & -E^{-1}G \\ -FE^{-1} & D^{-1} + FE^{-1}G \end{pmatrix} \\
=&\begin{pmatrix} AE^{-1} - BFE^{-1} & -AE^{-1}G + BD^{-1} + BFE^{-1}G \\ CE^{-1} - DFE^{-1} & -CE^{-1}G + DD^{-1} + DFE^{-1}G \end{pmatrix} \\
=&\begin{pmatrix} (A - BF)E^{-1} & G - (A - BF)E^{-1}G \\ (C - DF)E^{-1} & I + (DF - C)E^{-1}G \end{pmatrix} \\
=&\begin{pmatrix} EE^{-1} & G - EE^{-1}G \\ (C - DD^{-1}C)E^{-1} & I + (DD^{-1}C - C)E^{-1}G \end{pmatrix} \\
=&\begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}
\end{aligned}$$

Let $\boldsymbol{Q} = \boldsymbol{\Sigma}^{-1}$ be partitioned like $\boldsymbol{\Sigma}$. Using that the joint and conditional density is proportional and that $\boldsymbol{x}_2$ is known, we get that

$$\begin{aligned}
f(\boldsymbol{x}_1|\boldsymbol{x}_2) &\propto f(\boldsymbol{x}) \\
&= (2\pi)^{-d/2}|\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}\left(\begin{pmatrix}\boldsymbol{x}_1 \\ \boldsymbol{x}_2\end{pmatrix} - \begin{pmatrix}\boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2\end{pmatrix}\right)^T \begin{pmatrix}\boldsymbol{Q}_{11} & \boldsymbol{Q}_{12} \\ \boldsymbol{Q}_{21} & \boldsymbol{Q}_{22}\end{pmatrix}\left(\begin{pmatrix}\boldsymbol{x}_1 \\ \boldsymbol{x}_2\end{pmatrix} - \begin{pmatrix}\boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2\end{pmatrix}\right)\right) \\
&\propto \exp\left(-\frac{1}{2}(\boldsymbol{x}_1 - \boldsymbol{\mu}_1)^T\boldsymbol{Q}_{11}(\boldsymbol{x}_1 - \boldsymbol{\mu}_1) - (\boldsymbol{x}_1 - \boldsymbol{\mu}_1)^T\boldsymbol{Q}_{12}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)\right) \\
&\propto \exp\left(-\frac{1}{2}\boldsymbol{x}_1^T\boldsymbol{Q}_{11}\boldsymbol{x}_1 + \boldsymbol{x}_1^T\left(\boldsymbol{Q}_{11}\boldsymbol{\mu}_1 - \boldsymbol{Q}_{12}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)\right)\right)
\end{aligned}$$

From this we see that $\boldsymbol{\Sigma}_{1|2} = \boldsymbol{Q}_{11}^{-1}$. Knowing this and taking the linear term for $\boldsymbol{x}_1$ we get

$$\begin{aligned}
\boldsymbol{Q}_{11}\boldsymbol{\mu}_{1|2} =& \boldsymbol{Q}_{11}\boldsymbol{\mu}_1 - \boldsymbol{Q}_{12}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2) \\
\boldsymbol{Q}_{11}\boldsymbol{\mu}_{1|2} =& \boldsymbol{Q}_{11}(\boldsymbol{\mu}_1 - \boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{12}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)) \\
\boldsymbol{\mu}_{1|2} =& \boldsymbol{\mu}_1 - \boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{12}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)
\end{aligned}$$

Using (3.1) we get the result

$$\begin{aligned}
\boldsymbol{\Sigma}_{1|2} &= \boldsymbol{Q}_{11}^{-1} \\
&= (E^{-1})^{-1} \\
&= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{\mu}_{1|2} =& \boldsymbol{\mu}_1 - \boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{12}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2) \\
=& \boldsymbol{\mu}_1 + (E^{-1})^{-1}E^{-1}G(\boldsymbol{x}_2 - \boldsymbol{\mu}_2) \\
=& \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)
\end{aligned}$$

□

3.1 Maximum Likelihood Estimation

[Seber, 1984][sec:3.2.1]

Let $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ be i.i.d. with $\boldsymbol{x}_i \sim \mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $n > d$. Now we will show that the Maximum Likelihood Estimates (MLE) of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the same as the observed sample mean and sample covariance matrix

$$\begin{aligned}
\widehat{\boldsymbol{\mu}} =& \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i \\
\widehat{\boldsymbol{\Sigma}} =& \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})^T
\end{aligned}$$

The joint distribution of $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ taken as a function of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ is the likelihood

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^{nd}|\boldsymbol{\Sigma}|^n}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})\right)$$

We rewrite the log likelihood by using the trace property $\mathrm{tr}(ABC) = \mathrm{tr}(BCA)$ and $k = -\frac{1}{2}nd\log(2\pi)$

$$\begin{aligned}
l(\boldsymbol{\mu}, \boldsymbol{\Sigma}) =& k - \frac{1}{2}n\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}) \\
=& k - \frac{1}{2}n\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{n}\mathrm{tr}\left((\boldsymbol{x}_i - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})\right) \\
=& k - \frac{1}{2}n\log|\boldsymbol{\Sigma}| - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T\right)
\end{aligned}$$

Now we want to find the $\boldsymbol{\mu}$ that maximizes $l(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Looking at the remaining part $\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T$, which is none negative, and knowing $\boldsymbol{\Sigma}^{-1} > \mathbf{0}$, since $\boldsymbol{\Sigma} > \mathbf{0}$, we want to minimize this. We write

$$\begin{aligned}
&\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T \\
=&\sum_{i=1}^{n}(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}} + \widehat{\boldsymbol{\mu}} - \boldsymbol{\mu})(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}} + \widehat{\boldsymbol{\mu}} - \boldsymbol{\mu})^T \\
=&\sum_{i=1}^{n}(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})^T + \sum_{i=1}^{n}(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu})(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu})^T + 2(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu})\underbrace{\sum_{i=1}^{n}(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})}_{\mathbf{0}} \\
=&\sum_{i=1}^{n}(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})^T + n(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu})(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu})^T
\end{aligned}$$

Looking at this along with $\boldsymbol{\Sigma}^{-1}$ we see that

$$\begin{aligned}
\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})^T\right) &= (\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})^T \geq 0 \\
\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu})(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu})^T\right) &= (\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu})^T \geq 0
\end{aligned}$$

which is minimized, along with $l(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ being maximized, when $\boldsymbol{\mu} := \widehat{\boldsymbol{\mu}}$. Next we want to find what $\boldsymbol{\Sigma} > \mathbf{0}$ will maximize

$$\begin{aligned} l(\widehat{\boldsymbol{\mu}}, \boldsymbol{\Sigma}) &= k - \frac{1}{2} n \log |\boldsymbol{\Sigma}| - \frac{1}{2} \mathrm{tr} \left(\boldsymbol{\Sigma}^{-1} \sum_{i=1}^{n} (\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})^T \right) \\ &= k - \frac{1}{2} n \left(\log |\boldsymbol{\Sigma}| + \mathrm{tr} \left(\boldsymbol{\Sigma}^{-1} \frac{\sum_{i=1}^{n} (\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})^T}{n} \right) \right) \end{aligned}$$

From [Seber, 1984][A.7.1] we get that this is uniquely maximized when

$$\boldsymbol{\Sigma} := \frac{\sum_{i=1}^{n} (\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})^T}{n} = \widehat{\boldsymbol{\Sigma}} \tag{3.2}$$

This method is called ordinary least squares. When $\boldsymbol{\mu}$ is unknown and $\widehat{\boldsymbol{\mu}}$ is used instead we get a biased estimate. In which case we use $\frac{n}{n-1}\widehat{\boldsymbol{\Sigma}}$ to obtain an unbiased estimate.

3.1.1 Conditional MLE

Let $\boldsymbol{x} = (\boldsymbol{x}_A^T, \boldsymbol{x}_B^T)^T \sim \mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then the MLE are

$$\widehat{\boldsymbol{\mu}} = \begin{pmatrix} \widehat{\boldsymbol{\mu}}_A \\ \widehat{\boldsymbol{\mu}}_B \end{pmatrix} \quad \text{and} \quad \widehat{\boldsymbol{\Sigma}} = \begin{pmatrix} \widehat{\boldsymbol{\Sigma}}_{AA} & \widehat{\boldsymbol{\Sigma}}_{AB} \\ \widehat{\boldsymbol{\Sigma}}_{BA} & \widehat{\boldsymbol{\Sigma}}_{BB} \end{pmatrix}$$

This is used when looking at $\boldsymbol{X}_A | \boldsymbol{X}_B \sim \mathcal{N}(\boldsymbol{\mu}_{A|B}, \boldsymbol{\Sigma}_{A|B})$ and determining the MLE of $\boldsymbol{\mu}_{A|B}$ and $\boldsymbol{\Sigma}_{A|B}$, as this can be done by estimating each part, which were found by the MLE from a sample and properties of the multivariate normal.

$$\begin{aligned} \widehat{\boldsymbol{\mu}}_{A|B} &= \widehat{\boldsymbol{\mu}}_A + \widehat{\boldsymbol{\Sigma}}_{AB} \widehat{\boldsymbol{\Sigma}}_{BB}^{-1} (\boldsymbol{x}_B - \widehat{\boldsymbol{\mu}}_B) \\ \widehat{\boldsymbol{\Sigma}}_{A|B} &= \widehat{\boldsymbol{\Sigma}}_{AA} - \widehat{\boldsymbol{\Sigma}}_{AB} \widehat{\boldsymbol{\Sigma}}_{BB}^{-1} \widehat{\boldsymbol{\Sigma}}_{BA} \end{aligned}$$

Here $\widehat{\boldsymbol{\Sigma}}_{A|B}$ can be rewritten to obtain a similar result as in the unconditional case. Let the estimated regression coefficients be $\widehat{\boldsymbol{\Sigma}}_{AB} \widehat{\boldsymbol{\Sigma}}_{BB}^{-1} = \widehat{\boldsymbol{\beta}}^T$

$$\begin{aligned}
\widehat{\boldsymbol{\Sigma}}_{A|B} &= \widehat{\boldsymbol{\Sigma}}_{AA} - \widehat{\boldsymbol{\beta}}^T \widehat{\boldsymbol{\Sigma}}_{BA} \\
&= \widehat{\boldsymbol{\Sigma}}_{AA} + \widehat{\boldsymbol{\beta}}^T \widehat{\boldsymbol{\Sigma}}_{BB} \widehat{\boldsymbol{\Sigma}}_{BB}^{-1} \widehat{\boldsymbol{\Sigma}}_{BA} - 2\widehat{\boldsymbol{\beta}}^T \widehat{\boldsymbol{\Sigma}}_{BA} \\
&= \widehat{\boldsymbol{\Sigma}}_{AA} + \widehat{\boldsymbol{\beta}}^T \widehat{\boldsymbol{\Sigma}}_{BB} \widehat{\boldsymbol{\beta}} - 2\widehat{\boldsymbol{\beta}}^T \widehat{\boldsymbol{\Sigma}}_{BA} \\
&= \frac{1}{n} \sum_{i=1}^{n} \left[(\boldsymbol{x}_{Ai} - \widehat{\boldsymbol{\mu}}_A)(\boldsymbol{x}_{Ai} - \widehat{\boldsymbol{\mu}}_A)^T + \widehat{\boldsymbol{\beta}}^T (\boldsymbol{x}_{Bi} - \widehat{\boldsymbol{\mu}}_B)(\boldsymbol{x}_{Bi} - \widehat{\boldsymbol{\mu}}_B)^T \widehat{\boldsymbol{\beta}} - 2\widehat{\boldsymbol{\beta}}^T (\boldsymbol{x}_{Bi} - \widehat{\boldsymbol{\mu}}_B)(\boldsymbol{x}_{Ai} - \widehat{\boldsymbol{\mu}}_A)^T \right] \\
&= \frac{1}{n} \sum_{i=1}^{n} \left[(\boldsymbol{x}_{Ai} - \widehat{\boldsymbol{\mu}}_A) - \widehat{\boldsymbol{\beta}}^T (\boldsymbol{x}_{Bi} - \widehat{\boldsymbol{\mu}}_B) \right] \left[(\boldsymbol{x}_{Ai} - \widehat{\boldsymbol{\mu}}_A) - \widehat{\boldsymbol{\beta}}^T (\boldsymbol{x}_{Bi} - \widehat{\boldsymbol{\mu}}_B) \right]^T \\
&= \frac{1}{n} \sum_{i=1}^{n} \left[\boldsymbol{x}_{Ai} - \widehat{\boldsymbol{\mu}}_{A|B} \right] \left[\boldsymbol{x}_{Ai} - \widehat{\boldsymbol{\mu}}_{A|B} \right]^T
\end{aligned} \tag{3.3}$$

Chapter 4

Obtaining conditional distributions

When needing to find the exact conditional distribution of a multivariate normal distribution different methods can be used. One can try to calculate it directly using

$$\begin{aligned}\boldsymbol{\mu}_{1|2} &= \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)\\ \boldsymbol{\Sigma}_{1|2} &= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\end{aligned}$$

This method suffers when the dimension of $\boldsymbol{x}_2$ becomes larger, as computing the inverse of $\boldsymbol{\Sigma}_{22}$ has polynomial growth. Using a simple naive method it takes $O(n^3)$ time to compute the inverse of a $n \times n$ matrix. Other methods have been developed and today it can be done in $O(n^{2.3728639})$ [Gall, 2014].

Another way is to use the Junction Tree Algorithm (JTA) [Paskin, 2003], which requires the variables to be represented by a junction tree that they can if and only if they admit a perfect numbering[Cowell et al., 1999][Ch.4].

We will now present a new method that takes the same requirements as the JTA. After that a general case will be presented, that covers both the new method and the JTA.

4.1 A new method

In this section, we will show a method to obtain conditional distributions of unobserved variables given observed variables. The dependencies between the variables must be such that they may be represented by a DAG with a perfect numbering. This method is divided into three main steps; Initialization, forward propagation and backward propagation.

4.1.1 Initialization

Given the variables $X_1, \dots, X_n$ let them be ordered such that they form a perfect sequence when we assume a decomposable model. This can be obtained by running the maximum cardinality search algorithm. As such we may represent the model by a DAG.

Thus the joint density is given by

$$f(\boldsymbol{x}) = \prod_{i=1}^{n} f(x_i | \boldsymbol{x}_{pa(i)})$$

Let $\mathrm{Cov}(X_i, X_{pa(i)}) = \boldsymbol{\Sigma}_{\{i\},pa(i)} := \boldsymbol{\Sigma}_{i,pa(i)}$, then $f(x_i|\boldsymbol{x}_{pa(i)})$ have mean $\bar{\mu}_i$ and variance τ_i^2:

$$\begin{aligned}\bar{\mu}_i &= \mu_i + \boldsymbol{\Sigma}_{i,pa(i)} \boldsymbol{\Sigma}^{-1}_{pa(i),pa(i)} (\boldsymbol{x}_{pa(i)} - \boldsymbol{\mu}_{pa(i)}) \\ &= \underbrace{\mu_i - \boldsymbol{\Sigma}_{i,pa(i)} \boldsymbol{\Sigma}^{-1}_{pa(i),pa(i)} \boldsymbol{\mu}_{pa(i)}}_{\alpha_i} + \underbrace{\boldsymbol{\Sigma}_{i,pa(i)} \boldsymbol{\Sigma}^{-1}_{pa(i),pa(i)}}_{\boldsymbol{\beta}_i^T} \boldsymbol{x}_{pa(i)}\end{aligned}$$

$$\tau_i^2 = \mathrm{Var}(X_i) - \boldsymbol{\Sigma}_{i,pa(i)} \boldsymbol{\Sigma}^{-1}_{pa(i),pa(i)} \boldsymbol{\Sigma}_{pa(i),i}$$

Then it is seen that X_i is a linear regression on $X_{pa(i)}$:

$$\log(f(x_i|\boldsymbol{x}_{pa(i)})) = -\frac{1}{2}\log(2\pi\tau_i^2) - \frac{1}{2\tau_i^2}(x_i - \alpha_i - \boldsymbol{\beta}_i^T \boldsymbol{x}_{pa(i)})^2$$

The regression coefficients vector $\boldsymbol{\beta}_i = (\beta_{i,j})_{j \in pa(i)}$ is indexed by $pa(i)$. Note that by standardising all variables, i.e. $\mu_i = 0$ $\boldsymbol{\Sigma}_{i,i} = 1$ $\forall$ $i \in \{1, \dots, n\}$, then $\alpha_i = 0$ $\forall$ $i \in \{1, \dots, n\}$.

The Maximum Likelihood Estimates(MLE) of $\alpha_i, \boldsymbol{\beta}_i, \tau_i$ are obtained by ordinary least squares as was shown earlier in section 3.1.

Note that the estimate of τ_i^2 is obtained by the sum of squared errors divided by the sample size if $\bar{\mu}_i$ is known or sample size minus one if $\bar{\mu}_i$ is estimated.

That this estimate is the same as above can be seen by rewriting τ_i^2, as in (3.3).

Expanding $\log(f(x_i|\boldsymbol{x}_{pa(i)})$ gives:

$$\begin{aligned}
&\log(f(x_i|\boldsymbol{x}_{pa(i)}) \\
=& -\frac{1}{2}\log(2\pi\tau_i^2) - \frac{1}{2\tau_i^2}(x_i - \alpha_i - \boldsymbol{\beta}_i^T\boldsymbol{x}_{pa(i)})^2 \\
=& -\frac{1}{2}\log(2\pi\tau_i^2) - \frac{1}{2\tau_i^2}(\alpha_i^2 + x_i^2 - 2x_i\alpha_i - 2x_i\boldsymbol{\beta}_i^T\boldsymbol{x}_{pa(i)} + 2\alpha_i\boldsymbol{\beta}_i^T\boldsymbol{x}_{pa(i)} + \boldsymbol{\beta}_i^T\boldsymbol{x}_{pa(i)}\boldsymbol{x}_{pa(i)}^T\boldsymbol{\beta}_i) \\
=& \underbrace{-\frac{1}{2}(\log(2\pi\tau_i^2) + \frac{\alpha_i^2}{\tau_i^2})}_{\kappa_i} - \frac{1}{2\tau_i^2}(x_i^2 - 2x_i\alpha_i - 2x_i\boldsymbol{\beta}_i^T\boldsymbol{x}_{pa(i)} + 2\alpha_i\boldsymbol{\beta}_i^T\boldsymbol{x}_{pa(i)} + \boldsymbol{\beta}_i^T\boldsymbol{x}_{pa(i)}\boldsymbol{x}_{pa(i)}^T\boldsymbol{\beta}_i)
\end{aligned}$$

Therefore the logarithm of the normalizing constant of $f(\boldsymbol{x})$ is:

$$\kappa = \sum_{i=1}^{n}\kappa_i = \sum_{i=1}^{n} -\frac{1}{2}(\log(2\pi\tau_i^2) + \frac{\alpha_i^2}{\tau_i^2})$$

The remaining contribution of $\log(f(x_i|\boldsymbol{x}_{pa(i)})$ to $\log(f(\boldsymbol{x}))$ is

$$g(x_i|\boldsymbol{x}_{pa(i)})) = \log(f(x_i|\boldsymbol{x}_{pa(i)})) - \kappa_i$$

We shall use the fact that the joint distribution, without the normalizing constant κ, may be written as product of potentials $p(x_i, \boldsymbol{x}_{pa(i)})$, which can be represented as

$$\log(p(x_i, \boldsymbol{x}_{pa(i)})) = -\frac{1}{2}\lambda_i x_i^2 + \delta_i x_i - x_i\boldsymbol{x}_{pa(i)}^T\boldsymbol{\gamma}_i$$

where $\boldsymbol{\gamma}_i = (\gamma_{i,j})_{j\in pa(i)}$ is indexed by $pa(i)$.

If we initialize $\lambda_i, \delta_i, \boldsymbol{\gamma}_i$ by zeroes, then we obtain these parameters by looping $i = 1\ldots, n$ and collecting terms from $g(x_i|\boldsymbol{x}_{pa(i)})$:

$$\begin{aligned}
&g(x_i|\boldsymbol{x}_{pa(i)}))\\
=&-\frac{1}{2\tau_i^2}(x_i^2 - 2x_i\alpha_i - 2x_i\boldsymbol{\beta}_i^T\boldsymbol{x}_{pa(i)} + 2\alpha_i\boldsymbol{\beta}_i^T\boldsymbol{x}_{pa(i)} + \boldsymbol{\beta}_i^T\boldsymbol{x}_{pa(i)}\boldsymbol{x}_{pa(i)}^T\boldsymbol{\beta}_i)\\
=&-\frac{1}{2}\underbrace{\frac{1}{\tau_i^2}}_{+\lambda_i}x_i^2 + \underbrace{\frac{\alpha_i}{\tau_i^2}}_{+\delta_i}x_i - x_i\boldsymbol{x}_{pa(i)}^T\underbrace{\frac{-\boldsymbol{\beta}_i}{\tau_i^2}}_{+\gamma_i} - \frac{\alpha_i\boldsymbol{\beta}_i^T}{\tau_i^2}\boldsymbol{x}_{pa(i)} - \frac{\boldsymbol{\beta}_i^T\boldsymbol{x}_{pa(i)}\boldsymbol{x}_{pa(i)}^T\boldsymbol{\beta}_i}{2\tau_i^2}
\end{aligned}$$

The parameters of $\log(p(x_i, \boldsymbol{x}_{pa(i)}))$ is updated:

$$\lambda_i := \lambda_i + \frac{1}{\tau_i^2}\ ,\ \delta_i := \delta_i + \frac{\alpha_i}{\tau_i^2} \text{ and } \boldsymbol{\gamma}_i := \boldsymbol{\gamma}_i - \frac{\boldsymbol{\beta}_i}{\tau_i^2}$$

That leaves:

$$\begin{aligned}
&-\frac{\alpha_i\boldsymbol{\beta}_i^T}{\tau_i^2}\boldsymbol{x}_{pa(i)} - \frac{\boldsymbol{\beta}_i^T\boldsymbol{x}_{pa(i)}\boldsymbol{x}_{pa(i)}^T\boldsymbol{\beta}_i}{2\tau_i^2}\\
=&\sum_{j\in pa(i)} -\frac{\alpha_i\beta_{i,j}}{\tau_i^2}x_j - \sum_{j,h\in pa(i)}\frac{\beta_{i,j}\beta_{i,h}x_jx_h}{2\tau_i^2}\\
=&\sum_{j\in pa(i)}\underbrace{-\frac{\alpha_i\beta_{i,j}}{\tau_i^2}}_{+\delta_j}x_j - \frac{1}{2}\sum_{j\in pa(i)}\underbrace{\frac{\beta_{i,j}^2}{\tau_i^2}}_{+\lambda_j}x_j^2 - \sum_{j\in pa(i)}\ \sum_{h\in pa(i)|h>j} x_jx_h\underbrace{\frac{\beta_{i,j}\beta_{i,h}}{\tau_i^2}}_{+\gamma_{j,h}}
\end{aligned}$$

which is used to update parameters for the parents:

$$\text{For } j\in pa(i) : \lambda_j := \lambda_j + \frac{\beta_{i,j}^2}{\tau_i^2} \text{ and } \delta_j := \delta_j - \frac{\alpha_i\beta_{i,j}}{\tau_i^2}$$

$$\text{For } j\in pa(i) : \text{for } h\in pa(i) :\ \text{ if } h > j : \gamma_{j,h} := \gamma_{j,h} + \frac{\beta_{i,j}\beta_{i,h}}{\tau_i^2}$$

Note that when $j, h \in pa(i)$ and $h > j$ then $h \in pa(j)$, since $pa(i)$ is a complete set in the graph.

4.1.2 Forward propagation

After the initialization, we then have the representation

$$\log(f(\boldsymbol{x})) = \kappa + \sum_{i=1}^{n} \log(p(x_i, \boldsymbol{x}_{pa(i)}))$$

Let e be a subset of $\{1, \ldots, n\}$ corresponding to the set of observed variables and let $\bar{e}$ correspond to the set of unobserved variables. We then intend to calculate $f(\boldsymbol{x}_e) = \int_{\mathbb{R}^{dim(\bar{e})}} f(\boldsymbol{x}_e, \boldsymbol{x}_{\bar{e}})\, dx_{\bar{e}}$, i.e. the likelihood of the actual observation. We do this by a forward propagation, where we loop $i = 1, \ldots, n$ and perform one of the following.

Option 1: If $i \in e$
We use the fact that x_i is now just a number, so from the i'th potential we get

$$\begin{aligned}\log(p(x_i, \boldsymbol{x}_{pa(i)})) &= -\frac{1}{2}\lambda_i x_i^2 + \delta_i x_i - x_i \boldsymbol{x}_{pa(i)}^T \boldsymbol{\gamma}_i \\ &= \underbrace{-\frac{1}{2}\lambda_i x_i^2 + \delta_i x_i}_{+\kappa} - \sum_{j \in pa(i)} \underbrace{x_i \gamma_{i,j}}_{+\delta_j} x_j\end{aligned}$$

which means that the log-normalizing constant is updated as

$$\kappa := \kappa - \frac{1}{2}\lambda_i x_i^2 + \delta_i x_i$$

and the parents receives an parameter update:

$$\text{For } j \in pa(i): \;\; \delta_j := \delta_j - x_i \gamma_{i,j}$$

Option 2: If $i \in \bar{e}$

First rewriting the log-potential

$$\begin{aligned}
&\log(p(x_i, \boldsymbol{x}_{pa(i)})) \\
=& -\frac{1}{2}\lambda_i x_i^2 + \delta_i x_i - x_i \boldsymbol{x}_{pa(i)}^T \boldsymbol{\gamma}_i \\
=& -\frac{1}{2}\lambda_i \left(x_i^2 - \frac{2\delta_i x_i}{\lambda_i} + 2\frac{x_i \boldsymbol{x}_{pa(i)}^T \boldsymbol{\gamma}_i}{\lambda_i} + \left(-\frac{\delta_i}{\lambda_i}\right)^2 + \left(\frac{\boldsymbol{x}_{pa(i)}^T \boldsymbol{\gamma}_i}{\lambda_i}\right)^2 - 2\frac{\delta_i \boldsymbol{x}_{pa(i)}^T \boldsymbol{\gamma}_i}{\lambda_i^2} \right) \\
& -\frac{1}{2}\lambda_i \left(-\left(-\frac{\delta_i}{\lambda_i}\right)^2 - \left(\frac{\boldsymbol{x}_{pa(i)}^T \boldsymbol{\gamma}_i}{\lambda_i}\right)^2 + 2\frac{\delta_i \boldsymbol{x}_{pa(i)}^T \boldsymbol{\gamma}_i}{\lambda_i^2} \right) \\
=& -\frac{1}{2\lambda_i^{-1}} \left(x_i - \left(\frac{\delta_i}{\lambda_i} - \frac{\boldsymbol{x}_{pa(i)}^T \boldsymbol{\gamma}_i}{\lambda_i} \right) \right)^2 + \frac{1}{2\lambda_i} \left(\delta_i^2 + \left(\boldsymbol{x}_{pa(i)}^T \boldsymbol{\gamma}_i\right)^2 - 2\delta_i \boldsymbol{x}_{pa(i)}^T \boldsymbol{\gamma}_i \right) \\
=& -\frac{1}{2}\log(\frac{2\pi}{\lambda_i}) - \frac{1}{2\lambda_i^{-1}} \left(x_i - \left(\frac{\delta_i}{\lambda_i} - \frac{\boldsymbol{x}_{pa(i)}^T \boldsymbol{\gamma}_i}{\lambda_i} \right) \right)^2 \\
& + \frac{1}{2}\log(\frac{2\pi}{\lambda_i}) + \frac{1}{2\lambda_i} \left(\delta_i^2 + \left(\boldsymbol{x}_{pa(i)}^T \boldsymbol{\gamma}_i\right)^2 - 2\delta_i \boldsymbol{x}_{pa(i)}^T \boldsymbol{\gamma}_i \right)
\end{aligned}$$

Then integrate out x_i in the potential

$$\begin{aligned}
\int_{\mathbb{R}} p(x_i, \boldsymbol{x}_{pa(i)})\, dx_i =& \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\lambda_i^{-1}}} \exp\left(-\frac{1}{2\lambda_i^{-1}} \left(x_i - \left(\frac{\delta_i}{\lambda_i} - \frac{\boldsymbol{x}_{pa(i)}^T \boldsymbol{\gamma}_i}{\lambda_i} \right) \right)^2 \right) dx_i \\
& \times \exp\left(\frac{1}{2}\log(\frac{2\pi}{\lambda_i}) + \frac{1}{2\lambda_i} \left(\delta_i^2 + \left(\boldsymbol{x}_{pa(i)}^T \boldsymbol{\gamma}_i\right)^2 - 2\delta_i \boldsymbol{x}_{pa(i)}^T \boldsymbol{\gamma}_i \right) \right) \\
=& \exp\left(\frac{1}{2}\log(\frac{2\pi}{\lambda_i}) + \frac{1}{2\lambda_i} \left(\delta_i^2 + \left(\boldsymbol{x}_{pa(i)}^T \boldsymbol{\gamma}_i\right)^2 - 2\delta_i \boldsymbol{x}_{pa(i)}^T \boldsymbol{\gamma}_i \right) \right)
\end{aligned}$$

which yields the log-message:

$$\begin{aligned}\log \int_{\mathbb{R}} p(x_i, \boldsymbol{x}_{pa(i)})\, dx_i =& \frac{1}{2}\log(\frac{2\pi}{\lambda_i}) + \frac{1}{2\lambda_i}(\delta_i^2 - 2\delta_i \boldsymbol{x}_{pa(i)}^T \boldsymbol{\gamma}_i + \boldsymbol{\gamma}_i^T \boldsymbol{x}_{pa(i)} \boldsymbol{x}_{pa(i)}^T \boldsymbol{\gamma}_i) \\ =& \frac{1}{2}(\log(\frac{2\pi}{\lambda_i}) + \frac{1}{\lambda_i}\delta_i^2) - \frac{\delta_i \boldsymbol{x}_{pa(i)}^T \boldsymbol{\gamma}_i}{\lambda_i} + \frac{\boldsymbol{\gamma}_i^T \boldsymbol{x}_{pa(i)} \boldsymbol{x}_{pa(i)}^T \boldsymbol{\gamma}_i}{2\lambda_i} \\ =& \underbrace{\frac{1}{2}(\log(\frac{2\pi}{\lambda_i}) + \frac{1}{\lambda_i}\delta_i^2)}_{+\kappa} + \sum_{j \in pa(i)} \underbrace{-\frac{\delta_i \gamma_{i,j}}{\lambda_i}}_{+\delta_j} x_j \\ &+ \sum_{j \in pa(i)} -\frac{1}{2} \underbrace{\frac{-\gamma_{i,j}^2}{\lambda_i}}_{+\lambda_j} x_j^2 + \sum_{j \in pa(i)} \sum_{h \in pa(i) | h > j} -x_j x_h \underbrace{\frac{-\gamma_{i,j}\gamma_{i,h}}{\lambda_i}}_{+\gamma_{j,h}}\end{aligned}$$

Hence we update the normalizing constant and the parent parameters accordingly:

$$\kappa := \kappa + \frac{1}{2}(\log(\frac{2\pi}{\lambda_i}) + \frac{1}{\lambda_i}\delta_i^2)$$

$$\text{For } j \in pa(i) : \lambda_j := \lambda_j - \frac{\gamma_{i,j}^2}{\lambda_i} \text{ and } \delta_j := \delta_j - \frac{\delta_i \gamma_{i,j}}{\lambda_i}$$

$$\text{For } j \in pa(i) : \text{ for } h \in pa(i) : \text{ if } h > j : \gamma_{j,h} := \gamma_{j,h} - \frac{\gamma_{i,j}\gamma_{i,h}}{\lambda_i}$$

After this forward propagation we have that $\log(f(\boldsymbol{x}_e)) = \kappa$ and $p(x_i, \boldsymbol{x}_{pa(i)})$ is proportional to $f(x_i | \boldsymbol{x}_{pa(i) \cup e})$ for $i \in \bar{e}$. The last part may not be obvious, but remark that for $i \in \bar{e}$:

- Let b index the observed variables preceding X_i and let d index the observed variables descending X_i in the DAG, i.e. $b \cup d = e$
- The potential $p(x_i, \boldsymbol{x}_{pa(i)})$ of X_i after forward propagation is proportional to the distribution of X_i given $X_{b \cup pa(i)}$. As X_i given $X_{b \cup pa(i)}$ is independent of X_d due to the property of a DAG, we can conclude that $p(x_i, \boldsymbol{x}_{pa(i)})$ is also proportional to the distribution of X_i given $X_{pa(i) \cup e}$.

4.1.3 Backward propagation

Next, we want to determine $f(x_i|\boldsymbol{x}_e)$ when $i \in \bar{e}$. First we rewrite the i'th potential

$$\begin{aligned}\log(p(x_i, \boldsymbol{x}_{pa(i)})) &= -\frac{1}{2}\lambda_i x_i^2 + \delta_i x_i - x_i \boldsymbol{x}_{pa(i)}^T \boldsymbol{\gamma}_i \\ &= -\frac{1}{2}\lambda_i x_i^2 + \delta_i x_i - x_i \boldsymbol{x}_{pa(i)}^T \boldsymbol{\gamma}_i - \frac{\delta_i^2}{2\lambda_i} - \frac{(\boldsymbol{x}_{pa(i)}^T \boldsymbol{\gamma}_i)^2}{2\lambda_i} + \frac{\delta_i \boldsymbol{x}_{pa(i)}^T \boldsymbol{\gamma}_i}{\lambda_i} \\ &\quad + \frac{\delta_i^2}{2\lambda_i} + \frac{(\boldsymbol{x}_{pa(i)}^T \boldsymbol{\gamma}_i)^2}{2\lambda_i} - \frac{\delta_i \boldsymbol{x}_{pa(i)}^T \boldsymbol{\gamma}_i}{\lambda_i} \\ &= -\frac{\lambda_i}{2}(x_i - \frac{1}{\lambda_i}(\delta_i - \boldsymbol{x}_{pa(i)}^T \boldsymbol{\gamma}_i))^2 + \frac{1}{2\lambda_i}(\delta_i - \boldsymbol{x}_{pa(i)}^T \boldsymbol{\gamma}_i)^2\end{aligned}$$

This means, that after the forward propagation we may from the i'th potential parameters calculate:

- $M_i := \mathbb{E}(X_i|X_{pa(i)\cup e} = \boldsymbol{x}_{pa(i)\cup e}) = \dfrac{1}{\lambda_i}(\delta_i - \boldsymbol{x}_{pa(i)}^T \boldsymbol{\gamma}_i)$
- $V_i := \mathrm{Var}(X_i|X_{pa(i)\cup e}) = \dfrac{1}{\lambda_i}$

Next we do a backward propagation, where we loop $i = n, \dots, 1$ and exploit that for $i \in e$ then

- $\tilde{\mu}_i := \mathbb{E}(X_i|X_e = \boldsymbol{x}_e) = x_i$
- $\tilde{\sigma}_i^2 := \mathrm{Var}(X_i|X_e) = 0$
- $\tilde{\boldsymbol{\varsigma}}_i := \mathrm{Cov}(X_i, X_{pa(i)}|X_e) = 0$

This is used to calculate $\tilde{\mu}_i$, $\tilde{\sigma}_i^2$ and $\tilde{\boldsymbol{\varsigma}}_i$ when $i \in \bar{e}$.

For $\tilde{\mu}_i$ we first use the fact that $\mathbb{E}(X|Z) = \mathbb{E}(\mathbb{E}(X|Y)|Z)$ and then that $\mathbb{E}(\alpha X + \beta|Z) = \alpha\mathbb{E}(X|Z) + \beta$:

$$\begin{aligned}\tilde{\mu}_i :=&\mathbb{E}(X_i|X_e = \boldsymbol{x}_e)\\ =&\mathbb{E}(M_i|X_e = \boldsymbol{x}_e)\\ =&\frac{1}{\lambda_i}(\delta_i - \boldsymbol{\gamma}_i^T\mathbb{E}(X_{pa(i)}|X_e = \boldsymbol{x}_e))\\ =&\frac{1}{\lambda_i}(\delta_i - \boldsymbol{\gamma}_i^T\tilde{\mu}_{pa(i)})\end{aligned}$$

For $\tilde{\sigma}_i^2$ we use $\text{Var}(X|Z) = \mathbb{E}(\text{Var}(X|Y)|Z) + \text{Var}(\mathbb{E}(X|Y)|Z)$ for the first equality then $\mathbb{E}(\alpha|Z) = \alpha$ and $\text{Var}(\alpha X + \beta|Z) = \alpha^2\text{Var}(X|Z)$ for the third equality:

$$\begin{aligned}\tilde{\sigma}_i^2 :=&\text{Var}(X_i|X_e)\\ =&\mathbb{E}(V_i|X_e = \boldsymbol{x}_e) + \text{Var}(M_i|X_e)\\ =&\mathbb{E}(\frac{1}{\lambda_i}|X_e = \boldsymbol{x}_e) + \text{Var}(\frac{\delta_i}{\lambda_i} - \frac{X_{pa(i)}^T\boldsymbol{\gamma}_i}{\lambda_i})|X_e)\\ =&\frac{1}{\lambda_i} + \frac{1}{\lambda_i^2}\boldsymbol{\gamma}_i^T\text{Var}(X_{pa(i)}|X_e)\boldsymbol{\gamma}_i\end{aligned}$$

For $\tilde{\boldsymbol{\varsigma}}_i$ we again use that $\mathbb{E}(X|Z) = \mathbb{E}(\mathbb{E}(X|Y)|Z)$ along with

- $\text{Cov}(X, Y|Z) = \mathbb{E}(XY|Z) - \mathbb{E}(X|Z)\mathbb{E}(Y|Z)$
- $\mathbb{E}(XY|X) = X\mathbb{E}(Y|X)$
- $\text{Cov}(\alpha X + \beta, Y|Z) = \alpha\text{Cov}(X, Y|Z)$

$$\begin{aligned}\tilde{\boldsymbol{\varsigma}}_i :=&\text{Cov}(X_i, X_{pa(i)}|X_e)\\ =&\mathbb{E}(X_{pa(i)}X_i|X_e = \boldsymbol{x}_e) - \mathbb{E}(X_{pa(i)}|X_e = \boldsymbol{x}_e)\mathbb{E}(X_i|X_e = \boldsymbol{x}_e)\\ =&\mathbb{E}(\mathbb{E}(X_{pa(i)}X_i|X_{pa(i)\cup e} = \boldsymbol{x}_{pa(i)\cup e})|X_e = \boldsymbol{x}_e) - \mathbb{E}(X_{pa(i)}|X_e = \boldsymbol{x}_e)\mathbb{E}(M_i|X_e = \boldsymbol{x}_e)\\ =&\mathbb{E}(X_{pa(i)}M_i|X_e = \boldsymbol{x}_e) - \mathbb{E}(X_{pa(i)}|X_e = \boldsymbol{x}_e)\mathbb{E}(M_i|X_e = \boldsymbol{x}_e)\\ =&\text{Cov}(M_i, X_{pa(i)}|X_e)\\ =&\text{Cov}(\frac{\delta_i}{\lambda_i} - \frac{X_{pa(i)}^T\boldsymbol{\gamma}_i}{\lambda_i}, X_{pa(i)}|X_e)\\ =&-\frac{1}{\lambda_i}\boldsymbol{\gamma}_i^T\text{Var}(X_{pa(i)}|X_e)\end{aligned}$$

where $\text{Var}(X_{pa(i)}|X_e)$ is determined from $(\tilde{\sigma}_j^2, \tilde{\boldsymbol{\varsigma}}_j)$, $j \in pa(i)$.
After the forward and backward propagation we have that

$$f(\boldsymbol{x}_e) = \exp(\kappa)$$

$$(X_i|X_e = \boldsymbol{x}_e) \sim N(\tilde{\mu}_i, \tilde{\sigma}_i^2) \text{ for } i = 1, \ldots, n$$

For all index variables pairs $i < j$ that are connected in the graph, then the propagation algorithm also has determined

$$\text{Cov}(X_i, X_j|X_e) = \tilde{\varsigma}_{i,j}$$

Thus, for any complete subset a of the graph, we can determine the distribution of $X_a|X_e$.

4.2 General case

In the previous section, we showed how to do the estimation when updating one vertex at the time. We will now show the case where we update complete clusters of vertices.

4.2.1 Initialization

Given the variables $X_1, \ldots, X_n$ let them be ordered such that they form a perfect sequence and represent them by a DAG. From that we obtain an ordering of the cliques $C_1, \ldots, C_q$ such that they have the running intersection property

$$S_i = C_i \cap (C_{i+1} \cup \ldots \cup C_q) \subset C_j \text{ for } j \in \{i+1, \ldots, q\}$$

Let $C_i \setminus S_i = C_i^*$ and $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m$ be vectors of variables such that

$$\boldsymbol{Y}_1 = (X_1, \ldots, X_{k_1})^T \text{ , } \boldsymbol{Y}_2 = (X_{k_1+1}, \ldots, X_{k_2})^T \text{ , } \ldots, \text{ } \boldsymbol{Y}_m = (X_{k_{m-1}+1}, \ldots, X_n)^T$$

where $\boldsymbol{Y}_i \subset C_j^*$, i.e. each C_j^* is split into disjoint subsets. For convenience let $\boldsymbol{Y}_i = (X_{k_{i-1}+1}, \ldots, X_{k_i})^T = (Y_{k_{i-1}+1}, \ldots, Y_{k_i})^T$

The joint density is now given by

$$f(\boldsymbol{x}) = f(\boldsymbol{y}) = \prod_{i=1}^{m} f(\boldsymbol{y}_i|\boldsymbol{y}_{pa(i)})$$

where $\boldsymbol{y}_{pa(i)} = \boldsymbol{x}_{pa(k_i)}$.

Let $\mathrm{Cov}(\boldsymbol{Y}_i, \boldsymbol{Y}_{pa(i)}) = \boldsymbol{\Sigma}_{\{i\},pa(i)} := \boldsymbol{\Sigma}_{i,pa(i)}$, then $f(\boldsymbol{y}_i|\boldsymbol{y}_{pa(i)})$ have mean $\bar{\boldsymbol{\mu}}_i$ and co-variance matrix $\bar{\boldsymbol{\Sigma}}_{i,i}$:

$$\begin{aligned}\bar{\boldsymbol{\mu}}_i &= \boldsymbol{\mu}_i + \boldsymbol{\Sigma}_{i,pa(i)}\boldsymbol{\Sigma}^{-1}_{pa(i),pa(i)}(\boldsymbol{y}_{pa(i)} - \boldsymbol{\mu}_{pa(i)}) \\ &= \underbrace{\boldsymbol{\mu}_i - \boldsymbol{\Sigma}_{i,pa(i)}\boldsymbol{\Sigma}^{-1}_{pa(i),pa(i)}\boldsymbol{\mu}_{pa(i)}}_{\boldsymbol{\alpha}_i} + \underbrace{\boldsymbol{\Sigma}_{i,pa(i)}\boldsymbol{\Sigma}^{-1}_{pa(i),pa(i)}}_{\boldsymbol{\beta}_i^T}\boldsymbol{y}_{pa(i)}\end{aligned}$$

$$\bar{\boldsymbol{\Sigma}}_{i,i} = \boldsymbol{\Sigma}_{i,i} - \boldsymbol{\Sigma}_{i,pa(i)}\boldsymbol{\Sigma}^{-1}_{pa(i),pa(i)}\boldsymbol{\Sigma}_{pa(i),i}$$

Let $K_i = \{k_{i-1}+1, \dots, k_i\}$ and $\bar{\boldsymbol{\Sigma}}^{-1}_{i,i} = \boldsymbol{Q}^i$ which gives us:

$$\log(f(\boldsymbol{y}_i|\boldsymbol{y}_{pa(i)}) = -\frac{1}{2}\log(|\bar{\boldsymbol{\Sigma}}_{i,i}|) - \frac{|K_i|}{2}\log(2\pi) - \frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{\alpha}_i - \boldsymbol{\beta}_i^T\boldsymbol{y}_{pa(i)})^T\boldsymbol{Q}^i(\boldsymbol{y}_i - \boldsymbol{\alpha}_i - \boldsymbol{\beta}_i^T\boldsymbol{y}_{pa(i)})$$

Here the regression coefficients matrix $\boldsymbol{\beta}_i = (\beta_{i(s,h)})$, $\boldsymbol{\alpha}_i = (\alpha_{i(h)})$ where $h \in K_i$, $s \in K_{pa(i)}$ and the Maximum Likelihood Estimates(MLE) of $\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i, \bar{\boldsymbol{\Sigma}}_{i,i}$ are obtained.

Expanding $\log(f(\boldsymbol{y}_i|\boldsymbol{y}_{pa(i)})$ gives:

$$\begin{aligned}&\log(f(\boldsymbol{y}_i|\boldsymbol{y}_{pa(i)}) \\ =&-\frac{1}{2}\log(|\bar{\boldsymbol{\Sigma}}_{i,i}|) - \frac{|K_i|}{2}\log(2\pi) - \frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{\alpha}_i - \boldsymbol{\beta}_i^T\boldsymbol{y}_{pa(i)})^T\boldsymbol{Q}^i(\boldsymbol{y}_i - \boldsymbol{\alpha}_i - \boldsymbol{\beta}_i^T\boldsymbol{y}_{pa(i)}) \\ =&-\frac{1}{2}\log(|\bar{\boldsymbol{\Sigma}}_{i,i}|) - \frac{|K_i|}{2}\log(2\pi) \\ &-\frac{1}{2}\left(\boldsymbol{\alpha}_i^T\boldsymbol{Q}^i\boldsymbol{\alpha}_i + \boldsymbol{y}_i^T\boldsymbol{Q}^i\boldsymbol{y}_i - 2\boldsymbol{\alpha}_i^T\boldsymbol{Q}^i\boldsymbol{y}_i - 2\boldsymbol{y}_{pa(i)}^T\boldsymbol{\beta}_i\boldsymbol{Q}^i\boldsymbol{y}_i + 2\boldsymbol{\alpha}_i^T\boldsymbol{Q}^i\boldsymbol{\beta}_i^T\boldsymbol{y}_{pa(i)} + \boldsymbol{y}_{pa(i)}^T\boldsymbol{\beta}_i\boldsymbol{Q}^i\boldsymbol{\beta}_i^T\boldsymbol{y}_{pa(i)}\right) \\ =&\underbrace{-\frac{1}{2}\left(\log(|\bar{\boldsymbol{\Sigma}}_{i,i}|) + |K_i|\log(2\pi) + \boldsymbol{\alpha}_i^T\boldsymbol{Q}^i\boldsymbol{\alpha}_i\right)}_{\kappa_i} \\ &-\frac{1}{2}\boldsymbol{y}_i^T\boldsymbol{Q}^i\boldsymbol{y}_i + \boldsymbol{\alpha}_i^T\boldsymbol{Q}^i\boldsymbol{y}_i + \boldsymbol{y}_{pa(i)}^T\boldsymbol{\beta}_i\boldsymbol{Q}^i\boldsymbol{y}_i - \boldsymbol{\alpha}_i^T\boldsymbol{Q}^i\boldsymbol{\beta}_i^T\boldsymbol{y}_{pa(i)} - \frac{1}{2}\boldsymbol{y}_{pa(i)}^T\boldsymbol{\beta}_i\boldsymbol{Q}^i\boldsymbol{\beta}_i^T\boldsymbol{y}_{pa(i)}\end{aligned}$$

Therefore the logarithm of the normalizing constant is:

$$\kappa = \sum_{i=1}^{m} \kappa_i = \sum_{i=1}^{m} -\frac{1}{2}\left(\log(|\bar{\boldsymbol{\Sigma}}_{i,i}|) + |K_i|\log(2\pi) + \boldsymbol{\alpha}_i^T \boldsymbol{Q}^i \boldsymbol{\alpha}_i\right)$$

The remaining contribution of $\log(f(\boldsymbol{y}_i|\boldsymbol{y}_{pa(i)})$ to $\log(f(\boldsymbol{y}))$ is

$$g(\boldsymbol{y}_i|\boldsymbol{y}_{pa(i)})) = \log(f(\boldsymbol{y}_i|\boldsymbol{y}_{pa(i)})) - \kappa_i$$

The representation of the potentials $p(\boldsymbol{y}_i, \boldsymbol{y}_{pa(i)})$ looks slightly different now:

$$\log(p(\boldsymbol{y}_i, \boldsymbol{y}_{pa(i)})) = -\frac{1}{2}\boldsymbol{y}_i^T \boldsymbol{\lambda}^i \boldsymbol{y}_i + \boldsymbol{\delta}_i \boldsymbol{y}_i - \boldsymbol{y}_{pa(i)}^T \boldsymbol{\gamma}^i \boldsymbol{y}_i \tag{4.1}$$

where $\boldsymbol{\lambda}^i$ is a $|K_i| \times |K_i|$ matrix, $\boldsymbol{\delta}_i$ is a row vector for length $|K_i|$ and $\boldsymbol{\gamma}^i$ is a $|K_{pa(i)}| \times |K_i|$ matrix.

If we initialize $\boldsymbol{\lambda}^i, \boldsymbol{\delta}_i, \boldsymbol{\gamma}^i$ by zeroes, then obtain these parameters by looping $i = 1 \ldots, m$ and collecting terms from $g(\boldsymbol{y}_i|\boldsymbol{y}_{pa(i)})$:

$$\begin{aligned}
&g(\boldsymbol{y}_i|\boldsymbol{y}_{pa(i)})) \\
=& -\frac{1}{2}\boldsymbol{y}_i^T \boldsymbol{Q}^i \boldsymbol{y}_i + \boldsymbol{\alpha}_i^T \boldsymbol{Q}^i \boldsymbol{y}_i + \boldsymbol{y}_{pa(i)}^T \boldsymbol{\beta}_i \boldsymbol{Q}^i \boldsymbol{y}_i - \boldsymbol{\alpha}_i^T \boldsymbol{Q}^i \boldsymbol{\beta}_i^T \boldsymbol{y}_{pa(i)} - \frac{1}{2}\boldsymbol{y}_{pa(i)}^T \boldsymbol{\beta}_i \boldsymbol{Q}^i \boldsymbol{\beta}_i^T \boldsymbol{y}_{pa(i)} \\
=& -\frac{1}{2}\boldsymbol{y}_i^T \underbrace{\boldsymbol{Q}^i}_{+\boldsymbol{\lambda}^i} \boldsymbol{y}_i + \underbrace{\boldsymbol{\alpha}_i^T \boldsymbol{Q}^i}_{+\boldsymbol{\delta}_i} \boldsymbol{y}_i - \boldsymbol{y}_{pa(i)}^T \underbrace{(-\boldsymbol{\beta}_i \boldsymbol{Q}^i)}_{+\boldsymbol{\gamma}^i} \boldsymbol{y}_i - \boldsymbol{\alpha}_i^T \boldsymbol{Q}^i \boldsymbol{\beta}_i^T \boldsymbol{y}_{pa(i)} - \frac{1}{2}\boldsymbol{y}_{pa(i)}^T \boldsymbol{\beta}_i \boldsymbol{Q}^i \boldsymbol{\beta}_i^T \boldsymbol{y}_{pa(i)}
\end{aligned}$$

The parameters of $\log(p(\boldsymbol{y}_i, \boldsymbol{y}_{pa(i)}))$ is updated:

$$\boldsymbol{\lambda}^i := \boldsymbol{\lambda}^i + \boldsymbol{Q}^i \text{ , } \boldsymbol{\delta}_i := \boldsymbol{\delta}_i + \boldsymbol{\alpha}_i^T \boldsymbol{Q}^i \text{ and } \boldsymbol{\gamma}^i := \boldsymbol{\gamma}^i - \boldsymbol{\beta}_i \boldsymbol{Q}^i \tag{4.2}$$

Let $h \in K_{pa(i)}$, then $h \in K_{h^*}$ for some cluster $\boldsymbol{y}_{h^*}$. The remaining is rewritten:

$$\begin{aligned}
&-\underbrace{\boldsymbol{\alpha}_i^T \boldsymbol{Q}^i \boldsymbol{\beta}_i^T}_{\boldsymbol{\omega}_i^T} \boldsymbol{y}_{pa(i)} - \frac{1}{2}\boldsymbol{y}_{pa(i)}^T \underbrace{\boldsymbol{\beta}_i \boldsymbol{Q}^i \boldsymbol{\beta}_i^T}_{\boldsymbol{\Omega}^i} \boldsymbol{y}_{pa(i)} = -\boldsymbol{\omega}_i^T \boldsymbol{y}_{pa(i)} - \frac{1}{2}\boldsymbol{y}_{pa(i)}^T \boldsymbol{\Omega}^i \boldsymbol{y}_{pa(i)} \\
&= \sum_{h \in K_{pa(i)}} -\omega_{i(h)} y_h - \frac{1}{2}\sum_{h \in K_{pa(i)}} \sum_{s \in K_{pa(i)}} y_h \Omega_{h,s}^i y_s \\
&= \sum_{h \in K_{pa(i)}} \underbrace{-\omega_{i(h)}}_{+\delta_h} y_h - \frac{1}{2}\sum_{h \in K_{pa(i)}} \sum_{s \in K_{pa(i)}|h^*=s^*} y_h \underbrace{\Omega_{h,s}^i}_{+\lambda_{h,s}^{h^*}} y_s - \sum_{h \in K_{pa(i)}} \sum_{s \in K_{pa(i)}|h^*<s^*} y_s \underbrace{\Omega_{s,h}^i}_{+\gamma_{s,h}^{h^*}} y_h
\end{aligned}$$

which is used to update parameters for the parents:

$$\begin{aligned} &\text{For } h,s \in K_{pa(i)}: \\ &\qquad \text{if } h^* = s^*: \qquad \lambda^{h^*}_{h,s} := \lambda^{h^*}_{h,s} + \Omega^i_{h,s} \\ &\qquad \text{if } h^* < s^*: \qquad \gamma^{h^*}_{s,h} := \gamma^{h^*}_{s,h} + \Omega^i_{s,h} \\ &\text{For } h \in K_{pa(i)}: \delta_h := \delta_h - \omega_{i(h)} \end{aligned}$$

4.2.2 Forward propagation

After the initialization, we then have the representation

$$\log(f(\boldsymbol{y})) = \kappa + \sum_{i=1}^{n} \log(p(\boldsymbol{y}_i, \boldsymbol{y}_{pa(i)}))$$

Let e be a subset of $\{1,\dots,m\}$ corresponding to the set of observed variables and let $\bar{e}$ correspond to the set of unobserved variables. We then intend to calculate $f(\boldsymbol{x}_e) = \int_{\mathbb{R}^{dim(\bar{e})}} f(\boldsymbol{y}_e, \boldsymbol{y}_{\bar{e}})\, dx_{\bar{e}}$. We do this by a forward propagation, where we loop $i = 1,\dots,m$ and perform step 1 and 2 depending on whether cluster i contains any vertices from e, $\bar{e}$ or both.

Step 1: If $\boldsymbol{y}_{ie} = \boldsymbol{y}_i \cap \boldsymbol{y}_e \neq \emptyset$
The i'th potential is now rewritten, using that $\boldsymbol{y}_{ie}$ are known and $\boldsymbol{y}_i \setminus \boldsymbol{y}_{ie} = \boldsymbol{y}_{i\bar{e}}$, into a new potential $\log(p(\boldsymbol{y}_{i\bar{e}}, \boldsymbol{y}_{pa(i)}))$

$$\begin{aligned} \log(p(\boldsymbol{y}_i, \boldsymbol{y}_{pa(i)})) &= -\frac{1}{2}\boldsymbol{y}_i^T \boldsymbol{\lambda}^i \boldsymbol{y}_i + \boldsymbol{\delta}_i \boldsymbol{y}_i - \boldsymbol{y}_{pa(i)}^T \boldsymbol{\gamma}^i \boldsymbol{y}_i \\ &= \underbrace{-\frac{1}{2}\boldsymbol{y}_{ie}^T \boldsymbol{\lambda}^i_{ie,ie} \boldsymbol{y}_{ie}}_{+\kappa} - \frac{1}{2}\boldsymbol{y}_{i\bar{e}}^T \underbrace{\boldsymbol{\lambda}^i_{i\bar{e},i\bar{e}}}_{+\boldsymbol{\lambda}^i_{i\bar{e},i\bar{e}}} \boldsymbol{y}_{i\bar{e}} \underbrace{- \boldsymbol{y}_{ie}^T \boldsymbol{\lambda}^i_{ie,i\bar{e}}}_{+\boldsymbol{\delta}_{i\bar{e}}} \boldsymbol{y}_{i\bar{e}} \\ &\quad + \underbrace{\boldsymbol{\delta}_{ie} \boldsymbol{y}_{ie}}_{+\kappa} + \underbrace{\boldsymbol{\delta}_{i\bar{e}}}_{+\boldsymbol{\delta}_{i\bar{e}}} \boldsymbol{y}_{i\bar{e}} \\ &\quad - \boldsymbol{y}_{pa(i)}^T \underbrace{\boldsymbol{\gamma}^i_{pa(i),ie} \boldsymbol{y}_{ie}}_{-\boldsymbol{\delta}^T_{pa(i)}} - \boldsymbol{y}_{pa(i)}^T \underbrace{\boldsymbol{\gamma}^i_{pa(i),i\bar{e}}}_{+\boldsymbol{\gamma}^i_{pa(i),i\bar{e}}} \boldsymbol{y}_{i\bar{e}} \end{aligned}$$

which means that the log-normalizing constant is updated as

$$\kappa := \kappa - \frac{1}{2}\boldsymbol{y}_{ie}^T \boldsymbol{\lambda}^i_{ie,ie} \boldsymbol{y}_{ie} + \boldsymbol{\delta}_{ie} \boldsymbol{y}_{ie}$$

The parents receives:

$$\text{For } h \in K_{pa(i)} : \delta_h := \delta_h - \boldsymbol{\gamma}^i_{h,ie}\boldsymbol{y}_{ie}$$

If $\boldsymbol{y}_{ie} \neq \boldsymbol{y}_i$ we are left with the potential

$$\log(p(\boldsymbol{y}_{i\bar{e}}, \boldsymbol{y}_{pa(i)})) = -\frac{1}{2}\boldsymbol{y}^T_{i\bar{e}}\boldsymbol{\lambda}^i_{i\bar{e},i\bar{e}}\boldsymbol{y}_{i\bar{e}} + \tilde{\boldsymbol{\delta}}_{i\bar{e}}\boldsymbol{y}_{i\bar{e}} - \boldsymbol{y}^T_{pa(i)}\boldsymbol{\gamma}^i_{pa(i),i\bar{e}}\boldsymbol{y}_{i\bar{e}}$$

Note that $\tilde{\boldsymbol{\delta}}_{i\bar{e}} = -\boldsymbol{y}^T_{ie}\boldsymbol{\lambda}^i_{ie,i\bar{e}} + \boldsymbol{\delta}_{i\bar{e}}$

Step 2: If $\boldsymbol{y}_{i\bar{e}} \neq \emptyset$
While rewriting the log-potential a simple notation will be used to ease reading and $\dim(\boldsymbol{y}_{i\bar{e}}) = |K_{i\bar{e}}|$

$$\begin{aligned}
&\log(p(\boldsymbol{y}_{i\bar{e}}, \boldsymbol{y}_{pa(i)}))\\
=&-\frac{1}{2}\boldsymbol{y}^T\boldsymbol{\lambda}\boldsymbol{y} + \boldsymbol{\delta}\boldsymbol{y} - \boldsymbol{y}_p^T\boldsymbol{\gamma}\boldsymbol{y}\\
=&-\frac{1}{2}\Big[\boldsymbol{y}^T\boldsymbol{\lambda}\boldsymbol{y} - 2\boldsymbol{\delta}\boldsymbol{\lambda}^{-1}\boldsymbol{\lambda}\boldsymbol{y} + 2\boldsymbol{y}_p^T\boldsymbol{\gamma}\boldsymbol{\lambda}^{-1}\boldsymbol{\lambda}\boldsymbol{y}\\
&\quad + (-\boldsymbol{\delta}\boldsymbol{\lambda}^{-1})\boldsymbol{\lambda}(-\boldsymbol{\delta}\boldsymbol{\lambda}^{-1})^T + (\boldsymbol{y}_p^T\boldsymbol{\gamma}\boldsymbol{\lambda}^{-1})\boldsymbol{\lambda}(\boldsymbol{y}_p^T\boldsymbol{\gamma}\boldsymbol{\lambda}^{-1})^T - 2(\boldsymbol{\delta}\boldsymbol{\lambda}^{-1})\boldsymbol{\lambda}(\boldsymbol{y}_p^T\boldsymbol{\gamma}\boldsymbol{\lambda}^{-1})^T\Big]\\
&+\frac{1}{2}\Big[(-\boldsymbol{\delta}\boldsymbol{\lambda}^{-1})\boldsymbol{\lambda}(-\boldsymbol{\delta}\boldsymbol{\lambda}^{-1})^T + (\boldsymbol{y}_p^T\boldsymbol{\gamma}\boldsymbol{\lambda}^{-1})\boldsymbol{\lambda}(\boldsymbol{y}_p^T\boldsymbol{\gamma}\boldsymbol{\lambda}^{-1})^T - 2(\boldsymbol{\delta}\boldsymbol{\lambda}^{-1})\boldsymbol{\lambda}(\boldsymbol{y}_p^T\boldsymbol{\gamma}\boldsymbol{\lambda}^{-1})^T\Big]\\
=&-\frac{1}{2}\Big(\boldsymbol{y} - \underbrace{((\boldsymbol{\delta}\boldsymbol{\lambda}^{-1})^T - (\boldsymbol{y}_p^T\boldsymbol{\gamma}\boldsymbol{\lambda}^{-1})^T)}_{\boldsymbol{\mu}^*}\Big)^T\boldsymbol{\lambda}\Big(\boldsymbol{y} - ((\boldsymbol{\delta}\boldsymbol{\lambda}^{-1})^T - (\boldsymbol{y}_p^T\boldsymbol{\gamma}\boldsymbol{\lambda}^{-1})^T)\Big)\\
&+\frac{1}{2}\Big(\boldsymbol{\delta}(\boldsymbol{\delta}\boldsymbol{\lambda}^{-1})^T + \boldsymbol{y}_p^T\gamma(\boldsymbol{\lambda}^{-1})^T\boldsymbol{\gamma}^T\boldsymbol{y}_p - 2\boldsymbol{\delta}(\boldsymbol{\lambda}^{-1})^T\boldsymbol{\gamma}^T\boldsymbol{y}_p\Big)\\
=&-\frac{1}{2}\log(|\boldsymbol{\lambda}^{-1}|) - \frac{|K_{i\bar{e}}|}{2}\log(2\pi) - \frac{1}{2}\left(\boldsymbol{y} - \boldsymbol{\mu}^*\right)^T\boldsymbol{\lambda}\left(\boldsymbol{y} - \boldsymbol{\mu}^*\right)\\
&+\frac{1}{2}\log(|\boldsymbol{\lambda}^{-1}|) + \frac{|K_{i\bar{e}}|}{2}\log(2\pi) + \frac{1}{2}\Big(\boldsymbol{\delta}(\boldsymbol{\delta}\boldsymbol{\lambda}^{-1})^T + \boldsymbol{y}_p^T\gamma(\boldsymbol{\lambda}^{-1})^T\boldsymbol{\gamma}^T\boldsymbol{y}_p - 2\boldsymbol{\delta}(\boldsymbol{\lambda}^{-1})^T\boldsymbol{\gamma}^T\boldsymbol{y}_p\Big)
\end{aligned} \tag{4.3}$$

Then integrate out $\boldsymbol{y}_{i\bar{e}}$ in the potential

$$
\begin{aligned}
&\int_{\mathbb{R}^{|K_{i\bar{e}}|}} p(\boldsymbol{y}_{i\bar{e}}, \boldsymbol{y}_{pa(i)})\, d\boldsymbol{y}_{i\bar{e}} \\
=&\int_{\mathbb{R}^{|K_{i\bar{e}}|}} \frac{1}{\sqrt{(2\pi)^{|K_{i\bar{e}}|}\boldsymbol{\lambda}^{-1}}} \exp\left(-\frac{1}{2}\left(\boldsymbol{y}-\boldsymbol{\mu}^*\right)^T \boldsymbol{\lambda}\left(\boldsymbol{y}-\boldsymbol{\mu}^*\right)\right)\, d\boldsymbol{y} \\
\times&\exp\left(\frac{1}{2}\log(|\boldsymbol{\lambda}^{-1}|) + \frac{|K_{i\bar{e}}|}{2}\log(2\pi) + \frac{1}{2}\Big(\boldsymbol{\delta}(\boldsymbol{\delta}\boldsymbol{\lambda}^{-1})^T + \boldsymbol{y}_p^T\gamma(\boldsymbol{\lambda}^{-1})^T\boldsymbol{\gamma}^T\boldsymbol{y}_p - 2\boldsymbol{\delta}(\boldsymbol{\lambda}^{-1})^T\boldsymbol{\gamma}^T\boldsymbol{y}_p\Big)\right) \\
=&\exp\left(\frac{1}{2}\log(|\boldsymbol{\lambda}^{-1}|) + \frac{|K_{i\bar{e}}|}{2}\log(2\pi) + \frac{1}{2}\Big(\boldsymbol{\delta}(\boldsymbol{\delta}\boldsymbol{\lambda}^{-1})^T + \boldsymbol{y}_p^T\gamma(\boldsymbol{\lambda}^{-1})^T\boldsymbol{\gamma}^T\boldsymbol{y}_p - 2\boldsymbol{\delta}(\boldsymbol{\lambda}^{-1})^T\boldsymbol{\gamma}^T\boldsymbol{y}_p\Big)\right)
\end{aligned}
$$

We now rewrite the result and go back to the normal notation

$$
\begin{aligned}
&\log\int_{\mathbb{R}^{|K_{i\bar{e}}|}} p(\boldsymbol{y}_{i\bar{e}}, \boldsymbol{y}_{pa(i)})\, d\boldsymbol{y}_{i\bar{e}} \\
=&\frac{1}{2}\log(|\boldsymbol{\lambda}^{-1}|) + \frac{|K_{i\bar{e}}|}{2}\log(2\pi) + \frac{1}{2}\Big(\boldsymbol{\delta}(\boldsymbol{\delta}\boldsymbol{\lambda}^{-1})^T + \boldsymbol{y}_p^T\gamma(\boldsymbol{\lambda}^{-1})^T\boldsymbol{\gamma}^T\boldsymbol{y}_p - 2\boldsymbol{\delta}(\boldsymbol{\lambda}^{-1})^T\boldsymbol{\gamma}^T\boldsymbol{y}_p\Big) \\
=&\frac{1}{2}\left(\log(|\boldsymbol{\lambda}^{-1}|) + |K_{i\bar{e}}|\log(2\pi) + \boldsymbol{\delta}(\boldsymbol{\delta}\boldsymbol{\lambda}^{-1})^T\right) + \frac{1}{2}\boldsymbol{y}_p^T\gamma(\boldsymbol{\lambda}^{-1})^T\boldsymbol{\gamma}^T\boldsymbol{y}_p - \boldsymbol{\delta}(\boldsymbol{\lambda}^{-1})^T\boldsymbol{\gamma}^T\boldsymbol{y}_p \\
=&\underbrace{\frac{1}{2}\left(\log(|{\boldsymbol{\lambda}^{i}_{i\bar{e},i\bar{e}}}^{-1}|) + |K_{i\bar{e}}|\log(2\pi) + \tilde{\boldsymbol{\delta}}_{i\bar{e}}(\tilde{\boldsymbol{\delta}}_{i\bar{e}}{\boldsymbol{\lambda}^{i}_{i\bar{e},i\bar{e}}}^{-1})^T\right)}_{\kappa^*} \\
&+\frac{1}{2}\boldsymbol{y}_{pa(i)}^T\underbrace{\boldsymbol{\gamma}^i_{pa(i),i\bar{e}}({\boldsymbol{\lambda}^{i}_{i\bar{e},i\bar{e}}}^{-1})^T{\boldsymbol{\gamma}^i_{pa(i),i\bar{e}}}^T}_{\boldsymbol{\lambda}^*}\boldsymbol{y}_{pa(i)} - \underbrace{\tilde{\boldsymbol{\delta}}_{i\bar{e}}({\boldsymbol{\lambda}^{i}_{i\bar{e},i\bar{e}}}^{-1})^T{\boldsymbol{\gamma}^i_{pa(i),i\bar{e}}}^T}_{\boldsymbol{\delta}^*}\boldsymbol{y}_{pa(i)} \\
=&\kappa^* + \frac{1}{2}\boldsymbol{y}_{pa(i)}^T\boldsymbol{\lambda}^*\boldsymbol{y}_{pa(i)} - \boldsymbol{\delta}^*\boldsymbol{y}_{pa(i)} \\
=&\kappa^* + \sum_{h\in K_pa(i)}\underbrace{-\delta^*_h}_{+\delta_h} y_h + \frac{1}{2}\sum_{h\in K_{pa(i)}}\sum_{s\in K_{pa(i)}|h^*=s^*} y_h\underbrace{\lambda^*_{h,s}}_{+\lambda^{h^*}_{h,s}} y_s + \sum_{h\in K_{pa(i)}}\sum_{s\in K_{pa(i)}|h^*<s^*} y_s\underbrace{\lambda^*_{s,h}}_{+\gamma^{h^*}_{s,h}} y_h
\end{aligned}
$$

Hence we update the normalizing constant and the parent parameters accordingly:

$$
\begin{aligned}
\kappa &:= \kappa + \kappa^* \\
\text{For } h, s \in K_{pa(i)} &: \\
\text{if } h^* = s^* :& \qquad \lambda^{h^*}_{h,s} := \lambda^{h^*}_{h,s} - \lambda^*_{h,s} \\
\text{if } h^* < s^* :& \qquad \gamma^{h^*}_{s,h} := \gamma^{h^*}_{s,h} - \lambda^*_{s,h} \\
\text{For } h \in K_{pa(i)} &: \delta_h := \delta_h - \delta^*_h
\end{aligned}
$$

As in the simple case we have that $\log(f(\boldsymbol{y}_e)) = \kappa$ and $p(\boldsymbol{y}_{i\bar{e}}, \boldsymbol{y}_{pa(i)})$ is proportional to $f(\boldsymbol{y}_{i\bar{e}}|\boldsymbol{y}_{pa(i)\cup e})$.

4.2.3 Backward propagation

Next, we want to determine $f(\boldsymbol{y}_{i\bar{e}}|\boldsymbol{y}_e)$ when $\boldsymbol{y}_{i\bar{e}} \neq \emptyset$. We rewrite the i'th potential like in equation (4.3)

$$\begin{aligned}
&\log(p(\boldsymbol{y}_{i\bar{e}}, \boldsymbol{y}_{pa(i)})) \\
=& -\frac{1}{2}\boldsymbol{y}_{i\bar{e}}^T\boldsymbol{\lambda}_{i\bar{e},i\bar{e}}^i\boldsymbol{y}_{i\bar{e}} + \boldsymbol{\delta}_{i\bar{e}}\boldsymbol{y}_{i\bar{e}} - \boldsymbol{y}_{pa(i)}^T\gamma_{pa(i),i\bar{e}}^i\boldsymbol{y}_{i\bar{e}} \\
=& -\frac{1}{2}\Big(\boldsymbol{y}_{i\bar{e}} - {\boldsymbol{\lambda}_{i\bar{e},i\bar{e}}^i}^{-1}(\boldsymbol{\delta}_{i\bar{e}}^T - {\gamma_{pa(i),i\bar{e}}^i}^T\boldsymbol{y}_{pa(i)})\Big)^T\boldsymbol{\lambda}_{i\bar{e},i\bar{e}}^i\Big(\boldsymbol{y}_{i\bar{e}} - {\boldsymbol{\lambda}_{i\bar{e},i\bar{e}}^i}^{-1}(\boldsymbol{\delta}_{i\bar{e}}^T - {\gamma_{pa(i),i\bar{e}}^i}^T\boldsymbol{y}_{pa(i)})\Big) \\
&+\frac{1}{2}\Big(\boldsymbol{\delta}_{i\bar{e}}(\boldsymbol{\delta}_{i\bar{e}}{\boldsymbol{\lambda}_{i\bar{e},i\bar{e}}^i}^{-1})^T + \boldsymbol{y}_{pa(i)}^T\gamma_{pa(i),i\bar{e}}^i({\boldsymbol{\lambda}_{i\bar{e},i\bar{e}}^i}^{-1})^T{\boldsymbol{\gamma}_{pa(i),i\bar{e}}^i}^T\boldsymbol{y}_{pa(i)} - 2\boldsymbol{\delta}_{i\bar{e}}({\boldsymbol{\lambda}_{i\bar{e},i\bar{e}}^i}^{-1})^T{\boldsymbol{\gamma}_{pa(i),i\bar{e}}^i}^T\boldsymbol{y}_{pa(i)}\Big)
\end{aligned}$$

This means, that after the forward propagation we may from the i'th potential parameters calculate:

- $M_i := \mathbb{E}(\boldsymbol{Y}_{i\bar{e}}|\boldsymbol{Y}_{pa(i)\cup e} = \boldsymbol{y}_{pa(i)\cup e}) = {\boldsymbol{\lambda}_{i\bar{e},i\bar{e}}^i}^{-1}(\boldsymbol{\delta}_{i\bar{e}}^T - {\boldsymbol{\gamma}_{pa(i),i\bar{e}}^i}^T\boldsymbol{y}_{pa(i)})$
- $V_i := \mathrm{Cov}(\boldsymbol{Y}_{i\bar{e}}, \boldsymbol{Y}_{i\bar{e}}|\boldsymbol{Y}_{pa(i)\cup e}) = \mathrm{Var}(\boldsymbol{Y}_{i\bar{e}}|\boldsymbol{Y}_{pa(i)\cup e}) = {\boldsymbol{\lambda}_{i\bar{e},i\bar{e}}^i}^{-1}$

Next we do a backward propagation, where we loop $i = m, \dots, 1$ and exploit that for $i \in e$ then

- $\tilde{\boldsymbol{\mu}}_{ie} := \mathbb{E}(\boldsymbol{Y}_{ie}|\boldsymbol{Y}_e = \boldsymbol{y}_e) = \boldsymbol{y}_{ie}$
- $\tilde{\boldsymbol{\sigma}}_{ie} := \mathrm{Var}(\boldsymbol{Y}_{ie}|\boldsymbol{Y}_e) = \boldsymbol{0}$
- $\tilde{\boldsymbol{\Sigma}}_{ie} := \mathrm{Cov}(\boldsymbol{Y}_{ie}, \boldsymbol{Y}_{pa(i)}|\boldsymbol{Y}_e) = \boldsymbol{0}$

This is used to calculate $\tilde{\boldsymbol{\mu}}_{i\bar{e}}$, $\tilde{\boldsymbol{\sigma}}_{i\bar{e}}$ and $\tilde{\boldsymbol{\Sigma}}_{i\bar{e}}$.

For $\tilde{\boldsymbol{\mu}}_{i\bar{e}}$ we first use the same method as in the simple case:

$$\begin{aligned}
\tilde{\boldsymbol{\mu}}_{i\bar{e}} :=& \mathbb{E}(\boldsymbol{Y}_{i\bar{e}}|\boldsymbol{Y}_e = \boldsymbol{y}_e) \\
=& \mathbb{E}(M_i|\boldsymbol{Y}_e = \boldsymbol{y}_e) \\
=& {\boldsymbol{\lambda}_{i\bar{e},i\bar{e}}^i}^{-1}(\boldsymbol{\delta}_{i\bar{e}}^T - {\boldsymbol{\gamma}_{pa(i),i\bar{e}}^i}^T\mathbb{E}(\boldsymbol{Y}_{pa(i)}|\boldsymbol{Y}_e = \boldsymbol{y}_e)) \\
=& {\boldsymbol{\lambda}_{i\bar{e},i\bar{e}}^i}^{-1}(\boldsymbol{\delta}_{i\bar{e}}^T - {\boldsymbol{\gamma}_{pa(i),i\bar{e}}^i}^T\tilde{\boldsymbol{\mu}}_{pa(i)})
\end{aligned}$$

For $\tilde{\boldsymbol{\sigma}}_{i\bar{e}}$ we use $\mathrm{Var}(X|Z) = \mathbb{E}(\mathrm{Var}(X|Y)|Z)+\mathrm{Var}(\mathbb{E}(X|Y)|Z)$ for the first equality then $\mathbb{E}(\alpha|Z) = \alpha$ and $\mathrm{Var}(\boldsymbol{A}\boldsymbol{X} + \boldsymbol{\beta}|\boldsymbol{Z}) = \boldsymbol{A}\mathrm{Var}(\boldsymbol{X}|\boldsymbol{Z})\boldsymbol{A}^T$ for the third equality:

$$\begin{aligned}
\tilde{\boldsymbol{\sigma}}_{i\bar{e}} =&\mathrm{Var}(\boldsymbol{Y}_{i\bar{e}}|\boldsymbol{Y}_e)\\
=&\mathbb{E}(V_i|\boldsymbol{Y}_e = \boldsymbol{y}_e) + \mathrm{Var}(M_i|\boldsymbol{Y}_e)\\
=&\mathbb{E}({\boldsymbol{\lambda}^i_{i\bar{e},i\bar{e}}}^{-1}|\boldsymbol{Y}_e = \boldsymbol{y}_e) + \mathrm{Var}({\boldsymbol{\lambda}^i_{i\bar{e},i\bar{e}}}^{-1}\boldsymbol{\delta}^T_{i\bar{e}} - {\boldsymbol{\lambda}^i_{i\bar{e},i\bar{e}}}^{-1}{\boldsymbol{\gamma}^i_{pa(i),i\bar{e}}}^T\boldsymbol{Y}_{pa(i)}|\boldsymbol{Y}_e)\\
=&{\boldsymbol{\lambda}^i_{i\bar{e},i\bar{e}}}^{-1} + {\boldsymbol{\lambda}^i_{i\bar{e},i\bar{e}}}^{-1}{\boldsymbol{\gamma}^i_{pa(i),i\bar{e}}}^T\mathrm{Var}(\boldsymbol{Y}_{pa(i)}|X_e)\boldsymbol{\gamma}^i_{pa(i),i\bar{e}}{\boldsymbol{\lambda}^i_{i\bar{e},i\bar{e}}}^{-1}
\end{aligned}$$

For $\tilde{\boldsymbol{\Sigma}}_{i\bar{e}}$ we again use that $\mathbb{E}(X|Z) = \mathbb{E}(\mathbb{E}(X|Y)|Z)$ along with

- $\mathrm{Cov}(X, Y|Z) = \mathbb{E}(XY|Z) - \mathbb{E}(X|Z)\mathbb{E}(Y|Z)$
- $\mathbb{E}(XY|X) = X\mathbb{E}(Y|X)$
- $\mathrm{Cov}(\boldsymbol{A}\boldsymbol{X} + \boldsymbol{\beta}, \boldsymbol{Y}|\boldsymbol{Z}) = \boldsymbol{A}\mathrm{Cov}(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{Z})$

$$\begin{aligned}
\tilde{\boldsymbol{\Sigma}}_{i\bar{e}} :=&\mathrm{Cov}(\boldsymbol{Y}_{i\bar{e}}, \boldsymbol{Y}_{pa(i)}|\boldsymbol{Y}_e)\\
=&\mathbb{E}(\boldsymbol{Y}_{pa(i)}\boldsymbol{Y}_{i\bar{e}}|\boldsymbol{Y}_e = \boldsymbol{y}_e) - \mathbb{E}(\boldsymbol{Y}_{pa(i)}|\boldsymbol{Y}_e = \boldsymbol{y}_e)\mathbb{E}(\boldsymbol{Y}_{i\bar{e}}|Y_e = \boldsymbol{y}_e)\\
=&\mathbb{E}(\mathbb{E}(\boldsymbol{Y}_{pa(i)}\boldsymbol{Y}_{i\bar{e}}|\boldsymbol{Y}_{pa(i)\cup e} = \boldsymbol{y}_{pa(i)\cup e})|\boldsymbol{Y}_e = \boldsymbol{y}_e) - \mathbb{E}(\boldsymbol{Y}_{pa(i)}|\boldsymbol{Y}_e = \boldsymbol{y}_e)\mathbb{E}(M_i|\boldsymbol{Y}_e = \boldsymbol{y}_e)\\
=&\mathbb{E}(\boldsymbol{Y}_{pa(i)}M_i|\boldsymbol{Y}_e = \boldsymbol{y}_e) - \mathbb{E}(\boldsymbol{Y}_{pa(i)}|\boldsymbol{Y}_e = \boldsymbol{y}_e)\mathbb{E}(M_i|\boldsymbol{Y}_e = \boldsymbol{y}_e)\\
=&\mathrm{Cov}(M_i, \boldsymbol{Y}_{pa(i)}|\boldsymbol{Y}_e)\\
=&\mathrm{Cov}({\boldsymbol{\lambda}^i_{i\bar{e},i\bar{e}}}^{-1}\boldsymbol{\delta}^T_{i\bar{e}} - {\boldsymbol{\lambda}^i_{i\bar{e},i\bar{e}}}^{-1}{\boldsymbol{\gamma}^i_{pa(i),i\bar{e}}}^T\boldsymbol{Y}_{pa(i)}, \boldsymbol{Y}_{pa(i)}|\boldsymbol{Y}_e)\\
=&-{\boldsymbol{\lambda}^i_{i\bar{e},i\bar{e}}}^{-1}{\boldsymbol{\gamma}^i_{pa(i),i\bar{e}}}^T\mathrm{Var}(\boldsymbol{Y}_{pa(i)}|\boldsymbol{Y}_e)
\end{aligned}$$

where $\mathrm{Var}(\boldsymbol{Y}_{pa(i)}|\boldsymbol{Y}_e)$ is determined from $(\tilde{\boldsymbol{\sigma}}_j, \tilde{\boldsymbol{\Sigma}}_j)$, $j \in pa(i)$.
After the forward and backward propagation we have that

$$f(\boldsymbol{y}_e) = \exp(\kappa)$$

$$(\boldsymbol{Y}_i|\boldsymbol{Y}_e = \boldsymbol{y}_e) \sim N(\tilde{\boldsymbol{\mu}}_i, \tilde{\boldsymbol{\sigma}}_i) \ \text{ for } \ i = 1, \ldots, m$$

Note that the matrix $\tilde{\boldsymbol{\sigma}}_i$ is 0 for entry (c, d) if $y_{ic} \notin \boldsymbol{y}_{i\bar{e}}$ or $y_{id} \notin \boldsymbol{y}_{i\bar{e}}$.

4.3 Special cases

Looking at the new method, we see that it is an extreme of the general case where each $\boldsymbol{Y}_i$ is as small a possible

$$\boldsymbol{Y}_1 = X_1 \ , \ \boldsymbol{Y}_2 = X_2 \ , \ldots , \ \boldsymbol{Y}_m = X_n$$

The other extreme is the JTA for which each $\boldsymbol{Y}_i$ is made as large as possible, i.e. $\boldsymbol{Y}_i = C_j^*$.

After initialization the size of $\boldsymbol{Y}_i$ matter whenever matrix inversion done. The only matrix inversion is ${\boldsymbol{\lambda}^i_{i\bar{e},i\bar{e}}}^{-1}$, which size is determined by the number of unobserved variables in $\boldsymbol{Y}_i$.

Chapter 5

Testing and comparison

The new method was implemented, tested and compared with the simple method. The following describes the code and the comparison of the two methods.

5.1 Code

The code is written in R and can be found on https://github.com/cosius/Master-Thesis. It contains an implementation of the new method along with a test script for test data generating and testing.

The implementation is divided into initialization, forward propagation, backward propagation, and a script that calls the two propagations.

Only the propagation's where tested. The data for testing was generated so all cliques were of the same size and had the same overlap, i.e. separator size. The cliques were generated so they only overlapped the previous and the next clique.

The test script uses the same data to test runtime for the simple method, which is done by using the package *condMVNorm* that can be found at https://cran.r-project.org/web/packages/condMVNorm/index.html.

5.2 Comparison

The new method and the simple method were compared by running the same data calculation ten times and recording the runtime. The same test was done several times, and a small variation was found. This is most likely due to other software running on the computer used.

Looking at the simple method, different clique and separator sizes gave no change in runtime. This is to be expected as these have no influence on the size of the covariance matrix, which is the determining factor for the runtime of a matrix inversion.

The number of vertices did, however, affect the runtime for the simple method. Tests were done for 500, 1000 and 2000 vertices. The runtime came to 3.8, 26.5, and 186.8 seconds. Assuming a polynomial complexity, i.e. $(a500)^x = 3.8$, we get:

$$\frac{\log\left(\frac{26.5}{3.8}\right)}{\log(2)} = 2.80 \qquad \frac{\log\left(\frac{186.8}{26.5}\right)}{\log(2)} = 2.82$$

This suggests that condMVNorm uses the Strassen algorithm for inverting the matrix, as its complexity is $O(n^{2.81})$ [Gall, 2014].

For the new method, clique and separator sizes did have an effect on runtime. This is also expected as increasing both, adds more edges thereby increasing the number of updates needed to be sent. Going from 3 to 8 in separator size with 1000 vertices, increased the runtime from 3.6 to 7 seconds, whereas the simple method stayed the same at around 26.5 seconds. However, it was not expected that for 1000 vertices and a clique size of 200, the runtime was 43.5 seconds for the new method, as opposed to the 26.6 seconds for the simple method. The reason of this is unknown and may be the result of a fault in the code.

Another problem with the new method arises when looking at the number of vertices. Here the runtime for 500, 1000, 2000, and 4000 vertices were 1.3, 3.6, 21.,3 and 93.7 seconds accordingly. Compared with the simple method these are still significantly lower, but no pattern is found. Runtime was expected to double when the number of vertices doubles, as clique and separator sizes were kept the same. This is because the propagation only depends on the number of parents a vertex has, which is determined by clique and separator sizes.

Whether these anomalies are due to coding or an oversight in the algorithm remains to be determined. However, the new method is proven faster than the simple method when the clique sizes are small.

Bibliography

Robert G. Cowell, A. Philip Dawid, Steffen L. Lauritzen, and David J. Spiegelhalter. *Probabilistic Networks and Expert Systems.* Springer, 1999. ISBN 0-387-98767-3.

Franc¸ois Le Gall. Powers of tensors and fast matrix multiplication. *Proceedings of the 39th International Symposium on Symbolic and Algebraic Computation*, 2014.

Steffen L. Lauritzen. *Graphical models.* Oxford University Press Inc., 2004. ISBN 0-19-852219-3.

Mark Paskin. A short course on graphical models : 3. the junction tree algorithms, 2003. URL `https://web.archive.org/web/20150319085443/http://ai.stanford.edu/~paskin/gm-short-course/lec3.pdf`. Slides from BAYES2013.

George Arthur Frederick Seber. *Multivariate Observations.* John Wiley & Sons, Inc, 1984. ISBN 0-471-88104-X.

Robert E. Tarjan and Mihalis Yannakakis. Simple linear-time algrithms to test chordality of graphs, test acyclicity of hypergraphs, and selectively reduce acyclic hypergraphs. *SIAM Journal on Computing*, 13:566–579, 1984.