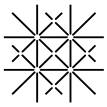


University of Basel

Sampling and Annealing for Dependency Subnetwork Estimation

Post-Handin Fixed Version of the Master Thesis

Natural Science Faculty of the University of Basel
Department of Mathematics and Computer Science
Biomedical Data Analysis
http://bmda.dmi.unibas.ch/

Examiner: Prof. Dr. Volker Roth
Second Examiner: Sonali Parbhoo
Supervisor: Mario Wieser, Sonali Parbhoo

Fabricio Arend Torres
fabricio.arendtorres@unibas.ch
2012-051-934

Last Change: 19th November 2018

Note about this Version

This file corresponds to an error-corrected version of the handed in Master Thesis and is updated as more mistakes are found. It has not been subject to the evaluation of the Thesis and simply serves as a reference for future readers.
The changes are mostly due to (minor) mistakes in formulas, figures or orthography and are not distinctly marked.

Acknowledgments

Special thanks go to my supervisors Mario Wieser and Sonali Parbhoo for continuously providing support and feedback to my work in our many meetings.

Likewise, I want to thank Professor Volker Roth not only for allowing me to work on an interesting project, but also for offering insight when I was seemingly stuck in the course of this thesis.

Further, many thanks go to everyone involved in proofreading my work and helping me shape it to its current form.

Last but not least, I'm grateful to my family and friends for their encouragement and support through the process of this thesis and all the previous years leading to it.

Abstract

Learning sparse dependency structures between multiple variables is very crucial in many application areas of biology and personalized medicine. In this thesis, we focus particularly on HIV, where the aim is to identify interactions between the viral genotype and multiple phenotype variables. Bayesian Markov Blanket (BMB) estimation allows us to only reconstruct a sparse dependency subnetwork of interest, namely the interactions between the viral genotype and corresponding phenotype variables. Via Markov Chain Monte Carlo sampling, the BMB provides the full posterior distribution. However, we are mostly interested in the mode of the posterior distribution. Finding the mode of a high dimensional empirical distribution would require multidimensional density estimation, which is known to suffer from the curse of dimensionality. In order to correct for this shortcoming, we extend the Markov Blanket estimation by introducing Simulated Annealing to the models Gibbs sampler, which allows us to approximately sample from the set of global maxima. Compared to a non-Bayesian approach, it has the advantage of still being able to provide information about the underlying distribution.

Our approach is evaluated on artificial data and compared to both the original BMB and the Graphical LASSO. In addition, we applied it on data from the SystemsX.ch HIV-X cohort. While the Annealing is not able to compete with the GLASSO on synthetic data, it shows similar performance to the BMB while being more robust & convenient for model selection when applied on the HIV-X data.

Furthermore, the behavior and modality of the models posterior distribution in relation to the hyperparameter is experimentally analyzed. We demonstrate that the marginal posterior of the Markov Blanket gets multi-modal if not enough sparsity is enforced.

Table of Contents

1 Introduction

Systems in biology, medicine and many other areas are subject to their underlying and usually unknown dependency structures. Clearly, learning these structures from available data could be invaluable, with the possibility of providing insight into a wide range of domain-specific problems. This proves to be a difficult problem and especially so in high-dimensional settings, i.e. when there are many variables in relation to the available number of data samples. This thesis focuses on the estimation of a subnetwork of such a dependency structure, namely the Markov Blanket (MB) of a few selected nodes of interest. The MB is the set of nodes in a network that render a node of interest independent of the remaining network when being conditioned on. An example is shown in Figure 1.1.

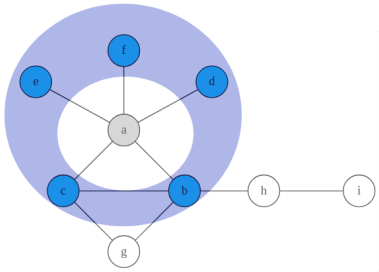


Figure 1.1: The blue nodes form a *blanket* around node 'a', rendering it conditionally independent of the remaining network.

In statistics the general problem of inferring the dependence structure of (assumed) Gaussian distributed data is classically known as *Covariance Selection* [Price, 1972], because the zero-pattern of the inverse covariance matrix encodes conditional independencies. Since dependence structures can be conveniently expressed in the form of a network graph (see Figure 1.2), the problem now tends to be referred to as the estimation of a Gaussian Graphical Model (GGM).

Many popular approaches for estimating such GGMs are based on enforcing sparsity in the network. A sparse network is generally favorable for its interpretability, since large and densely connected graphs are difficult to analyze. Prominent examples of sparse network

$$\text{sign}(\mathbf{\Sigma^{-1}}) = \begin{array}{c} \begin{array}{cccc} \mathbf{a} & \mathbf{b} & \mathbf{c} & \mathbf{h} \end{array} \\ \left(\begin{array}{cccc} 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & -1 \\ 1 & 1 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{array} \right) \end{array} \begin{array}{c} \mathbf{a} \\ \mathbf{b} \\ \mathbf{c} \\ \mathbf{h} \end{array} \Rightarrow$$

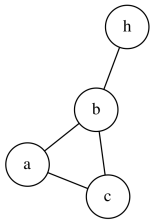


Figure 1.2: Example for a dependency network encoded by the zero-pattern of the inverse covariance matrix.

estimation are the Graphical LASSO [Friedman et al., 2008] and its Bayesian perspective [Wang, 2012]. As with many alternative approaches, they infer the full network, which is not always required for the problem at hand. In many real-world problems there are often only a few variables of interest (e.g. clinical factors) which might interact with a great number of other variables (e.g. genetic markers). The interactions between these few variables and the others are often the ones of highest interest. In such cases inferring the whole network might be unnecessary and computationally expensive (especially for models that rely on simulation, such as Markov Chain Monte Carlo (MCMC) methods). Consequently, it would be advantageous to limit estimation to the subnetwork of interest, and the BMB estimation by Kaufmann et al. [2016] offers such an approach. By avoiding the inference of the whole network, the method allows for a more efficient estimation than related Bayesian methods (e.g. Wang [2012]) when the Markov Blanket of a few variables is of importance. Kaufmann et al. [2016] infer the whole posterior distribution. However, in practice a point estimate is often desired, among others for the sake of interpretability and convenience. While we could base this estimate on statistics such as the mean or median of the empirical posterior, the most probable point, name the maximum a posteriori probability (MAP), would be more sensible in many cases. However, estimating this mode from the empirical posterior distribution acquired by MCMC sampling poses some problems in high dimensional settings. The estimation would require some sort of binning or kernel density estimation, which generally does not scale well for higher dimensions due to the 'curse of dimensionality' [Scott, 1991, Stone, 1980]. Furthermore, if the mode is not surrounded by a lot of probability mass, the majority of the simulation time will be spent exploring regions of lower interest. We address this issue by introducing Simulated Annealing (SA) to the Bayesian Markov Blanket estimation, allowing for a direct and more robust estimation of the MAP. Because Simulated Annealing is seamlessly added on top of an already existing MCMC sampler, small changes of the original model, such as the introduction of (hyper) priors, can be integrated into the MAP estimation. Compared to the non-Bayesian approach, SA also gives the advantage of having both the full posterior and a point estimate for the same model available, if it is desired. Performance of the Annealing in comparison to the BMB and the Graphical LASSO is evaluated on artificial data. Furthermore, we investigate the modality of the posterior as well as the general behavior of the model on both artificial and real data.

As already mentioned, learning dependencies inherent to the data can be very useful in

medical domains. In context of this thesis, we will focus on the area of HIV-1 treatment responses for individuals that have been successfully treated for more than five years. While current treatments are capable of suppressing the virus, it cannot be fully exterminated and dormant viral populations will multiply again after treatment stop. Furthermore, mutations of the virus might result in resistances that can affect current and future treatments. We aim to find potential dependencies between resistance relevant mutations in latent viral populations and clinical factors that quantify the success of the treatments (e.g. the viral load and CD4 cell counts). Additionally, dependencies between the clinical factors and the prior & current treatment of the patients are of interest. Discovering these dependencies might for example serve as a preliminary feature selection step for predicting the success of antiretroviral treatments for individual patients. This serves as our motivation for estimating the Markov Blankets of the clinical factors. The data used for this originates from the SystemsX.ch HIV-X project [Huldrych et al.] and the Swiss HIV Cohort Study [SHCS et al., 2010].
The thesis is divided into 6 chapters. After a short overview of existing related work, the general background required for the BMB and SA is given. On top of a basic overview of used statistical tools, we elaborate on the need for enforcing sparsity in the process of network estimation. In the third chapter, the BMB model itself as well as its extension are presented. This is mostly limited to a theoretic overview. Details regarding the implementation, as well as the setup for the artificially created data are discussed in the fourth chapter. Subsequently, we present the results obtained from the application on both synthetic data and HIV. Finally, a conclusion as well as an outlook for possible future work is given.

1.1 Contributions

With the general goals outlined, the contributions of this thesis can be summarized to the following points:

- Introduction of Simulated Annealing to Bayesian Markov Blanket estimation [Kaufmann et al., 2016]
- Evaluation and Comparison of BMB, SA and the Graphical LASSO on artificial data
- Experimental study regarding the modality of the underlying model
- Analysis of potential dependencies in the SystemsX.ch HIV-X cohort data [Huldrych et al., SHCS et al., 2010] with Markov Blanket estimation

2

Background and Related Work

In the following chapter the background necessary for the Bayesian Markov Blanket Estimation and its extension with Simulated Annealing is provided. First, the statistical methods used for the inference of our models are covered. This includes MCMC sampling for the estimation of the whole posterior distribution and Simulated Annealing for a MAP estimate. Subsequently, copulas and more specifically the semi parametric Gaussian copula are presented. Copulas allows generalizing the BMB and similar models to mixed data by introducing latent variables and relaxing the assumptions about the marginal distributions of the data. At last we introduce Gaussian Graphical Models and the matters of their estimation in a high-dimensional setting.

2.1 Markov Chain Monte Carlo

This section is mainly based on Andrieu et al. [2003], which provides an overview and introduction to MCMC methods.

Bayesian inference allows us to update our prior beliefs when observing data to produce our posterior belief. Let θ be the parameters of a statistical model and X the observed data distributed according to the (unknown) probability density function (pdf) $p(X)$. The Likelihood $p(X|\theta)$ then quantifies how 'likely' it is, that data X was created according to the statistical model parametrized by θ. Preexisting and arguably subjective beliefs we have about the parameters are captured in the prior pdf $p(\theta)$. When observing data X, the prior can then be updated to the posterior according to the well known Bayes' Theorem [Bishop, 2006, Chapter 1.2.3]:

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{p(X)} = \frac{p(X|\theta)p(\theta)}{\int_\Theta p(X|\theta')p(\theta')d\theta'}$$

The posterior pdf then expresses our uncertainty about θ, while incorporating both the observed data and our prior beliefs. Although the posterior can be given as a closed-form expression in special cases (e.g. by using conjugate priors), this is generally not the case. The main difficulty arises from the unknown $p(X)$. Although it can be calculated by marginalization, the integral usually offers no analytical solution, and numerical integration is computationally intractable for even moderately large parameter spaces. For tackling

this problem, MCMC methods can be used. MCMC techniques provide a framework for sampling from target distributions which can be evaluated up to a normalizing constant. As such, the posterior distribution can be empirically estimated without the need to explicitly compute the prohibitively large normalization.
Because the posterior need be known only up to a constant, knowing the likelihood $p(X|\theta)$ and the prior $p(\theta)$ is sufficient:

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{\int_\Theta p(X|\theta')p(\theta')d\theta'} \propto p(X|\theta)p(\theta)$$

Sampling from a target distribution is achieved by constructing homogeneous, 'irreducible and aperiodic Markov chains that have the target distribution as the invariant distribution' [Andrieu et al., 2003]. That is, the chain must not have any cycles (*aperiodicity*) and every state with non-zero probability must be reachable within a finite number of steps (*irreducibility*). A sufficient (but not necessary) condition for having a specific distribution $p(x)$ as the invariant distribution of a discrete Markov chain with transition matrix T is the *detailed balance* property ([Bishop, 2006, Chapter 11.2.1]):

$$p(x^{(i)})T(x^{(i-1)}|x^{(i)}) = p(x^{(i-1)})T(x^{(i)}|x^{(i-1)}) \tag{2.1}$$

By constructing a Markov chain that satisfies these conditions, a corresponding sampler can then be constructed. The empirical distribution of the samples drawn from the chain then converges *almost surely* to the target distribution, which in our case is the posterior distribution. But while convergence itself might be a.s. guaranteed for $n \to \infty$, the convergence rate strongly depends on the specific Markov chain used. So while fulfilling the necessary conditions might be sufficient in theory, a Markov chain with quick convergence is required in practice. In the following subsections we will present two well known MCMC techniques for sampling from a target distribution.

2.1.1 Metropolis Hastings

Metropolis Hastings (MH) is a MCMC method introduced by Hastings [1970]. While we do not directly make use of a Metropolis Hastings (MH) sampler in the presented models, MH can be viewed as a general case of other samplers. As we'll show in the following sections, this will be useful for extending a Gibbs Sampler for Simulated Annealing.
The MH procedure is based on drawing samples from a proposal function and then randomly accepting or rejecting them as draws from the target distribution. Let $p(x)$ be the invariant distribution we are interested in (known up to a constant) and $q(x^*|x)$ some proposal function. The proposal function suggests a new candidate for our sampler, given a current value. Unlike the Metropolis sampler [Metropolis et al., 1953], MH does not require the proposals to be symmetric.

Algorithm 1 Metropolis-Hastings Sampler

Initialize: $x^{(0)}$
for $i = 0$ to $N - 1$ **do**
  Sample $u \sim \mathcal{U}_{[0,1]}$
  Sample $x^* \sim q(x^*|x^{(i)})$
  if $u < \mathcal{A}(x^{(i)}, x^*) = \min\left\{1, \frac{p(x^*)q(x^{(i)}|x^*)}{p(x^{(i)})q(x^*|x^{(i)})}\right\}$ **then**
    $x^{(i+1)} = x^*$
  else
    $x^{(i+1)} = x^{(i)}$
  end if
end for

In Algorithm 1 the general sampler for which the Markov chain admits $p(x)$ as the invariant distribution is shown. By construction, the transition probabilities of the MH sampler satisfy Equation 2.1. The aperiodicity of the chain is given by the possibility of rejection in the sampler. For guaranteeing irreducibility, it need only be ensured that the proposal function $q(x * |x)$ has the same support as $p(x)$ (i.e. can lead to all states x^* that have non-zero probability $p(x^*)$). It should be noted, that $p(x)$ need be known only up to a constant. This follows from the fraction $\frac{p(x^*)q(x^{(i)}|x^*)}{p(x^{(i)})q(x^*|x^{(i)})}$ which is not affected by a constant factor in $p(x)$. While MH is a very general method, its performance largely depends on the proposal function. If the proposals are too conservative (i.e. if there are only very small changes compared to the previous sample), exploring the whole posterior will take a long time. Conversely a proposal function that jumps around a lot may suffer from a low acceptance ratio, as many improbable values are considered.

2.1.2 Gibbs Sampling

Gibbs sampling [Geman and Geman, 1984] is an MCMC method that can be applied if the target distribution is multivariate. Instead of the joint distribution, it relies on the individual conditional distributions and is especially useful if the conditionals are known and easy to sample from.

Let $p(\mathbf{x}) = p(x_1, \ldots, x_n)$ be the distribution we wish to sample from, where the conditionals

$$p(x_j|x_{-j}) = p(x_j|x_1, \ldots x_{j-1}, x_{j+1}, \ldots, x_n) \qquad \forall j \in \{1, \ldots, n\}$$

are known. By now iteratively drawing samples from the individual distributions while conditioning on the latest drawn samples of the other parameters, we can approximate the joint posterior $p(x_1, \ldots, x_n)$. This is convenient since the posterior conditionals can often be expressed in terms of a known closed form distribution (by using semi-conjugate priors), even if the full posterior cannot.

The Gibbs sampler can be viewed as a special case of Metropolis Hastings with the following distribution for the proposal function [Andrieu et al., 2003]:

$$q(x^*|x^{(i)}) = \begin{cases} p(x_j^*|x_{-j}^{(i)}) & \text{If } x_{-j}^* = x_{-j}^{(i)} \\ 0 & \text{otherwise} \end{cases}$$

Substituting this into Algorithm 1 and using $p(x) = p(x_j|x_{-j})p(x_{-j})$ leads to [Bishop, 2006,

p. 544]

$$\mathcal{A}(x^{(i)}, x^*) = \min\left\{1, \frac{p(x_j^*|x_{-j}^*)p(x_{-j}^*)p(x_j|x_{-j}^*)}{p(x_j|x_{-j})p(x_{-j})p(x_j^*|x_{-j})}\right\} = 1$$

Thus, the Gibbs sampler is a MH sampler with acceptance probability 1. This perspective can now be used for incorporating Simulated Annealing on top of a Gibbs sampler, for which the posterior conditionals are available in closed forms.

The general sampler is shown in Algorithm 2. Each iteration of it corresponds to a Gibbs sweep, where a new sample is drawn for each variable while conditioning on the most recent samples of the other variables. Thus, one draw from the joint distribution is obtained with each Gibbs sweep. It should be noted, that the order of the variables does not influence the sample distribution for a sufficiently large number of iterations.

Algorithm 2 Gibbs Sampler

Initialize: $x_{0,1:n}$
for $i = 0$ to $N - 1$ **do**
  Sample $x_1^{(i+1)} \sim p(x_1|x_2^{(i)}, x_3^{(i)}, \ldots, x_n^{(i)})$
  Sample $x_2^{(i+1)} \sim p(x_2|x_1^{(i+1)}, x_3^{(i)}, \ldots, x_n^{(i)})$

  $\vdots$

  Sample $x_j^{(i+1)} \sim p(x_j|x_1^{(i+1)}, x_{j-1}^{(i+1)}, x_{j+1}^{(i)}, \ldots, x_n^{(i)})$

  $\vdots$

  Sample $x_n^{(i+1)} \sim p(x_n|x_1^{(i+1)}, x_2^{(i+1)}, \ldots, x_{n-1}^{(i+1)})$
end for

2.1.3 Simulated Annealing

If we are only interested in a point estimation of the parameters, inferring the whole posterior distribution might not be the best solution. Unless the area surrounding the mode of the posterior has a high amount of probability mass, the sampler will mainly explore areas that are of low interest. Instead we can try to directly find the mode of the posterior, i.e. the MAP. Simulated Annealing [Kirkpatrick et al., 1983] is an optimization framework that can be used for finding the global maximum of a given function. The concept stems from metallurgy, where slowly cooling a previously heated material leads to it reaching an 'optimal' minimum energy state.

For applying SA in the context of MCMC sampling, we switch to a inhomogeneous Markov chain with the invariant distribution

$$p_i(x) \propto p^{1/T_i}$$

where T_i is the temperature at iteration i.

Initially the temperature is very high, such that large jumps between the drawn samples are possible (as the invariant distribution will be very flat). The temperature is then slowly decreased, such that the distribution becomes more peaked and less jumping is possible. Eventually the invariant $p^{1/T_i}(x)$ tends to $p^{\infty}(x)$ for $lim_{i\to\infty}T_i = 0$, which concentrates on

the global maxima. The concrete value of T_i and its speed of decrease are defined by a cooling schedule $f_T(i)$. Convergence to the set of global maxima has been shown to hold for logarithmic cooling schedules in discrete domains [Geman and Geman, 1984]. Unfortunately, a logarithmic cooling schedule is not a computationally feasible option for realistic problem sizes. In the literature, a range of cooling schedules has been analyzed (e.g. Abramson et al. [1999]), but to our knowledge there is not one distinct cooling method that works best in practice. In general, slower cooling of the system is preferred and the decrease in temperature should slow down with the temperature going towards 0. A simple method that is quite popular in practice is a geometric cooling schedule (see for example Andrieu and Doucet [2000], Yuan et al. [2004]).

Geometric Cooling Schedule (1)

Let T_0 be the initial temperature, $a \in \mathbb{R}_{>0}$ be a constant. Then the temperature T_i at iteration i is:

$$T_i = f_T(i) = T_0 \; a^i \qquad i = 0, 1, \ldots, n$$

Furthermore, as $\frac{1}{T}$ can never reach infinity in practice, one can set a (low) target temperature T_n, which is reached after n iterations. Adapting the geometric cooling schedule to reach a final temperature T_n leads to:

Geometric Cooling Schedule (2)

$$f_T(i) = T_0 \Big(\frac{T_n}{T_0}\Big)^{i/n} \qquad i = 0, 1, \ldots, n \tag{2.2}$$

The form of the geometric cooling schedule can be seen in Figure 2.1.

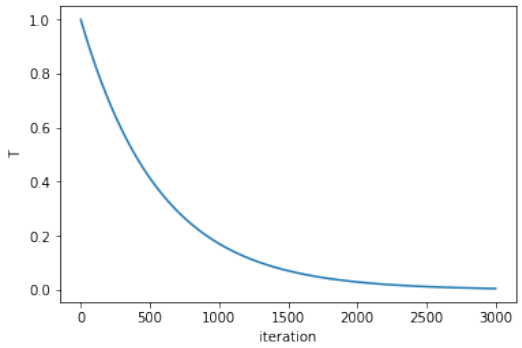


Figure 2.1: Geometric cooling schedule for $T_n = 0.005$, $T_0 = 1$ and $n = 3000$

At the target temperature we can continue sampling from the invariant distribution $p^{1/T_n}(x)$. If convergence was achieved (i.e. if the cooling was slow enough) samples from this target temperature will be close to the true mode with high probability. For estimating the final value, we can then for example look at statistics such as the median, mean or the credible intervals of the samples. The Annealing can be implemented by slightly altering the MH sampler, such that the invariant distribution at each step is $p^{1/T_i}(x)$, where the temperature T_i is set according to the cooling schedule $f_T(i)$ (see Algorithm 3).

Algorithm 3 Simulated Annealing: Metropolis-Hastings

Initialize: $x^{(0)}$, $T_0 = 1$
for $i = 0$ to $N - 1$ **do**
  $T_i = f_T(i)$

  Sample $u \sim \mathcal{U}_{[0,1]}$
  Sample $x^* \sim q(x^*|x^{(i)})$
  if $u < \mathcal{A}(x^{(i)}, x^*) = \min\left\{1, \frac{p(x^*)^{1/T_i} q(x^{(i)}|x^*)}{p(x^{(i)})^{1/T_i} q(x^*|x^{(i)})}\right\}$ **then**
    $x^{(i+1)} = x^*$
  else
    $x^{(i+1)} = x^{(i)}$
  end if
end for

Similar to subsection 2.1.2, it is possible to use the conditional distributions as proposals:

$$q(x^*|x^{(i)}) = \begin{cases} p(x_j^*|x_{-j}^{(i)})^{1/T_i} & \text{If } x_{-j}^* = x_{-j}^{(i)} \\ 0 & \text{otherwise} \end{cases}$$

As before, one can show that this leads to an acceptance probability of 1 when input into Algorithm 3. In consequence we can formulate the Simulated Annealing for a Gibbs sampler as shown in Algorithm 4.

Algorithm 4 Simulated Annealing: Gibbs Sampler

Initialize: $x_{0,1:n}$, $T_0 = 1$
for $i = 0$ to $N - 1$ **do**
  $T_i = f_T(i)$

  Sample $x_1^{(i+1)} \sim p(x_1|x_2^{(i)}, x_3^{(i)}, \ldots, x_n^{(i)})^{1/T_i}$
  Sample $x_2^{(i+1)} \sim p(x_2|x_1^{(i+1)}, x_3^{(i)}, \ldots, x_n^{(i)})^{1/T_i}$

  $\vdots$

  Sample $x_j^{(i+1)} \sim p(x_j|x_1^{(i+1)}, x_{j-1}^{(i+1)}, x_{j+1}^{(i)}, \ldots, x_n^{(i)})^{1/T_i}$

  $\vdots$

  Sample $x_n^{(i+1)} \sim p(x_n|x_1^{(i+1)}, x_2^{(i+1)}, \ldots, x_{n-1}^{(i+1)})^{1/T_i}$
end for

2.2 Semi-Parametric Gaussian Copula

In many problem settings it is common to have mixed data for which we cannot assume a specific (closed form) marginal distributions, while we still would like to assume a joint distribution of the data. However, a model that would for example be based on multivariate Gaussian data also relies on univariate Gaussian marginals. A solution for this is provided by Sklar's theorem [Sklar, 1959], which says that a p-dimensional joint distribution can be represented in two parts: The p marginal distributions and a *copula* function, which

describes the dependencies between the variables.

Sklar's Theorem

Theorem (extracted from Nadarajah et al. [2018])
Let F be a p-dimensional cumulative distribution function with marginal distribution functions $F_i, i = 1, \dots, p$. Then there exists a copula C such that

$$F(x_1, \dots, x_p) = C(F_1(x_1), \dots, F_p(x_p))$$

Conversely, for any univariate cumulative distribution functions $F_1, \dots, F_p$ and any copula C, the function F is a p-dimensional cumulative distribution function (cdf) with marginals $F_1, \dots, F_p$. Furthermore, if $F_1, \dots, F_p$ are continuous, then C is unique.

So with the use of copulas, it is possible to parametrize the marginal distributions separately from their joint distribution. Consequently the dependency structure of the data can be assumed to follow a specific joint distribution (e.g. a multivariate Gaussian) without having to consider the unknown marginals. It should be noted that Sklar's Theorem guarantees the uniqueness of the copula only where marginals are continuous.
Let $y_{i,j}$ be the i-th observation of variable $\mathbf{y}_j$, with $i = 1, \dots, n$ and $j = 1, \dots, p$. For modeling the copula, latent variables $\mathbf{z}_j$ that are distributed according to the assumed joint distribution are introduced. The observed variables $y_{i,j}$ can then be represented as the result of a non-decreasing transformation $g_j(z_{i,j})$ of the latent data. In the case of Gaussian copula models we assume that the latent variables are distributed according to a multivariate Normal distribution and can be modeled as follows [Hoff, 2009].

Gaussian Copula: Latent Normal Model

Let $\mathbf{z}_1, \dots, \mathbf{z}_n$ be *i.i.d.* (latent) random variables drawn from a p-variate normal distribution, with $\mathbf{C}$ being a $p \times p$ correlation matrix. Furthermore, let $\mathbf{y}_j$ be the observed variables with marginal cdf F_j. With $g_1, \dots, g_p$ being non-decreasing functions, the Gaussian copula model is

$$\mathbf{z}_1, \dots, \mathbf{z}_n | \mathbf{C} \overset{i.i.d.}{\sim} N(0, \mathbf{C}) \tag{2.3}$$

$$y_{i,j} = g_j(z_{i,j}) = F_j^{-1}(\Phi(z_{i,j}))$$

The aim now is to estimate $\mathbf{C}$, as we are interested in the interactions of the data. If the marginal distributions are known and continuous, the computation is trivial. The $z_{i,j}$ could be directly computed from the observed values. Inference of C could then be done by using standard estimation methods suitable for Gaussian data.
While many approaches do use parametric distributions for modeling the marginals, Hoff et al. [2007] introduce a semi-parametric Gaussian copula estimation dealing with both unknown and discrete marginals. It is semi-parametric in the sense that the model still parametrizes the joint pdf as a multivariate Gaussian, but does not make any assumptions about the marginals. This is achieved by basing the estimation of $\mathbf{C}$ on an 'extended rank

likelihood', which is based only on an order preserving statistic of the observations $y_{i,j}$, independent of the marginals F_j.

2.2.1 The extended rank likelihood

As it is given that the F_js are non-decreasing, the order of the observations of $\mathbf{y}_j$ will always be preserved for any transformation by F_j.
Let $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)^T$ and $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)^T$ then $\mathbf{Z}$ must lie in the set

$$\{\mathbf{Z} \in \mathbb{R}^{n \times p} : max\{z_{k,j} : y_{k,j} < y_{i,j}\} < z_{i,j} < min\{z_{k,j} : y_{i,j} < y_{k,j}\}\} \tag{2.4}$$

So the latent value $z_{i,j}$ for an observation $y_{i,j}$ is always constricted by the latent values corresponding to the two neighboring observations $max\{y_{k,j} : y_{k,j} < y_{i,j}\}$ and $min\{y_{k,j} : y_{i,j} < y_{k,j}\}$. Figure 2.2 provides a visualization for this relationship.

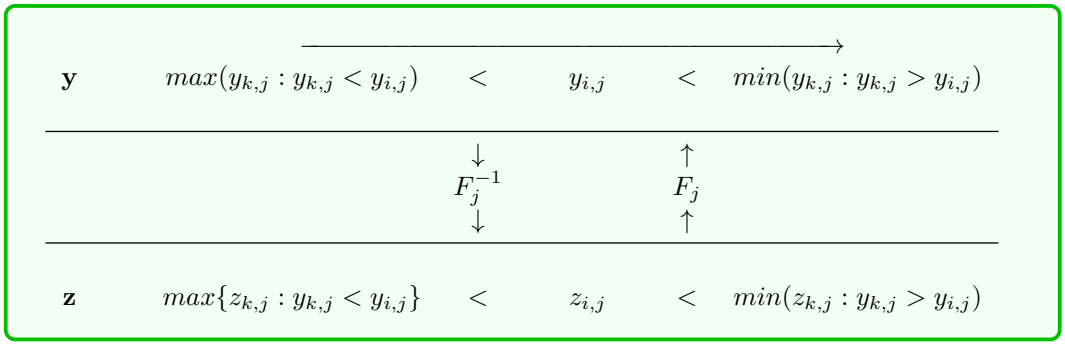


Figure 2.2: Relationship between latent values and observations.

Since $\mathbf{Z}$ always lies in this set for any observed $\mathbf{Y}$, we can write the likelihood as:

$$\begin{aligned} p(\mathbf{Y}|\mathbf{C}, F_1, \ldots, F_p) &= p(\mathbf{Z} \in D, \mathbf{Y}|\mathbf{C}, F_1, \ldots, F_p) \\ &= p(\mathbf{Z} \in D|\mathbf{C}, F_1, \ldots, F_p) p(\mathbf{Y}|\mathbf{Z} \in D, \mathbf{C}, F_1, \ldots, F_p) \end{aligned}$$

As $\{\mathbf{Z} \in D\}$ is independent of F_j due to the order preservation, it follows that

$$P(\mathbf{Z} \in D|\mathbf{C}, F_1, \ldots, F_p) = P(\mathbf{Z} \in D|\mathbf{C}) \tag{2.5}$$

Equation 2.5 is then called the 'extended rank likelihood'. Using this likelihood, a Gibbs sampler for the posterior

$$p(\mathbf{C}|\mathbf{Z} \in D) \propto p(\mathbf{C}) \times p(\mathbf{Z} \in D|\mathbf{C}) \tag{2.6}$$

can be constructed by choosing a semi-conjugate prior for $\mathbf{C}$.

2.2.2 Gibbs Sampler

Because of our model's assumption of a joint normal distribution in Equation 2.3, and the constriction due to the extended rank likelihood by Equation 2.4, the conditional posteriors

of $z_{i,j}$ follow a truncated normal distribution. Furthermore we can base the inference on the covariance matrix $\mathbf{V}$ instead of the correlation matrix $\mathbf{C}$ for easing the construction of a sampler. As the correlation matrix $\mathbf{C}$ can be calculated from $\mathbf{V}$, we will get an equivalent model. For the prior $p(\mathbf{V})$ an inverse Wishart prior with expected value $E[\mathbf{V}^{-1}] = \mathbf{V}_0^{-1}$ is suggested by Hoff et al. [2007].
This results in the model

$$\begin{aligned} V &\sim \mathcal{W}^{-1}(v_0, v_0\mathbf{V}_0) \\ \mathbf{C} &= \text{diag}(V)^{-\frac{1}{2}} V \text{diag}(V)^{-\frac{1}{2}} \\ \mathbf{z}_1 \ldots, \mathbf{z}_n &\overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{C}) \\ \tilde{z}_{i,j} &= z_{i,j} / \sqrt{\mathbf{V}_{j,j}} \\ y_{i,j} &= F_j^{-1}(\tilde{z}_{i,j}) \end{aligned}$$

with posterior conditionals

$$\begin{aligned} z_{i,j}|(\mathbf{V}, \mathbf{Z}_{-i,-j}, \mathbf{Z} \in D) \sim & \mathcal{TN}(\mu_{i,j}, \sigma_j, a_j, b_j) \\ \mu_{i,j} &= \mathbf{Z}_{i,-j}\Big(\mathbf{V}_{j,-j}\mathbf{V}_{-j,-j}^{-1}\mathbf{V}_{-j,j}\Big)^T \\ \sigma_j^2 &= \mathbf{V}_{j,j} - \mathbf{V}_{j,-j}\mathbf{V}_{-j,-j}^{-1}\mathbf{V}_{-j,j} \\ a_j &= \sup(z_{k,j} : y_{k,j} < y_{i,j}) \\ b_j &= \inf(z_{k,j} : y_{i,j} < y_{k,j}) \\ \mathbf{V}|(\mathbf{Z} \in D) \sim & \mathcal{W}^{-1}(v_0 + n, v_0\mathbf{V}_0 + \mathbf{Z}^T\mathbf{Z}) \end{aligned}$$

It should be noted that there are no bounds for missing values $y_{i,j}$ and the latent values $z_{i,j}$ are then imputed by using the (non-truncated) Gaussian distribution. In the present context we use a more efficient formulation suggested by Adametz et al. [2014]. It avoids matrix inversions in the inner-most loop by using a Wishart prior for sampling from the inverse covariance matrix, instead of an inverse Wishart prior for the covariance matrix. Furthermore, $\mathbf{Z}$ is initialized with its normal scores (see Rey and Roth [2012]):

$$z_{i,j} = \Phi^{-1}\Big(\frac{\text{rank}_j(y_{i,j})}{n_j + 1}\Big)$$

where n_j denotes the number of levels of variable j, i.e. the number of unique observed values of $\mathbf{y}_j$. In the following chapters it will be clear that this formulation works well in our context due to each of the BMBs Gibbs sweeps effectively being a draw from the precision matrix. The full sampler of this formulation for the semi-parametric copula is shown in Algorithm 5.

Algorithm 5 Gibbs Sampler: Semi-Parametric Gaussian Copula

Input: $n \times (p+q)$ data matrix $\mathbf{Y}$ containing observations from $p+q$ rvs
Set: $V \leftarrow I_{p+q}$, $V_0 \leftarrow I_{p+q}$, $v \leftarrow (p+q+1)$
Initialize: $Z = \left\{ z_{i,j} \leftarrow \Phi^{-1}\left(\frac{\text{rank}_i(y_{i,j})}{n_i+1}\right) \text{ for } j \in \{1,\dots,(p+q)\}, i \in \{1,\dots,n\} \right\}$
repeat
  for $j = 1,\dots,(p+q)$ **do**
    $\sigma_j^2 \leftarrow 1/V_{j,j}$
    for $y \in unique\{y_{1,j},\dots,y_{n,j}\}$ **do**
      Find lower bound $a \leftarrow \sup(z_{i,j} : y_{i,j} < y)$
      Find upper bound $b \leftarrow \inf(z_{i,j} : y < y_{i,j})$
      for each i such that $y_{i,j} = y$ **do**
        $\mu_{i,j} \leftarrow -\sigma_j^2 z_{i,-j} V_{-j,j}$
        Sample $z_{i,j} \sim \mathcal{TN}(\mu_{i,j}, \sigma_j^2, a, b)$
      end for
    end for
  end for
  Sample $V \sim \mathcal{W}_{p+q}(v+n, [V_0 + Z^T Z]^{-1})$
  Compute $\mathbf{C} = \left\{ \mathbf{C}_{i,j} \leftarrow (V_{i,j}^{-1} / \sqrt{(V^{-1})_{u,u}(V^{-1})_{j,j}}) \text{ for } (i,j) \in \{1,\dots,(p+q)\} \right\}$
  Compute $\tilde{Z} = \left\{ \tilde{z}_{i,j} \leftarrow z_{i,j} / \sqrt{(V^{-1})_{j,j}} \text{ for } j \in \{1,\dots,(p+q)\}, i \in \{1,\dots,n\} \right\}$
until converged

2.3 Graphical Models

Undirected Graphical Models provide a convenient framework for representing dependencies between random variables in the form of network graphs. In the literature they are also commonly referred to as Markov Random Fields. Let a graph $G = (V, E)$ be denoted by a set of nodes V and a set of edges E. Each node $v \in V$ then corresponds to a random variable X_v and each edge connects two nodes. Additionally, the graph fulfills the following properties [Murphy, 2013, Chapter 19.2]:

- **Global Markov property** Iff node v separates nodes s and t in the graph, s and t are conditionally independent given v:

$$v_A \perp\!\!\!\perp v_B | v_C$$

- **Local Markov property** A node v is conditionally independent of the remaining network given its immediate neighbors $N(v)$:

$$v \perp\!\!\!\perp V \setminus (N(v) \cup v) | N(v)$$

- **Pairwise Markov property** Two nodes v and t are conditionally independent given the remaining network if there is no direct edge between them:

$$v \perp\!\!\!\perp t | V \setminus s, t$$

In general the graph structure is unknown and we wish to estimate it from observed data. That is, we want to learn the dependencies inherent to the data. The following section provides a short introduction to Gaussian Graphical Models and highlights some of the difficulties that arise when estimating them from high-dimensional data.

2.3.1 Gaussian Graphical Models

This chapter mainly follows Chapter 9.2 & 9.3 of Hastie et al. [2015]. In the setting of GGMs, it is assumed that the p-dimensional data follows a Normal distribution:

$$\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \qquad \boldsymbol{\mu} \in \mathbb{R}^p \quad \boldsymbol{\Sigma} \in \mathbb{R}^{p \times p} \tag{2.7}$$

$$P(\mathbf{x})_{\boldsymbol{\mu},\boldsymbol{\Sigma}} = \frac{1}{(2\pi)^{p/2} \det(\boldsymbol{\Sigma})^{1/2}} \exp\Big(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\Big)$$

Alternatively the Gaussian can be represented in terms of the canonical parameters $\mathbf{W} \in \mathbb{R}^{p \times p}$ and $\boldsymbol{\gamma} \in \mathbb{R}^p$, where $\mathbf{W} = \boldsymbol{\Sigma}^{-1}$ is the precision matrix:

$$P(\mathbf{x})_{\boldsymbol{\gamma},\mathbf{W}} = \exp\Big(\sum_{s=1}^{p} \lambda_s x_s - \frac{1}{2} \sum_{s,t=1}^{p} w_{st} x_s x_t + \frac{1}{2} \log\det[\mathbf{W}/(2\pi)]\Big)$$

In this formulation the pdf factorizes according to the zero entries of $\mathbf{W}$. That is, if and exactly if (*iff*) $w_{st} = 0$ then x_s and x_t are conditionally independent of each other. Therefore the zero pattern of the precision matrix encodes the conditional independencies of our model. Combining this with the Markov properties of a graphical Model, a Graph can be constructed by connecting all pairs of nodes with non-zero entries in the precision matrix. Aside from the conditional independencies, a measure of dependence between the variables is of importance. It was shown by Lauritzen [1996] that the partial correlation between two variables given all the other variables can be written in terms of the precision matrix.

Partial Correlation Coefficient

Let $\mathbf{X} = (X^{(1)}, \ldots, X^{(p)})$ be Normal distributed with

$$\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \mathbf{W}^{-1})$$

and let $V = \{1, \ldots, p\}$.
Then the partial correlation coefficient $\rho_{i,j}$ between $X^{(i)}$ and $X^{(j)}$, given all the other variables is

$$\rho_{i,j|V\setminus\{i,j\}} = -\frac{\mathbf{W}_{i,j}}{\sqrt{\mathbf{W}_{i,i}\mathbf{W}_{j,j}}} \qquad i \neq j \tag{2.8}$$

[Lauritzen, 1996]

With this relationship established, it is clear that estimating the structure of a GGM is equivalent to finding precision matrix. This problem is also commonly known in classical statistics, namely as covariance selection (see for example Dempster [1972]).

Estimation

By going back to Equation 2.7 the model can be further simplified. As the interest lies only in the dependency structure, $\mu = \mathbf{0}$ (i.e. mean-free data) is assumed. The log likelihood of the observed data $X = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$ with $\mathbf{x}_i \in \mathbb{R}^n$ is then:

$$
\begin{aligned}
\mathcal{L}(\mathbf{W}; X) &= \log\left(\prod_{i=1}^{n}(2\pi)^{-p/2}\det(\mathbf{W})^{1/2}\exp\Big(-\frac{1}{2}\mathbf{x}_i^T\mathbf{W}\mathbf{x}_i\Big)\right) \\
&= \log\left((2\pi)^{-np/2}\det(\mathbf{W})^{n/2}\exp\Big(-\frac{1}{2}\sum_{i=1}^{n}\mathbf{x}_i^T\mathbf{W}\mathbf{x}_i\Big)\right) \\
&\propto \frac{n}{2}\log\det(\mathbf{W}) - \frac{1}{2}\sum_{i=1}^{n}\mathbf{x}_i^T\mathbf{W}\mathbf{x}_i \\
&= \frac{n}{2}\log\det(\mathbf{W}) - \frac{1}{2}\operatorname{tr}\Big(\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T\mathbf{W}\Big) \\
&\propto \log\det(\mathbf{W}) - \operatorname{tr}\Big(\mathbf{S}\mathbf{W}\Big) \qquad\qquad \mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T
\end{aligned}
$$

A simple and straightforward approach for estimating $\boldsymbol{W}$ would be (unregularized) Maximum Likelihood Estimation (MLE). As it turns out, the MLE is

$$\hat{\mathbf{W}}_{MLE} = \arg\max\{\log\det(\mathbf{W}) - \operatorname{tr}(\mathbf{S}\mathbf{W})\} = \mathbf{S}^{-1}$$

Although this MLE indeed converges to the true precision matrix as n goes towards infinity, it fails to exist for high dimensional data [Hastie et al., 2015, Chapter 9.3.1]. The cause for this is the rank-degeneracy of the sample covariance for $p > n$. This can be problematic for real world applications in the aforementioned areas, where very high dimensional data ($p \gg n$) is common. Furthermore we might want to estimate a GGM that is interpretable (i.e. a GMM corresponding to a network with few edges). This leads to the notion of estimating a sparse $\mathbf{W}$, as regularizing the likelihood would deal with both of these problems. For enforcing a sparse solution, a L_0 penalty (i.e. a penalty based on the number of non-zero edges) would seem to be a logical choice. But due to the highly non-convex nature of L_0, an approximation with the L_1 constraint is often preferred:

$$\mathbf{W} = \arg\max\{\log\det(\mathbf{W}) - \operatorname{tr}(\mathbf{S}\mathbf{W}) - \rho||\mathbf{W}||_1\} \tag{2.9}$$

This convex problem can be solved by various methods. A popular approach is the "Graphical Lasso" [Friedman et al., 2008] which solves it by using a first-order block wise coordinate descent. We will not go into further details of this approach, as we are interested in the Bayesian perspective of the problem.

2.3.2 Bayesian Graphical Lasso

It is well known that an L1 penalized maximum likelihood estimation is equivalent to a MAP with a double exponential prior [Tibshirani, 1996]. The aforementioned GLASSO estimator can therefore be seen as equivalent to a maximum posterior estimation with a double exponential prior on the non-diagonal entries of $\mathbf{W}$. Depending on the formulation, an exponential prior on the diagonal entries is additionally used.

Wang [2012] formulate it as follows:

$$p(\mathbf{x}_i|\mathbf{W}) = \mathcal{N}(\mathbf{0}, \mathbf{W}) \qquad i = 1, \ldots, n \tag{2.10}$$
$$p(\mathbf{W}|\lambda) = C^{-1} \prod_{i<j} \{DE(w_{ij}|\lambda)\} \prod_{i=1}^{p} \{EXP(w_{ii}|\lambda/2)\} 1_{\mathbf{W}\in M^+}$$

It should be noted that due to the constraint of $\mathbf{W}$ being positive-definite, the marginals of $\mathbf{W}$ do not follow a double-exponential distribution.
The prior Equation 2.10 can then be represented as a scale mixture of normals as introduced by West [1987].

$$p(\mathbf{w}|\tau, \lambda) = C_\tau^{-1} \prod_{i<j} \Big\{ \frac{1}{\sqrt{2\pi\tau_{ij}}} \exp(-\frac{w_{ij}^2}{2\tau_{ij}}) \Big\} \prod_{i=1}^{p} \Big\{ \frac{\lambda}{2} \exp(-\frac{\lambda}{2} w_{ii}) \Big\} 1_{\mathbf{W}\in M^+} \tag{2.11}$$

But the normals of the scale mixtures are not independent, given the scale parameters, due to the positive definite constraint. This leads to the intractable normalizing factor C_τ which is dependent on τ. The independence is necessary for a hierarchical representation similar to the Bayesian LASSO by Park and Casella [2008]. Wang [2012] fixes this issue by proposing the following mixing density:

$$p(\tau|\lambda) \propto C_\tau \prod_{i<j} \frac{\lambda^2}{2} \exp(-\frac{\lambda^2}{2}\tau_{ij})$$

With this mixture density, the C_τ cancel each other out in the marginals of $\mathbf{W}$. Wang [2012] then derives a (block) Gibbs sampler by partitioning $\mathbf{W}$, $\mathbf{S}$ and τ in a block-wise manner, with a subsequent reparameterization of the posterior conditionals using the Schur complement of $\mathbf{W}$. The BMB estimation presented by Kaufmann et al. [2016] is based on the model by Wang [2012]. The main difference is that the BMB exploits a block-wise conditional independence structure for constructing a sampler that is capable of reconstructing a subnetwork of interest efficiently.

3 Model

The estimation methods mentioned up to this point all estimate the whole network. One problem with this approach is that, for real world problems, the network size may be rather large. Especially so in the biomedical domain, where we might look at problems having only a few variables of interest (e.g. clinical factors such as the viral load of an HIV patient), though these might interact with a large number of other variables (e.g. known relevant mutations of patients). Instead of the whole network, we are often only concerned with the interactions that affect a few variables. More specifically, we want to infer the Markov Blanket of the few nodes. The Markov Blanket of a node is the set of nodes that when conditioned on, render its distribution independent of all the other nodes in the network. For GGMs (and other undirected graphical models) this is equivalent to the local Markov property. As such, the Markov Blanket of a node consists of the set of its direct neighbors in the graph (see section 2.3). An example is given in Figure 3.1.

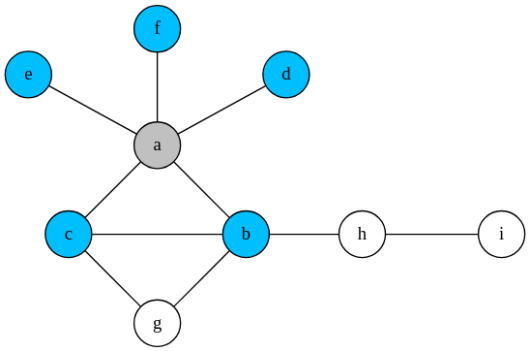


Figure 3.1: The Markov Blanket of node 'a' consists of all blue nodes, as they are its direct neighbors.

Kaufmann et al. [2016] present an efficient Gibbs sampler for inferring the Markov Blanket for a small number of variables. While following an approach similar to Wang [2012], a conditional independence property in the likelihood is used for separating the inference of the Markov Blanket from the rest of the network. Due to the reliance on computationally expensive MCMC methods for the estimation, this is a highly useful property and makes

the Bayesian approach to Covariance Selection feasible for bigger problem sizes (in terms of total network size).
In this Chapter we will provide an overview of Kaufmann et al. [2016] and its most important properties. The full model and its derivation are omitted for the sake of brevity and can be found in section A.2. Furthermore, we extend the sampler with Simulated Annealing for estimating the MAP instead of the whole posterior marginals.

3.1 Bayesian Markov Blanket

We stay in the setting of Gaussian Graphical Models. Let $\mathbf{W} \in \mathbb{R}^{(p+q)}$ be a symmetric positive definite matrix. Data $\mathbf{X}$ is assumed to be normally distributed with

$$\mathbf{X} \sim \mathcal{N}_{(p+q)}(0, \mathbf{W}^{-1})$$

Additionally, the observed data $X = (\mathbf{x_1}, \ldots, \mathbf{x_{(p+q)}})$ with $\mathbf{x}_i \in \mathbb{R}^n$ is ordered such that the first p columns correspond to query variables we are interested in, with q corresponding to the remaining variables. It is assumed that $p \ll q$. Furthermore, $\mathbf{S}$ and $\mathbf{W}$ are partitioned into block matrices according to p and q:

$$\mathbf{W} = \begin{array}{c} \begin{matrix} p & q \end{matrix} \\ \begin{pmatrix} \mathbf{W}_{11} & \mathbf{W}_{12} \\ \mathbf{W}_{12}^T & \mathbf{W}_{22} \end{pmatrix} \end{array} \begin{matrix} p \\ q \end{matrix} \qquad \mathbf{S} = \begin{array}{c} \begin{matrix} p & q \end{matrix} \\ \begin{pmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{12}^T & \mathbf{S}_{22} \end{pmatrix} \end{array} \begin{matrix} p \\ q \end{matrix}$$

The Markov Blanket of the query variables p is now given by the block $\mathbf{W}_{12}$. As such, the aim is to estimate the precision matrix block $\mathbf{W}_{12}$. It is important to note that $\mathbf{S}_{11}$ and $\mathbf{W}_{11}$ are much smaller than $\mathbf{S}_{22}$ and $\mathbf{W}_{22}$ respectively (given $p \ll q$).
The main motivation of the BMB sampler is a block-wise factorization of the likelihood that effectively allows decoupling the inference of $\mathbf{W}_{12}$ from the large matrix $\mathbf{W}_{22}$. By then constructing a prior admitting a similar factorization, the factorization can be propagated to the posterior which then reveals a block-wise conditional independence structure.

3.1.1 Likelihood

Let the sample Covariance be $\mathbf{S} = \mathbf{X}^T\mathbf{X}$. As shown in subsection 2.3.1, the likelihood is of form

$$p(\mathbf{S}|\mathbf{W}) \propto \det(\mathbf{W}^{n/2}) \exp \operatorname{tr}\Big(-\frac{1}{2}\mathbf{W}\mathbf{S}\Big) \tag{3.1}$$

which corresponds to a Wishart distribution:

$$p(\mathbf{S}|\mathbf{W}) \propto \mathcal{W}_{p+q}(n, \mathbf{W})$$

Kaufmann et al. [2016] then show that the likelihood factorizes as shown in Lemma 1.

Lemma 1 (Kaufmann et al. [2016])
Let $\mathbf{W}_{22.1} = \mathbf{W}_{22} - \mathbf{W}_{21}\mathbf{W}_{11}^{-1}\mathbf{W}_{12}$ *be the Schur complement of block* $\mathbf{W}_{11}$ *in* $\mathbf{W}$. *Then the likelihood of the covariance matrix factorizes in terms of* $\mathbf{W}$ *as follows:*

$$\mathcal{L}_S(\mathbf{W}) \propto \mathcal{L}_1(\mathbf{W}_{11}, \mathbf{W}_{12})\mathcal{L}_2(\mathbf{W}_{22.1})$$

It should be noted that when seen as a function of $\mathbf{W}$, the likelihood is no proper pdf [1] and the factorization is as such only a functional statement.
We will not prove this factorization here but rather assume it, for constructing a prior. Subsequently the factorization can be shown in the full posterior. This results in an arguably less convoluted derivation.

3.1.2 Prior

The prior for $\mathbf{W}$ has to ensure symmetry and positive-definiteness. Consequently, a Wishart prior is chosen. It is the natural conjugate prior to $p(\mathbf{S}|\mathbf{W})$ with its support being symmetric positive definite matrices. But the Wishart alone is insufficient, as sparsity is to be enforced. For this, a double exponential (DE) prior for the non-diagonal entries of $\mathbf{W}$ can be used. To admit a block-wise factorization similar to the likelihood, the DE prior is only placed on the non-diagonal entries of the $\mathbf{W}_{12}$ block. Analogous to Wang [2012], the DE prior is represented as a scale mixture of Gaussians (see subsection 2.3.2) with inverse-Gaussian distributed scale parameters $\mathbf{T} = \{t_{ij}\}$.

$$\begin{aligned} P(\mathbf{W}|\mathbf{T}) &= \mathcal{W}_{p+q}\big(p+q+1, \mathbf{I}\big) p(\mathbf{W}_{12}|\mathbf{T}) \\ &\propto \exp \operatorname{tr}\Big(-\frac{1}{2}\mathbf{W}\Big) \prod_{\substack{i=1,\dots,p \\ j=1,\dots,q}} \frac{1}{\sqrt{2\pi\mathbf{T}_{ij}}} \exp\Big(-\frac{(\mathbf{W}_{12})_{ij}^2}{2\mathbf{T}_{ij}}\Big) \end{aligned} \tag{3.2}$$

$$P(\mathbf{T}|\lambda) \propto \prod_{\substack{i=1,\dots,p \\ j=1,\dots,q}} \frac{\lambda^2}{2} \exp\Big(-\frac{\lambda^2}{2}\mathbf{T}_{ij}\Big)$$

3.1.3 Factorization of the Posterior

The joint distribution of the model arises from combining the compound prior in Equation 3.2 with the likelihood given in Equation 3.1.

$$\begin{aligned} p(\mathbf{W}, \mathbf{S}, \mathbf{T}|\lambda) =& p(\mathbf{W}_{11}, \mathbf{W}_{12}, \mathbf{W}_{22}, \mathbf{S}, \mathbf{T}|\lambda) \\ &\propto \det(\mathbf{W})^{\frac{n}{2}} \det(\mathbf{S})^{\frac{n-(p+q)-1}{2}} \\ &\quad \exp\left(-\frac{1}{2}\operatorname{tr}[\mathbf{W}\mathbf{S} + \mathbf{W}] - \frac{1}{2}\sum_{\substack{i=1,\dots,p \\ j=1,\dots,q}} \frac{(\mathbf{W}_{12})_{ij}^2}{\mathbf{T}_{ij}}\right) \\ &\quad \times \Big[\prod_{\substack{i=1,\dots,p \\ j=1,\dots,q}} \frac{1}{\sqrt{2\pi\mathbf{T}_{ij}}}\Big] p(\mathbf{T}|\lambda) \end{aligned}$$

Let the Schur complement of $\mathbf{W}_{11}$ in $\mathbf{W}$ be:

$$\mathbf{W}_{22.1} = \mathbf{W}_{22} - \mathbf{W}_{12}^T \mathbf{W}_{11}^{-1} \mathbf{W}_{12}$$

[1] The integral of the likelihood over $\mathbf{W}$ is not equal to one: $\int \mathcal{L}_S(\mathbf{W})\, \mathrm{d}\mathbf{W} \neq 1$

We can reparameterize the joint distribution by $(\mathbf{W}_{11}, \mathbf{W}_{12}, \mathbf{W}_{22.1})$, with the change of variable admitting a constant Jacobian:

$$J\big((\mathbf{W}_{11}, \mathbf{W}_{12}, \mathbf{W}_{22}) \to (\mathbf{W}_{11}, \mathbf{W}_{12}, \mathbf{W}_{22.1})\big) = \mathbf{1}$$

With the new parameterization, the determinants and traces can be written as:

$$\begin{aligned}
\det(\mathbf{W}) &= \det(\mathbf{W}_{11}) \det(\mathbf{W}_{22} - \mathbf{W}_{12}^T \mathbf{W}_{11}^{-1} \mathbf{W}_{12}) \\
&= \det(\mathbf{W}_{11}) \det(\mathbf{W}_{22.1})
\end{aligned}$$

$$\begin{aligned}
\operatorname{tr}(\mathbf{WS}) &= \operatorname{tr}\left[\mathbf{W}_{11}\mathbf{S}_{11} + \mathbf{W}_{12}\mathbf{S}_{21} + \mathbf{W}_{21}\mathbf{S}_{12} + \mathbf{W}_{22}\mathbf{S}_{22}\right] \\
&= \operatorname{tr}\left[\mathbf{W}_{11}\mathbf{S}_{11} + \mathbf{W}_{12}\mathbf{S}_{12}^T + \mathbf{W}_{12}^T\mathbf{S}_{12} + (\mathbf{W}_{22.1} + \mathbf{W}_{12}^T\mathbf{W}_{11}^{-1}\mathbf{W}_{12})\mathbf{S}_{22}\right] \\
&= \operatorname{tr}\left[\underline{\mathbf{W}_{11}\mathbf{S}_{11}} + \underline{\mathbf{W}_{12}\mathbf{S}_{12}^T + \mathbf{W}_{12}^T\mathbf{S}_{12}} + \underline{\mathbf{W}_{22.1}\mathbf{S}_{22}} + \underline{\mathbf{W}_{12}^T\mathbf{W}_{11}^{-1}\mathbf{W}_{12}\mathbf{S}_{22}}\right]
\end{aligned}$$

$$\begin{aligned}
\operatorname{tr}(\mathbf{W}) &= \operatorname{tr}(\mathbf{W}_{11}) + \operatorname{tr}(\mathbf{W}_{22}) \\
&= \operatorname{tr}(\underline{\mathbf{W}_{11}}) + \operatorname{tr}(\underline{\mathbf{W}_{22.1}}) + \operatorname{tr}(\underline{\mathbf{W}_{12}^T\mathbf{W}_{11}^{-1}\mathbf{W}_{12}})
\end{aligned}$$

When plugged into the joint distribution, this results in:

$$\begin{aligned}
p(\mathbf{W}_{11}, \mathbf{W}_{12}, \mathbf{W}_{22.1}, \mathbf{S}, \mathbf{T}|\lambda) \propto \\
&\det(\mathbf{W}_{11})^{n/2} \det(\mathbf{W}_{22.1})^{n/2} \det(\mathbf{S})^{\frac{n-(p+q)-1}{2}} \\
&\times \exp\Big(-\frac{1}{2}\operatorname{tr}\big[\underline{\mathbf{W}_{11}(\mathbf{S}_{11}+\mathbf{I})} + \underline{\mathbf{W}_{22.1}(\mathbf{S}_{22}+\mathbf{I})} + \\
&\qquad \underline{2(\mathbf{W}_{12}^T\mathbf{S}_{12})} + \underline{\mathbf{W}_{12}^T\mathbf{W}_{11}^{-1}\mathbf{W}_{12}(\mathbf{S}_{22}+\mathbf{I})}\big]\Big) \\
&\times \exp\Big(-\frac{1}{2}\sum_{\substack{i=1,\dots,p \\ j=1,\dots,q}} \frac{(\mathbf{W}_{12})_{ij}^2}{\mathbf{T}_{ij}}\Big) \quad \Big[\prod_{\substack{i=1,\dots,p \\ j=1,\dots,q}} \frac{1}{\sqrt{2\pi\mathbf{T}_{ij}}}\Big] \quad p(\mathbf{T}|\lambda)
\end{aligned}$$

$$\begin{aligned}
&= \det(\mathbf{W}_{11})^{n/2} \det(\mathbf{S})^{\frac{n-(p+q)-1}{2}} \\
&\times \exp\Big(-\frac{1}{2}\operatorname{tr}\big[\underline{\mathbf{W}_{11}(\mathbf{S}_{11}+\mathbf{I})} + \underline{2(\mathbf{W}_{12}^T\mathbf{S}_{12})} + \underline{\mathbf{W}_{12}^T\mathbf{W}_{11}^{-1}\mathbf{W}_{12}(\mathbf{S}_{22}+\mathbf{I})}\big]\Big) \\
&\times \exp\Big(-\frac{1}{2}\sum_{\substack{i=1,\dots,p \\ j=1,\dots,q}} \frac{(\mathbf{W}_{12})_{ij}^2}{\mathbf{T}_{ij}}\Big) \quad \Big[\prod_{\substack{i=1,\dots,p \\ j=1,\dots,q}} \frac{1}{\sqrt{2\pi\mathbf{T}_{ij}}}\Big] \quad p(\mathbf{T}|\lambda) \\
&\times \det(\mathbf{W}_{22.1})^{n/2} \exp\left(-\frac{1}{2}\operatorname{tr}\left[\underline{\mathbf{W}_{22.1}(\mathbf{S}_{22}+\mathbf{I})}\right]\right)
\end{aligned}$$

In this formulation it's clear to see that by conditioning on $\mathbf{S}_{22}$, $\mathbf{W}_{22.1}$ is (conditionally) independent of $\mathbf{W}_{11}$ and $\mathbf{W}_{12}$.

Conditional Independence Property

The posterior distribution of $\mathbf{W}$ factorizes as follows:

$$p(\mathbf{W}_{11}, \mathbf{W}_{12}, \mathbf{W}_{22.1}, \mathbf{T}|\mathbf{S}, \lambda) = p(\mathbf{W}_{11}, \mathbf{W}_{12}, \mathbf{T}|\mathbf{S}, \lambda)p(\mathbf{W}_{22.1}|\mathbf{S}, \lambda) \tag{3.3}$$

This implies that the posterior of $(\mathbf{W}_{11}, \mathbf{W}_{12})$ is conditionally independent of $\mathbf{W}_{22.1}$, given $\mathbf{S}$:

$$(\mathbf{W}_{11},\mathbf{W}_{12}) \perp\!\!\!\perp \mathbf{W}_{22.1}|\mathbf{S} \tag{3.4}$$

As a consequence, the Markov Blanket can be inferred independently of the prohibitively large $\mathbf{W}_{22}$, using only the joint posterior over $\mathbf{W}_{11}$, $\mathbf{W}_{12}$ and $\mathbf{T}$.

Joint Posterior Distribution

$$\begin{aligned} p(\mathbf{W}_{11}, \mathbf{W}_{12}, \mathbf{T}|\mathbf{S}, \lambda) \propto \quad & \det(\mathbf{W}_{11})^{n/2} \exp\Big(-\frac{1}{2} \operatorname{tr}\big[\mathbf{W}_{11}(\mathbf{S}_{11} + \mathbf{I})\big]\Big) \\ & \times \exp\Bigg(-\frac{1}{2}\operatorname{vec}(\mathbf{W}_{12})^T \Big[(\mathbf{S}_{22} + \mathbf{I}) \otimes \mathbf{W}_{11}^{-1} + \mathbf{D}\Big]\operatorname{vec}(\mathbf{W}_{12}) \\ & \qquad - \operatorname{vec}(\mathbf{W}_{12})^T \operatorname{vec}(\mathbf{S}_{12})\Bigg) \\ & \times \Big[\prod_{\substack{i=1,\dots,p \\ j=1,\dots,q}} \frac{1}{\sqrt{2\pi \mathbf{T}_{ij}}}\Big] p(\mathbf{T}|\lambda) \end{aligned} \tag{3.5}$$

Where $\mathbf{D} = \operatorname{diag}(\operatorname{vec}(\mathbf{T}))^{-1}$, i.e. $\mathbf{D}$ is a diagonal matrix with entries $(\mathbf{T}_{ij})^{-1}$.

3.1.4 Posterior Conditionals

It can now be shown that each of the posterior conditionals of $\mathbf{W}_{11}$, $\mathbf{W}_{12}$, $\mathbf{T}$ and $\mathbf{W}_{22.1}$ follows a known closed-form distribution. While we may not be interested in $\mathbf{W}_{22.1}$ (or $\mathbf{W}_{22}$) directly, it is required for embedding the BMB in the copula sampler.

The resulting distributions are shown in the following overview, while the full derivation can be found in subsection A.2.5. They are omitted at this point of the thesis for the sake of compactness.

Posterior Conditional of $\mathbf{W}_{12}$

1. $\mathbf{W}_{12}$ follows a multivariate Normal distribution:

$$\begin{aligned} \text{vec}(\mathbf{W}_{12})|\mathbf{W}_{11}, \mathbf{T}, \mathbf{S}, \lambda &\sim \mathcal{N}(-\mathbf{C}^{-1}\text{vec}(\mathbf{S}_{12}), \mathbf{C}^{-1}) \\ \mathbf{C} &= \Big[(\mathbf{S}_{22} + \mathbf{I}) \otimes \mathbf{W}_{11}^{-1} + \mathbf{D}\Big] \end{aligned} \tag{3.6}$$

   With $\mathbf{D} = \text{diag}(\text{vec}(\mathbf{T}))^{-1}$
   It should be noted, that as a special case for $\mathbf{T}_{ij} \to \infty$, i.e. $\mathbf{D} \to \mathbf{0}$ the posterior simplifies to:

$$\mathcal{N}\Big(- \text{vec}(\mathbf{W}_{11}\mathbf{S}_{12}(\mathbf{S}_{22} + \mathbf{I})^{-1}), (\mathbf{S}_{22} + \mathbf{I})^{-1} \otimes \mathbf{W}_{11}\Big)$$

2. $\mathbf{W}_{11}$ follows a Matrix Generalized Inverse Gaussian (MGIG) distribution:

$$\mathbf{W}_{11}|\mathbf{W}_{12}, \mathbf{T}, \mathbf{S}, \lambda \sim \mathcal{MGIG}_B\Big(\frac{n+p+1}{2},\ \mathbf{S}_{11} + \mathbf{I},\ \mathbf{W}_{12}(\mathbf{S}_{22} + I)\mathbf{W}_{12}^T\Big) \tag{3.7}$$

   (Using the parametrization of the MGIG as presented by Butler [1998])

3. $\mathbf{W}_{22.1}$ follows a Wishart distribution:

$$\mathbf{W}_{22.1}|\mathbf{S}, \lambda \sim \mathcal{W}\Big(n - \frac{p-1}{3}, \mathbf{S}_{22}\Big) \tag{3.8}$$

4. $(\mathbf{T}_{ij})^{-1}$ follow an inverse Gaussian distribution:

$$(\mathbf{T}_{ij})^{-1} \sim \mathcal{IG}\Big(\mu' = \sqrt{\frac{\lambda^2}{(\mathbf{W}_{12})_{ij}^2}}, \lambda' = \lambda^2\Big) \tag{3.9}$$

While $\mathbf{W}_{12}$, $\mathbf{W}_{22.1}$ and $\mathbf{T}$ follow distributions that can be easily sampled, the MGIG is more difficult. Bernadac [1995] shows that the MGIG can be represented and sampled as a continued fraction of Wishart distributed Matrices. But by its very nature, continued fractions of matrices are numerically unstable. Consequently, finding an alternative representation of the posterior that does not rely on the MGIG is favored.

3.1.4.1 Posterior Marginal of $\mathbf{W}_{11}$

The joint posterior of $\mathbf{W}_{11}$ and $\mathbf{W}_{12}$ can be split up using the chain rule:

$$p(\mathbf{W}_{11}, \mathbf{W}_{12}|\mathbf{T}, \mathbf{S}, \lambda) = p(\mathbf{W}_{12}|\mathbf{W}_{11}, \mathbf{T}, \mathbf{S}, \lambda)p(\mathbf{W}_{11}|\mathbf{T}, \mathbf{S}, \lambda)$$

From the previous section, it is known that the posterior conditional of $\mathbf{W}_{12}|\mathbf{W}_{11}, \mathbf{T}, \mathbf{S}, \lambda$ is proportional to the exponential part of a normal distribution (for details see subsection A.2.5). The marginal $p(\mathbf{W}_{11}|\mathbf{T}, \mathbf{S}, \lambda)$ can then be computed by integrating out $\mathbf{W}_{12}$. For this step we utilize the fact that integrating over the unnormalized part of a Gaussian distribution results in its normalization constant. This is given by construction, since a pdf

has to integrate to one.

$$\int \mathcal{N}(x;\mu,\mathbf{\Sigma})\,\mathrm{d}x = \int \frac{1}{\sqrt{(2\pi)^n \det(\mathbf{\Sigma})}} \exp\left(-\frac{1}{2}(x-\mu)^T\mathbf{\Sigma}^{-1}(x-\mu)\right)\mathrm{d}x = 1$$
$$\Rightarrow \int \exp\left(-\frac{1}{2}(x-\mu)^T\mathbf{\Sigma}^{-1}(x-\mu)\right)\mathrm{d}x = \sqrt{(2\pi)^n \det(\mathbf{\Sigma})}$$
$$\propto \sqrt{\det(\mathbf{\Sigma})}$$

Now plugging the joint posterior Equation 3.5 into the marginalization results in:

$$\begin{aligned}
p(\mathbf{W}_{11}|\mathbf{T},\mathbf{S},\lambda) &= \int p(\mathbf{W}_{11},\mathbf{W}_{12}|\mathbf{T},\mathbf{S},\lambda)\,\mathrm{d}\mathbf{W}_{12} \\
&\propto \int \det(\mathbf{W}_{11})^{n/2}\exp\Big(-\frac{1}{2}\operatorname{tr}\left[\mathbf{W}_{11}(\mathbf{S}_{11}+\mathbf{I})\right]\Big) \\
&\times \exp\left(-\frac{1}{2}\mathrm{vec}(\mathbf{W}_{12})^T\Big[(\mathbf{S}_{22}+\mathbf{I})\otimes\mathbf{W}_{11}^{-1}+\mathbf{D}\Big]\mathrm{vec}(\mathbf{W}_{12}) - \mathrm{vec}(\mathbf{W}_{12})^T\mathrm{vec}(\mathbf{S}_{12})\right)\mathrm{d}\mathbf{W}_{12} \\
&= \det(\mathbf{W}_{11})^{n/2}\exp\Big(-\frac{1}{2}\operatorname{tr}\left[\mathbf{W}_{11}(\mathbf{S}_{11}+\mathbf{I})\right]\Big) \\
&\times \int \underbrace{\exp\left(-\frac{1}{2}\mathrm{vec}(\mathbf{W}_{12})^T\Big[(\mathbf{S}_{22}+\mathbf{I})\otimes\mathbf{W}_{11}^{-1}+\mathbf{D}\Big]\mathrm{vec}(\mathbf{W}_{12}) - \mathrm{vec}(\mathbf{W}_{12})^T\mathrm{vec}(\mathbf{S}_{12})\right)}_{\propto\mathcal{N}\left(-\mathbf{C}^{-1}\mathrm{vec}(\mathbf{W}_{12}),\mathbf{C}^{-1}\right)}\mathrm{d}\mathbf{W}_{12} \\
&\propto \det(\mathbf{W}_{11})^{n/2}\exp\Big(-\frac{1}{2}\operatorname{tr}\left[\mathbf{W}_{11}(\mathbf{S}_{11}+\mathbf{I})\right]\Big)\det(\mathbf{C})^{\frac{1}{2}}
\end{aligned}$$

Posterior Marginal $\mathbf{W}_{11}|\mathbf{T},\mathbf{S},\lambda$

$$p(\mathbf{W}_{11}|\mathbf{T},\mathbf{S},\lambda) = \det(\mathbf{W}_{11})^{n/2}\exp\left(-\frac{1}{2}\operatorname{tr}\left[\mathbf{W}_{11}(\mathbf{S}_{11}+\mathbf{I})\right]\right)\det\left((\mathbf{S}_{22}+\mathbf{I})\otimes\mathbf{W}_{11}^{-1}+\mathbf{D}\right)^{\frac{1}{2}}$$

To our knowledge, this offers no possibility of reformulation into some known closed-form distribution, as the determinant arising from the normalization of $p(\mathbf{W}_{12}|\mathbf{W}_{11},\mathbf{T},\mathbf{S},\lambda)$ cannot be split up.
For the special case $\mathbf{T}_{ij}\to 0$, i.e. $\mathbf{D}\to\mathbf{0}$:

$$\mathbf{C}\overset{D\to\mathbf{0}}{\to}\left((\mathbf{S}_{22}+I)\otimes\mathbf{W}_{11}^{-1}\right)$$

$$\begin{aligned}
\det\left((\mathbf{S}_{22}+\mathbf{I})\otimes\mathbf{W}_{11}^{-1}\right)^{\frac{1}{2}} &\overset{\text{(516 Matr. Cookbook)}}{=} \left(\det(\mathbf{S}_{22}+\mathbf{I})^{\mathrm{rank}(\mathbf{W}_{11}^{-1})}\det(\mathbf{W}_{11}^{-1})^{\mathrm{rank}(\mathbf{S}_{22}+\mathbf{I})}\right)^{\frac{1}{2}} \\
&= \left(\det(\mathbf{S}_{22}+\mathbf{I})^{p}\det(\mathbf{W}_{11}^{-1})^{q}\right)^{\frac{1}{2}} \\
&= \det(\mathbf{S}_{22}+\mathbf{I})^{\frac{p}{2}}\det(\mathbf{W}_{11})^{-\frac{q}{2}}
\end{aligned}$$

As $\mathbf{W}_{11}$ and $\mathbf{S}_{22}$ are symmetric positive definite matrices, they have full rank p and q respectively. It then also follows that $\mathrm{rank}(\mathbf{S}_{22}+\mathbf{I}) = \mathrm{rank}(\mathbf{S}_{22})$.

Then

$$p(\mathbf{W}_{11}|\mathbf{T},\mathbf{S},\lambda) \overset{\mathbf{D}\to\mathbf{0}}{\propto} \det(\mathbf{W}_{11})^{n/2} \exp\Big(-\frac{1}{2}\operatorname{tr}\big[\mathbf{W}_{11}(\mathbf{S}_{11}+\mathbf{I})\big]\Big) \det(\mathbf{W}_{11})^{-q/2}$$
$$= \det(\mathbf{W}_{11})^{(n-q)/2} \exp\Big(-\frac{1}{2}\operatorname{tr}\big[\mathbf{W}_{11}(\mathbf{S}_{11}+\mathbf{I})\big]\Big)$$

Comparing the determinants to the Wishart Distribution($\mathbf{W}_{11}$ is known to be of size p):

$$\frac{1}{2}(df-p-1) = \frac{1}{2}(n-q)$$
$$\Rightarrow df = n-q+p-1$$

$$\Downarrow$$

$\mathbf{W}_{11}|\mathbf{T}$,$\mathbf{S}$,$\lambda$ for special case $T_{ij} \to \infty$

$$p(\mathbf{W}_{11}|\mathbf{T},\mathbf{S},\lambda) \overset{\mathbf{D}\to\mathbf{0}}{\propto} \mathcal{W}(n-q+p-1, \mathbf{S_{11.2}}) \tag{3.10}$$

3.1.5 Gibbs Sampler

With the posterior conditionals available, constructing a Gibbs sampler is straightforward. Furthermore, the sampler can be expanded by the semi-parametric Gaussian copula for relaxing the assumption of Gaussian marginals. For the extension, the Wishart draw of the precision matrix in the copula sampler (see Algorithm 5) is to be replaced by a draw of $\mathbf{W}$ from the BMB. At this step, the whole precision matrix and thus the $\mathbf{W}_{22}$ block is needed. Due to the factorization, $\mathbf{W}_{22.1}$ can be sampled from a Wishart (See Equation 3.4) and subsequently be used to calculate $\mathbf{W}_{22}$.

After obtaining the empirical posterior distribution with the Gibbs sampler, the zero-edges in the precision matrix are estimated by thresholding. As sparse networks are wanted, we estimate entries of $\mathbf{W}$ as zero if they are deemed insignificant according to the Credible Interval (i.e. if the Credible Interval (CI) of a certain level includes 0). Significant entries of the precision matrix are then estimated by the median.

The full Gibbs sampler (including thresholding and the copula) is shown in Algorithm 6.

Algorithm 6 Gibbs Sampler using Semi-Parametric Gaussian Copula

function DRAW_MB($\mathbf{W}, \mathbf{S}, \mathbf{T}, \lambda$)
  Sample ▷ BMB Gibbs sweep
    $(\mathbf{T}_{ij})^{-1} \sim \mathcal{IG}\Big(\mu' = \sqrt{\frac{\lambda^2}{(\mathbf{W}_{12})_{ij}^2}}, \lambda' = \lambda^2\Big)$
    $\mathbf{W}_{11}|\mathbf{W}_{12}, \mathbf{T}, \mathbf{S}, \lambda \sim \mathcal{MGIG}_B\Big(\frac{n+p+1}{2},\ \mathbf{S}_{11} + \mathbf{I},\ \mathbf{W}_{12}(\mathbf{S}_{22} + \mathbf{I})\mathbf{W}_{12}^T\Big)$
    $\text{vec}(\mathbf{W}_{12})|\mathbf{W}_{11}, \mathbf{T}, \mathbf{S}, \lambda \sim \mathcal{N}(-\mathbf{C}^{-1}\text{vec}(\mathbf{S}_{12}), \mathbf{C}^{-1})$
    $\mathbf{W}_{22.1}|\mathbf{S}, \lambda \sim \mathcal{W}\big(n - \frac{(p-1)}{3}, \mathbf{S}_{22}\big)$
  $\mathbf{W}_{22} \leftarrow \mathbf{W}_{22.1} + \mathbf{W}_{12}^T\mathbf{W}_{11}^{-1}\mathbf{W}_{12}$
  return $(\mathbf{W}, \mathbf{T})$
end function

function DRAW_LATENT_SCORES($\mathbf{X}, \mathbf{W}, \mathbf{Z}, \mathbf{S}$) ▷ Draw random scores
  for $j = 1, \ldots, (p+q)$ **do**
    $\sigma_j^2 \leftarrow 1/\mathbf{W}_{j,j}$
    for $x \in unique\{x_{1,j}, \ldots, x_{n,j}\}$ **do**
      Find lower bound $a \leftarrow \sup(z_{i,j} : x_{i,j} < x)$
      Find upper bound $b \leftarrow \inf(z_{i,j} : x < x_{i,j})$
      for each i such that $x_{i,j} = x$ **do**
        $\mu_{i,j} \leftarrow -\sigma_j^2 \mathbf{z}_{i,-j}\mathbf{S}_{-j,j}$
        Sample $z_{i,j} \sim \mathcal{TN}(\mu_{i,j}, \sigma_j^2, a, b)$
      end for
    end for
  end for
  return $(\mathbf{Z})$
end function

BMB Gibbs Sampler

Input:

$\mathbf{X}$	$n \times (p+q)$ data matrix containing observations from $(p+q)$ rvs
$\lambda > 0$	sparsity inducing hyper parameter
thresh_low	defines $[thresh\_low, 1 - thresh\_low]$ Credible Interval
NSCAN	total number of Gibbs sweeps
BURNIN	number of samples discarded for final estimation

Initialize:

$\mathbf{W}^{(0)} \leftarrow \lambda\mathbf{I}_{p+q}$
$\mathbf{Z}^{(0)} \leftarrow \Big\{z_{i,j} \leftarrow \Phi^{-1}\Big(\frac{\text{rank}_i(x_{i,j})}{n_i+1}\Big) \text{ for } j \in \{1, \ldots, (p+q)\}, i \in \{1, \ldots, n\}\Big\}$
$\mathbf{S}^{(0)} \leftarrow (\mathbf{Z}^{(0)})^T\mathbf{Z}^{(0)}$
Sample $\mathbf{W} \sim \mathcal{W}_{p+q}\Big(n, (S + W_0)^{-1}\Big)$

for $k = 1, \ldots, NSCAN$ **do**
  $\mathbf{Z}^{(k)} \leftarrow$ DRAW_LATENT_SCORES$(\mathbf{X}, \mathbf{W}^{(k-1)}, \mathbf{Z}^{(k-1)}, \mathbf{S}^{(k-1)})$ ▷ Draw random scores
  $\mathbf{S}^{(k)} \leftarrow (\mathbf{Z}^{(k)})^T\mathbf{Z}^{(k)}$ ▷ Update S
  $(\mathbf{W}^{(k)}, \mathbf{T}^{(k)}) \leftarrow$ DRAW_MB$(\mathbf{W}^{(k-1)}, \mathbf{S}^{(k)}, \mathbf{T}^{(k-1)}, \lambda)$ ▷ Draw BMB Posterior Conditional
end for

for i,j=1,...,(p+q) **do** ▷ Threshold and estimate
  CI $\leftarrow$ CREDIBLEINTERVAL$(\mathbf{W}_{i,j}^{(burnin:NSCAN)}, thresh_low)$
  if $0 \in$ CI **then**
    $\hat{\mathbf{W}}_{i,j} \leftarrow 0$
  else
    $\hat{\mathbf{W}}_{i,j} \leftarrow$ MEDIAN$(\mathbf{W}_{i,j}^{(burnin:NSCAN)})$
  end if
end for

Output:
  Estimated Precision Matrix $\hat{\mathbf{W}}$

3.2 Simulated Annealing

The Gibbs sampler of Kaufmann et al. [2016] allows us to investigate the whole posterior distribution of the BMB model, making it possible to compute CIs and other statistics of interest. But ultimately, the MAP estimate of the network is wanted, which corresponds to the point in the posterior having highest probability. Even though MCMC provides the full posterior probability, it is still an empirical one, which leads to the need of multivariate density estimation for finding the mode. However, it has been shown that multivariate density estimation is not feasible for high dimensional distributions due to the 'curse of dimensionality' [Scott, 1991]. In addition, most of the samples drawn will not explore the region of the mode unless it is closely surrounded by the majority of probability mass. Simulated Annealing solves this by allowing to approximately sample from the set of global maxima. Thus, it can be used for estimating the MAP.
It should be noted here, that the MAP is not guaranteed to lead to the 'best' estimate for any problem but actually depends on the posterior distribution at hand. If the posterior distribution is multi-modal or the majority of probability mass is far away from the mode, one might argue that the MAP is not representative for the distribution. On the other hand, if the posterior distribution is unimodal with a lot of mass surrounding it, the MAP might be preferable. While we cannot claim this to be true in general, experiments on artificial data indicate that the $\mathbf{W}_{12}$ block of interest is unimodal for sufficiently high lambda. We will expand on this in subsection 5.1.3.

3.2.1 The Joint MAP

First of all, it is important to distinguish which MAP we are actually estimating with the Simulated Annealing. The posterior distribution that is sampled from in the original BMB is $p(\mathbf{W}, \mathbf{T}|\mathbf{S}, \lambda)$, where λ denotes the sparsity inducing hyperparameter. Extending it with the semi-parametric Gaussian copula (section 2.2) leads to:

$$p(\mathbf{W}, \mathbf{T}, \mathbf{S}|\mathbf{Z} \in D, \lambda) \tag{3.11}$$

As MCMC sampling provides the full posterior distribution, inference on $\mathbf{W}$ and more specifically $\mathbf{W}_{12}$ is made on the marginals:

$$p((\mathbf{W})_{i,j}|\mathbf{Z} \in D, \lambda)$$

On the other hand, with Simulated Annealing the auxiliary variables $\mathbf{T}$ and $\mathbf{Z}$ influence the MAP problem and the following joint MAP is estimated:

$$\arg\max_{\mathbf{W},\mathbf{T},\mathbf{S}} p(\mathbf{W}, \mathbf{T}, \mathbf{S}|\mathbf{Z} \in D, \lambda)$$

It should be noted that the $\mathbf{W}$ resulting from this joint MAP estimate is different from the MAP of the marginal:

$$\arg\max_{\mathbf{W}} p(\mathbf{W}|\mathbf{Z} \in D, \lambda)$$

In our approach for the Annealing, the latent scores $\mathbf{Z}$ are first estimated by posterior mean with the Gibbs sampler:

$$\hat{Z} = E[\mathbf{Z}|\mathbf{Z} \in D, \lambda]$$

Subsequently, $\mathbf{S}$ is fixed to

$$\hat{S} = \hat{Z}^T \hat{Z}$$

As $\mathbf{W}_{22}$ was only necessary for the copula, the joint MAP we get when fixing the latent variables is:

$$(\hat{W}_{11}, \hat{W}_{12}, \hat{T}) = \underset{(\mathbf{W}_{11}, \mathbf{W}_{12}, \mathbf{T})}{\operatorname{argmax}} p(\mathbf{W}_{11}, \mathbf{W}_{12}, \mathbf{T} | \hat{S}, \lambda)$$

Aside from the unimodality mentioned before, it is now also not guaranteed that this joint MAP estimate leads to a sparse estimate of $\mathbf{W}_{12}$. A similar problem in context of SA for generalized linear regression has been addressed by Raman and Roth [2012]. Through a variational formulation the joint MAP was shown to be identical to the marginal posterior for certain choices of a sparsity parameter (note, that this parameter has a different effect than the λ). While we can offer no such proof, our experimental results suggest that the inferred networks are indeed sparse for sufficiently high λ.

3.2.2 Model

For implementing the SA, we make use of the method for extending a Gibbs sampler shown in subsection 2.1.3. As such, a cooling parameter T_i is introduced to the posterior conditionals:

$$p(x_i | x_{-i})^{1/T_i}$$

With distributions that are part of the exponential family, the cooling only results in a shift of the parameters while retaining the general form. This is given for all posterior conditionals of the BMB. The following subsections show the derivation of the respective cooled down distributions. To avoid confusion between the variables, the temperature parameter is from now on referred to as T_c.

3.2.2.1 Cooling of the Posterior Conditional $\mathbf{W}_{11}$

Let the Matrix Variate $\mathbf{X}$ be distributed according to the MGIG (using the parameterization of Butler [1998])

$$\begin{aligned} \boldsymbol{X} &\sim \mathcal{MGIG}_B(\lambda, A, B) \\ p(X) &\propto \det(X)^{\lambda - \frac{1}{2}(p+1)} \exp \operatorname{tr} \Big(-\frac{1}{2}(\mathbf{A}X + \mathbf{B}X^{-1}) \Big) \end{aligned}$$

Then $p(X)^{1/T}$ can be written as

$$
\begin{aligned}
& T_c \in \mathbb{R}_{>0} \\
p(X)^{\frac{1}{T_c}} &\propto det(X)^{\frac{\lambda-\frac{1}{2}(p+1)}{T_c}} exp[tr(-\frac{1}{2}(AX+BX^{-1}))]^{\frac{1}{T}} \\
&\propto det(X)^{\frac{\lambda}{T_c}-\frac{1}{2T_c}(p+1)} exp[tr(-\frac{1}{2T_c}(AX+BX^{-1}))] \\
&\propto det(X)^{\frac{\lambda}{T_c}-\frac{1}{2T_c}(p+1)} exp[tr(-\frac{1}{2}(\frac{A}{T_c}X+\frac{B}{T_c}X^{-1}))] \\
&\propto det(X)^{\left(\frac{\lambda}{T_c}-\frac{1}{2T_c}(p+1)+\frac{1}{2}(p+1)\right)-\frac{1}{2}(p+1)} exp[tr(-\frac{1}{2}(\frac{A}{T_c}X+\frac{B}{T_c}X^{-1}))] \\
&\propto det(X)^{\left(\frac{\lambda}{T_c}+\frac{p+1}{2}(1-\frac{1}{T_c})\right)-\frac{1}{2}(p+1)} exp[tr(-\frac{1}{2}(\frac{A}{T_c}X+\frac{B}{T_c}X^{-1}))] \\
&\propto p(Y) \\
\boldsymbol{Y} &\sim \mathcal{MGIG}_B\left(\left(\frac{\lambda}{T_c}+\frac{p+1}{2}(1-\frac{1}{T_c})\right),\ \frac{A}{T_c},\ \frac{B}{T_c}\right)
\end{aligned}
$$

Now we can express the posterior conditional of $\mathbf{W}_{11}$ as follows.

Cooled Posterior Conditional of $\mathbf{W}_{11}$

$$
p(\mathbf{W}_{11}|\mathbf{W}_{12},\mathbf{T},\mathbf{S},\lambda)^{1/T_c} \propto \mathcal{MGIG}_B\left(\frac{\frac{n}{T_c}+p+1}{2},\ \frac{\mathbf{S}_{11}+\mathbf{I}}{T_c},\ \frac{\mathbf{W}_{12}(\mathbf{S}_{22}+I)\mathbf{W}_{12}^T}{T_c}\right) \tag{3.12}
$$

3.2.2.2 Cooling of the Posterior Conditional $\mathbf{W}_{12}$

Let $\mathbf{X}$ be a normally distributed (vector) random variable:

$$
\begin{aligned}
\boldsymbol{X} &\sim N(\boldsymbol{\mu},\boldsymbol{\Sigma}) \\
p(\boldsymbol{x}) &= \left((2\pi)^k \det(\boldsymbol{\Sigma})\right)^{-\frac{1}{2}} exp[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})]
\end{aligned}
$$

Then introducing the cooling temperature T_c leads to:

$$
\begin{aligned}
& T_c \in \mathbb{R}_{>0} \\
p(x)^{\frac{1}{T_c}} &= \left((2\pi)^k \det(\Sigma)\right)^{-\frac{1}{2T_c}} exp[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})]^{\frac{1}{T_c}} \\
&\propto exp[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})]^{\frac{1}{T_c}} \\
&\propto exp[-\frac{1}{2T_c}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})] \\
&\propto exp[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\frac{\boldsymbol{\Sigma}^{-1}}{T_c}(\boldsymbol{x}-\boldsymbol{\mu})] \\
&\propto exp[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T(T_c\boldsymbol{\Sigma})^{-1}(\boldsymbol{x}-\boldsymbol{\mu})] \\
&\propto p(y) \\
Y &\sim \mathcal{N}(\boldsymbol{\mu},T_c\boldsymbol{\Sigma})
\end{aligned}
$$

Cooled Posterior Conditional of $\mathbf{W}_{12}$

$$p\big(\text{vec}(\mathbf{W}_{12})|\mathbf{W}_{11},\mathbf{T},\mathbf{S},\lambda\big)^{1/T_c} \propto \mathcal{N}(-\mathbf{C}^{-1}\text{vec}(\mathbf{S}_{12}), T_c\mathbf{C}^{-1}) \tag{3.13}$$
$$\mathbf{C} = \Big[(\mathbf{S}_{22}+\mathbf{I})\otimes\mathbf{W}_{11}^{-1}+\mathbf{D}\Big]$$

3.2.2.3 Cooling of the Posterior Conditional $\mathbf{T}$

Let the Inverse Gaussian (IG) distribution be denoted by

$$X \sim \mathcal{IG}(\mu,\rho)$$
$$p(x) = \Big(\frac{\rho}{2\pi x^3}\Big)^{1/2}\exp\Big[\frac{-\rho(x-\mu)^2}{2\mu^2 x}\Big]$$

and the Generalized Inverse Gaussian (GIG) distribution by

$$Y \sim \mathcal{GIG}(a,b,p)$$
$$p(y) = \frac{(a/b)^{p/2}}{2K_p(\sqrt{ab})}y^{(p-1)}\exp\Big[-\frac{1}{2}\big(ay+\frac{b}{y}\big)\Big]$$

The IG distribution can be written as a special case of the GIG:
With

$$Y \sim \mathcal{GIG}(a=\frac{\rho}{\mu^2}, b=\rho, p=-\frac{1}{2}) \tag{3.14}$$

we get

$$p(x \geq \boldsymbol{X}) = p(x \geq \boldsymbol{Y})$$

As a consequence, the cooled $\mathcal{IG}$ can be written in terms of the $\mathcal{GIG}$:

$$p(y) \propto y^{(p-1)}\exp\Big[-\frac{1}{2}\big(ay+\frac{b}{y}\big)\Big]$$

$$\begin{aligned} T_c &\in \mathbb{R}_{>0} \\ p(y)^{\frac{1}{T_c}} &\propto y^{\frac{(p-1)}{T_c}}\exp\Big[-\frac{1}{2T_c}\big(ay+\frac{b}{y}\big)\Big] \\ &\propto y^{(\frac{p-1}{T_c}+1)-1}\exp\Big[-\frac{1}{2}\Big(\frac{a}{T_c}y+\frac{b/T_c}{y}\Big)\Big] \\ &\propto p(z) \end{aligned}$$

$$Z \sim \mathcal{GIG}\Big(a'=\frac{a}{T_c}, b'=\frac{b}{T_c}, p'=\Big(\frac{p-1}{T_c}+1\Big)\Big)$$

Plugging in the parameters of the IG (Equation 3.14) then leads to:

$$Z \sim \mathcal{GIG}\Big(a'=\frac{\lambda/\mu^2}{T_c}, b'=\frac{\lambda}{T_c}, p'=\Big(-\frac{1.5}{T_c}+1\Big)\Big) \tag{3.15}$$

Finally, the cooled down IG posterior can be expressed as a GIG distribution.

Cooled Posterior Conditional of **T**

$$p\big((\mathbf{T}_{ij})^{-1}\big)^{1/T_c} \propto \mathcal{GIG}\Big(a = \frac{(\mathbf{W}_{12})_{ij}^2}{T_c}, b = \frac{\lambda^2}{T_c}, p = \big(-\frac{1.5}{T_c} + 1\big)\Big) \tag{3.16}$$

3.2.3 Cooling Schedule

For the sake of simplicity, a geometric cooling schedule was chosen. While different (non-adaptive) schedules were examined at first, we could not observe any significant differences. The cooling schedule used corresponds to the one presented in subsection 2.1.3:

Geometric Cooling Schedule (2)

$$f_T(i) = T_0\Big(\frac{T_n}{T_0}\Big)^{i/n} \qquad i = 0, 1, \ldots, n \tag{3.17}$$

3.2.4 Sampler

With the cooled down posterior conditionals and a cooling schedule available, the sampler can be constructed.
The sampler consists of three phases:

1. Running the Gibbs sampler for the BMB with the copula.

   This part serves both as burn-in and for the estimation of the latent scores. The sample covariance **S** is estimated from the mean of the copula draws.

2. The Annealing itself.

   Starting at temperature $T_0 = 1$, the system is slowly cooled down to the target temperature T_n according to the cooling function f_T.

3. Sampling from the cooled down system.

   At the final temperature T_n the sampling is continued for a small number of iterations, to sample from an area closely surrounding the mode. The CIs of these samples are then finally used for estimating **W**.

The complete Simulated Annealing process for the BMB is given in Algorithm 7.

Algorithm 7 Simulated Annealing of the Bayesian Markov Blanket

function DRAW_MB_COOLED($\mathbf{W}_{11}, \mathbf{W}_{12}, \mathbf{S}, \mathbf{T}, \lambda, T_c$)

Sample

$$(\mathbf{T}_{ij})^{-1} \sim \mathcal{GIG}\Big(a = \tfrac{(\mathbf{W}_{12})_{ij}^2}{T_c}, b = \tfrac{\lambda^2}{T_c}, p = \big(-\tfrac{1.5}{T_c} + 1\big)\Big)$$

$$\mathbf{W}_{11}|\mathbf{W}_{12}, \mathbf{T}, \mathbf{S}, \lambda \sim \mathcal{MGIG}_B\Big(\tfrac{\frac{n}{T_c}+p+1}{2},\ \tfrac{\mathbf{S}_{11}+\mathbf{I}}{T_c},\ \tfrac{\mathbf{W}_{12}(\mathbf{S}_{22}+I)\mathbf{W}_{12}^T}{T_c}\Big)$$

$$\text{vec}(\mathbf{W}_{12})|\mathbf{W}_{11}, \mathbf{T}, \mathbf{S}, \lambda \sim \mathcal{N}(-\mathbf{C}^{-1}\text{vec}(\mathbf{S}_{12}), \mathbf{C}^{-1})$$

return $(\mathbf{W}_{11}, \mathbf{W}_{12}, \mathbf{T})$

end function

BMB Simulated Annealing

Input:

$\mathbf{X}$	$n \times (p+q)$ data matrix containing observations from $(p+q)$ rvs
$\lambda > 0$	sparsity inducing hyper parameter
$thresh\_low$	defines $[thresh\_low, 1 - thresh\_low]$ Credible Interval
N_{BMB}, N_{SA}	number of BMB and SA iterations
N_{CD}	number of draws from cooled down posterior
$BURNIN$	number of samples discarded for final estimation
T_0, T_n	initial and final temperature

Initialize:

$$\mathbf{W}^{(0)} \leftarrow \lambda \mathbf{I}_{p+q}$$

$$\mathbf{Z}^{(0)} \leftarrow \Big\{ z_{i,j} \leftarrow \Phi^{-1}\Big(\tfrac{\text{rank}_i(x_{i,j})}{n_i+1}\Big) \text{ for } j \in \{1, \ldots, (p+q)\}, i \in \{1, \ldots, n\} \Big\}$$

$$\mathbf{S}^{(0)} \leftarrow (\mathbf{Z}^{(0)})^T \mathbf{Z}^{(0)}$$

Sample $\mathbf{W} \sim \mathcal{W}_{p+q}\Big(n, (S + W_0)^{-1}\Big)$

▷ BMB Gibbs Sampler (See Algorithm 6)

for $k = 1, \ldots, N_{BMB}$ **do**

$\mathbf{Z}^{(k)} \leftarrow$ DRAW_LATENT_SCORES$(\mathbf{X}, \mathbf{W}^{(k-1)}, \mathbf{Z}^{(k-1)}, \mathbf{S}^{(k-1)})$

$\mathbf{S}^{(k)} \leftarrow (\mathbf{Z}^{(k)})^T \mathbf{Z}^{(k)}$

$(\mathbf{W}^{(k)}, \mathbf{T}^{(k)}) \leftarrow$ DRAW_MB$(\mathbf{W}^{(k-1)}, \mathbf{S}^{(k)}, \mathbf{T}^{(k-1)}, \lambda)$

end for

$\hat{Z} \leftarrow \text{MEAN}(\mathbf{Z}^{(burnin:N)})$ ▷ Estimate and fix S

$\hat{S} \leftarrow \hat{Z}^T \hat{Z}$

▷ Simulated Annealing

for $k = (N_{BMB} + 1), \ldots, (N_{BMB} + N_{SA} - N_{CD})$ **do**

$T_c^{(k)} = f_T(k - N_{BMB} - 1)$

$(\mathbf{W}_{11}^{(k)}, \mathbf{W}_{12}^{(k)}, \mathbf{T}^{(k)}) \leftarrow$ DRAW_MB_COOLED$(\mathbf{W}_{11}^{(k-1)}, \mathbf{W}_{12}^{(k-1)}, \hat{S}, \mathbf{T}^{(k-1)}, \lambda, T_c^{(k)})$

end for

▷ Draw from cooled distribution

for $i = (k+1), \ldots, (N_{BMB} + N_{SA} + N_{CD})$ **do**

$(\mathbf{W}_{11}^{(i)}, \mathbf{W}_{12}^{(i)}, \mathbf{T}^{(i)}) \leftarrow$ DRAW_MB_COOLED$(\mathbf{W}_{11}^{(i-1)}, \mathbf{W}_{12}^{(i-1)}, \hat{S}, \mathbf{T}^{(i-1)}, \lambda, T_c^{(k)})$

end for

▷ Threshold and estimate

for i,j=1,...,(p+q) **do**

CI $\leftarrow$ CREDIBLEINTERVAL$(\mathbf{W}_{i,j}^{(k:END)}, thresh_low)$

if $0 \in$ CI **then**

$\hat{\mathbf{W}}_{i,j} \leftarrow 0$

else

$\hat{\mathbf{W}}_{i,j} \leftarrow$ MEDIAN$(\mathbf{W}_{i,j}^{(k:END)})$

end if

end for

Output:

Estimated Precision Matrix $\hat{\mathbf{W}}$

3.2.5 Relation between Sparsity Hyperparameters

For relating the regularization parameter ρ of the Graphical LASSO to the λ of the BMB we refer to the result of Chapter 4.5.3 in Kaufmann [2017]. By comparing the posterior of the BMB with the regularized likelihood of the GLASSO, it is concluded that:

$$\lambda = \frac{n}{2}\rho$$

where n refers to the number of data samples. Because of the same model underlying both the BMB and the SA, it is reasonable to assume that the respective λ are similar.

4 Implementation

In this chapter, details regarding the implementation of the samplers and the general setup of the artifical data used for evaluating the models are presented. The Bayesian Markov Blanket and its SA extension were both implemented in Python 3 with the SciPy stack [Jones et al., 2001–]. The code is openly available at

https://github.com/FabricioArendTorres/BayesianMarkovBlanket/

4.1 Sampling

While SciPys statistics library includes many distributions to sample from, more uncommon ones such as the GIG and MGIG are not covered. Furthermore, SciPy relies on inverse transform sampling as a default sampling procedure, which might not be optimal in terms of speed or numerical precision. For the truncated normal distribution new sampling approaches have emerged that improve those aspects. As such they are preferable for the implementation of an MCMC sampler. Last but not least, even for simple distributions such as the Normal distribution it can be beneficial to exploit the structure of the parameters for more efficient sampling procedures. In this section we will expand on the methods used for sampling from the respective distributions in the presented Gibbs sampler.

4.1.1 Normal Distribution: $\mathbf{W}_{12}$

We want to draw samples from:

$$\text{vec}(\mathbf{W}_{12})|\mathbf{W}_{11}, \mathbf{T}, \mathbf{S}, \lambda \sim \mathcal{N}(-\mathbf{C}^{-1}\text{vec}(\mathbf{S}_{12}), \mathbf{C}^{-1})$$
$$\mathbf{C} = \left[(\mathbf{S}_{22} + \mathbf{I}) \otimes \mathbf{W}_{11}^{-1} + \mathbf{D}\right]$$

A naive approach would require the inversion of $\mathbf{C}$, a $(pq) \times (pq)$ matrix, where q is known to be rather large. This should be avoided, as the inversion is computationally expensive. Instead we can sample from the transpose using Theorem 2.3.1.

Theorem 2.3.1 (Gupta and Nagar [2018], Chapter 2.3)
If $\mathbf{X} \sim \mathcal{N}_{p,n}(\mathbf{M}, \mathbf{\Sigma} \otimes \mathbf{\Psi})$ *then* $\mathbf{X}^T \sim \mathcal{N}_{n,p}(\mathbf{M}^T, \mathbf{\Psi} \otimes \mathbf{\Sigma})$

This leads to:

$$\text{vec}(\mathbf{W}_{12}^T)|\mathbf{W}_{11}, \mathbf{T}, \mathbf{S}, \lambda \sim \mathcal{N}(-(\mathbf{C}')^{-1}\text{vec}(\mathbf{S}_{12}^T), (\mathbf{C}')^{-1})$$
$$\mathbf{C}' = \left[\mathbf{W}_{11}^{-1} \otimes (\mathbf{S}_{22} + \mathbf{I}) + \mathbf{D}\right]$$

where $\mathbf{C}'$ now admits a block-wise structure that can be exploited for an efficient Cholesky decomposition (as shown by Kaufmann et al. [2016]).
Let $\mathbf{Q} = \mathbf{W}_{11}^{-1}$ and $\mathbf{D} = \text{diag}(\mathbf{D}_1, \ldots, \mathbf{D}_p)$. then:

$$\mathbf{C}' = \begin{bmatrix} Q_{11}(\mathbf{S}_{22} + \mathbf{I}) + \mathbf{D}_1 & \ldots & Q_{1j}(\mathbf{S}_{22} + \mathbf{I}) & \ldots & Q_{1p}(\mathbf{S}_{22} + \mathbf{I}) \\ \vdots & & & & \vdots \\ Q_{i1}(\mathbf{S}_{22} + \mathbf{I}) & & \ddots & & Q_{ip}(\mathbf{S}_{22} + \mathbf{I}) \\ \vdots & & & & \vdots \\ Q_{p1}(\mathbf{S}_{22} + \mathbf{I}) & \ldots & Q_{pj}(\mathbf{S}_{22} + \mathbf{I}) & \ldots & Q_{pp}(\mathbf{S}_{22} + \mathbf{I}) + \mathbf{D}_p \end{bmatrix}$$

A block-wise Cholesky decomposition of the form

$$\mathbf{C}' = \mathbf{L}\mathbf{L}^T$$

is then given by

$$\mathbf{L}^T = \begin{bmatrix} \mathbf{K}_1^{\frac{1}{2}} & Q_{12}\mathbf{B}_1 & \cdots & Q_{1j}\mathbf{B}_1 & \cdots & Q_{1p}\mathbf{B}_1 \\ \mathbf{0} & \mathbf{K}_2^{\frac{1}{2}} & \cdots & Q_{2j}\mathbf{B}_2 & \cdots & Q_{2p}\mathbf{B}_2 \\ \vdots & \vdots & \ddots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{K}_j^{\frac{1}{2}} & \cdots & Q_{jp}\mathbf{B}_j \\ \vdots & \vdots & & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \cdots & \mathbf{K}_p^{\frac{1}{2}} \end{bmatrix}$$

with

$$\mathbf{K}_i = \left(Q_{ii}(\mathbf{S}_{22} + \mathbf{I}) + \mathbf{D}_i\right)$$
$$\mathbf{B}_i = \mathbf{K}_i^{-\frac{1}{2}}(\mathbf{S}_{22} + \mathbf{I})$$

Non-diagonal entries only differ by a (scalar) factor Q_{ij}, which allows for an efficient computation by reusing the $\mathbf{B}_i$. The $\mathbf{K}_i^{\frac{1}{2}}$ can be calculated by a naive Cholesky decomposition as $\mathbf{K}_i$ are symmetric positive definite.
For drawing from the Normal distribution we can simply draw from a standard Normal Distribution and then transform it by multiplying with the standard deviation and adding the mean [Gentle, 2009, Chapter 7.4]. In our context this corresponds to Algorithm 8, where the mean and standard deviation are calculated from the above mentioned block-wise Cholesky decomposition.

Algorithm 8 Draw Sample from $\mathcal{N}\big(-(\mathbf{C}')^{-1}\text{vec}(\mathbf{S}_{12}^T), \mathbf{C}'^{-1}\big)$

function DRAW_W12($\mathbf{W}_{11}, \mathbf{S}_{22}, \mathbf{D}$)

Compute $\mathbf{L} \leftarrow BlockChol(\mathbf{W}_{11}^{-1}, \mathbf{S}_{22}, \mathbf{D})$

Solve for y: $\mathbf{L}y = -\text{vec}(\mathbf{S}_{12}^T)$ ▷ using forward substitution
Solve for μ: $\mathbf{L}^T\mu = y$ ▷ using backward substitution

Sample $r \sim \mathcal{N}(0, \mathbf{I}_{pq})$
Solve for b: $\mathbf{L}^T b = r$ ▷ using backward substitution
return $(b + \mu)$ ▷ $(b+\mu) \sim \mathcal{N}\big(-(\mathbf{C}')^{-1}\text{vec}(\mathbf{S}_{12}^T), \mathbf{C}'^{-1}\big)$
end function

4.1.2 Matrix Generalized Inverse Gaussian: $\mathbf{W}_{11}$

For the $\mathbf{W}_{11}$ block we have to sample from

$$\mathbf{W}_{11}|\mathbf{W}_{12}, \mathbf{T}, \mathbf{S}, \lambda \sim \mathcal{MGIG}_B\Big(\lambda' = \frac{n+p+1}{2},\ \mathbf{A} = \mathbf{S}_{11} + \mathbf{I},\ \mathbf{B} = \mathbf{W}_{12}(\mathbf{S}_{22} + I)\mathbf{W}_{12}^T\Big)$$

Bernadac [1995] has shown that the MGIG can be represented as a continued fraction of Wishart random variables. Similar to Kaufmann [2017] we make use of this result for the sampling process.

Theorem 5.1 Bernadac [1995]. *Let* $\mathbf{X}, \mathbf{Y}_1, \mathbf{Y}_2$ *be independent random variables, such that*

$$\mathcal{L}(\mathbf{Y}_1) = \mathcal{W}_p(2k, \mathbf{B}'^{-1}) \qquad \mathcal{L}(\mathbf{Y}_2) = \mathcal{W}_p(2k, \mathbf{A}'^{-1}) \qquad k > \frac{p-1}{2}$$

$$\mathcal{L}(\mathbf{X}) = \mathcal{MGIG}_B\big(-k, \mathbf{A}', \mathbf{B}'\big)$$

$$\textit{iff} \quad \mathcal{L}(\mathbf{X}) = \mathcal{L}((\mathbf{Y}_1 + (\mathbf{Y}_2 + \mathbf{X})^{-1})^{-1})$$

However, the condition for λ' does not hold in our given MGIG distribution:

$$k = -\frac{n+p+1}{2} < \frac{p-1}{2} \qquad \lightning$$

It is important to note at this point, that the relation between the degrees of freedom $-\lambda' = 2df$ arises from comparing the determinant of the Wishart to the determinant of the MGIG reparameterized by its inverse in Equation 4.1.

Proposition 3.4 Bernadac [1995].

$$\mathcal{L}(\mathbf{X}) = \mathcal{MGIG}_B(\lambda', \mathbf{A}', \mathbf{B}') \Leftrightarrow \mathcal{L}(\mathbf{X}^{-1}) = \mathcal{MGIG}_B(-\lambda', \mathbf{B}', \mathbf{A}') \tag{4.1}$$

For addressing the constraint issue, the MGIG is written in terms of the parameterization by Letac (similar to the GIG formulation in Letac and Seshadri [1983]):

$$\mathbf{X} \sim \mathcal{MGIG}_L(n', \mathbf{A}, \mathbf{B})$$

$$p(\mathbf{X}) \propto \det(\mathbf{X})^{-n'-1} \exp \text{tr}\Big(-\frac{1}{2}(\mathbf{AX} + \mathbf{BX}^{-1})\Big) \tag{4.2}$$

with the condition now being [Kaufmann, 2017, Appendix 4]:

$$n' > \frac{p-1}{2}$$

A change of variable to its inverse with $\mathbf{W} = \mathbf{X}^{-1}$ with Jacobian $det(\mathbf{W})^{-(p+1)}$ leads to:

$$\begin{aligned} p(\mathbf{W}) &\propto \det(\mathbf{W}^{-1})^{-n'-1} \exp \operatorname{tr}\Big(-\frac{1}{2}(\mathbf{A}\mathbf{W}^{-1} + \mathbf{B}\mathbf{W})\Big) \\ &= \det(\mathbf{W})^{n'-p} \exp \operatorname{tr}\Big(-\frac{1}{2}(\mathbf{A}\mathbf{W}^{-1} + \mathbf{B}\mathbf{W})\Big) \end{aligned} \tag{4.3}$$

If the posterior conditional is now expressed in terms of Equation 4.3, the corresponding inverse formulated as in Equation 4.2 fulfills the constraint and we can sample from this inverse.
With the posterior conditional being (see subsection A.2.5 for details)

$$p(\mathbf{W}_{11}|\mathbf{W}_{12}, \mathbf{T}, \mathbf{S}, \lambda) \propto \det(\mathbf{W}_{11})^{n/2} \exp\Big(-\frac{1}{2}\operatorname{tr}\left[(\mathbf{S}_{11}+\mathbf{I})\mathbf{W}_{11} + \mathbf{W}_{12}(\mathbf{S}_{22}+\mathbf{I})\mathbf{W}_{12}^T\mathbf{W}_{11}^{-1}\right]\Big)$$

a comparison with Equation 4.3 gives

$$\begin{aligned} A &= \mathbf{W}_{12}(\mathbf{S}_{22}+\mathbf{I})\mathbf{W}_{12}^T \\ B &= \mathbf{S}_{11} + \mathbf{I} \\ \frac{n}{2} &= n' - p \\ \Rightarrow n' &= \frac{n}{2} + p \end{aligned}$$

So the inverse of our posterior conditional is

$$\mathbf{W}_{11}^{-1}|\mathbf{W}_{12}, \mathbf{T}, \mathbf{S}, \lambda \sim \mathcal{MGIG}_L\Big(n' = \frac{n}{2} + p, A = (\mathbf{W}_{12}(\mathbf{S}_{22}+\mathbf{I})\mathbf{W}_{12}^T), B = (\mathbf{S}_{11}+\mathbf{I})\Big)$$

For this inverse, the required constraint is satisfied:

$$n' = \frac{n}{2} + p > \frac{(p-1)}{2}$$

For sampling from the inverse, the determinant of this MGIG is compared to the determinant of the Wishart distribution.

$$\begin{aligned} n' - 1 = \frac{n}{2} + p - 1 &= \frac{df - p - 1}{2} \\ \Rightarrow df &= n + 3p - 1 \end{aligned}$$

Finally, a sample can be drawn by the following continued fraction

$$X = \Big((Y_1)_1 + \Big((Y_2)_1 + \Big((Y_1)_2 + \Big((Y_2)_2 + \dots\Big)^{-1}\Big)^{-1}\Big)^{-1}\Big)^{-1}$$

$$(\mathbf{Y}_1)_i \sim \mathcal{W}_p\Big(n+3p-1, (\mathbf{W}_{12}(\mathbf{S}_{22}+\mathbf{I})\mathbf{W}_{12}^T)^{-1}\Big) \qquad (\mathbf{Y}_2)_i \sim \mathcal{W}_p\Big(n+3p-1, (\mathbf{S}_{11}+\mathbf{I})^{-1}\Big)$$

and then inverting the result:

$$\mathbf{W}_{11} = X^{-1}$$

As we are drawing the inverse of the r.v. we are interested in, simply omitting the last inversion in the continued fraction leads to the desired result.

Simulated Annealing

In case of the Simulated Annealing, the cooled posterior conditional is

$$p(\mathbf{W}_{11}|\mathbf{W}_{12},\mathbf{T},\mathbf{S},\lambda)^{1/T_c} \propto \det(\mathbf{W}_{11})^{n/(2T_c)} \exp\left(-\frac{1}{2}\operatorname{tr}\left[\frac{(\mathbf{S}_{11}+\mathbf{I})}{T_c}\mathbf{W}_{11}+\frac{\mathbf{W}_{12}(\mathbf{S}_{22}+\mathbf{I})\mathbf{W}_{12}^T}{T_c}\mathbf{W}_{11}^{-1}\right]\right)$$

Consequently the comparison with the inverted MGIG in Equation 4.3 is slightly different:

$$\begin{aligned} A &= \frac{\mathbf{W}_{12}(\mathbf{S}_{22}+\mathbf{I})\mathbf{W}_{12}^T}{T_c} \\ B &= \frac{\mathbf{S}_{11}+\mathbf{I}}{T_c} \\ \frac{n}{2T_c} &= n' - p \\ \Rightarrow n' &= \frac{n}{2T_c} + p \end{aligned}$$

Note, that the constraint is still satisfied for any $T_c > 0$:

$$n' = \frac{n}{2T_c} + p > \frac{(p-1)}{2}$$

Comparing the determinant with the Wishart:

$$\begin{aligned} n' - 1 = \frac{n}{2T_c} + p - 1 &= \frac{df - p - 1}{2} \\ \Rightarrow df &= \frac{n}{T_c} + 3p - 1 \end{aligned}$$

Finally, a sample can be drawn by the following continued fraction

$$X = \Big((Y_1)_1 + \Big((Y_2)_1 + \Big((Y_1)_2 + \Big((Y_2)_2 + \dots\Big)^{-1}\Big)^{-1}\Big)^{-1}\Big)^{-1}$$

$$(\mathbf{Y}_1)_i \sim \mathcal{W}_p\left(\frac{n}{T_c}+3p-1, \Big(\frac{\mathbf{W}_{12}(\mathbf{S}_{22}+\mathbf{I})\mathbf{W}_{12}^T}{T_c}\Big)^{-1}\right) \qquad (\mathbf{Y}_2)_i \sim \mathcal{W}_p\left(\frac{n}{T_c}+3p-1, \Big(\frac{\mathbf{S}_{11}+\mathbf{I}}{T_c}\Big)^{-1}\right)$$

and then inverting the result:

$$\mathbf{W}_{11} = X^{-1}$$

4.1.3 Truncated Normal Distribution

The inverse transform sampling implemented in SciPy fails and returns infinity for bounds that are far away from the mean. Dependent on the data, this can break the copula Gibbs sampler, so an alternative method is required. In the original implementation of the BMB by Kaufmann et al. [2016], the rejection sampler by Geweke [1991] was used. We follow a more recent approach presented by Chopin [2011]. Compared to the rejection sampler, it offers slightly higher precision and speed. Furthermore it does not require a table setup for sampling values from a new set of parameters. In context of the copula, this is obviously a useful property as the scores are sampled from a new set of parameters each iteration.

4.1.4 Generalized Inverse Gaussian

For the GIG we utilize the sampler of the R package *GIGrvg* [2]. It is based on three different algorithms, each of which has performance advantages for specific ranges of the parameter values. The algorithms have been introduced by Hörmann and Leydold [2014] and Lehner [1989] (which is a version of Dagpunar [1989] with a faster setup).
Alternative approaches may be faster for sampling multiple values from fixed parameters, but require a much longer setup. As mentioned before, this is disadvantageous in our setting. The code was included by wrapping a C++ implementation[3] of the R code in Cython. It is openly available via

https://github.com/FabricioArendTorres/gig

4.2 Artificial Data: Setup and Metrics

For evaluating the performance of the different network reconstruction methods, a data set that reflect the properties of the networks encountered in real applications is required.
We base the creation of the artificial data on the setup of Kaufmann et al. [2016], which will be described in the subsequent section. For measuring how well the networks are reconstructed, a simple approach based on counting correctly predicted edges is followed.

4.2.1 Data Setup

Each artificial data set is created by drawing 1000 samples from a zero-mean normal distribution $\mathcal{N}_{p+q}(\mathbf{0}, \mathbf{W}^{-1})$ with $p = 10$ and $q = 90$. The structure of $\mathbf{W}$ is created according to a beta-binomial model. That is, we first draw from the beta distribution

$$\pi_i \sim Beta(\alpha = 0.01, \beta = 1) \qquad i = 1, \ldots, (p+q)$$

and then use the π for drawing from a binomial

$$n_{ij} \sim Bin(n = 1, p = \pi_i) \qquad j = 1, \ldots, (p+q)$$

where $n_{ij} = 1$ indicates that $\mathbf{W}_{ij}$ is non-zero and vice-versa. As $E[\pi] = \frac{\alpha}{(\alpha+\beta)} = \frac{1}{101}$, the resulting network will be mostly sparse (i.e. have few edges) with only few nodes having a high π. The few that have a higher π will then exhibit a lot of edges to other nodes, resulting in a small-world structure of the network.
This sampling process can be repeated until a neighborhood structure fulfilling specific requirements is created. In our case, those requirements are limits for the number of neighbors in $\mathbf{W}_{12}$ and $\mathbf{W}_{22}$:

1. $5p < \sum_{i=1}^{p} \sum_{j=p+1}^{p+q} n_{ij} < q$
2. $q < \sum_{i=p+1}^{p+q} \sum_{j=p+1}^{p+q} n_{ij} < 2q$

[2] https://cran.r-project.org/web/packages/GIGrvg/
[3] https://github.com/horta/gig

The first requirement implies, that the fraction of non-zero edges lies between $\frac{5p}{pq}$ and $\frac{1}{p}$. For $p = 10$ and $q = 10$ this means that 90% to $94.\overline{4}$% of the edges in $\mathbf{W}_{12}$ are zero.
Finally, the edge weights are calculated by drawing uniformly from $[0.3, 1]$ and then randomly flipping the sign of the weight:

$$w_{ij} \sim \mathcal{U}(0.3, 1)$$
$$f_{ij} \sim \mathcal{U}(-1, 1)$$
$$\mathbf{W}_{ij} = n_{ij} * w_{ij} * sgn(f)$$

The resulting networks exhibits many connections to a few selected nodes, similarly to the real-world problems we target. An example network is shown in Figure 4.1.

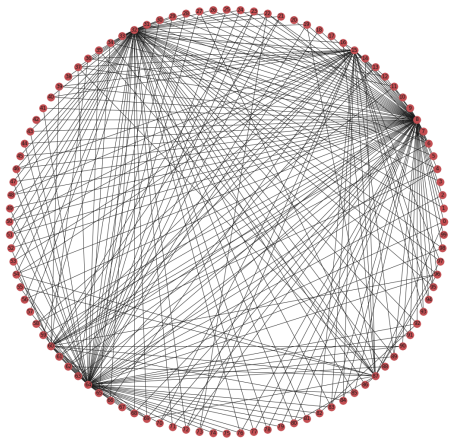


Figure 4.1: Example network for $p = 10$ and $q = 90$, ordered such that the first 10 nodes correspond to p. It can be seen that the network exhibits small-world properties in form of a few highly connected nodes, both in the p as well as in the q variables.

4.2.2 Performance Metrics

The metrics used for measuring the performance should be focused on how the structure of the predicted network graph compares to the ground truth. But the problem of quantifying dissimilarities between graphs is in itself a difficult and computationally expensive one. Instead we follow the simple approach of counting the prediction errors. Similarly to Kaufmann et al. [2016], prediction of an (existing) edge with the wrong sign is counted as a false positive, as shown in the following confusion matrix.

		Actual		
		$(\mathbf{W}_{12})_{i,j} < 0$	$(\mathbf{W}_{12})_{i,j} > 0$	$(\mathbf{W}_{12})_{i,j} = 0$
Predicted	$(\mathbf{W}_{12})_{i,j} < 0$	TP	FP	FP
	$(\mathbf{W}_{12})_{i,j} > 0$	FP	TP	FP
	$(\mathbf{W}_{12})_{i,j} = 0$	FN	FN	TN

On this basis, commonly known metrics such as the sensitivity (recall), specificity and precision can be calculated:

$$sensitivity = \frac{TP}{TP + FN}$$
$$specificity = \frac{TN}{TN + FP}$$
$$precision = \frac{TP}{TP + FP}$$

As there is always a trade-off between these, we also look at the F-score and the Matthews correlation coefficient. The F score is the harmonic mean between precision and recall, while the MCC is a measure of correlation between the predicted and the actual values [Baldi et al., 2000].

$$F = \frac{2TP}{2TP + FP + FN}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

5 Results

With the models established, we focus on their application. First, the performance and general behavior of the BMB, the Simulated Annealing and the Graphical LASSO will be compared on artificial data. In the subsequent section we apply the BMB and SA in context of HIV-1. The SystemsX.ch HIV-X cohort [Huldrych et al.] provides data of long-term patients whose viral reservoir were successfully suppressed with antiretroviral therapy. Our main goal is to find possible interactions between the resistance relevant mutations gathered from the viral genotype and multiple clinical factors that quantify the success of the treatment.

5.1 Application on Artificial Data

We compare the Graphical LASSO, the Bayesian Markov Blanket and its Simulated Annealing variant on multiple artificial data sets, created as described in subsection 4.2.1. In addition, we look at the behavior of the posterior distribution of $\mathbf{W}_{12}$ experimentally for different values of λ.

In Figure 5.1, a small example for two synthetic data sets with $p = 5$ and $q = 15$ is shown. Both the SA and the BMB were used for reconstructing the Markov Blanket of the p query variables, with $lambda = 70$ and $\lambda = 1400$ respectively.

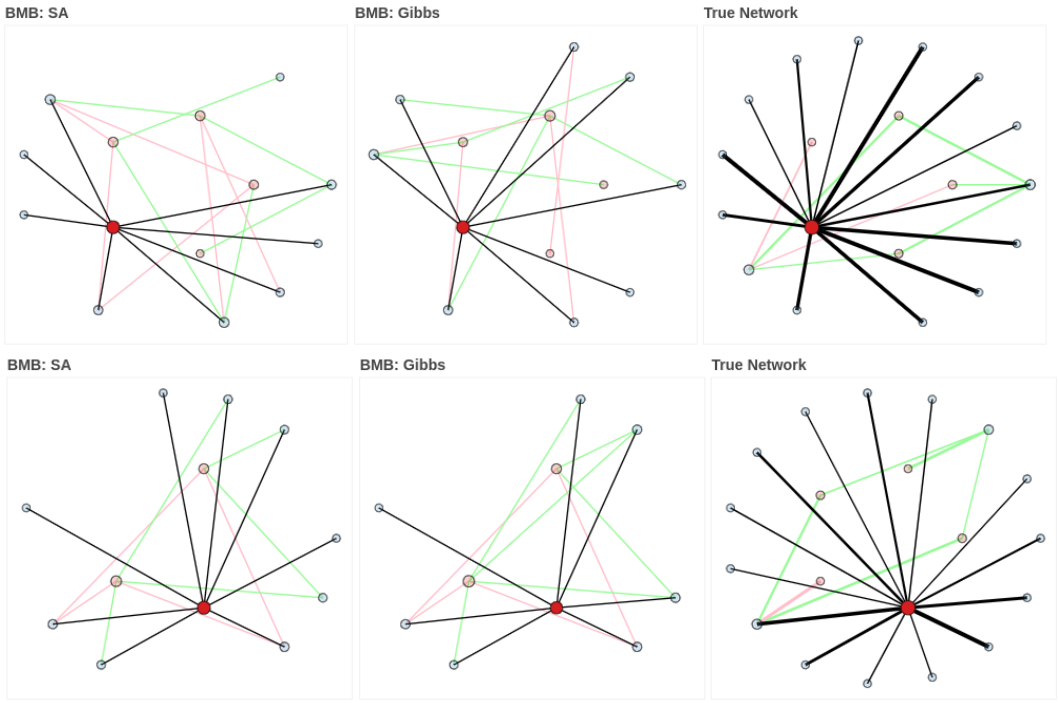


Figure 5.1: Comparison of the true W12 subnetwork with the one resulting from the BMB ($lambda = 70$) and SA ($\lambda = 1400$) for two example networks with a threshold of 0.2. The data was created similarly to the description in subsection 4.2.1, but with $p = 5$ and $q = 15$. The red nodes correspond to the p query variables.

It can be seen that both networks are very similar, even though widely different λ were used. This is unexpected as the underlying model for both methods is the same. In practice, finding an appropriate λ is the most relevant part of the model selection, as it is the only hyperparameter directly influencing the model. Because of this, our focus in all three models will be the sparsity inducing hyperparameter and its effect on the quality of the reconstructed networks.

5.1.1 Setup & Model Hyperparameters

Aim of the tests on the artificial data is a sound comparison between the original BMB and its Simulated Annealing variant. As such, the methods should be set up in a similar way to avoid performance advantages simply due to parameter fine-tuning or the use of more computing resources (in our case iterations/sweeps) for either one. The latter is especially important in the case of MCMC sampling and Simulated Annealing, where strong theoretical results for convergence exist.

We fix the number of total iterations in both samplers to 3000. In the case of SA this refers to the summed amount of initial Gibbs sweeps, cooling steps and the draws from the cooled down system at temperature T_n. As a consequence the computational effort is roughly equal

between the two methods [4]. The detailed settings are shown in Table 5.1. It should be noted that the burn-in is in both cases 0.3 times the number of Gibbs Sweeps.

Table 5.1: Model Parameters used for Testing

	BMB	**Simulated Annealing**
Total Iterations	3000	3000
Gibbs Sweeps	3000	900
Burn-In	900	270
Credible Interval	$[0.15, 0.85]$	$[0.15, 0.85]$
Cooling Steps	-	2100
Draws at T_n	-	300
T_0	-	1
T_n	-	0.01

With these settings, the models were applied to 40 artificially created data sets with $p = 10$ and $q = 90$.

5.1.2 Restrictions on Lambda

Figure 5.2 displays the average fraction of zero-edges (i.e. a measure of sparsity) of the reconstructed networks in relation to the chosen λ. The λ were restricted to ranges that

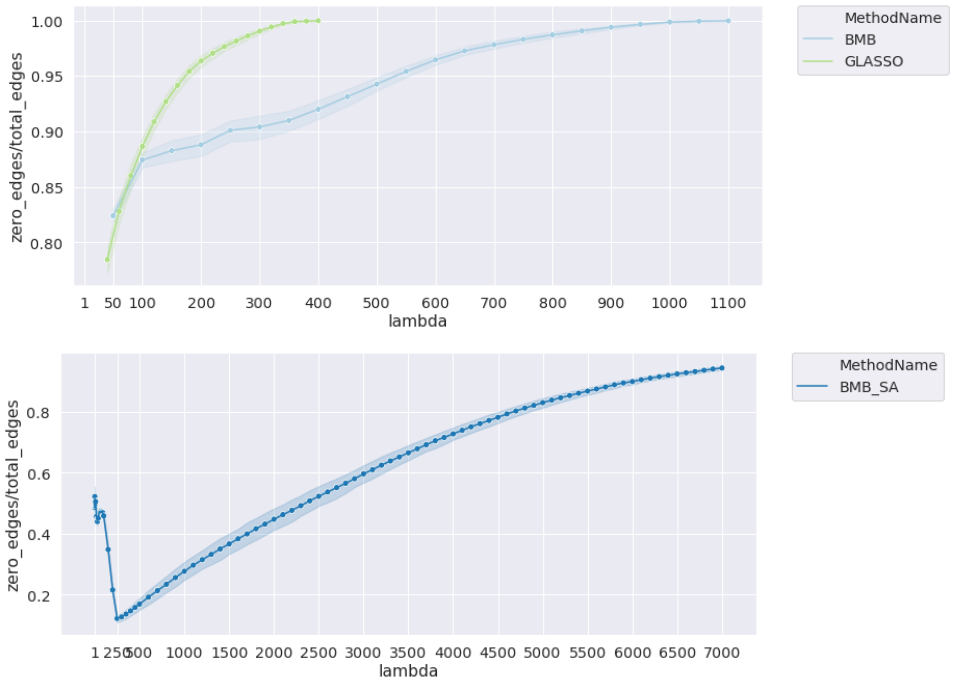


Figure 5.2: Sparsity of Networks for different λ over all 40 data sets. Plotted points correspond to the mean over all datasets and error bars to a 95% bootstrapping confidence interval for the mean.

[4] Only roughly, as the cooling part of the SA does not entail draws from the copula.

result in non-trivial (i.e. not empty) graphs. Furthermore, computational issues cause a lower bound for the GLASSO and an upper bound for the SA. Lambdas below 100 led to convergence problems for the Graphical LASSO, even in the case of a high number (50000) of iterations. Conversely, SA has an upper bound due to numerical instability. With increasingly high λ the MGIG draws of the $\mathbf{W}_{11}$ block (see subsection 4.1.2) and more specifically the inversion

$$A = (\mathbf{W}_{12}(\mathbf{S}_{22} + \mathbf{I})\mathbf{W}_{12}^T)^{-1}$$

fails as A^{-1} becomes nearly singular. We suspect the continued fraction of Wisharts to be a major cause of numerical instability in the sampler. Decreasing the number of fractions used for sampling the MGIG increases the maximum possible λ before failure, supporting this assumption. While all three methods generally follow the trend of getting sparser for higher λ, the scale on which this happens is vastly different. Especially the Simulated Annealing requires a much higher λ for the same level of sparsity as the BMB and GLASSO. Aside from that, the BMB and SA behave differently than expected for very low λ. Both recreate denser networks again, with the effect being more pronounced in the Annealing. A problem possibly underlying this can be identified by comparing the MCMC diagnostic plots for the different λ. Figure 5.3 shows the MCMC diagnostics for SA with a λ of 7000.

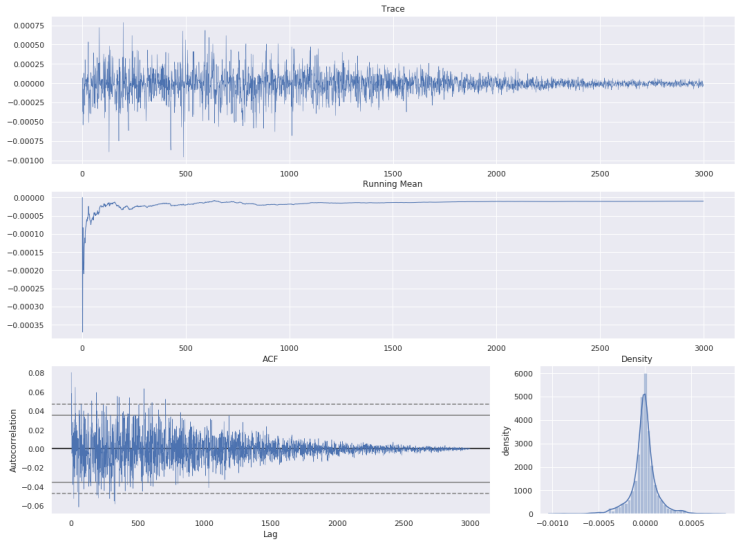


Figure 5.3: MCMC Diagnostics: Simulated Annealing with a λ of 7000 for the posterior marginal $(\mathbf{W}_{12})_{14}$.

In the trace plot we can clearly see the effect of the Annealing. Starting at iteration 1000, the system cools down and the samples slowly focus more on the region around the MAP. Furthermore, the density plot indicates the marginal being unimodal[5]. When switching to a low λ (see Figure 5.4), the cooled down chain starts fluctuating between two points, which is also reflected in the very high autocorrelation switching signs for each lag. As Simulated Annealing eventually samples from all global maxima, this behavior as well as the density

[5] Note that the density plots of the Annealing includes *all* drawn samples and only serves an diagnostic purpose. The actual estimation is only based on samples drawn at the target temperature.

plot indicates a multi-modality in the posterior marginals. This effect gets more pronounced for smaller λ as shown in Figure 5.4, where the distance between the modes also increases. Since the modes seemingly diverge further from each other for lower λ, the Credible Interval also increases and is more likely to cover the origin, estimating an edge as 0. This might explain the sparser networks for very small λ.

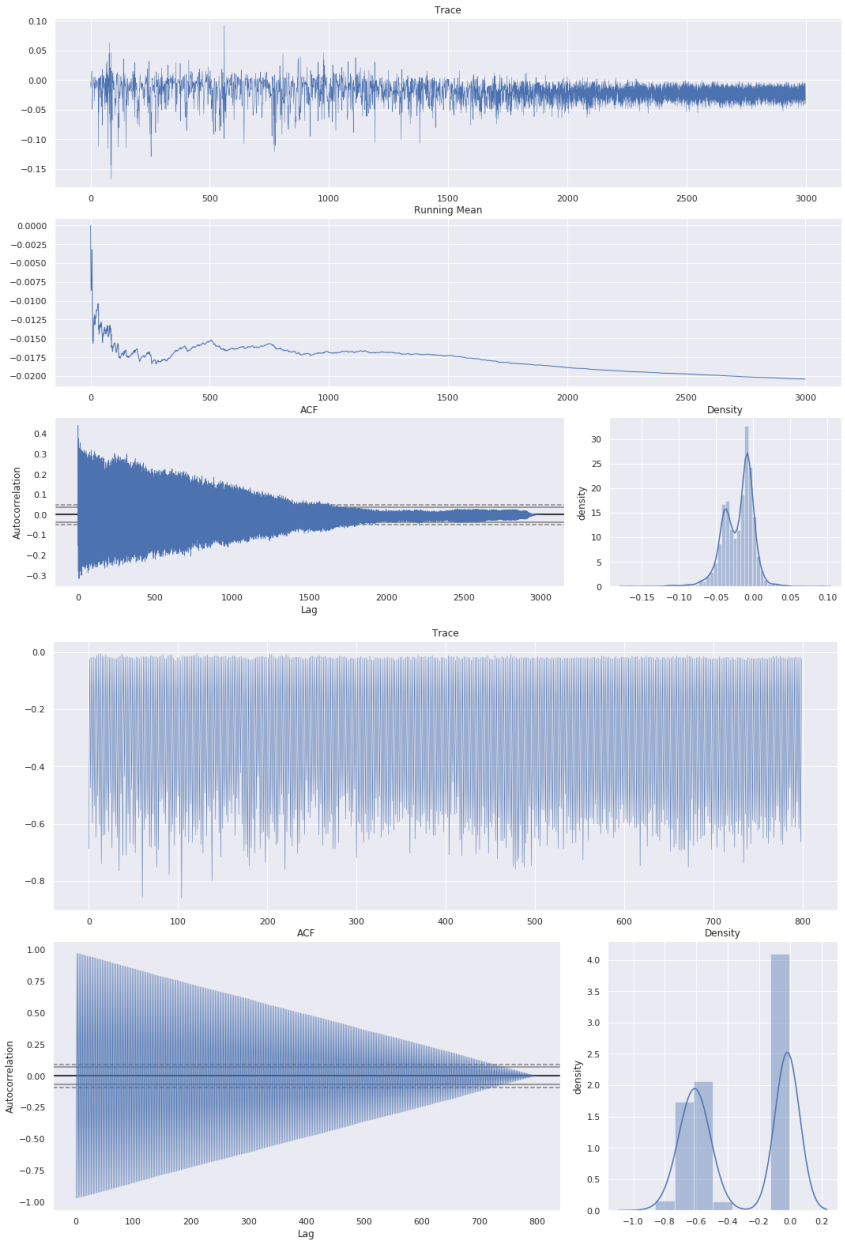


Figure 5.4: MCMC Diagnostics: Simulated Annealing with a λ of 90 (top) and 20 (bottom) for the posterior marginal $(\mathbf{W}_{12})_{14}$.

In case of the standard BMB this behavior appears more subtle, but is still existent and can be observed for extremely small λ (see Figure 5.5). But in this case, a high autocorrelation and jumping between values indicate mixture problems of the Markov chain. A possible

cause might be numerical problems for low λ. The hyperparameter λ indirectly affects the diagonal of $\mathbf{C}$, which has to be inverted for draws of the $\mathbf{W}_{12}$ block (see subsection 4.1.1). However, the Markov chains corresponding to λ in the range of 200 to 300 (where the bimodality in the SA marginals begins) did not exhibit similar problems, and neither could we observe any bimodal behavior in the non-cooled marginals.
So the question arises whether multi-modality in the posterior actually exists for low λ (< 300) or just occurs as a side-effect of the Markov chains convergence problems. In the next section we will expand on this by looking the empirical posteriors of $\mathbf{W}_{12}$ for relatively small λ that do not exhibit unusual behavior of the Markov chain.
Regardless of the outcome, it can already be seen that the λs below 300 do not lead to reasonable estimates for SA with the current thresholding method based on Credible Intervals. Because of this, they will be disregarded for the performance evaluation in subsection 5.1.4. Furthermore, λ below 50 will be ignored for the Gibbs BMB due to convergence problems.

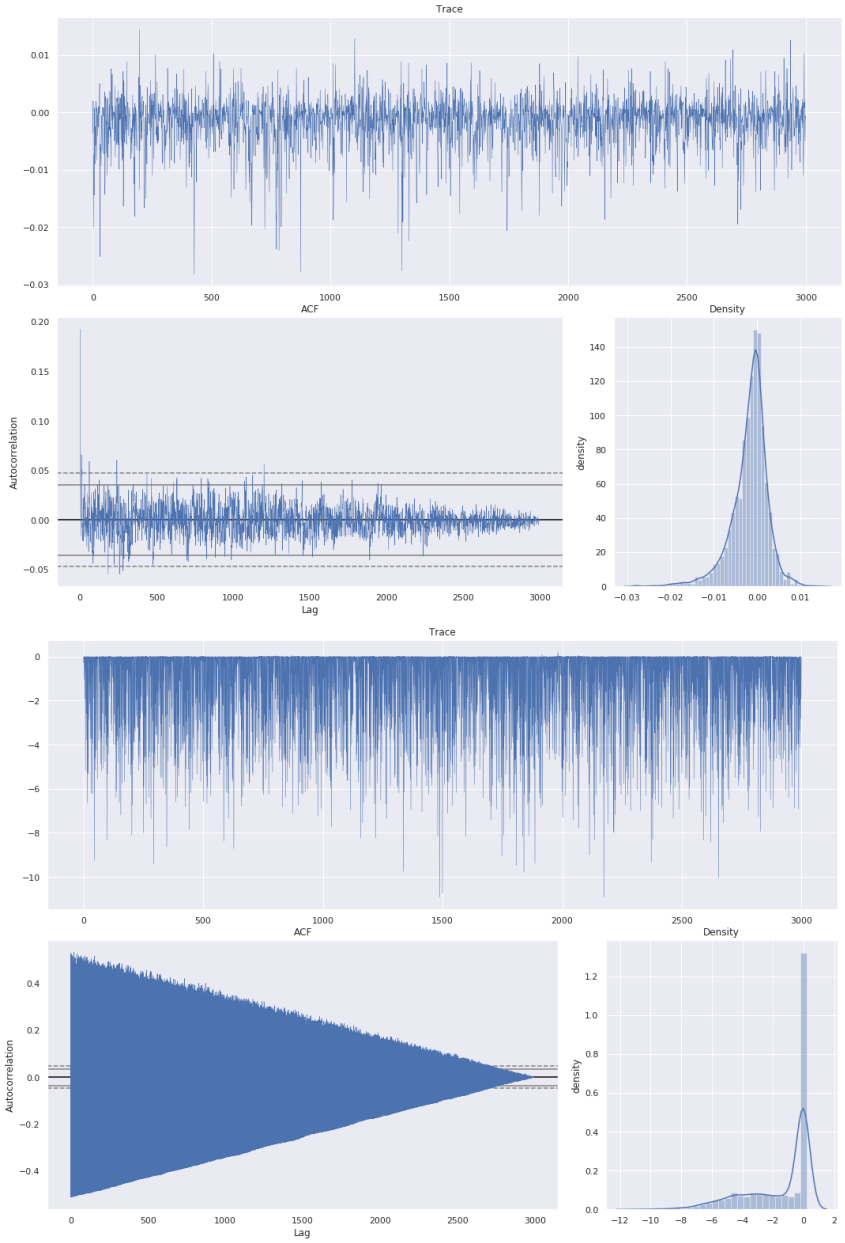


Figure 5.5: MCMC Diagnostics: BMB with a λ of 400 (top) and 1 (bottom) for the posterior marginal $(\mathbf{W}_{12})_{14}$.

5.1.3 Modality of the Posterior

In Chapter 3 we already mentioned that the modality of the posterior is important to consider when choosing an estimation method. Additionally, the results of the previous section indicated that the posterior marginals of the individual $\mathbf{W}_{12}$ edges get bimodal for very low λ. Instead of individual edges, we now look at the marginal posterior $p(\mathbf{W}_{12}|\mathbf{S}, \lambda)$ [6]. For this, linear Principal Component Analysis (PCA) was applied to reduce the dimensionality

[6] $p(\mathbf{W}_{12}|\mathbf{S}, \lambda)$ has shown to behave similarly to $p(\mathbf{W}_{12}|\hat{S}, \lambda)$ for the Gibbs sampler, presumably because the copula does not have a big influence on Gaussian test data.

of the posterior. The empirical distribution can then be observed alongside the first principal components.

In Figure 5.6 the empirical posterior marginal of $\mathbf{W}_{12}$ for one artificial data set is shown alongside the first principal components for $\lambda = 300$.

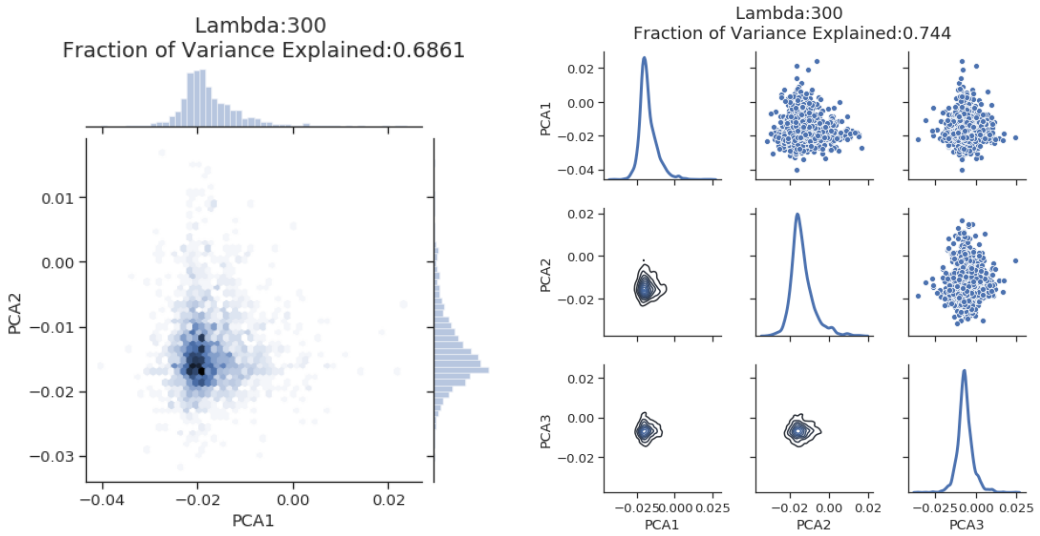


Figure 5.6: Empirical distribution of the $\mathbf{W}_{12}$ posterior marginal over the first principal components for one example data set with $\lambda = 300$.

It can be seen that the posterior is unimodal with a lot of mass surrounding the mode. In contrast, a slightly lower λ of 200 already gets bimodal, as shown in Figure 5.7.

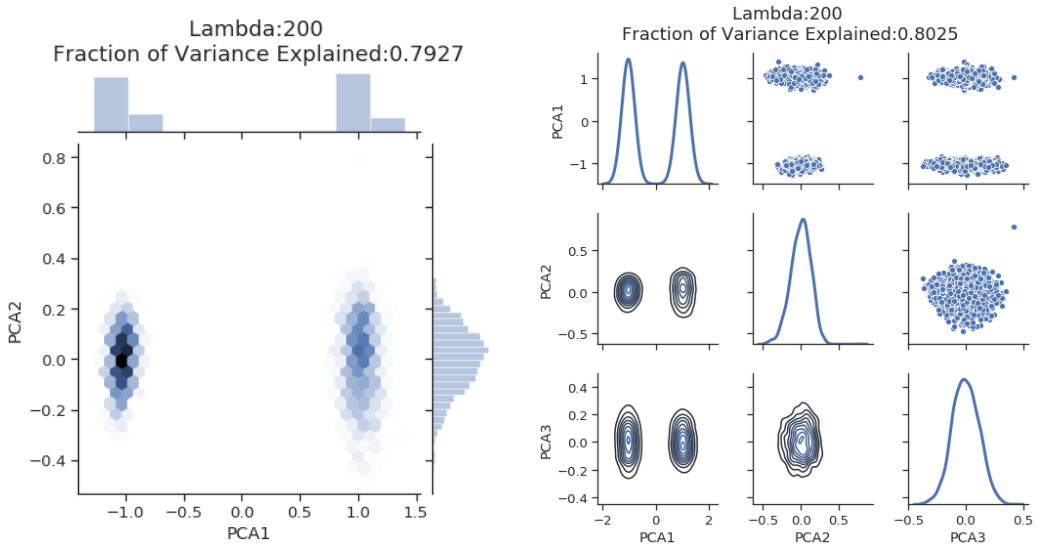


Figure 5.7: Empirical distribution of the $\mathbf{W}_{12}$ posterior marginal over the first principal components for one example data set with $\lambda = 200$.

Additionally, the distribution seems flatter alongside the first principal component. The corresponding MCMC diagnostics do not indicate any issues with the convergence in this

range of λ. As far as can be ascertained, the posterior stays unimodal for higher λ and multi-modal for lower ones. This change of modality aligns to the sudden change of sparsity of the Annealing in Figure 5.2, confirming that the fluctuation is indeed between two modes and not a result of numerical instability or bad mixing of the chain. Consequently we can assume that the posterior distribution of $\mathbf{W}_{12}$ actually is multi-modal for low λ.
An other interesting observation is that the variance explained by the first few principal components decreases for higher λ. This could be due to the regularizing effect of the λ. Aside from enforcing sparsity, the prior also decreases the values of the non-zero edges significantly. If variance in the non-zero edges decreases faster than the variance of the zero-edges (in terms of absolute values), the variance in the distribution is not anymore dominated by the non-zero edges. This is illustrated in Figure 5.8, where the distribution of a (truly) non-zero edge is compared to a zero edge for a specific data set. Initially, the variance of the non-zero dominates and the PCA would still be able to a lot of information with just one axis. For higher λ however, this changes and the joint distribution becomes more circular, leaving a linear PCA no possibility to summarize the data in a low-dimensional projection (in this case onto one axis). For higher dimensional data the behavior should be similar, with the joint distribution tending towards a sphere.

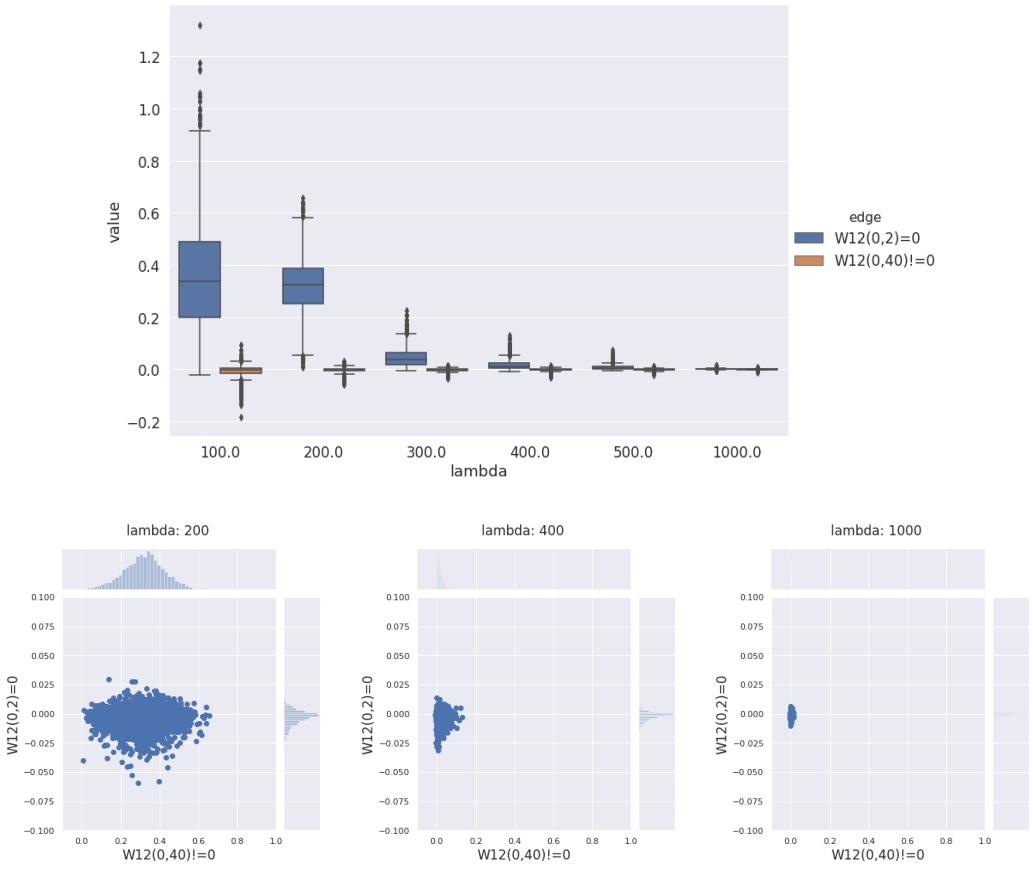


Figure 5.8: The effect of a higher λ on the joint distribution between two marginals of the $\mathbf{W}_{12}$ block. The true edge of $(\mathbf{W}_{12})_{(0,40)}$ (x-axis) is non-zero, while the true edge of $(\mathbf{W}_{12})_{(0,2)}$ (y-axis) is zero.

We now know the modality of $p(\mathbf{W}_{12}|\lambda)$, but Simulated Annealing estimates the MAP of the joint posterior $p(\mathbf{W}_{11}, \mathbf{W}_{12}, \mathbf{T}|\hat{S}, \lambda)$. Unfortunately we are unable infer anything significant about the joint posterior because the first few (2-3) principal components only explain a low amount of the variance ($< 10\%$"), rendering the 2 dimensional projection useless.

5.1.4 Quality of Reconstructed Networks

With the range of possible λ established, the predictive performance of the individual methods can be analyzed.

The performance metrics for the BMB, GLASSO and SA are visualized in Figure 5.9. We plotted the mean over all datasets as well as error bars corresponding to a 95% bootstrapping confidence interval for the mean. First of all, the effect of the λ on the specificity is of interest. It can be seen that the specificity generally increases with a higher λ for all three methods, which is to be expected due to its sparsity inducing property. Furthermore, there is a linear relationship between the fraction of nonzero edges and the specificity, as shown in Figure 5.10. In practical applications we can only observe the resulting network, so this relation is very useful. We can effectively control for the specificity by adjusting λ until a satisfactory density of the network graph is reached. While the BMB and GLASSO show a maximum average MCC and F-score at around 600 and 200 respectively, the limited range of the Annealing seems to not include it's maximum. But the drop in sensitivity for very high λ will also have to occur in the Annealing as the networks will presumably be empty for a high enough λ. With the specificity already getting close to 1, we can thus assume that we would not observe much of an improvement for slightly higher λ. Nonetheless, the Annealing reaches values of around 0.4 for both the F-Score and the MCC, which are comparable to the performance of the BMB. But in all cases, the SA and BMB are outperformed by the GLASSO. Aside from that, the Annealing is capable of covering a wider range of the sensitivity, allowing the estimation of less sparse networks.

To show the similarity between the methods, we fix λ such that the resulting network size for the test data is (close to) equal in all methods. In Figure 5.12 the performance for individual data sets is shown for these fixed λ. While the GLASSO always has the highest MCC and F-score, the difference between the Annealing and the BMB is not that well-defined. In addition, the methods mostly agree on the relative performance for the specific data sets. For example, all show low performance on data sets 160 and 180 while performing well on 178. Aside from confirming the previous results, the box-plots in Figure 5.11 furthermore show that the Annealing exhibits a higher variance among different data sets than the BMB for basically all metrics, excluding the specificity (with fixed λ).

Finally, we will look at how the performance of the methods relate to the sparsity for easing the search of a suitable λ in practice. While information criteria such as the (extended) BIC [Foygel and Drton, 2010] are popular for model selection of GGMs, they are based on the calculation of the (log) likelihood, which requires the complete matrix $\mathbf{W}$. In our current setting for the Simulated Annealing, the whole precision matrix is not available since cooling the $\mathbf{W}_{22.1}$ block would change the estimated joint MAP (see subsection 3.2.1). Consequently we have to look at networks resulting from a range of different λ and see

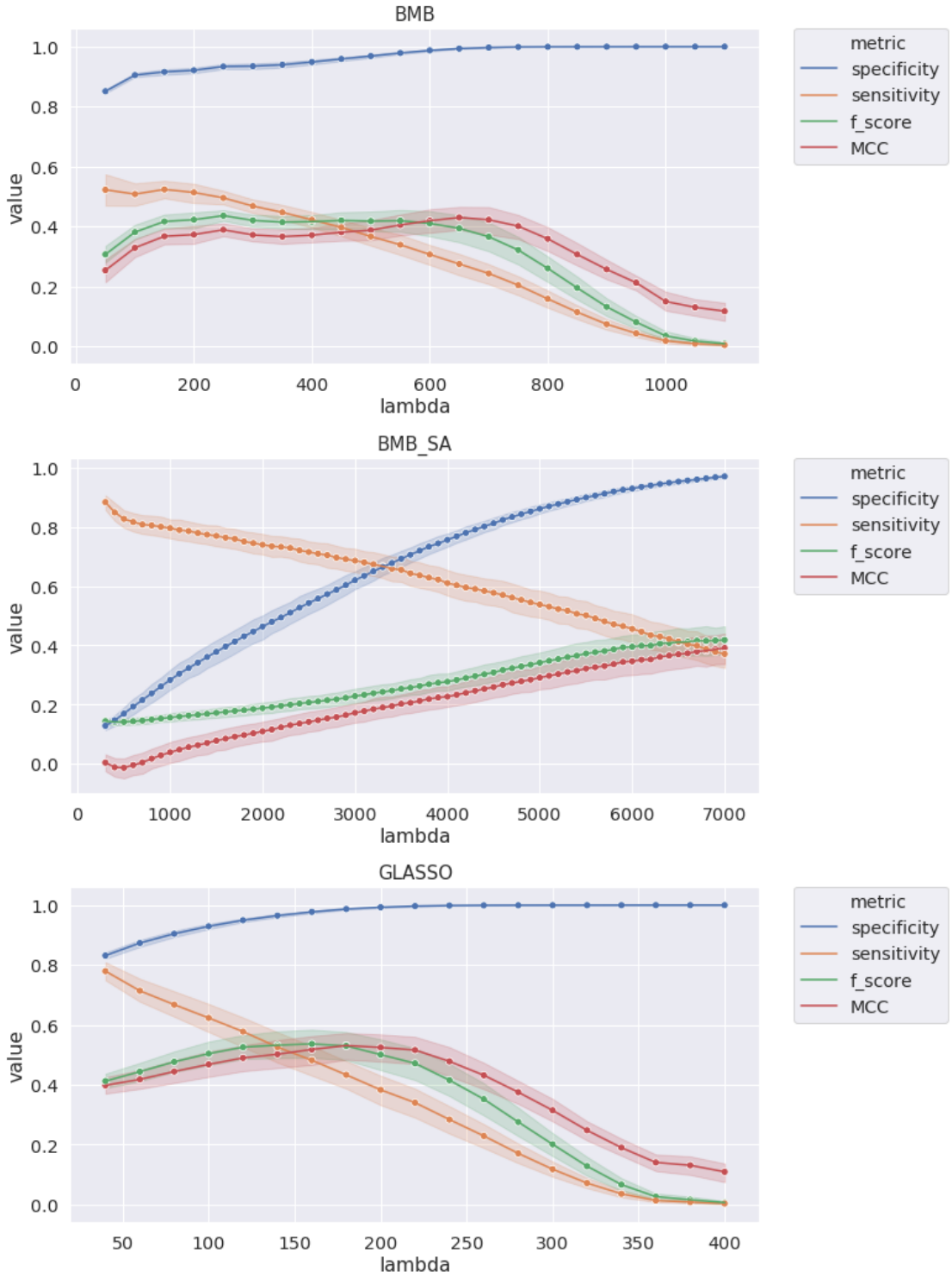


Figure 5.9: Average performance metrics for the Gibbs BMB, Simulated Annealing, and the GLASSO. The shaded area corresponds to a 95% bootstrapping confidence interval for the mean.

how they change for different λ. In this case, knowing how the model generally behaves in relation to the sparsity can then be helpful. Figure 5.13 shows the relation between the

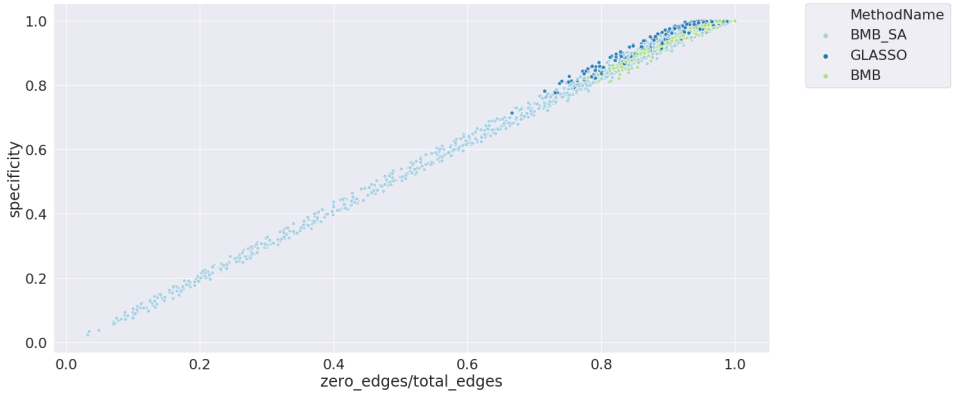


Figure 5.10: Relation between the sparsity in the reconstructed networks and the resulting specificity for all 40 data sets.

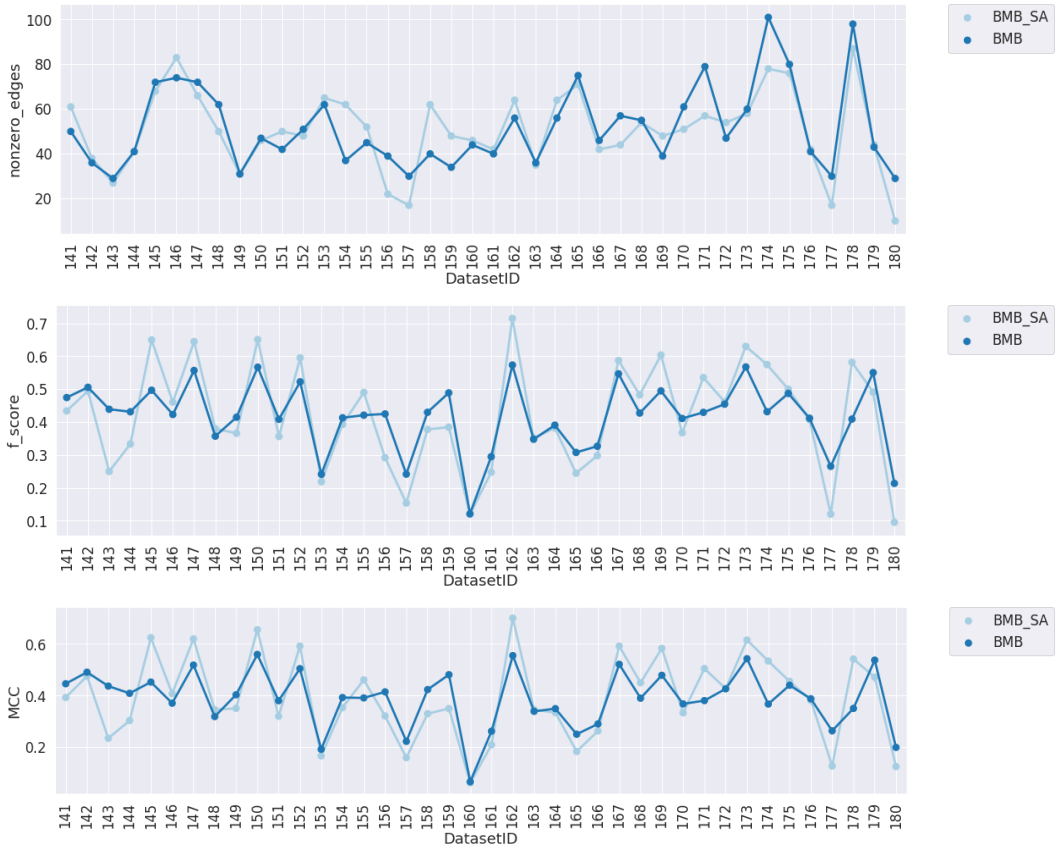


Figure 5.11: Different Metrics for fixed λ resulting in a roughly similar network size.
$\lambda_{BMB} = 500 \quad \lambda_{SA} = 7000 \quad \lambda_{GLASSO} = 150$

number of zero-edges and the metrics. We can see that both the F-Score and the MCC are highest when in the range corresponding to the true sparsity of the networks (which is between around 90% to 94.$\overline{4}$%). After reaching this point, the sensitivity drops sharply for sparser graphs while the specificity only increases slowly, thus resulting in a decrease of the F-score and MCC. However, even for higher sparsity the specificity does not always reach

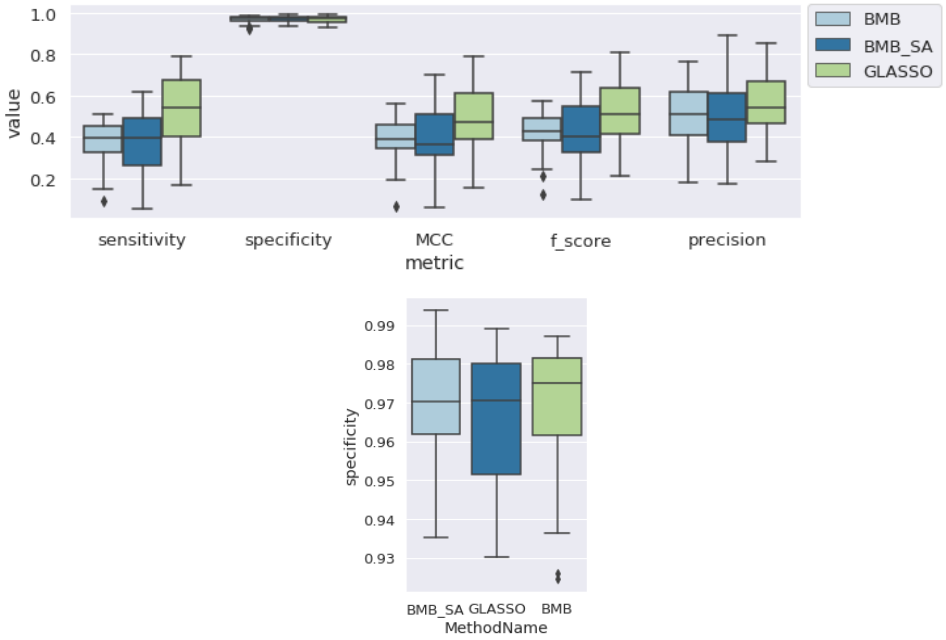


Figure 5.12: Different Metrics for fixed λ resulting in an approximately similar network size.

$$\lambda_{BMB} = 500 \quad \lambda_{SA} = 7000 \quad \lambda_{GLASSO} = 150$$

100% for non-trivial graphs. So for avoiding false positives it is advisable to overestimate the sparsity of the target network when selecting a λ. Aside from that, the plots agree with the previous results in relation to the λ.

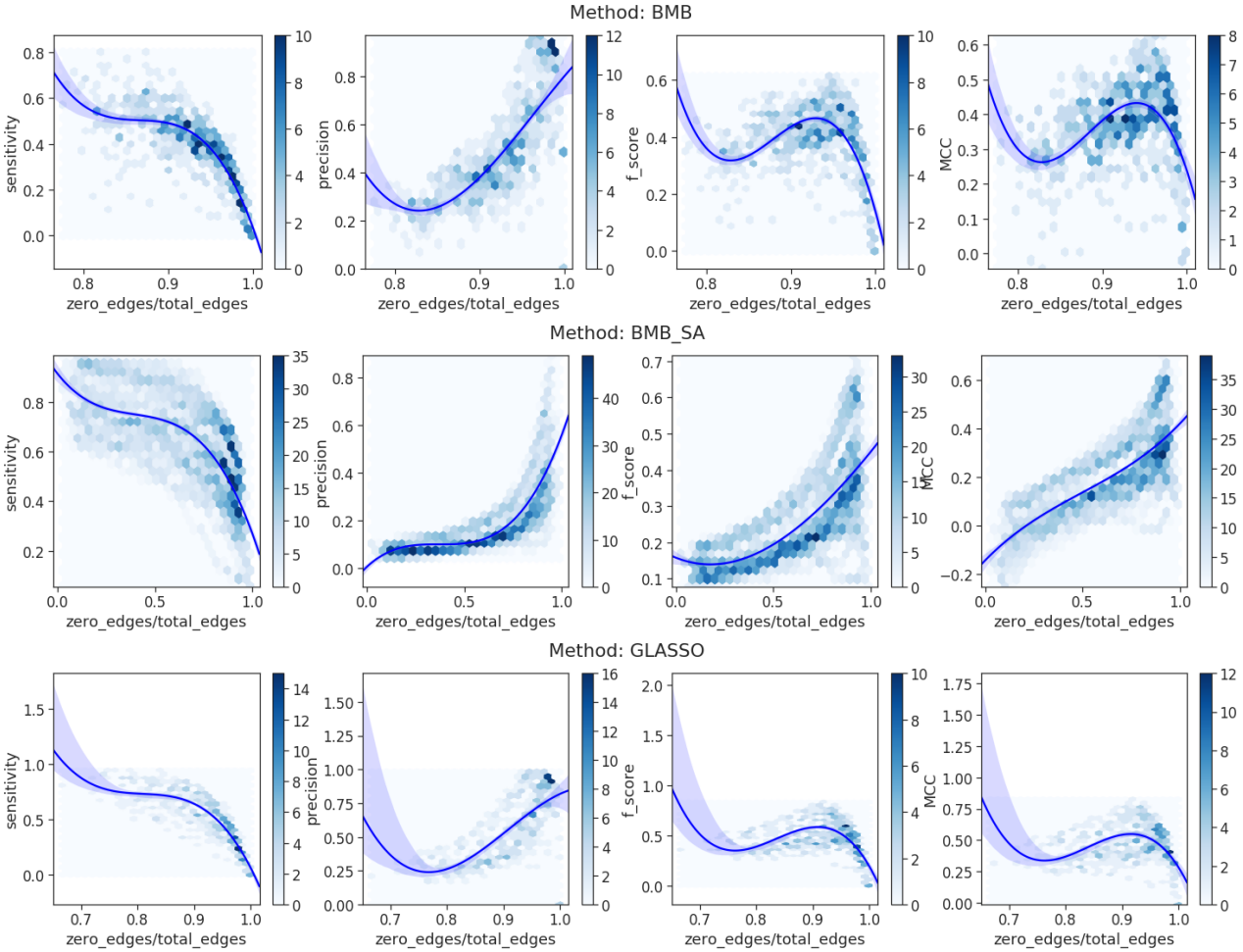


Figure 5.13: Relationship between the sparsity of the reconstructed Network and different performance measurements. The plots were created using 2D histograms with hexagon binning and a polynomial regression of degree 3.

The similarity of the resulting networks for the BMB and SA is illustrated in Figure 5.14 and Figure 5.15. While they still slightly differ, we could not find any consistent advantage in either of them (in terms of the used measures). More sophisticated graph comparison methods would be necessary for a better understanding of their actual differences. Attempts at combining the two methods by intersection of the resulting graphs has proven to be difficult due to the different scales of λ. Even when parameters resulting in the same network size were found, the combinations did not provide any significant improvement.

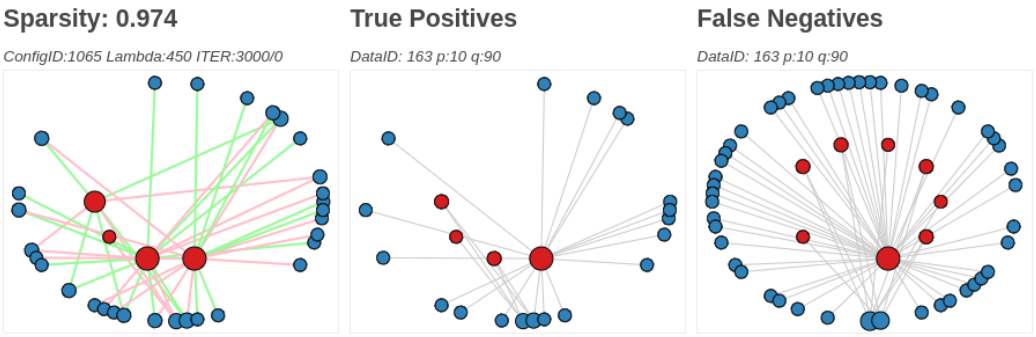


Figure 5.14: W12 subnetwork reconstructed with Gibbs BMB (for artificial data with id 162)

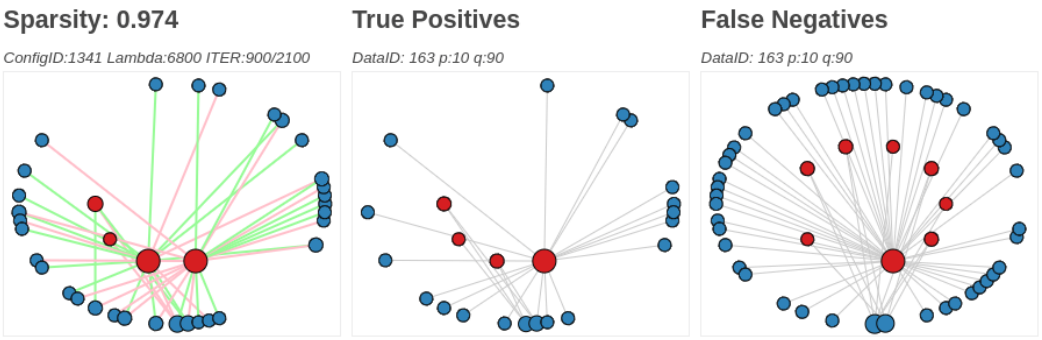


Figure 5.15: W12 subnetwork reconstructed with SA (for artificial data with id 162)

5.2 Application on Real Data: HIV-X

We now switch to the Markov Blanket estimation in context of the HIV-X MRD Project [Huldrych et al.].

The Human Immunodeficiency Virus (HIV) is one of the most widespread and harmful viruses in the world. If left untreated, it leads to the reduction of T helper cells expressing CD4, which are essential for the immune system. The final stage of the infection is accompanied by a complete shut down of the immune system and is referred to as Acquired Immunodeficiency Syndrome (AIDS) (e.g. Alimonti et al. [2003]). Lacking a working immune system, affected individuals become vulnerable to opportunistic diseases such as toxoplasmosis and pneumonia, which may ultimately lead to death [Chaisson et al., 1998]. While current treatments are effective at suppressing the virus by targeting stages of its lifecycle [Bartlett et al., 2001], no cure that completely eliminates the virus has yet been found. Even when the virus is suppressed by treatment, latent HIV reservoirs remain ineradicable and ready to multiply as soon as treatment is stopped. Another ongoing problem of HIV treatment is the high mutation rate of the virus; mutations can lead to resistances against specific antiretroviral drugs (and even multiple drugs) in the patient, rendering treatment with those drugs ineffective [Wainberg and Friedland, 1998]. Once developed, such resistances reduce the potency of the relevant drugs for a lifetime [Noë et al., 2005]. The swiss HIV-X MRD Project [Huldrych et al.] aims to identify factors in latent viral reservoirs pertinent to the future treatment of patients. Using the Markov Blanket estimation we in-

vestigate HIV-1 data available from the HIV-X Project for potential interactions significant to the therapy of patients. Among others, this includes the interactions between resistance relevant mutations and multiple clinical factors such as the viral load.

5.2.1 Data and Setup

The data of the HIV-X MRD Project [Huldrych et al.] is a subset of the Swiss HIV Cohort Study [SHCS et al., 2010], consisting only of treatment experienced HIV-1 patients. That is, all individuals have been successfully treated with Antiretroviral Therapy (ART) for at least five years. This implies that the viral load has been suppressed to non-detectable levels and the measured RNA levels of HIV in the data mostly remain very low. The exceptions are occasional so-called 'blips'. A blip is a transient but measurable spike of the viral load, in patients for whom the virus is effectively suppressed to undetectable levels. The significance of blips has not yet been established and remains controversial. While some studies found associations with the development of drug resistances [Stuart et al., 2001] and the CD4 cell count recovery [Martinez et al., 2005], others claim blips to be of limited significance [Lee et al., 2006].

Aside from the viral load measured by RNA levels in the plasma, the data contains CD4 and CD8 cell counts. CD8 T cells are an important part of the immune system and are responsible for controlling HIV in the early infection phase [Gulzar and Copeland, 2004]. While an infection initially leads to an increase in CD8 T cell count, they decrease again in later stages [Roederer et al., 1995]. Furthermore, a quantification of the HIV DNA obtained by droplet digital PCR (ddPCR) is provided, as viral load is undetectable for all the patients. Droplet Digital PCR is a method for accurately measuring the DNA copy number, and has since recently been applied in the context of HIV [Strain et al., 2013]. Additionally, two differently obtained sets of resistance relevant mutations of the patients are available. The first one is based on Genomic Resistance Testing (GRT). Even though resistances mostly arise in response to ART, they may also be transmitted to newly infected individuals. As this should be taken into account when choosing a treatment, new patients as well as patients experiencing treatment failure are subject to Genotypic (Antiretroviral) Resistance Testing (GRT) [Günthard et al., 2014, Shafer, 2002]. The second set of resistance relevant mutations is based on haplotype reconstruction of Next Generation Sequencing (NGS) data. The gag-pol region of the viral genome was reconstructed with PredictHaplo[7][Prabhakaran et al., 2014]. Subsequently the reconstructed haplotypes were translated into (possible) amino acids. With the amino acids of the haplotypes available, resistance relevant mutations were mapped according to the '2017 Update of the Drug Resistance Mutations in HIV-1' [Shafer, 2017].

An overview of the variables available for each individual patient is shown in the following table. The first $p = 5$ variables are query variables for which we wish to infer dependencies; the remaining $q = 225$ variables pertain to resistance relevant mutations, and the drugs used in the treatments.

[7] Source of PredictHaplo: https://bmda.dmi.unibas.ch/software.html

Variable	Domain	Description
CD4	$\mathbb{R}_{\geq 0}$	CD4 T cell count
CD8	$\mathbb{R}_{\geq 0}$	CD8 T cell count
RNA	$\mathbb{R}_{\geq 0}$	Viral Load / HIV RNA Level
HIV_copies	$\mathbb{R}_{\geq 0}$	Measured by HIV DNA cell count with ddPCR
SLOPE_ddPCR	$\mathbb{R}$	Error corrected change of HIV_copies to next time point
HIST_DRU_*	$\{0,1\}^{34}$	Administered ART drugs in any previous treatment (e.g. HIST_DRU_EFV for Enfuvirtide [8])
drug*	$\mathbb{N}^{34}$	Amount of days the drug has been administered in current treatment. Is 0, if the patient is currently not taking this drug (e.g. drugEFV)
MU_*	$\{0,1\}^{77}$	Resistance relevant mutations obtained with GRT (e.g. MU_RT103)
INT*/PR*/RT*	$\{0,1\}^{80}$	Resistance relevant mutations inferred from reconstructed haplotypes (e.g. RT103)

Tests for CD4/8 counts, viral load and ddPCR did (mostly) not occur at the same time. As the ddPCR offers the most reliable estimate of HIV persistence in successfully treated patients, the data has been merged such that each sample corresponds to one of a patient's ddPCR measurements; for each patient, there are between 2 to 5 such measurements available, each from a different point in time. Viral load and CD4/8 cell counts in each sample correspond to the measurements closest in time to the respective ddPCR value. Additionally, the rate of change between successive (patient-wise) ddPCR measurements is included as *SLOPE_ddPCR*.

In total there are 2760 samples with 1092 individual patients. For 268 individuals, resistance mutations from the haplotype reconstruction are available. In contrast to that, the GRT mutations as well as the prior and current treatments are available for all patients. A short summary of the variables of interest is given in Table 5.2.

Table 5.2: Summary of the (non-boolean) variables in the data set.

	CD4	CD8	RNA	SLOPE_ddPCR	HIV_copies	days_since_1stddPCR
count	2760.00	2754.00	2734.00	2760.00	2760.00	2760.00
mean	524.96	823.92	54.15	-0.02	4111.24	643.03
std	256.10	393.13	1155.23	27.67	104741.81	697.02
min	0.00	0.00	0.00	-585.95	0.00	0.00
25%	345.00	547.00	0.00	-0.01	215.49	0.00
50%	484.00	755.00	0.00	0.00	448.94	693.00
75%	663.00	1019.50	0.00	0.03	861.78	875.00
max	2101.00	3407.00	39178.00	899.29	4238988.63	5249.00

As expected for successfully treated patients, the viral load is small for most of the data, except for the occasional blips. This is also illustrated in Figure 5.16. The ddPCR in *HIV_copies* and the corresponding slopes furthermore exhibit a few extreme outliers, which can also be seen in Figure 5.17. The outliers should not be problematic for the estimation, since the extended likelihood of the copula depends only on an order preserving statistic.

In Figure 5.18 the time difference between the NGS and the GRT data is given. It can

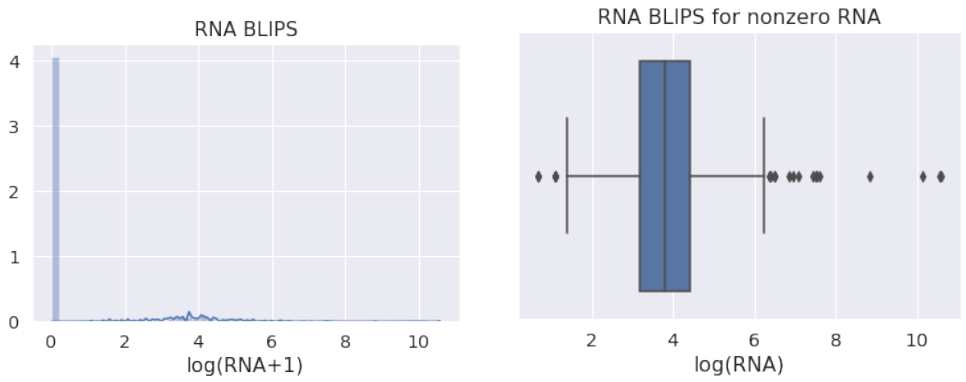


Figure 5.16: Distribution of the viral loads. The Boxplot only includes the non-zero values.

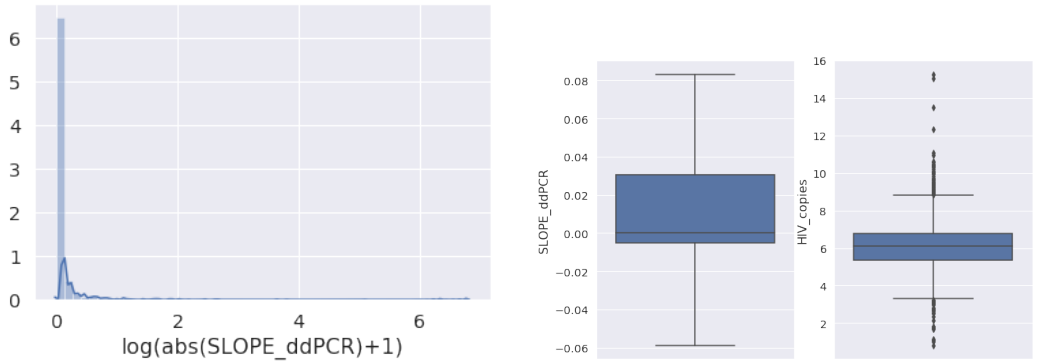


Figure 5.17: Distribution of ddPCR HIV_copies and ddPCR slope.

be seen that the majority of the GRT mutations are either from the same date as the NGS data (the start of the treatment) or a later point of time (presumably when treatment had to be switched).

The patients in the data set have been observed for up to 14 years (see Figure 5.19), while the NGS testing is only done once for each patient. It would be questionable to assume that the haplotypes of the virus remain unchanged, over such a long interval and during possibly multiple treatment changes. For this reason we will only use the first sample of each patient for the network estimation, as it is closest in time to the NGS data.

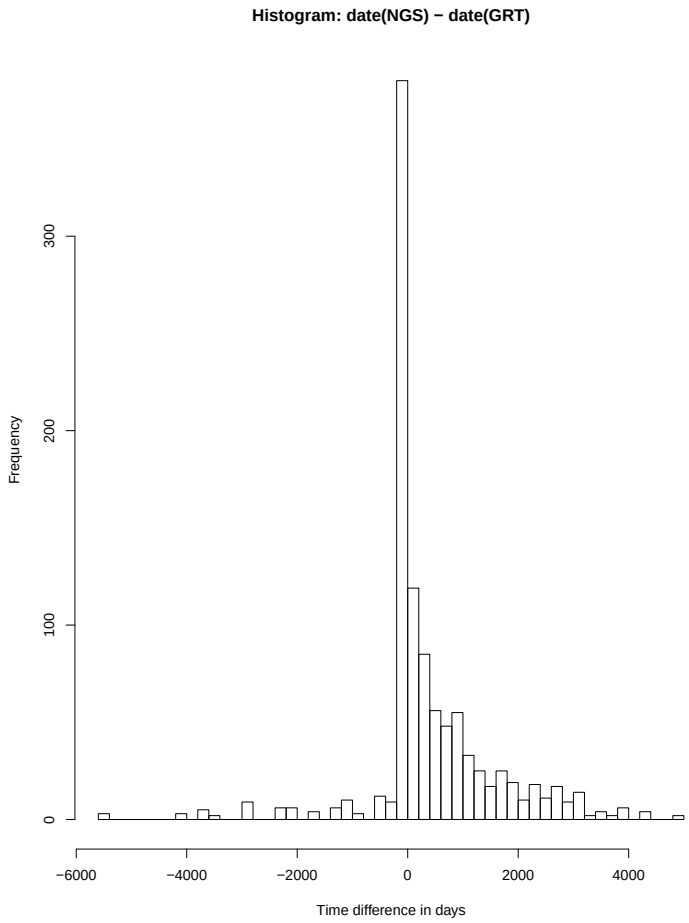


Figure 5.18: Histogram of patient wise time difference (in days) between the GRT mutations and the NGS data.

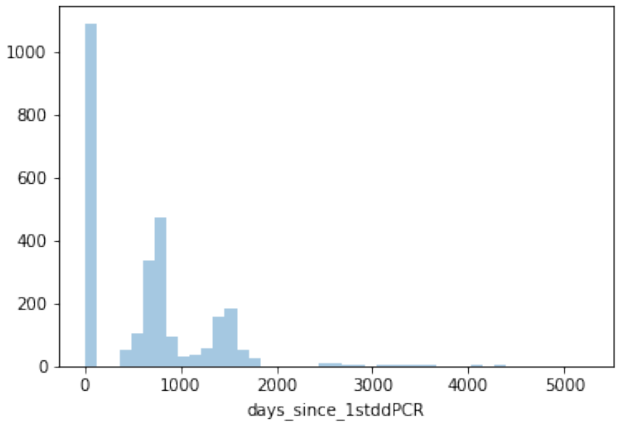


Figure 5.19: Histogram of time passed between current ddPCR sample and the first ddPCR test of the respective patient.

Setup

The settings for the Gibbs BMB and the Annealing run are given in Table 5.3. Note that in contrast to the test data setting, no thresholding Credible Interval was set in advance.

Table 5.3: Model Parameters used for the HIV-X Data

	BMB	**Simulated Annealing**
Total Iterations	7000	7000
Gibbs Sweeps	7000	2100
Burn-In	2100	630
Cooling Steps	-	4900
Draws at T_n	-	490
T_0	-	1
T_n	-	0.01

5.2.2 Modality of the Posterior

Analogous to the test setting, we look at a low dimensional projection of the marginal posterior distribution $p(\mathbf{W}_{12}|\mathbf{Z} \in D, \lambda)$ via PCA. As it turns out, the $\mathbf{W}_{12}$ posterior appears to be multi-modal for λ lower than 250, while being unimodal for higher λ. Figures 5.20 and 5.21 show the distributions for a λ of 200 and 300 respectively. It should be noted again that the first two principal components for $\lambda = 300$ only explain about 32% of the variance in the data, so the projection is rather inaccurate. Therefore the conclusion on unimodality should be taken with caution. When looking at the MCMC diagnostics, we could not find any indications of convergence problems for $\lambda = 200$, $\lambda = 300$. A representative example for the diagnostic plots is shown in Figure 5.22 and Figure 5.23. Both the autocorrelation and the trace plot seem acceptable. In contrast, λ of 50 and lower are to be avoided, as the autocorrelation indicate repeating fluctuation in the samples (see Figure 5.24).

But in general, the multi-modality for low λ is not an unexpected observation. A big part of the HIV data, namely the resistance mutations of the haplotypes and GRT, as well as current and previous treatments, are very sparse. In addition, there are a lot of missing values for the haplotype mutations and the data is high dimensional (230 variables for 1092 samples). Therefore it is reasonable that various configurations are capable of explaining the data, if the prior does not enforce a lot of sparsity in the parameters. As a consequence we will avoid the annealing in this λ region.

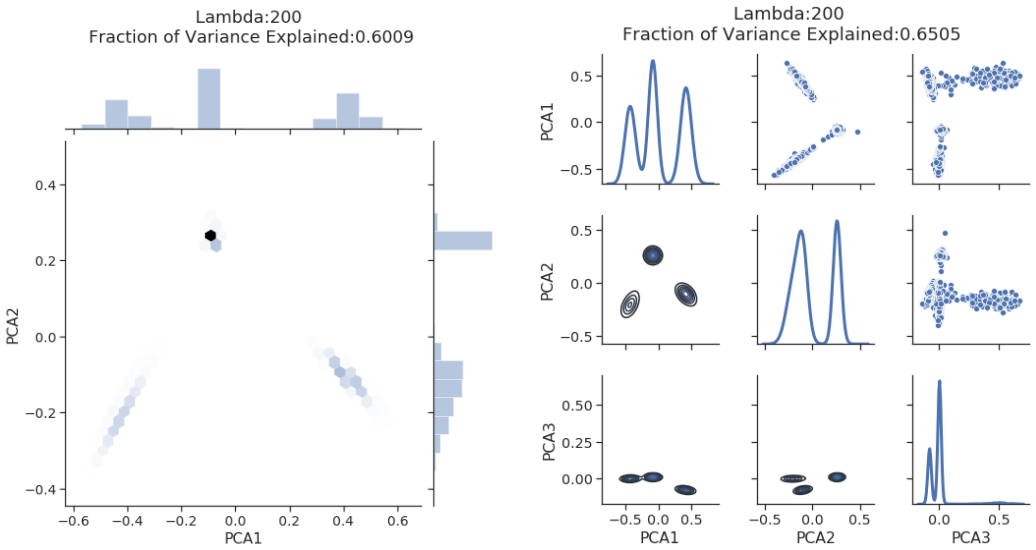


Figure 5.20: Empirical distribution of the $\mathbf{W}_{12}$ posterior marginal over the first principal components with $\lambda = 200$.

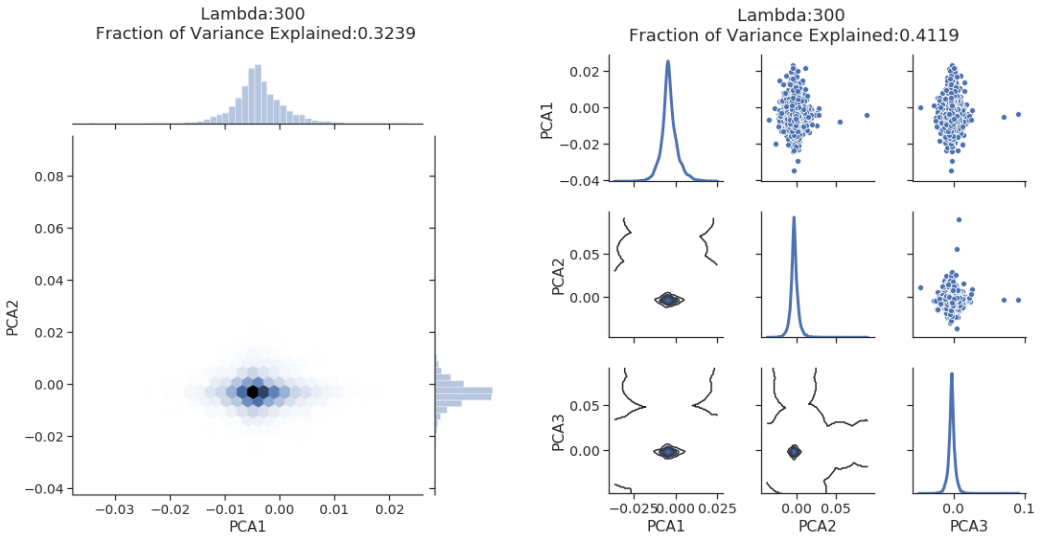


Figure 5.21: Empirical distribution of the $\mathbf{W}_{12}$ posterior marginal over the first principal components with $\lambda = 300$.

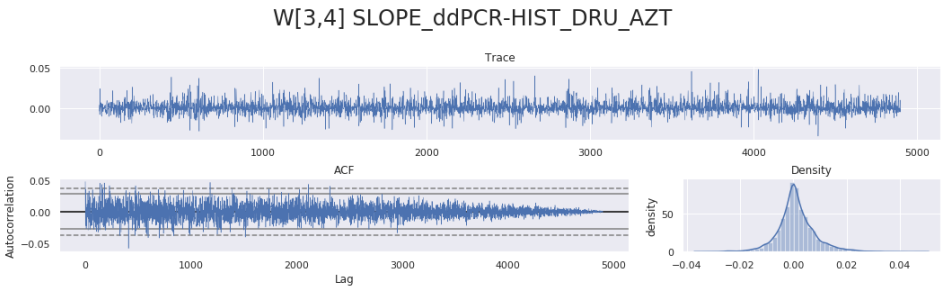


Figure 5.22: MCMC diagnostics for the Gibbs sampler with $\lambda = 200$. Note that W in this case refers to W_{12}.

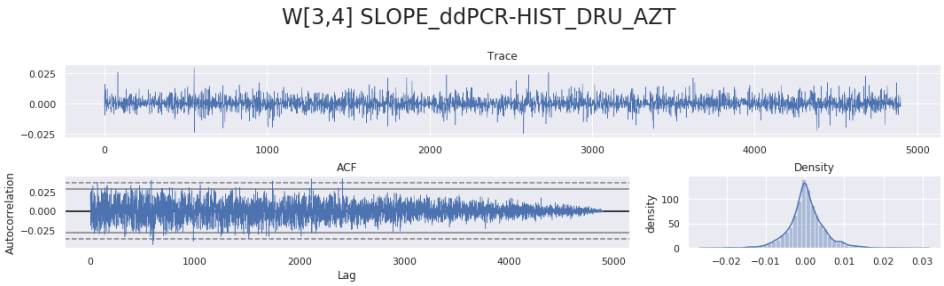


Figure 5.23: MCMC diagnostics for the Gibbs sampler with $\lambda = 300$. Note that W in this case refers to W_{12}.

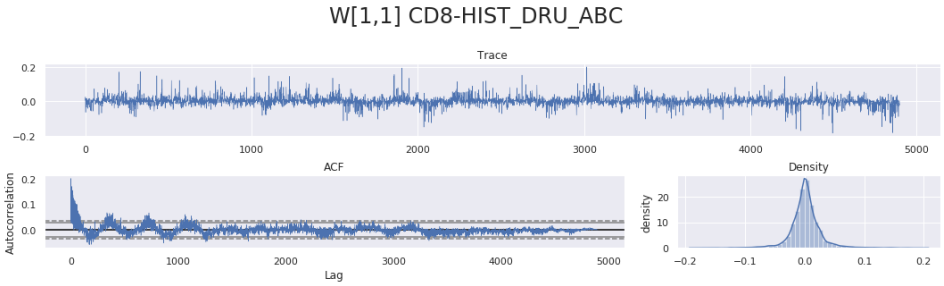


Figure 5.24: MCMC diagnostics for the Gibbs sampler with $\lambda = 50$. Note that W in this case refers to W_{12}.

5.2.3 Selection of Lambda and Threshold

First of all, the network sizes for different λ and thresholds were explored. As we can see in Figure 5.25, Simulated Annealing leads to a smooth change in network size over the range of λ and thresholds explored, which is what we would expect. This allows us to fix the thresholding while exploring the whole range of networks, from rather densely connected to completely sparse. As we know that both precision and specificity tend to increase with network sparsity, connections found in sparser networks (i.e. with higher λ) can be seen as more certain.

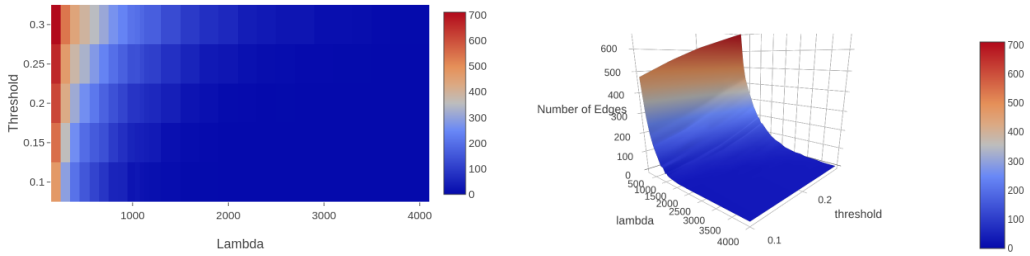


Figure 5.25: Heatmap and 3D surface plot showing the number of edges in the inferred networks with the SA BMB for different λ and Credible Intervals (CI defined by [threshold, 1-threshold]).

In contrast, finding a suitable range of λ for the Gibbs BMB is quite difficult. As shown in Figure 5.26, the resulting networks are non-trivial for reasonable Credible Intervals only in the multi-modal region ($\lambda < 200$). But even then, the resulting networks are extremely sparse for 80% Credible Intervals and above, with a non-smooth and non-monotonic decrease in the number of edges for higher λ. As with SA, we know the precision and specificity generally increases for higher λ and sparser networks. However, the usable range of λ is very close to the range subject to convergence problems. While we can clearly see that λ of 50 and lower are unusable, it is not certain that slightly higher λ are immune; it is possible, that similar problems are not obvious from mere inspection of the diagnostic plots. Combined with non-monotonic change of network size, network and model exploration is made considerably more difficult.

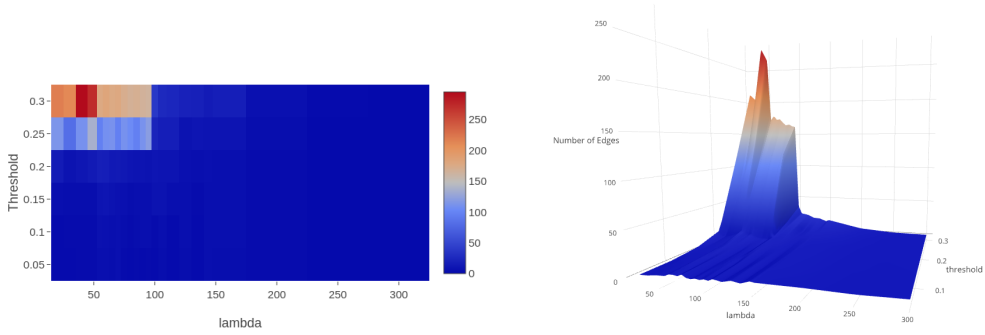


Figure 5.26: Heatmap and 3D surface plot showing the number of edges in the inferred networks with the Gibbs BMB for different λ and Credible Intervals (CI defined by [threshold, 1-threshold]).

5.2.4 Dependencies by Relevance

Starting from the sparsest graph we can decrease the λ in small steps and look at which edges are added. It is important to note, that edges can in some cases vanish again for smaller λ. That is, the network resulting from a higher λ are not always subsets of the denser networks. It is unclear, whether this is occurs due to convergence issues of the Annealing or simply is part of the models behavior. Nonetheless, the increase in specificity and precision for higher λ is still relevant. Figure 5.27 shows edges that have been observed for a λ of 1400 and higher with a threshold of 0.15. The plotted bars correspond to the highest λ they were observed at, i.e. the sparsest network that still contained them.

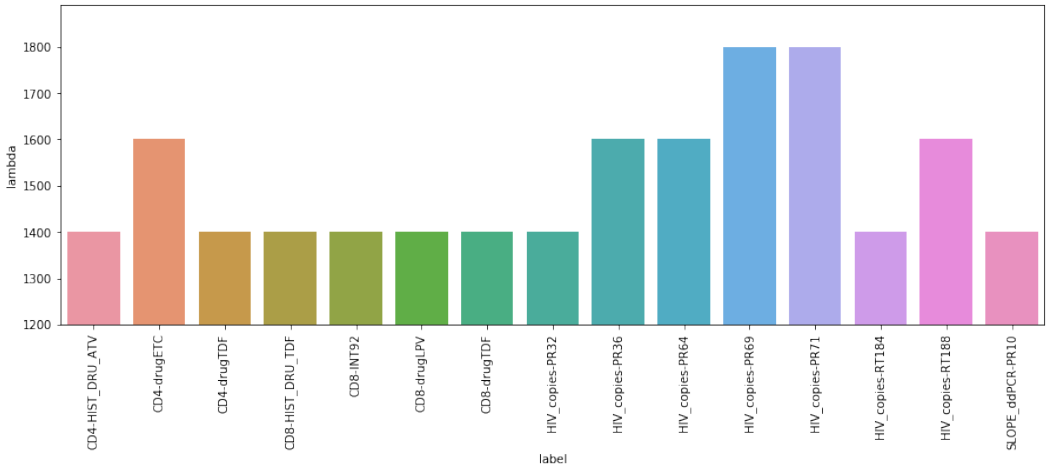


Figure 5.27: Maximum λ for which each edge was observed with threshold 0.15 and $\lambda >= 1400$.

We can see that about half of the dependencies there correspond to connections between the ddPCR counts (or its slope) and resistance relevant mutations from the haplotypes. The two first edges correspond to the haplotype mutations PR69 and PR71. According to Shafer [2017], PR69 is associated only with the Protease Inhibitor (PI) tipranavir, which has shown to be effective for patients being resistant to other PIs [Doyon et al., 2005]. Thus a mutation restricting the effectiveness of tipranavir being relevant for the HIV levels of treatment experienced patients seems reasonable. PR71 is associated with multi-drug resistance for five different PIs (of totally 8). Considering that, this resistance can limit the range of possible PI in a treatment considerably, making it an important factor for successful ART. In general we can see that a large portion of the first few dependencies correspond to resistances against protease inhibitors. For the interested reader Figure 5.28 additionally provides the edges found for λ greater than 1000 and smaller than 1400. We will not go into further details at this point and instead switch to the network graphs.

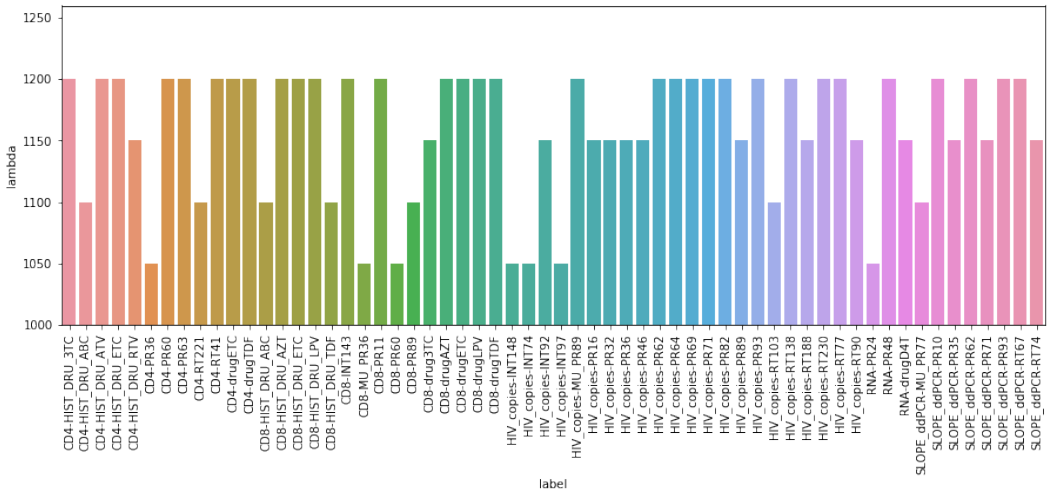
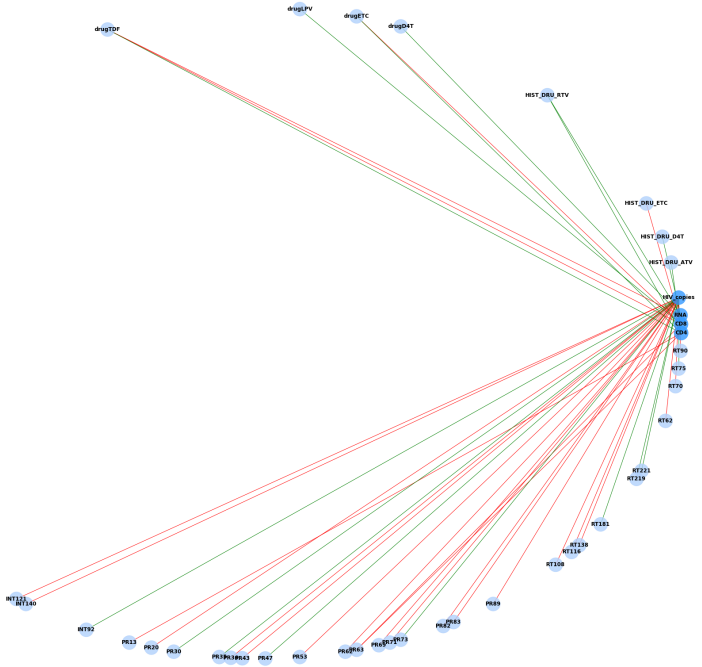


Figure 5.28: Maximum λ for which each edge was observed with threshold 0.15 and $1000 < \lambda < 1400$.

5.2.5 Dependency Networks

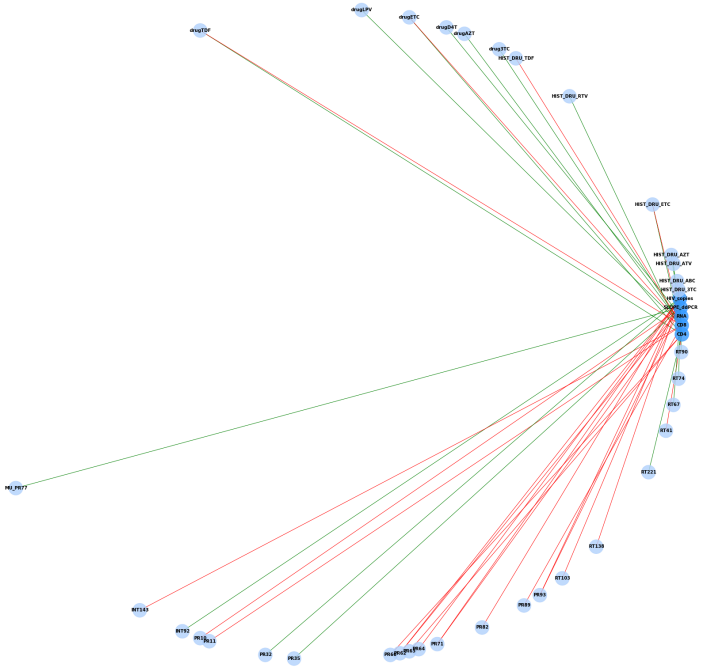
For comparing the networks resulting from the BMB and SA, we fix the graph size to 42 non-zero edges and then select a corresponding threshold and λ, which leads in our opinion to a network size that is still interpretable while not being trivial. In case of the Annealing, we left the threshold fixed at 0.15 with a λ of 1100. For the BMB, the threshold had to be finely tuned, resulting in a λ of 70 with the threshold 0.22. In Figure 5.29 and Figure 5.30 the network graphs are shown in a circular layout. It can be seen, that the networks resulting from the BMB and the SA are rather similar. Both exhibit a lot of connections with the lower section of the circle, which corresponds to the mutations calculated from the reconstructed haplotypes. In contrast, there are barely any connections to the left side, i.e. the GRT mutations. Simulated Annealing finds one dependency there, while the BMB does not show any. The found edge corresponds to a positive partial correlation between MU_PR77 affecting 3 different PIs and the ddPCR slope, indicating a deterioration caused by the mutation. However, it is unclear why barely any interactions with the GRT mutations are present in the networks. Examining GRT mutations for planning the next step in a antiretroviral therapy is currently a standard practice, due to the shown improvements in terms of therapy success Günthard et al. [2018]. Instead of assuming a lack of dependency between the clinical factors and the GRT mutations, this might indicate that the mutations inferred from the haplotypes are more accurate, thus making the GRT mutations mostly redundant when conditioning on the haplotypes.

Figure 5.31 and Figure 5.32 show the the same networks with a different layout. Here we can see that a lot of haplotype mutations form a cluster around the ddPCR values (in case of the SA also around the slope). While the connections themselves seem reasonable, the signs of the partial correlations are counter-intuitive. A red edge corresponds to a negative partial correlation. Consequently, a lot of haplotype mutations would seem to result in a decrease of the HIV levels. Instead, we would have expected an increase, as the resistance



Method: Gibbs, Lambda:70, Threshold: 0.22, #Edges: 42

Figure 5.29: W12 subnetwork estimated with Gibbs BMB.



Method: SA, Lambda:1100, Threshold: 0.15, #Edges: 42

Figure 5.30: W12 subnetwork estimated with SA.

relevant mutations generally inhibit the effectiveness of specific drugs. Furthermore, we can see that ETC and TDF have a positive edge to the CD4 values, while simultaneously being negatively correlated with the CD8 counts. In the initial phase of HIV infection, the CD8 counts increase significantly and are followed by a decrease of CD4 counts. In addition the ratio of CD4 to CD8 counts is known to be a marker for the clinical outcome in virologically suppressed HIV patients [Lu et al., 2015] and elevated CD8 counts during ART are associated with failure in antiretroviral therapy [Krantz et al., 2011]. As a consequence, we can assume that the positive correlation with CD4 and negative correlation with CD8 counts corresponds to the ETC and TDF having a positive impact on treatment success.
An interesting connection that is present in both networks (see Figure 5.33) is the positive partial correlation between RNA and the drug stavudine (D4T), with the Gibbs BMB also showing an edge to HIST_D4T. D4T is a nucleoside reverse transcriptase inhibitor (NRTI), meaning that it blocks the enzyme responsible for changing RNA into the form of DNA. While being largely used in the past as a initial treatment for patients with advanced immunodeficiency [9] the World Health Organization (WHO) no longer recommends its use due to side effects [10]. The connection might indicate that patients starting treatment at a later stage show a higher tendency for exhibiting blips in therapy. In any case, the connections between viral load and ART drugs do not shed any light on the controversial effects of the blips on developing new resistance mutations.

[9] http://www.aidsmap.com/d4T-stavudine-iZeriti/page/1730937/
[10] http://www.who.int/hiv/pub/guidelines/arv2013/arv2013supplement_to_chapter09.pdf

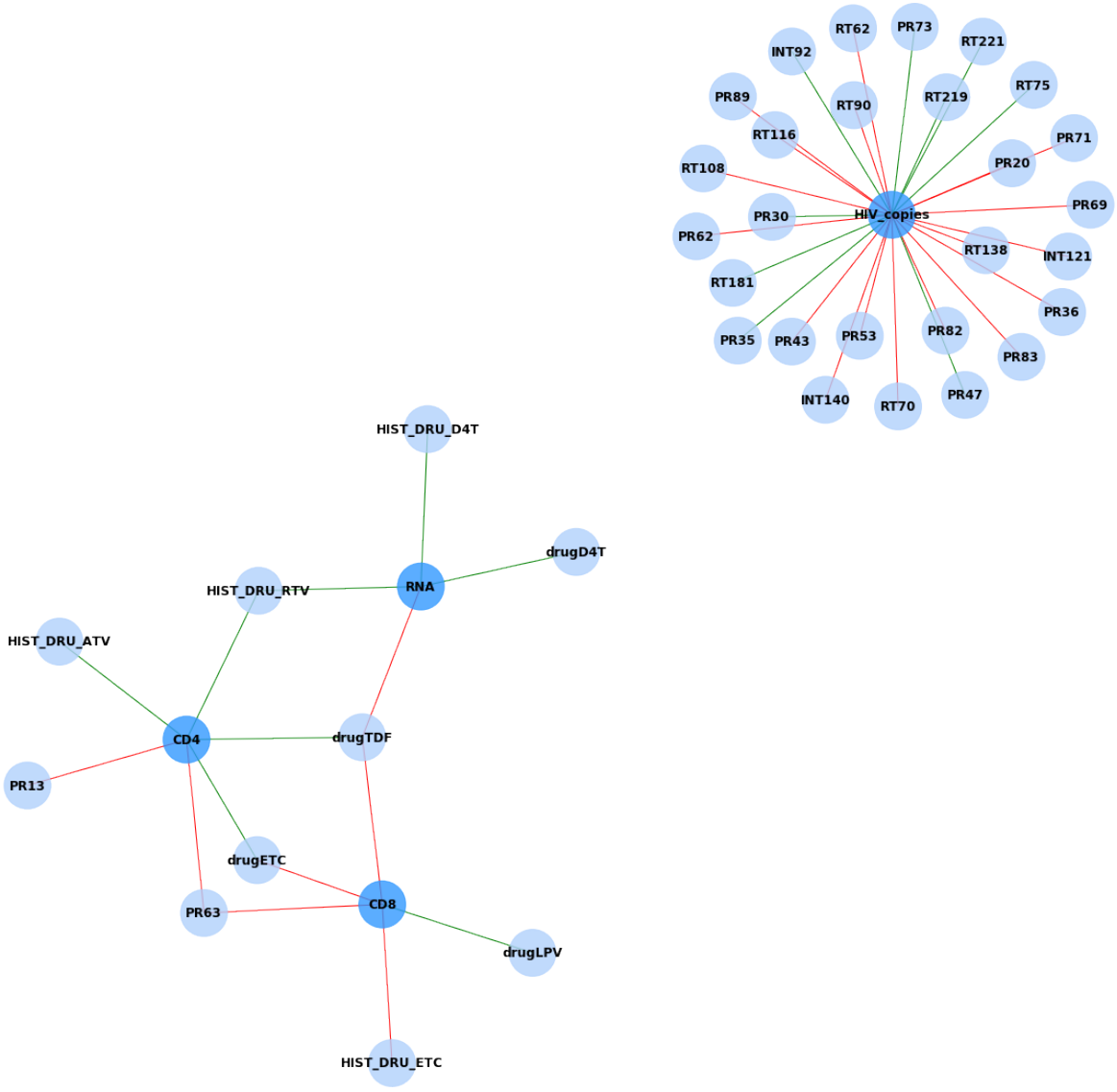
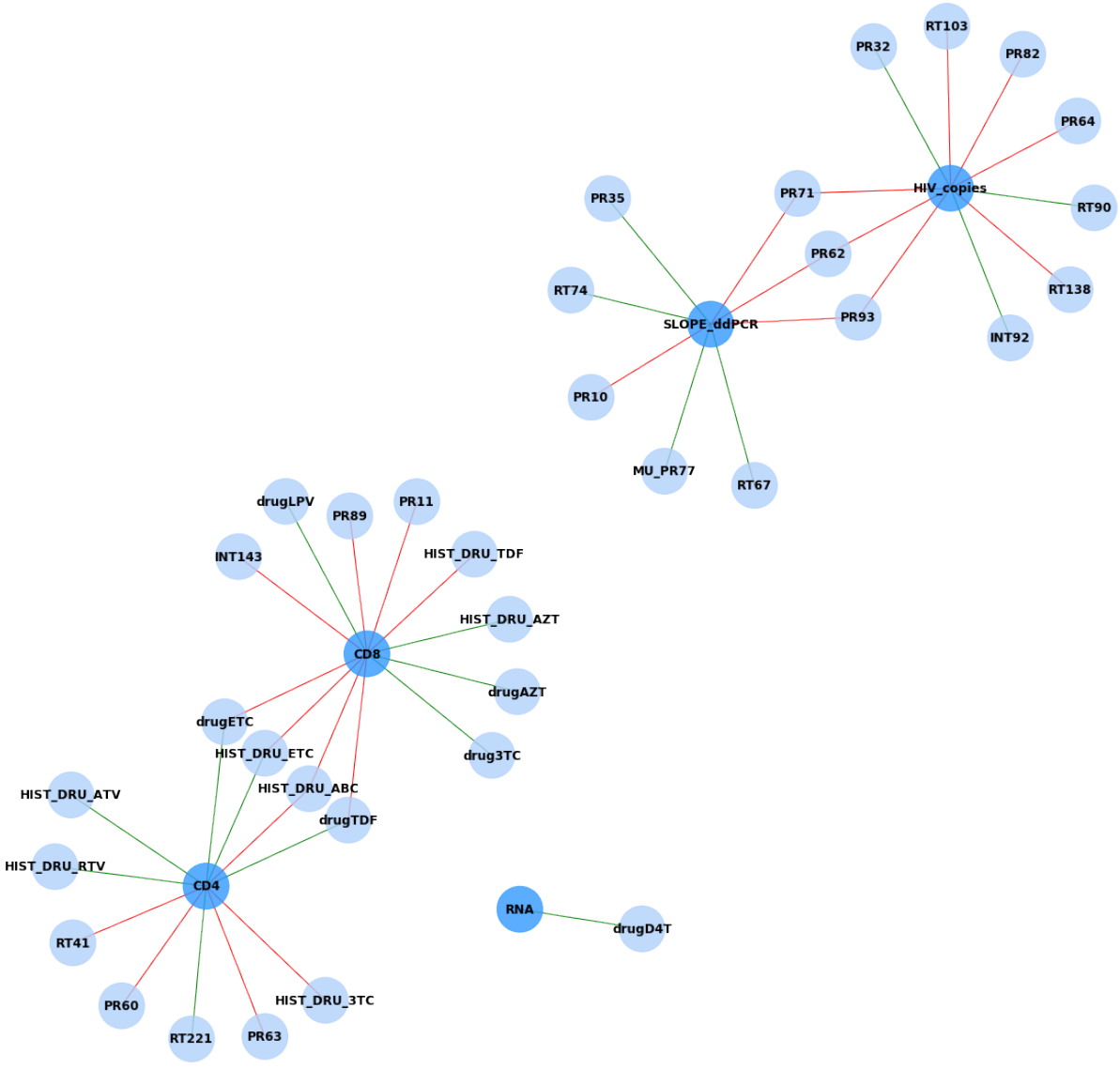


Figure 5.31: Inferred network with Gibbs BMB, using $\lambda = 70$ and threshold 0.22.



Method: SA, Lambda:1100, Threshold: 0.15, #Edges: 42

Figure 5.32: Inferred network with Simulated Annealing, using $\lambda = 1100$ and threshold 0.15.

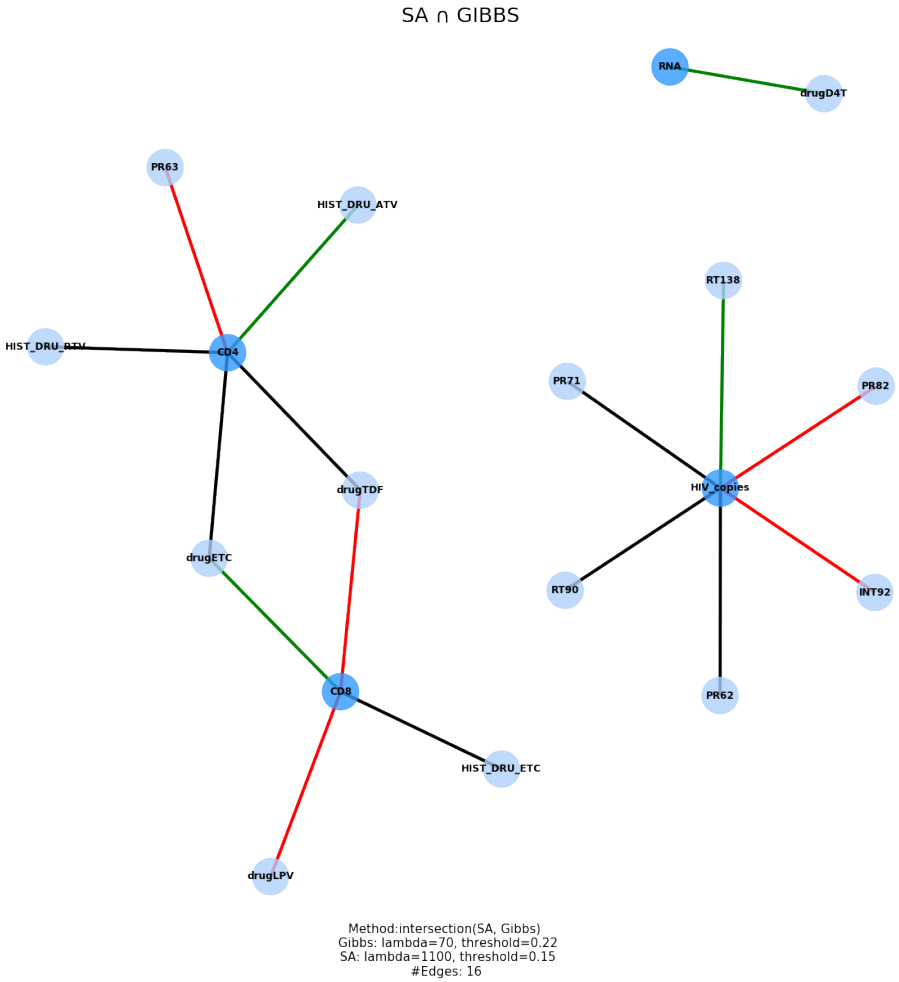


Figure 5.33: Intersection of the SA and BMB network. Black edges indicate a disagreement in the sign of the edge.

5.2.6 Latent Scores

Figure 5.34 and Figure 5.35 illustrate the effect of the copula transformation on the data for 2100 Gibbs sweeps with a burn-in of 630 and a λ of 1000. The first column shows the distribution of untransformed observations for one variable; the second, the distribution of the normal scores (which are used for initialization); the third column shows the distribution of the posterior mean of the latent values. That is, the histogram of the estimated $E[z_{i,j}|\mathbf{Z} \in D, \lambda = 1000]$ for all observations i with fixed variable j. The distributions did not show any significant change with further iteration of the sampler. The latent values of the remaining query variables are not displayed for the sake of compactness, since they did not show any unusual behavior.

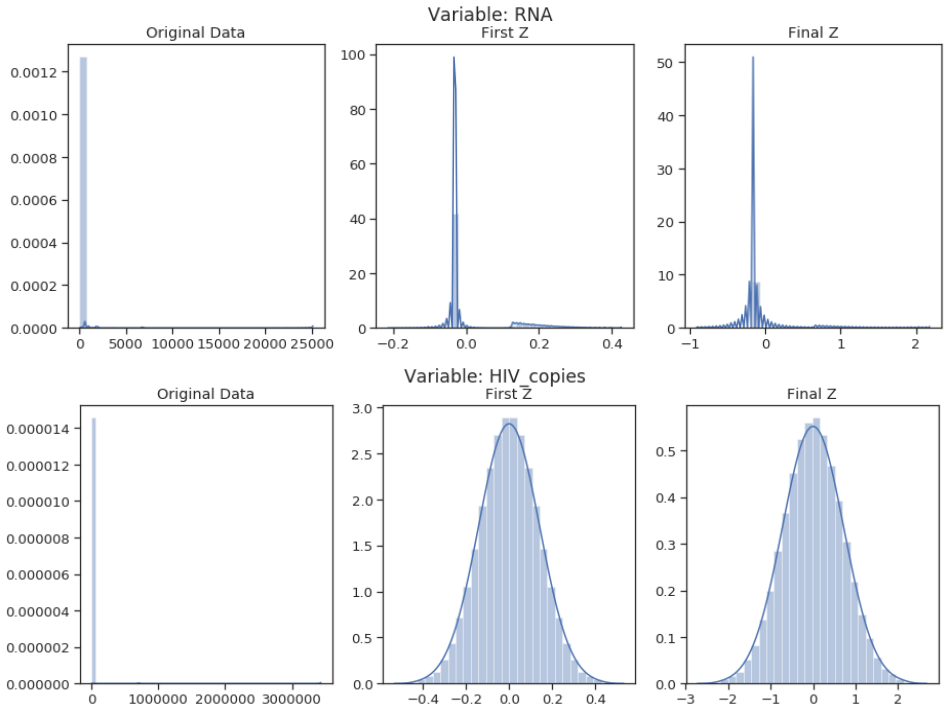


Figure 5.34: Initial distribution of the data compared to the distribution of the estimated latent values.

The semi-parametric copula seems to be troubled mainly when the data is highly imbalanced with a lot of identical values. While the estimated latent values for the ddPCR count and even the boolean HIST_DRU_3TC are relatively normal, the viral load (RNA) and the GRT mutations (in this case MU_RT70) seem to be flawed. The RNA is similar to the GRT mutations in the sense that it has the majority of samples at the same value 0, with the exception of a few blips. In contrast, the variables indicating the per-drug treatment-experience are more balanced. Additionally, the mutations inferred from the haplotypes contain a lot of missing values, which are filled by the copula sampler with normal draws. That the unbalanced GRT mutations do seemingly not converge to a normal distribution might be related to the failure to find dependencies shared with the query variables.

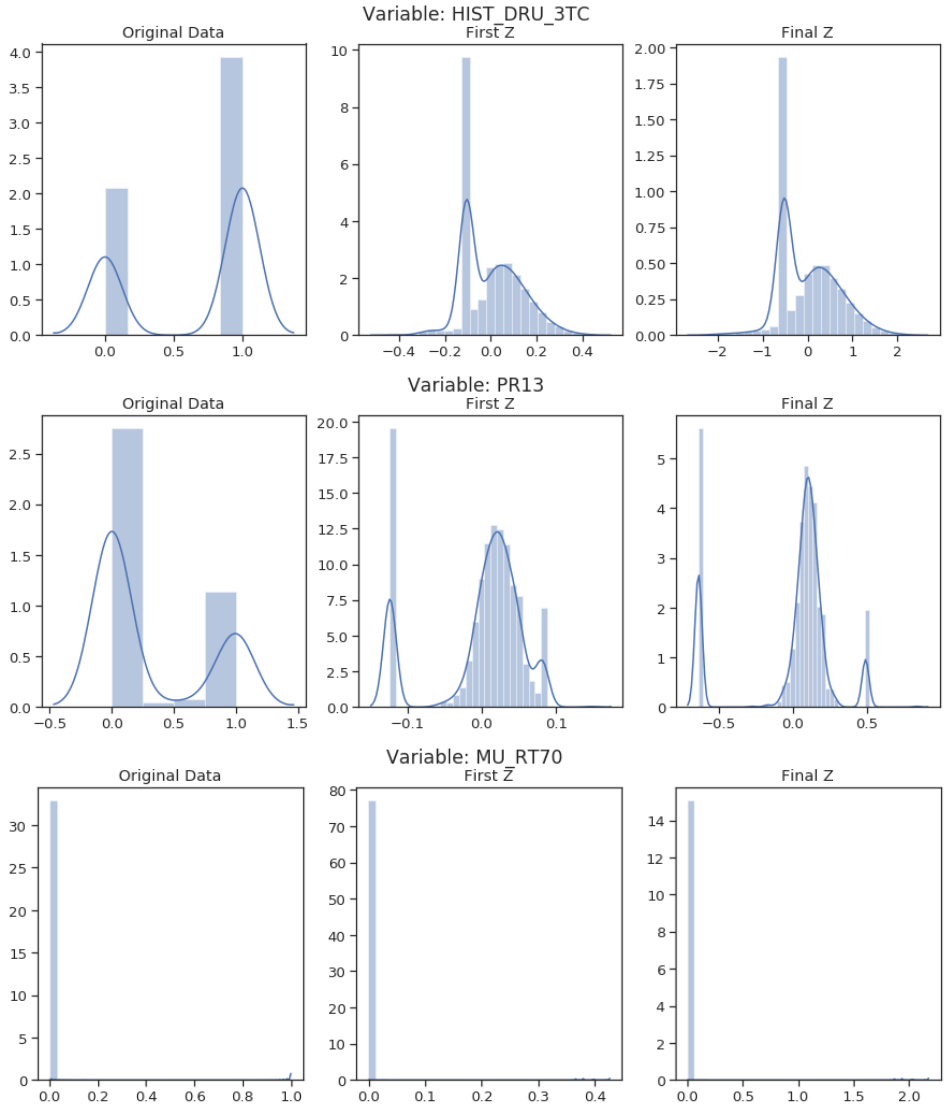


Figure 5.35: Initial distribution of the data compared to Z estimated by averaging over the draws. The second column corresponds to the distribution of the normal scores, the last column to the mean over all draws.

6
Conclusion and Future Work

In this work we have introduced Simulated Annealing as an alternative estimation procedure for the Markov Blanket. While the Annealing MAP did not result in performance competitive to the GLASSO, it was shown to be comparable to the previous approach by Kaufmann et al. [2016]. In addition, SA has proven to be more robust in its application on the HIV-X data set by allowing for a smoother process of network selection, while the BMB gave unexpected problems. The feasible λ region of the BMB lies closely to the region subject to convergence problems of the Markov chains, thus making the inferred networks less certain. In the test data however, SA had the disadvantage of an upper limit on the sparsity, presumably due to numerical instability. We assume that this instability mainly stems from the MGIG distributed $\mathbf{W}_{11}$ posterior conditional.
In applications of both models, we could identify multi-modality of the posterior marginal of $\mathbf{W}_{12}$ for small λ, while posterior marginals for higher λ seem to be unimodal. This is also reflected in the Annealing behavior, as the sampler starts fluctuating between multiple modes when in the corresponding λ region.
Finally, the application on the HIV-X dataset indicated that the GRT mutations may not have a significant effect on the clinical factors in comparison to the haplotype mutations. However, it is unclear whether the seemingly non-normal distributions of the latent values for the GRT mutations influenced this result. Although we could identify some connections in the networks that are agreeable with known results, a lot of edges express counter-intuitive partial correlations. For example, multiple resistance relevant mutations are negatively correlated with the ddPCR HIV cell count.

Future Work

One of the main shortcomings of the model underlying the BMB is the reliance on the MGIG distribution. In general, the MGIG proves to be difficult both in theory and practice due to the suboptimal methods available for sampling from it. Thus future work could be focused on finding an different sampling method. A possible alternative for directly sampling the MGIG would be substituting the draw of the MGIG with a draw from the $\mathbf{W}_{11}$ marginal $P(\mathbf{W}_{11}|\mathbf{T}, \mathbf{S}, \lambda)$. While the marginal does not offer a closed form solution in terms of a

known distribution, we can use a Metropolis Hastings sampler. Due to the similarity, we would suggest the Wishart arising from the special case of $P(\mathbf{W}_{11}|\mathbf{T}, \mathbf{S}, \lambda)$ for $\mathbf{D} \to \mathbf{0}$ (see subsubsection 3.1.4.1) as proposal function. The difference between the special case and the general case of the marginal posterior lies in the determinant:

$$\det\left((\mathbf{S}_{22} + \mathbf{I}) \otimes \mathbf{W}_{11}^{-1} + \mathbf{D}\right)$$

We know that $A = ((\mathbf{S}_{22} + \mathbf{I}) \otimes \mathbf{W}_{11}^{-1})$ is symmetric and positive definite (so all eigenvalues of A are positive), and $\mathbf{D}$ is a diagonal matrix with positive entries. With $\det(A) = \prod_i \lambda_i$ and $tr(A) = \sum_i \lambda_i$ it's clear to see that

$$\det\left((\mathbf{S}_{22} + \mathbf{I}) \otimes \mathbf{W}_{11}^{-1} + \mathbf{D}\right) > \det\left((\mathbf{S}_{22} + \mathbf{I}) \otimes \mathbf{W}_{11}^{-1}\right)$$

As only the normalization differs, we assume that the general case of the posterior marginal is a flatter version of the special case. Additionally, $\mathbf{D}$ tends to be quite small in practice, so the special case should serve as a good proposal function with a high acceptance rate.

Another interesting point would be adjusting the synthetic data to a more general setting. We have seen that the usability of the Annealing approach played a big role for the application on real data. However, the cause of this difference compared to the BMB is unclear, as the tests on artificial data did not indicate similar problems. While the artificial data does reflect the small world property of real problems, the already normally distributed data presumably obsoletes the semi-parametric copula. In contrast, many realistic use cases with mixed data rely on the copula transformation, which should motivate the use of a more appropriate data set. Because of this, we would suggest comparing the BMB and the Annealing on unbalanced and sparse data (similar to the mutations) that still exhibits small world properties.

Bibliography

D. Abramson, M. Krishnamoorthy, H. Dang, et al. Simulated annealing cooling schedules for the school timetabling problem. *Asia Pacific Journal of Operational Research*, 16: 1–22, 1999.

D. Adametz, M. Rey, and V. Roth. Information bottleneck for pathway-centric gene expression analysis. In *German Conference on Pattern Recognition*, pages 81–91. Springer, 2014.

J. B. Alimonti, T. B. Ball, and K. R. Fowke. Mechanisms of cd4+ t lymphocyte cell death in human immunodeficiency virus infection and aids. *Journal of general Virology*, 84(7): 1649–1661, 2003.

C. Andrieu and A. Doucet. Simulated annealing for maximum a posteriori parameter estimation of hidden markov models. *IEEE Transactions on Information Theory*, 46(3): 994–1004, 2000.

C. Andrieu, N. De Freitas, A. Doucet, and M. I. Jordan. An introduction to mcmc for machine learning. *Machine learning*, 50(1-2):5–43, 2003.

P. Baldi, S. Brunak, Y. Chauvin, C. A. Andersen, and H. Nielsen. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16(5):412–424, 2000.

J. A. Bartlett, R. DeMasi, J. Quinn, C. Moxham, and F. Rousseau. Overview of the effectiveness of triple combination therapy in antiretroviral-naive hiv-1 infected adults. *Aids*, 15(11):1369–1377, 2001.

E. Bernadac. Random continued fractions and inverse Gaussian distribution on a symmetric cone. *Journal of Theoretical Probability*, 8(2):221–259, Apr. 1995. ISSN 0894-9840, 1572-9230. doi: 10.1007/BF02212879.

C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0387310738.

R. W. Butler. Generalized Inverse Gaussian Distributions and their Wishart Connections. *Scandinavian Journal of Statistics*, 25(1):69–75, Mar. 1998. ISSN 1467-9469. doi: 10.1111/1467-9469.00089. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-9469.00089.

R. E. Chaisson, J. E. Gallant, J. C. Keruly, and R. D. Moore. Impact of opportunistic disease on survival in patients with hiv infection. *Aids*, 12(1):29–33, 1998.

N. Chopin. Fast simulation of truncated gaussian distributions. *Statistics and Computing*, 21(2):275–288, 2011.

J. Dagpunar. An easily implemented generalised inverse gaussian generator. *Communications in Statistics-Simulation and Computation*, 18(2):703–710, 1989.

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

L. Doyon, S. Tremblay, L. Bourgon, E. Wardrop, and M. G. Cordingley. Selection and characterization of hiv-1 showing reduced susceptibility to the non-peptidic protease inhibitor tipranavir. *Antiviral research*, 68(1):27–35, 2005.

F. Fazayeli and A. Banerjee. The matrix generalized inverse gaussian distribution: Properties and applications. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 648–664. Springer, 2016.

R. Foygel and M. Drton. Extended bayesian information criteria for gaussian graphical models. In *Advances in neural information processing systems*, pages 604–612, 2010.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, July 2008. ISSN 1465-4644, 1468-4357. doi: 10.1093/biostatistics/kxm045. URL https://academic.oup.com/biostatistics/article-lookup/doi/10.1093/biostatistics/kxm045.

S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6):721–741, 1984.

J. E. Gentle. *Computational statistics*, volume 308. Springer, 2009.

J. Geweke. Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints and the evaluation of constraint probabilities, 1991.

N. Gulzar and K. F. Copeland. Cd8+ t-cells: function and response to hiv infection. *Current HIV research*, 2(1):23–37, 2004.

H. F. Günthard, J. A. Aberg, J. J. Eron, J. F. Hoy, A. Telenti, C. A. Benson, D. M. Burger, P. Cahn, J. E. Gallant, M. J. Glesby, et al. Antiretroviral treatment of adult hiv infection: 2014 recommendations of the international antiviral society–usa panel. *Jama*, 312(4):410–425, 2014.

H. F. Günthard, V. Calvez, R. Paredes, D. Pillay, R. W. Shafer, A. M. Wensing, D. M. Jacobsen, and D. D. Richman. Human immunodeficiency virus drug resistance: 2018 recommendations of the international antiviral society–usa panel. *Clinical Infectious Diseases*, 2018.

A. K. Gupta and D. K. Nagar. *Matrix variate distributions*. Chapman and Hall/CRC, 2018.

T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.

W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. 1970.

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

P. D. Hoff et al. Extending the rank likelihood for semiparametric copula estimation. *The Annals of Applied Statistics*, 1(1):265–283, 2007.

W. Hörmann and J. Leydold. Generating generalized inverse gaussian random variates. *Statistics and Computing*, 24(4):547–557, 2014.

G. Huldrych, N. Beerenwinkel, B. Jasmina, S. Bonhoeffer, J. Fellay, R. Kouyos, K. Metzner, and V. Roth. The HIV-X MRD Project. URL http://www.systemsx.ch/projects/medical-research-and-development-projects/hiv-x/.

E. Jones, T. Oliphant, P. Peterson, et al. SciPy: Open source scientific tools for Python, 2001–. URL http://www.scipy.org/.

D. Kaufmann. *Semi-parametric Gaussian copula models for machine learning*. PhD thesis, University_of_Basel, 2017.

D. Kaufmann, S. Parbhoo, A. Wieczorek, S. Keller, D. Adametz, and V. Roth. Bayesian markov blanket estimation. In *Artificial Intelligence and Statistics*, pages 333–341, 2016.

S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *science*, 220(4598):671–680, 1983.

E. M. Krantz, K. H. Hullsiek, J. F. Okulicz, A. C. Weintrob, B. K. Agan, N. F. Crum-Cianflone, A. Ganesan, T. M. Ferguson, B. R. Hale, I. D. C. R. P. H. W. Group, et al. Elevated cd8 counts during haart are associated with hiv virologic treatment failure. *Journal of acquired immune deficiency syndromes (1999)*, 57(5):396, 2011.

S. L. Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.

P. K. Lee, T. L. Kieffer, R. F. Siliciano, and R. E. Nettles. Hiv-1 viral load blips are of limited clinical significance. *Journal of Antimicrobial Chemotherapy*, 57(5):803–805, 2006.

K. Lehner. *Erzeugung von Zufallszahlen für zwei exotische stetige Verteilungen*. na, 1989.

G. Letac and V. Seshadri. A characterization of the generalized inverse gaussian distribution by continued fractions. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 62(4):485–489, 1983.

W. Lu, V. Mehraj, K. Vyboh, W. Cao, T. Li, and J.-P. Routy. Cd4: Cd8 ratio as a frontier marker for clinical outcome, immune dysfunction and viral reservoir size in virologically suppressed hiv-positive patients. *Journal of the International AIDS Society*, 18(1):20052, 2015.

V. Martinez, A.-G. Marcelin, J.-P. Morini, J. Deleuze, A. Krivine, I. Gorin, S. Yerly, L. Perrin, G. Peytavin, V. Calvez, et al. Hiv-1 intermittent viraemia in patients treated by non-nucleoside reverse transcriptase inhibitor-based regimen. *Aids*, 19(10):1065–1069, 2005.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6): 1087–1092, 1953.

K. P. Murphy. *Machine learning : a probabilistic perspective*. MIT Press, 1 edition, Aug. 2013. ISBN 0262018020. URL http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0262018020.

S. Nadarajah, E. Afuecheta, and S. Chan. A compendium of copulas. *Statistica*, 77(4): 279–328, 2018.

A. Noë, J. Plum, and C. Verhofstede. The latent hiv-1 reservoir in patients undergoing haart: an archive of pre-haart drug resistance. *Journal of Antimicrobial Chemotherapy*, 55(4):410–412, 2005.

T. Park and G. Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

S. Prabhakaran, M. Rey, O. Zagordi, N. Beerenwinkel, and V. Roth. Hiv haplotype inference using a propagating dirichlet process mixture model. *IEEE/ACM transactions on computational biology and bioinformatics*, 11(1):182–191, 2014.

G. R. Price. Extension of covariance selection mathematics. *Annals of human genetics*, 35 (4):485–490, 1972.

S. Raman and V. Roth. Sparse Point Estimation for Bayesian Regression via Simulated Annealing. In A. Pinz, T. Pock, H. Bischof, and F. Leberl, editors, *Pattern Recognition*, pages 317–326, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-32717-9.

M. Rey and V. Roth. Meta-gaussian information bottleneck. In *Advances in Neural Information Processing Systems*, pages 1916–1924, 2012.

M. Roederer, J. G. Dubs, M. T. Anderson, P. A. Raju, and L. A. Herzenberg. Cd8 naive t cell counts decrease progressively in hiv-infected adults. *The Journal of clinical investigation*, 95(5):2061–2066, 1995.

D. W. Scott. Feasibility of multivariate density estimates. *Biometrika*, 78(1):197–205, 1991.

R. W. Shafer. Genotypic testing for human immunodeficiency virus type 1 drug resistance. *Clinical microbiology reviews*, 15(2):247–277, 2002.

R. W. Shafer. Human immunodeficiency virus type 1 drug resistance mutations update. *The Journal of infectious diseases*, 216(suppl_9):S843–S846, 2017.

T. S. H. C. S. SHCS, F. Schoeni-Affolter, B. Ledergerber, M. Rickenbach, C. Rudin, H. F. Günthard, A. Telenti, H. Furrer, S. Yerly, and P. Francioli. Cohort profile: The swiss hiv cohort study. *International Journal of Epidemiology*, 39(5):1179–1189, 2010. doi: 10.1093/ije/dyp321. URL http://dx.doi.org/10.1093/ije/dyp321.

M. Sklar. Fonctions de repartition an dimensions et leurs marges. *Publ. inst. statist. univ. Paris*, 8:229–231, 1959.

C. J. Stone. Optimal rates of convergence for nonparametric estimators. *The annals of Statistics*, pages 1348–1360, 1980.

M. C. Strain, S. M. Lada, T. Luong, S. E. Rought, S. Gianella, V. H. Terry, C. A. Spina, C. H. Woelk, and D. D. Richman. Highly precise measurement of hiv dna by droplet digital pcr. *PloS one*, 8(4):e55943, 2013.

J. Stuart, A. M. Wensing, C. Kovacs, M. Righart, D. de Jong, S. Kaye, R. Schuurman, C. J. Visser, and C. A. Boucher. Transient relapses (" blips") of plasma hiv rna levels during haart are associated with drug resistance. *JAIDS-HAGERSTOWN MD-*, 28(2):105–113, 2001.

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

M. A. Wainberg and G. Friedland. Public health implications of antiretroviral therapy and hiv drug resistance. *Jama*, 279(24):1977–1983, 1998.

H. Wang. Bayesian Graphical Lasso Models and Efficient Posterior Computation. *Bayesian Analysis*, 7(4):867–886, Dec. 2012. ISSN 1936-0975, 1931-6690. doi: 10.1214/12-BA729. URL https://projecteuclid.org/euclid.ba/1354024465.

M. West. On Scale Mixtures of Normal Distributions. *Biometrika*, 74(3):646, Sept. 1987. ISSN 00063444. doi: 10.2307/2336707. URL http://www.jstor.org/stable/2336707?origin=crossref.

C. Yuan, T.-C. Lu, and M. J. Druzdzel. Annealed map. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 628–635. AUAI Press, 2004.

A
Appendix

A.1 Distributions

A.1.1 Matrix Generalized Inverse Gaussian

The Matrix MGIG distribution is a probability distribution over positive definite symmetric (pxp) matrices $\{X : X > 0\}$ Butler [1998]. It has recently been shown to be unimodal [Fazayeli and Banerjee, 2016].

A.1.1.1 Butler Parameterization

In the models we mostly denote the MGIG by the parametrization of Butler [1998] as it is the most commonly used one.

$$\mathbf{X} \sim \mathcal{MGIG}_B(\lambda, \mathbf{A}, \mathbf{B}) \tag{A.1}$$

$$p(\mathbf{X}) \propto \det(\mathbf{X})^{\lambda - \frac{1}{2}(p+1)} \exp \operatorname{tr}\Big(-\frac{1}{2}(\mathbf{AX} + \mathbf{BX}^{-1})\Big)$$

A.1.1.2 Letac Parameterization

For sampling from the MGIG, we use a notation similar to the one used for the GIG in Letac and Seshadri [1983].

$$\mathbf{X} \sim \mathcal{MGIG}_L(n', \mathbf{A}, \mathbf{B}) \tag{A.2}$$

$$p(\mathbf{X}) \propto \det(\mathbf{X})^{-n'-1} \exp \operatorname{tr}\Big(-\frac{1}{2}(\mathbf{AX} + \mathbf{BX}^{-1})\Big) \qquad n' > \frac{p-1}{2}$$

A.2 Model

$$\begin{aligned} \mathbf{x}_{1,\dots,n} &\overset{iid}{\sim} \mathcal{N}_{(p+q)}(0, \mathbf{\Sigma}) \\ \mathbf{X} &= (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) \quad \mathbf{X} \in \mathbb{R}^{(p+q)\times n} \end{aligned}$$

$$\begin{aligned} \mathbf{S} &= \mathbf{X}\mathbf{X}^T \\ \mathbf{S} &\sim \mathcal{W}_{p+q}(n, \mathbf{\Sigma}) \end{aligned}$$

$$\boldsymbol{\Sigma}^{-1} = \mathbf{W} = \begin{matrix} \\ p \\ q \end{matrix} \begin{matrix} \begin{matrix} p & q \end{matrix} \\ \begin{pmatrix} \mathbf{W}_{11} & \mathbf{W}_{12} \\ \mathbf{W}_{12}^T & \mathbf{W}_{22} \end{pmatrix} \end{matrix} \qquad \mathbf{S} = \begin{matrix} \\ p \\ q \end{matrix} \begin{matrix} \begin{matrix} p & q \end{matrix} \\ \begin{pmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{12}^T & \mathbf{S}_{22} \end{pmatrix} \end{matrix}$$

Let

$$\mathbf{W}_{22.1} = \mathbf{W}_{22} - \mathbf{W}_{12}^T \mathbf{W}_{11}^{-1} \mathbf{W}_{12}$$

be the Schur complement of $\mathbf{W}_{22}$.

A.2.1 Likelihood

$$p(\mathbf{S}|\mathbf{W}) \propto \det(\mathbf{W})^{\frac{n}{2}} det(\mathbf{S})^{\frac{n-(p+q)-1}{2}} \exp \operatorname{tr} \big(-\frac{1}{2}\mathbf{W}\mathbf{S} \big)$$

A.2.2 Prior

$$\begin{aligned} P(\mathbf{W}|\mathbf{T}) &= \mathcal{W}_{p+q}\big(p+q+1, \mathbf{I}\big) p(\mathbf{W}_{12}|\mathbf{T}) \\ &\propto \exp \operatorname{tr} \Big(-\frac{1}{2}\mathbf{W} \Big) \prod_{\substack{i=1,\dots,p \\ j=1,\dots,q}} \frac{1}{\sqrt{2\pi \mathbf{T}_{ij}}} \exp \Big(-\frac{(\mathbf{W}_{12})_{ij}^2}{2\mathbf{T}_{ij}} \Big) \end{aligned}$$

$$P(\mathbf{T}|\lambda) \propto \prod_{\substack{i=1,\dots,p \\ j=1,\dots,q}} \frac{\lambda^2}{2} \exp \Big(-\frac{\lambda^2}{2} \mathbf{T}_{ij} \Big)$$

A.2.3 Joint Distribution

$$
\begin{aligned}
p(\mathbf{W},\mathbf{S},\mathbf{T}|\lambda) =& p(\mathbf{W}_{11},\mathbf{W}_{12},\mathbf{W}_{22},\mathbf{S},\mathbf{T}|\lambda) \\
\propto\quad & \det(\mathbf{W})^{\frac{n}{2}}\,\det(\mathbf{S})^{\frac{n-(p+q)-1}{2}} \\
& \times \exp\operatorname{tr}\Big(-\frac{1}{2}\mathbf{W}\mathbf{S}\Big)\exp\operatorname{tr}\Big(-\frac{1}{2}\mathbf{W}\Big) \\
& \times \prod_{\substack{i=1,\dots,p\\ j=1,\dots,q}} \frac{1}{\sqrt{2\pi\mathbf{T}_{ij}}}\exp\Big(-\frac{(\mathbf{W}_{12})_{ij}^2}{2\mathbf{T}_{ij}}\Big) \\
& \times \prod_{\substack{i=1,\dots,p\\ j=1,\dots,q}} \frac{\lambda^2}{2}\exp\Big(-\frac{\lambda^2}{2}\mathbf{T}_{ij}\Big)
\end{aligned}
$$

$$
\begin{aligned}
=& \det(\mathbf{W})^{\frac{n}{2}}\,\det(\mathbf{S})^{\frac{n-(p+q)-1}{2}} \\
& \exp\left(-\frac{1}{2}\operatorname{tr}[\mathbf{W}\mathbf{S}+\mathbf{W}] - \frac{1}{2}\sum_{\substack{i=1,\dots,p\\ j=1,\dots,q}}\frac{(\mathbf{W}_{12})_{ij}^2}{\mathbf{T}_{ij}}\right) \\
& \times \Big[\prod_{\substack{i=1,\dots,p\\ j=1,\dots,q}}\frac{1}{\sqrt{2\pi\mathbf{T}_{ij}}}\Big]p(\mathbf{T}|\lambda)
\end{aligned}
$$

A.2.3.1 Reparametrization with $\mathbf{W}_{11},\mathbf{W}_{12},\mathbf{W}_{22.1}$

$$
J\big((\mathbf{W}_{11},\mathbf{W}_{12},\mathbf{W}_{22})\to(\mathbf{W}_{11},\mathbf{W}_{12},\mathbf{W}_{22.1})\big) = \mathbf{1}
$$

$$
\begin{aligned}
\det(\mathbf{W}) &= \det(\mathbf{W}_{11})\det(\mathbf{W}_{22}-\mathbf{W}_{12}^T\mathbf{W}_{11}^{-1}\mathbf{W}_{12}) \\
&= \det(\mathbf{W}_{11})\det(\mathbf{W}_{22.1})
\end{aligned}
$$

$$
\begin{aligned}
\operatorname{tr}(\mathbf{W}\mathbf{S}) &= \operatorname{tr}\left[\mathbf{W}_{11}\mathbf{S}_{11}+\mathbf{W}_{12}\mathbf{S}_{21}+\mathbf{W}_{21}\mathbf{S}_{12}+\mathbf{W}_{22}\mathbf{S}_{22}\right] \\
&= \operatorname{tr}\left[\mathbf{W}_{11}\mathbf{S}_{11}+\mathbf{W}_{12}\mathbf{S}_{12}^T+\mathbf{W}_{12}^T\mathbf{S}_{12}+(\mathbf{W}_{22.1}+\mathbf{W}_{12}^T\mathbf{W}_{11}^{-1}\mathbf{W}_{12})\mathbf{S}_{22}\right] \\
&= \operatorname{tr}\left[\underline{\mathbf{W}_{11}\mathbf{S}_{11}}+\underline{\mathbf{W}_{12}\mathbf{S}_{12}^T}+\underline{\mathbf{W}_{12}^T\mathbf{S}_{12}}+\underline{\mathbf{W}_{22.1}\mathbf{S}_{22}}+\underline{\mathbf{W}_{12}^T\mathbf{W}_{11}^{-1}\mathbf{W}_{12}\mathbf{S}_{22}}\right]
\end{aligned}
$$

$$
\begin{aligned}
\operatorname{tr}(\mathbf{W}) &= \operatorname{tr}(\mathbf{W}_{11})+\operatorname{tr}(\mathbf{W}_{22}) \\
&= \operatorname{tr}(\underline{\mathbf{W}_{11}})+\operatorname{tr}(\underline{\mathbf{W}_{22.1}})+\operatorname{tr}(\underline{\mathbf{W}_{12}^T\mathbf{W}_{11}^{-1}\mathbf{W}_{12}})
\end{aligned}
$$

$$
\begin{aligned}
p(\mathbf{W}_{11}, \mathbf{W}_{12}, \mathbf{W}_{22.1}, \mathbf{S}, \mathbf{T}|\lambda) \propto \\
&\det(\mathbf{W}_{11})^{n/2} \det(\mathbf{W}_{22.1})^{n/2} \det(\mathbf{S})^{\frac{n-(p+q)-1}{2}} \\
&\times \exp\Big(-\frac{1}{2} \operatorname{tr}\big[\underline{\mathbf{W}_{11}(\mathbf{S}_{11}+\mathbf{I})} + \underline{\mathbf{W}_{22.1}(\mathbf{S}_{22}+\mathbf{I})} + \\
&\qquad \underline{2(\mathbf{W}_{12}^T \mathbf{S}_{12})} + \underline{\mathbf{W}_{12}^T \mathbf{W}_{11}^{-1} \mathbf{W}_{12}(\mathbf{S}_{22}+\mathbf{I})}\big]\Big) \\
&\times \exp\Big(-\frac{1}{2} \sum_{\substack{i=1,\dots,p \\ j=1,\dots,q}} \frac{(\mathbf{W}_{12})_{ij}^2}{\mathbf{T}_{ij}}\Big) \quad \Big[\prod_{\substack{i=1,\dots,p \\ j=1,\dots,q}} \frac{1}{\sqrt{2\pi \mathbf{T}_{ij}}}\Big] \quad p(\mathbf{T}|\lambda)
\end{aligned}
$$

A.2.4 Full Posterior

Let $\mathbf{D} = \operatorname{diag}(\operatorname{vec}(\mathbf{T}))^{-1}$, i.e. a diagonal matrix, where the elements are the inverses of the elements of $\mathbf{T}$

$$
\exp\Big(-\frac{1}{2} \sum_{\substack{i=1,\dots,p \\ j=1,\dots,q}} \frac{(\mathbf{W}_{12})_{ij}^2}{\mathbf{T}_{ij}}\Big) = \exp\Big(-\frac{1}{2} \underline{\operatorname{vec}(\mathbf{W}_{12})^T \ \mathbf{D} \ \operatorname{vec}(\mathbf{W}_{12})}\Big)
$$

$$
\operatorname{tr}(\underline{\mathbf{W}_{12}^T \mathbf{S}_{12}}) = \operatorname{vec}(\mathbf{W}_{12})^T \operatorname{vec}(\mathbf{S}_{12})
$$

$$
\begin{aligned}
p(\mathbf{W}_{11}, \mathbf{W}_{12}, \mathbf{W}_{22.1}, \mathbf{T}|\mathbf{S}, \lambda) \propto & \det(\mathbf{W}_{11})^{n/2} \exp\Big(-\frac{1}{2}\operatorname{tr}\left[\mathbf{W}_{11}(\mathbf{S}_{11}+\mathbf{I})\right]\Big) \\
& \times \det(\mathbf{W}_{22.1})^{n/2} \exp\Big(-\frac{1}{2}\operatorname{tr}\left[\mathbf{W}_{22.1}(\mathbf{S}_{22}+\mathbf{I})\right]\Big) \\
& \times \exp\Big(-\frac{1}{2}\operatorname{tr}\left[\mathbf{W}_{12}^T\mathbf{W}_{11}^{-1}\mathbf{W}_{12}(\mathbf{S}_{22}+\mathbf{I})\right]\Big) \\
& \times \exp\Big(-\frac{1}{2}\operatorname{vec}(\mathbf{W}_{12})^T\ \mathbf{D}\ \operatorname{vec}(\mathbf{W}_{12})\Big) \\
& \times \exp\Big(-\operatorname{vec}(\mathbf{W}_{12})^T\operatorname{vec}(\mathbf{S}_{12})\Big) \\
& \times \Big[\prod_{\substack{i=1,\dots,p \\ j=1,\dots,q}} \frac{1}{\sqrt{2\pi\mathbf{T}_{ij}}}\Big] p(\mathbf{T}|\lambda)
\end{aligned}
\tag{A.3}
$$

$$
\begin{aligned}
\operatorname{tr}\left[\mathbf{W}_{12}^T\mathbf{W}_{11}^{-1}\mathbf{W}_{12}(\mathbf{S}_{22}+\mathbf{I})\right] &= \operatorname{vec}(\mathbf{W}_{12})^T\operatorname{vec}(\mathbf{W}_{11}^{-1}\mathbf{W}_{12}(\mathbf{S}_{22}+\mathbf{I})) \\
&\overset{\text{(520 Matr. Cookbook)}}{=} \operatorname{vec}(\mathbf{W}_{12})^T\Big[(\mathbf{S}_{22}+\mathbf{I})^T\otimes\mathbf{W}_{11}^{-1}\Big]\operatorname{vec}(\mathbf{W}_{12}) \\
&= \operatorname{vec}(\mathbf{W}_{12})^T\Big[(\mathbf{S}_{22}+\mathbf{I})\otimes\mathbf{W}_{11}^{-1}\Big]\operatorname{vec}(\mathbf{W}_{12})
\end{aligned}
$$

$$
\begin{aligned}
p(\mathbf{W}_{11}, \mathbf{W}_{12}, \mathbf{W}_{22.1}, \mathbf{T}|\mathbf{S}, \lambda) \propto & \det(\mathbf{W}_{11})^{n/2} \exp\Big(-\frac{1}{2}\operatorname{tr}\left[\mathbf{W}_{11}(\mathbf{S}_{11}+\mathbf{I})\right]\Big) \\
& \times \det(\mathbf{W}_{22.1})^{n/2} \exp\Big(-\frac{1}{2}\operatorname{tr}\left[\mathbf{W}_{22.1}(\mathbf{S}_{22}+\mathbf{I})\right]\Big) \\
& \times \exp\Big(-\frac{1}{2}\operatorname{vec}(\mathbf{W}_{12})^T\Big[(\mathbf{S}_{22}+\mathbf{I})\otimes\mathbf{W}_{11}^{-1}+\mathbf{D}\Big]\operatorname{vec}(\mathbf{W}_{12}) - \operatorname{vec}(\mathbf{W}_{12})^T\operatorname{vec}(\mathbf{S}_{12})\Big) \\
& \times \Big[\prod_{\substack{i=1,\dots,p \\ j=1,\dots,q}} \frac{1}{\sqrt{2\pi\mathbf{T}_{ij}}}\Big] p(\mathbf{T}|\lambda)
\end{aligned}
\tag{A.4}
$$

A.2.4.1 Factorization of the Posterior

We can now factorize wrt. $W_{22.1}$.

$$
p(\mathbf{W}_{11}, \mathbf{W}_{12}, \mathbf{W}_{22.1}, \mathbf{T}|\mathbf{S}, \lambda) = p(\mathbf{W}_{11}, \mathbf{W}_{12}, \mathbf{T}|\mathbf{S}, \lambda)p(\mathbf{W}_{22.1}|\mathbf{S}, \lambda)
$$

Using

$$
\begin{aligned}
\det(\mathbf{S})^{\frac{n-(p+q)-1}{2}} &= \det(\mathbf{S}_{22})^{\frac{n-p-q-1}{2}}\det(\mathbf{S}_{11.2})^{\frac{n-p-q-1}{2}} \\
&= \det(\mathbf{S}_{22})^{\frac{n-p-2q-1}{2}}\det(\mathbf{S}_{22}^{-1})^{-\frac{q}{2}}\det(\mathbf{S}_{11.2})^{\frac{n-p-q-1}{2}}
\end{aligned}
$$

It follows that

$$
p(\mathbf{W}_{22.1}|\mathbf{S}, \lambda) \propto \det(\mathbf{S}_{22})^{\frac{n-p-q-1}{2}}\det(\mathbf{W}_{22.1})^{\frac{n}{2}}\exp\Big(-\frac{1}{2}\operatorname{tr}\left[\mathbf{W}_{22.1}(\mathbf{S}_{22}+\mathbf{I})\right]\Big) \tag{A.5}
$$

Comparing the determinants of Equation A.5 to the ones from the Wishart (denoted by df, furthermore we know the size of the Wishart distributed matrix is q):

$$\begin{aligned}\det(\mathbf{W}_{22.1})^{\frac{df-q-1}{2}} &= \det(\mathbf{W}_{22.1})^{\frac{n}{2}} \\ \det(\mathbf{S}_{22})^{\frac{df}{2}} &= \det(\mathbf{S}_{22})^{\frac{n-p-2q-1}{2}}\end{aligned}$$

$$\begin{aligned}\Rightarrow q &= df - n - 1 \\ df &= n - p - 2q - 1 \\ &= n - p - 2(df - n - 1) - 1 \\ \Rightarrow 3df &= 3n - p + 1 \\ \Rightarrow df &= n - \frac{(p-1)}{3}\end{aligned}$$

$$\Downarrow$$

$$\boxed{\mathbf{W}_{22.1}|\mathbf{S}, \lambda \sim \mathcal{W}\Big(n - \frac{(p-1)}{3}, \mathbf{S}_{22}\Big)} \tag{A.6}$$

Additionally, now the joint posterior for $(\mathbf{W}_{11}, \mathbf{W}_{12})$ is:

$$\begin{aligned}p(\mathbf{W}_{11}, \mathbf{W}_{12}, \mathbf{T}|\mathbf{S}, \lambda) \propto \quad & \det(\mathbf{W}_{11})^{n/2} \exp\Big(-\frac{1}{2}\operatorname{tr}\big[\mathbf{W}_{11}(\mathbf{S}_{11} + \mathbf{I})\big]\Big) \\ & \times \exp\Big(-\frac{1}{2}\operatorname{vec}(\mathbf{W}_{12})^T\Big[(\mathbf{S}_{22} + \mathbf{I}) \otimes \mathbf{W}_{11}^{-1} + \mathbf{D}\Big]\operatorname{vec}(\mathbf{W}_{12}) - \operatorname{vec}(\mathbf{W}_{12})^T\operatorname{vec}(\mathbf{S}_{12})\Big) \\ & \times \Big[\prod_{\substack{i=1,\dots,p \\ j=1,\dots,q}} \frac{1}{\sqrt{2\pi\mathbf{T}_{ij}}}\Big] p(\mathbf{T}|\lambda)\end{aligned} \tag{A.7}$$

A.2.5 Posterior Conditionals

A.2.5.1 $\mathbf{W}_{12}$

We condition Equation A.7 additionally on $\mathbf{T}$ and $\mathbf{W}_{11}$:

$$\begin{aligned}p(\mathbf{W}_{12}|\mathbf{W}_{11}, \mathbf{T}, \mathbf{S}, \lambda) &\propto \exp\Big(-\frac{1}{2}\operatorname{vec}(\mathbf{W}_{12})^T\Big[(\mathbf{S}_{22} + \mathbf{I}) \otimes \mathbf{W}_{11}^{-1} + \mathbf{D}\Big]\operatorname{vec}(\mathbf{W}_{12}) - \operatorname{vec}(\mathbf{W}_{12})^T\operatorname{vec}(\mathbf{S}_{12})\Big) \\ &= \exp\Big(-\frac{1}{2}x^T\mathbf{C}x - b^Tx\Big) \\ &\overset{\text{(8.16Matr. Cookbook)}}{=} \exp\Big(-\frac{1}{2}(x + \mathbf{C}^{-1}b)^T\mathbf{C}(x + \mathbf{C}^{-1}b) + \frac{1}{2}b^T\mathbf{C}^{-1}b\Big)\end{aligned}$$

with

$$x = \operatorname{vec}(\mathbf{W}_{12}) \qquad \mathbf{C} = \Big[(\mathbf{S}_{22} + \mathbf{I}) \otimes \mathbf{W}_{11}^{-1} + \mathbf{D}\Big] \qquad b = \operatorname{vec}(\mathbf{S}_{12})$$

$$\Downarrow$$

$$\boxed{\operatorname{vec}(\mathbf{W}_{12})|\mathbf{W}_{11}, \mathbf{T}, \mathbf{S}, \lambda \sim \mathcal{N}(-\mathbf{C}^{-1}\operatorname{vec}(\mathbf{S}_{12}), \mathbf{C}^{-1})} \tag{A.8}$$

A.2.5.2 $\mathbf{W}_{11}$

We take Equation A.3, marginalize wrt. $\mathbf{W}_{22.1}$ and then additionally condition on $\mathbf{W}_{12}, \mathbf{T}$:

$$\begin{aligned} p(\mathbf{W}_{11}|\mathbf{W}_{12},\mathbf{T},\mathbf{S},\lambda) &\propto \det(\mathbf{W}_{11})^{n/2} \exp\Big(-\frac{1}{2}\operatorname{tr}\big[\mathbf{W}_{11}(\mathbf{S}_{11}+\mathbf{I}) + \mathbf{W}_{12}(\mathbf{S}_{22}+\mathbf{I})\mathbf{W}_{12}^T\mathbf{W}_{11}^{-1}\big]\Big) \\ &= \det(\mathbf{W}_{11})^{n/2} \exp\Big(-\frac{1}{2}\operatorname{tr}\big[\underbrace{(\mathbf{S}_{11}+\mathbf{I})}_{\text{MGIG A}}\mathbf{W}_{11} + \underbrace{\mathbf{W}_{12}(\mathbf{S}_{22}+\mathbf{I})\mathbf{W}_{12}^T}_{\text{MGIG B}}\mathbf{W}_{11}^{-1}\big]\Big) \end{aligned}$$

$$\Downarrow$$

Comparing the determinants to the (Butler) MGIG defined in subsubsection A.1.1.1 leads to

$$\boxed{\mathbf{W}_{11}|\mathbf{W}_{12},\mathbf{T},\mathbf{S},\lambda \sim \mathcal{MGIG}_B\Big(\frac{n+p+1}{2},\ \mathbf{S}_{11}+\mathbf{I},\ \mathbf{W}_{12}(\mathbf{S}_{22}+\mathbf{I})\mathbf{W}_{12}^T\Big)} \tag{A.9}$$

A.2.5.3 $\mathbf{T}$

$$p(\mathbf{T}|\mathbf{W}_{11},\mathbf{W}_{12},\mathbf{S},\lambda) \propto \prod_{\substack{i=1,\dots,p\\ j=1,\dots,q}} \frac{1}{\sqrt{2\pi\mathbf{T}_{ij}}} \exp\Big(-\frac{(\mathbf{W}_{12})_{ij}^2}{2\mathbf{T}_{ij}}\Big) \prod_{\substack{i=1,\dots,p\\ j=1,\dots,q}} \frac{\lambda^2}{2}\exp\Big(-\frac{\lambda^2}{2}\mathbf{T}_{ij}\Big)$$

$$\begin{aligned} p(\mathbf{T}_{ij}|\mathbf{W}_{11},\mathbf{W}_{12},\mathbf{S},\lambda) \propto& \frac{1}{\sqrt{2\pi\mathbf{T}_{ij}}}\exp\Big(-\frac{(\mathbf{W}_{12})_{ij}^2}{2\mathbf{T}_{ij}}\Big)\frac{\lambda^2}{2}\exp\Big(-\frac{\lambda^2}{2}\mathbf{T}_{ij}\Big) \\ =& \frac{\lambda^2}{2}\Big(\frac{1}{2\pi\mathbf{T}_{ij}}\Big)^{1/2}\exp\Big(-\frac{(\mathbf{W}_{12})_{ij}^2}{2\mathbf{T}_{ij}} - \frac{\lambda^2}{2}\mathbf{T}_{ij}\Big) \end{aligned}$$

Substitute $\mathbf{T}_{ij}$ with $\mathbf{T}_{ij}^{-1}$: $\quad J(\mathbf{T}_{ij} \to \mathbf{T}_{ij}^{-1}) = -\mathbf{T}_{ij}^{-2}$

$$\begin{aligned} p\Big(\frac{1}{\mathbf{T}_{ij}}\Big) \propto& -\mathbf{T}_{ij}^{-2}\frac{\lambda^2}{2}\Big(\frac{1}{2\pi\mathbf{T}_{ij}^{-1}}\Big)^{1/2}\exp\Big(-\frac{(\mathbf{W}_{12})_{ij}^2}{2}\mathbf{T}_{ij} - \frac{\lambda^2}{2}\mathbf{T}_{ij}^{-1}\Big) \\ =& -\frac{\lambda^2}{2}\Big(\frac{1}{2\pi\mathbf{T}_{ij}^3}\Big)^{1/2}\exp\Big(-\frac{(\mathbf{W}_{12})_{ij}^2}{2\mathbf{T}_{ij}}\Big(\mathbf{T}_{ij}^2 - \frac{\lambda^2}{(\mathbf{W}_{12})_{ij}^2}\Big)\Big) \end{aligned}$$

$$\text{Substitute } (\mu')^2 = \frac{\lambda^2}{(\mathbf{W}_{12})_{ij}^2}$$

$$\begin{aligned} &= -\frac{\lambda^2}{2}\Big(\frac{1}{2\pi\mathbf{T}_{ij}^3}\Big)^{1/2}\exp\Big(-\frac{\lambda^2}{2(\mu')^2\mathbf{T}_{ij}}\big(\mathbf{T}_{ij}^2 - 2\mathbf{T}_{ij}\mu' + 2\mathbf{T}_{ij}\mu' + (\mu')^2\big)\Big) \\ &\propto -\frac{\lambda^2}{2}\Big(\frac{1}{2\pi\mathbf{T}_{ij}^3}\Big)^{1/2}\exp\Big(-\frac{\lambda^2}{2(\mu')^2\mathbf{T}_{ij}}\big(\mathbf{T}_{ij}-\mu'\big)^2 - \frac{\lambda^2}{\mu'}\Big) \\ &\propto \Big(\frac{\lambda^2}{2\pi\mathbf{T}_{ij}^3}\Big)^{1/2}\exp\Big(-\frac{\lambda^2}{2(\mu')^2\mathbf{T}_{ij}}\big(\mathbf{T}_{ij}-\mu'\big)^2\Big) \end{aligned}$$

$$\downarrow$$

$$\boxed{\mathbf{T}_{ij}^{-1} \sim \mathcal{IG}\Big(\mu' = \sqrt{\frac{\lambda^2}{(\mathbf{W}_{12})_{ij}^2}}, \lambda' = \lambda^2\Big)} \tag{A.10}$$

A.3 Cooling of the posterior conditionals

A.3.1 Inverse Gaussian

Let the Inverse Gaussian (IG) Distribution be denoted by

$$\boldsymbol{X} \sim \mathcal{IG}(\mu, \lambda)$$

$$p(x) = \Big(\frac{\lambda}{2\pi x^3}\Big)^{1/2} \exp\Big[\frac{-\lambda(x-\mu)^2}{2\mu^2 x}\Big]$$

and the Generalized Inverse Gaussian (GIG) Distribution by

$$\boldsymbol{Y} \sim \mathcal{GIG}(a, b, p)$$

$$p(y) = \frac{(a/b)^{p/2}}{2K_p(\sqrt{ab})} y^{(p-1)} \exp\Big[-\frac{1}{2}\big(ay + \frac{b}{y}\big)\Big]$$

We can write the $\mathcal{IG}$ distribution as a special case of the $\mathcal{GIG}$ distribution:
With

$$\boldsymbol{Y} \sim \mathcal{GIG}(a = \frac{\lambda}{\mu^2}, b = \lambda, p = -\frac{1}{2}) \tag{A.11}$$

we get

$$p(x \geq \boldsymbol{X}) = p(x \geq \boldsymbol{Y})$$

As a consequence, we can write the cooled $\mathcal{IG}$ in terms of the $\mathcal{GIG}$:

$$p(y) \propto y^{(p-1)} \exp\Big[-\frac{1}{2}\big(ax + \frac{b}{y}\big)\Big]$$

$$\begin{aligned}
&T \in \mathbb{R}_{>0} \\
p(y)^{\frac{1}{T}} &\propto y^{\frac{(p-1)}{T}} \exp\Big[-\frac{1}{2T}\big(ay + \frac{b}{y}\big)\Big] \\
&\propto y^{(\frac{p-1}{T}+1)-1} \exp\Big[-\frac{1}{2}\Big(\frac{a}{T}y + \frac{b/T}{y}\Big)\Big] \\
&\propto p(z)
\end{aligned}$$

$$\boldsymbol{Z} \sim \mathcal{GIG}\Big(a' = \frac{a}{T}, b' = \frac{B}{T}, p' = \Big(\frac{p-1}{T} + 1\Big)\Big)$$

using Equation A.11 we get:

$$\boldsymbol{Z} \sim \mathcal{GIG}\Big(a' = \frac{\lambda/\mu^2}{T}, b' = \frac{\lambda}{T}, p' = \Big(-\frac{1.5}{T} + 1\Big)\Big) \tag{A.12}$$

A.3.2 Matrix Generalized Inverse Gaussian

$$\begin{aligned}\boldsymbol{X} &\sim \mathcal{MGIG}_B(\lambda, A, B)\\ p(X) &\propto \det(X)^{\lambda-\frac{1}{2}(p+1)} \exp \operatorname{tr}\Big(-\frac{1}{2}(\mathbf{A}X+\mathbf{B}X^{-1})\Big)\end{aligned}$$

$$\begin{aligned}T &\in \mathbb{R}_{>0}\\ p(X)^{\frac{1}{T}} &\propto det(X)^{\frac{\lambda-\frac{1}{2}(p+1)}{T}} exp[tr(-\frac{1}{2}(AX+BX^{-1}))]^{\frac{1}{T}}\\ &\propto det(X)^{\frac{\lambda}{T}-\frac{1}{2T}(p+1)} exp[tr(-\frac{1}{2T}(AX+BX^{-1}))]\\ &\propto det(X)^{\frac{\lambda}{T}-\frac{1}{2T}(p+1)} exp[tr(-\frac{1}{2}(\frac{A}{T}X+\frac{B}{T}X^{-1}))]\\ &\propto det(X)^{\left(\frac{\lambda}{T}-\frac{1}{2T}(p+1)+\frac{1}{2}(p+1)\right)-\frac{1}{2}(p+1)} exp[tr(-\frac{1}{2}(\frac{A}{T}X+\frac{B}{T}X^{-1}))]\\ &\propto det(X)^{\left(\frac{\lambda}{T}+\frac{p+1}{2}(1-\frac{1}{T})\right)-\frac{1}{2}(p+1)} exp[tr(-\frac{1}{2}(\frac{A}{T}X+\frac{B}{T}X^{-1}))]\\ &\propto p(Y)\\ \boldsymbol{Y} &\sim \mathcal{MGIG}_B\left(\Big(\frac{\lambda}{T}+\frac{p+1}{2}(1-\frac{1}{T})\Big),\ \frac{A}{T},\ \frac{B}{T}\right)\end{aligned}$$

A.3.3 Normal Distribution

$$\begin{aligned}X &\sim N(\mu, \Sigma)\\ p(x) &= \left((2\pi)^k \det(\Sigma)\right)^{-\frac{1}{2}} exp[-\frac{1}{2}(x-\mu)^t\Sigma^{-1}(x-\mu)]\end{aligned}$$

$$\begin{aligned}T &\in \mathbb{R}_{>0}\\ p(x)^{\frac{1}{T}} &= \left((2\pi)^k \det(\Sigma)\right)^{-\frac{1}{2T}} exp[-\frac{1}{2}(x-\mu)^t\Sigma^{-1}(x-\mu)]^{\frac{1}{T}}\\ &\propto exp[-\frac{1}{2}(x-\mu)^t\Sigma^{-1}(x-\mu)]^{\frac{1}{T}}\\ &\propto exp[-\frac{1}{2T}(x-\mu)^t\Sigma^{-1}(x-\mu)]\\ &\propto exp[-\frac{1}{2}(x-\mu)^t\frac{\Sigma^{-1}}{T}(x-\mu)]\\ &\propto exp[-\frac{1}{2}(x-\mu)^t(T\Sigma)^{-1}(x-\mu)]\\ &\propto p(y)\\ Y &\sim \mathcal{N}(\mu, T\Sigma)\end{aligned}$$

A.3.4 Wishart Distribution

$$
\begin{aligned}
S &\sim W_p(n, \Sigma) \\
p(S) &\propto det(S)^{\frac{1}{2}(n-p-1)} etr\left(-\frac{1}{2}\Sigma^{-1}S\right) && S > 0, n \geq p \\
&\propto det(S)^{\frac{1}{2T}(n-p-1)} etr\left(-\frac{1}{2T}\Sigma^{-1}S\right) \\
&\propto det(S)^{\frac{1}{2}\left(\left(\frac{n-p-1}{T}+p+1\right)-p-1\right)} etr\left(-\frac{1}{2}(T\Sigma)^{-1}S\right) \\
&\propto p(y) \\
Y &\sim W_p\left(\frac{n-p-1}{T}+p+1, T\Sigma\right)
\end{aligned}
$$

A.4 Reconstructed Graphs from Artificial Data

In this section, multiple graphs reconstructed with the BMB and SA are shown and compared to the ground truth in terms of their graph intersection (i.e. the true positives) and the false positive predictions.

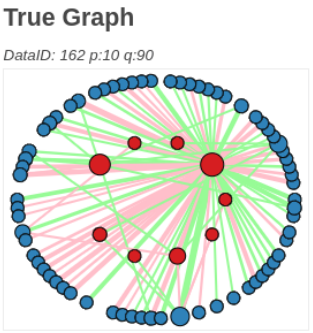


Figure A.1: True W12 subnetwork (for artificial data set with id 162)

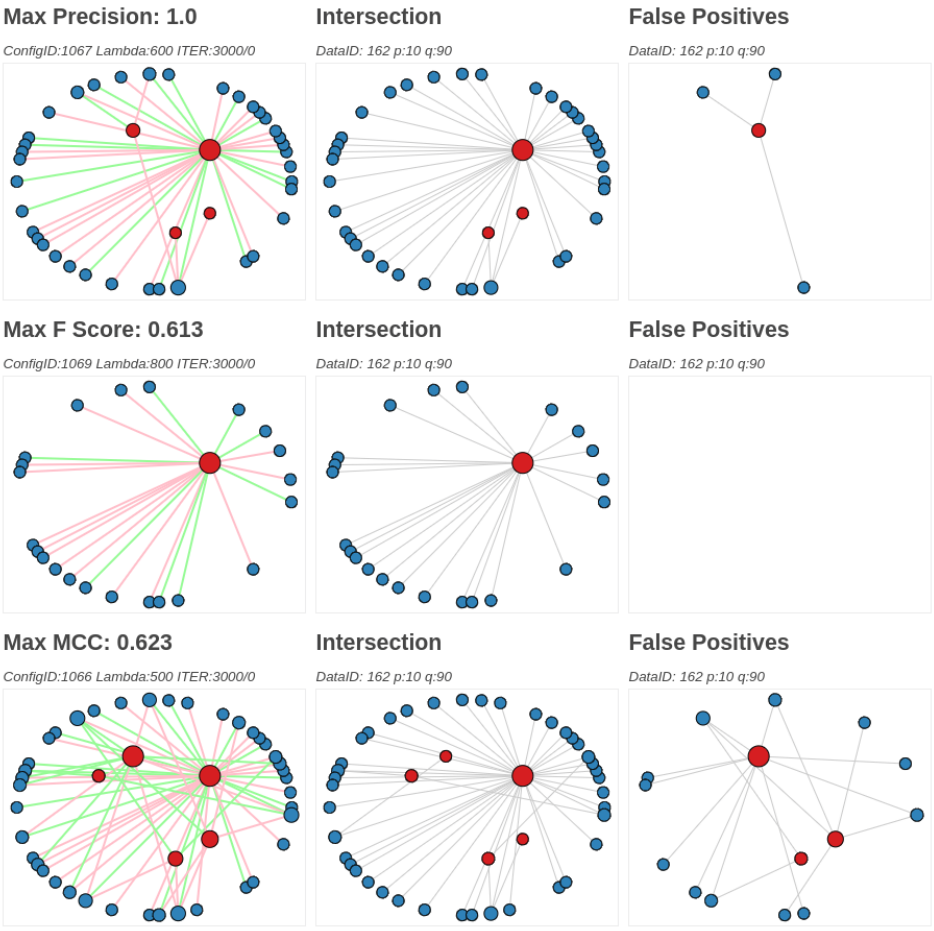


Figure A.2: W12 subnetwork reconstructed with Gibbs BMB (for artificial data set with id 162)

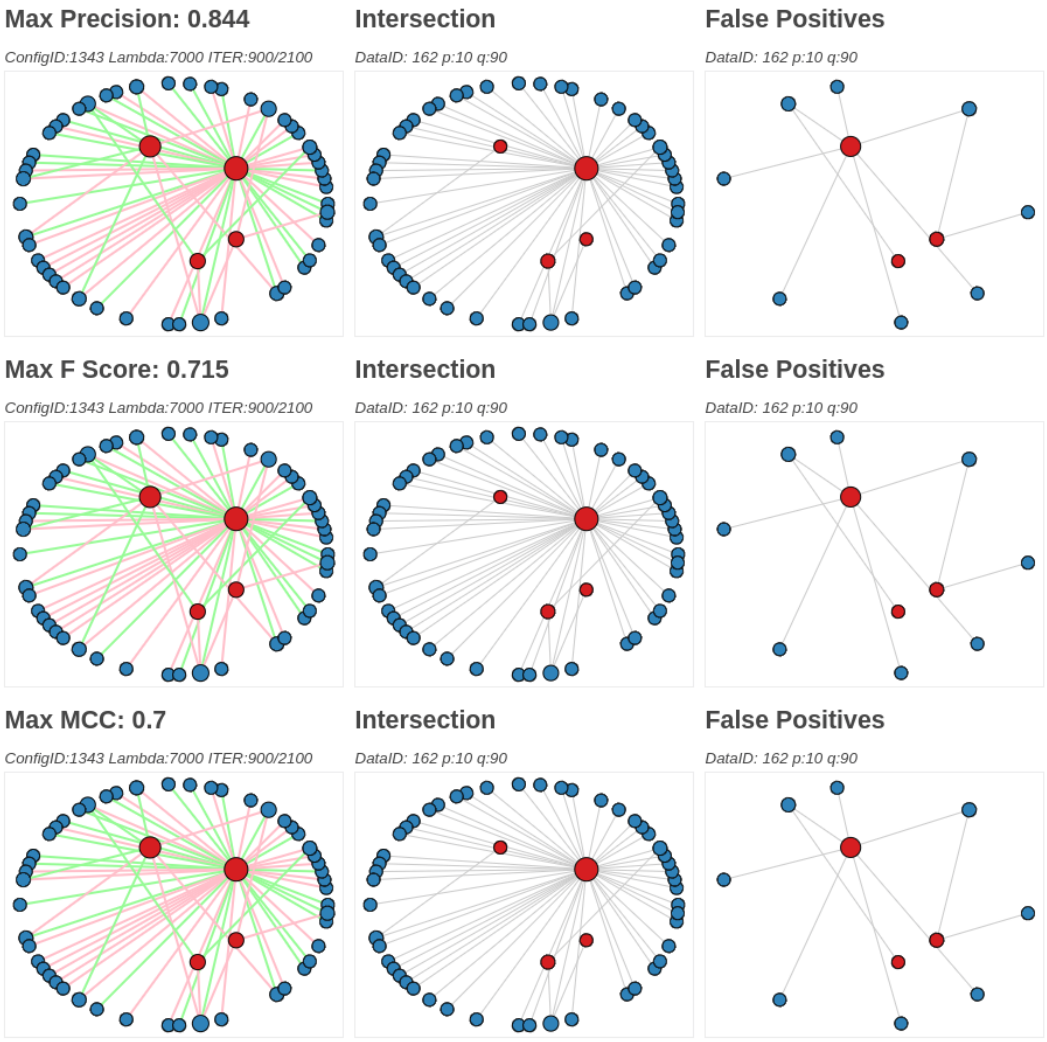


Figure A.3: W12 subnetwork reconstructed with Simulated Annealing (for artificial data set with id 162)

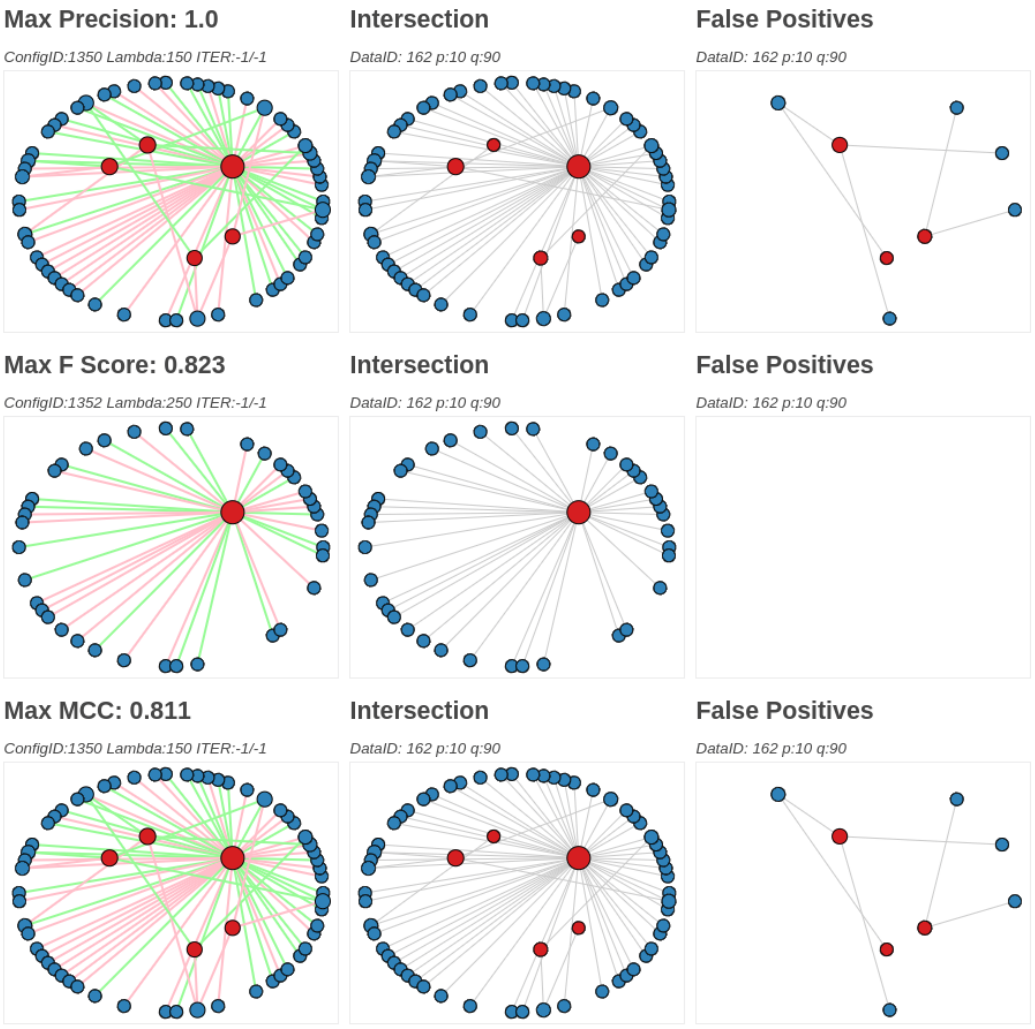


Figure A.4: W12 subnetwork reconstructed with GLASSO(for artificial data set with id 162)

Declaration on Scientific Integrity
Erklärung zur wissenschaftlichen Redlichkeit

includes Declaration on Plagiarism and Fraud
beinhaltet Erklärung zu Plagiat und Betrug

Author — Autor
Fabricio Arend Torres

Matriculation number — Matrikelnummer
2012-051-934

Title of work — Titel der Arbeit
Sampling and Annealing for Dependency Subnetwork Estimation

Type of work — Typ der Arbeit
Post-Handin Fixed Version of the Master Thesis

Declaration — Erklärung
I hereby declare that this submission is my own work and that I have fully acknowledged the assistance received in completing this work and that it contains no material that has not been formally acknowledged. I have mentioned all source materials used and have cited these in accordance with recognised scientific rules.

Hiermit erkläre ich, dass mir bei der Abfassung dieser Arbeit nur die darin angegebene Hilfe zuteil wurde und dass ich sie nur mit den in der Arbeit angegebenen Hilfsmitteln verfasst habe. Ich habe sämtliche verwendeten Quellen erwähnt und gemäss anerkannten wissenschaftlichen Regeln zitiert.

Basel, Last Change: 19th November 2018

Signature — Unterschrift