Invariances for Gaussian Models

Inauguraldissertation
zur
Erlangung der Würde eines Doktors der Philosophie
vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel

von

David Adametz
aus Reutlingen, Deutschland

Basel, 2015

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät
auf Antrag von

Prof. Dr. Volker Roth, Dissertationsleiter

Prof. Dr. Thomas Vetter, Korreferent

Basel, den 19. Mai 2015

Prof. Dr. Jörg Schibler, Dekan

Abstract

At the heart of a statistical analysis, we are interested in drawing conclusions about random variables and the laws they follow. For this we require a sample, therefore our approach is best described as learning from data. In many instances, we already have an intuition about the generating process, meaning the space of all possible models reduces to a specific class that is defined up to a set of unknown parameters. Consequently, learning becomes the task of inferring these parameters given observations. Within this scope, the thesis answers the following two questions:

Why are invariances needed? Among all parameters of the model, we often distinguish between those of interest and the so-called nuisance. The latter does not carry any meaning for our purposes, but may still play a crucial role in how the model supports the parameters of interest. This is a fundamental problem in statistics which is solved by finding suitable transformations such that the model becomes invariant against unidentifiable properties. Often, the application at hand already decides upon the necessary requirements: a Euclidean distance matrix, for example, does not carry translational information of the underlying coordinate system.

Why Gaussian models? The normal distribution constitutes an important class in statistics due to frequent occurrences in nature, hence it is highly relevant for many research disciplines including physics, astronomy, engineering, but also psychology and social sciences. Besides fundamental results like the central limit theorem, a significant part of its appeal is rooted in convenient mathematical properties which permit closed-form solutions to numerous problems.

In this work, we develop and discuss generalizations of three established models: a Gaussian mixture model, a Gaussian graphical model and the Gaussian information bottleneck. On the one hand, all of these are analytically convenient, but on the other hand they suffer from strict normality requirements which seriously limit their range of application. To this end, our focus is to explore solutions and relax these restrictions. We successfully show that with the addition of invariances, the aforementioned models gain a substantial leap forward while retaining their core concepts of the Gaussian foundation.

Acknowledgments

I am deeply grateful to my advisor Prof. Dr. Volker Roth without whom this work would not have come to fruition. Looking back at countless constructive and enlightening discussions, he influenced the thesis in a major fashion and helped me shape it into what it became today. His wholehearted dedication to machine learning was *the* driving force during my time in Basel.

Further, I am greatly indebted to Prof. Dr. Thomas Vetter for all the time and effort he put into the review of the thesis. His comments and suggestions enabled me to view the topics from a different angle and lead to significant improvements.

The whole Biomedical Data Analysis group played an important part in both theoretical and practical aspects of this work. This includes the strong line-up of Dinu Kaufmann, Sonali Parbhoo, Aleksander Wieczorek and Sebastian Keller, but also former colleagues Sandhya Prabhakaran, Melanie Rey, Behrouz Tajoddin, Julia Vogt and Sudhir Raman. The teamwork was nothing short of exceptional and everybody contributed to a friendly and encouraging atmosphere. Also, I would like to thank all colleagues for making my stay a very pleasant one, in particular Ghazi Bouabene, Nenad Stojnic, Filip-Martin Brinkmann, Manolis Sifalakis, Klara Spalek, Ivan Giangreco, Claudiu Tanase, Matthias Amberg, Marcel Lüthi, Bernhard Egger, Andreas Forster, Thomas Gerig, Clemens Blumer and everybody of GRAVIS.

Outside of the computer science unit, I am glad to have been part of numerous collaborations with Christian Vogler, Eva Dazert, Paul Jenö, Suzanne Moes, Pankaj Shende and Zuzanna Makowska. Thank you for opening my eyes to other research fields and giving me valuable insights into your area of expertise.

My wife Jing Sheng deserves a special mention for her love, dedication and never-ending patience during this journey. She has always been my source of inspiration.

Finally, I would like to express gratitude to my parents Gottfried and Elisabeth for all their ongoing support, my sisters Katrin and Dora for creative input and, of course, my good friend Matthias Eckert for technical advice and his knowledge in LaTeX typesetting.

Thank you all, indeed.

Symbols and Notation

	Symbol	Description	Format
Univariate Normal			
	X	random variable	1×1
	μ	mean	1×1
	σ^2	variance	1×1
	x	realization	1×1
Multivariate Normal			
	$\boldsymbol{X}$	vector of random variables	$p \times 1$
	$\boldsymbol{\mu}$	mean vector	$p \times 1$
	Σ	covariance matrix	$p \times p$
	$\boldsymbol{x}$	realization	$p \times 1$
Matrix-variate Normal			
	X	matrix of random variables	$p \times n$
	M	mean matrix	$p \times n$
	Σ	row covariance matrix	$p \times p$
	W	inverse row covariance matrix	$p \times p$
	Ψ	column covariance matrix	$n \times n$
	X	realization	$p \times n$
	$\boldsymbol{v}$	vector of row means	$p \times 1$
	$\boldsymbol{w}$	vector of column means	$n \times 1$
	S	inner product matrix	$p \times p$
	D	squared Euclidean distance matrix	$p \times p$
Vector and Matrix			
	$\mathbf{1}_p$	column vector of 1s	$p \times 1$
	$0_{p \times n}$	matrix of 0s	$p \times n$
	$X^\top$	transpose of matrix X	
	Σ^{-1}	inverse of matrix Σ	

Likelihood			
	$F(\bullet)$	distribution function	
	$f(\bullet)$	density function	
	$L(\bullet)$	likelihood function	
	$\ell(\bullet)$	log-likelihood function	
	ψ	parameter of interest	
	λ	nuisance parameter	
Distributions			
	$\mathcal{N}(\bullet)$	normal distribution	
	$\mathcal{G}(\bullet)$	gamma distribution	
	$\mathcal{W}(\bullet)$	Wishart distribution	
	$\mathcal{T}(\bullet)$	T distribution	
	$\mathcal{U}(\bullet)$	uniform distribution	
Sets			
	$\mathbb{R}$, $\mathbb{R}^p$, $\mathbb{R}^{p\times n}$	real numbers/vectors/matrices	
	diag	diagonal matrices	
	$\mathbb{S}_+$	symm. pos.-definite matrices	
	$\mathbb{S}_-$	symm. neg.-definite matrices	
Clustering			
	π	mixture weight	1×1
	A	inner product of cluster means	$k \times k$
	B	counterpart of A in the inverse covariance matrix	$k \times k$
	Z	cluster assignment matrix	$p \times k$
Compression			
	$C(\bullet)$, $C_R(\bullet)$	copula, Gaussian copula	
	$c(\bullet)$	copula density	
	R	correlation matrix	$p \times p$
	$\bar{X}$	standard normal random variable	1×1

Contents

Chapter 1
Introduction & Basic Concepts

In the field of machine learning, we are interested in drawing conclusions about *random variables* (*rvs*) and the principles and laws they follow. For this, we require *realizations*, also referred to as *observations*, or *the sample*. In essence, the approach can be described as learning from data.

Often, we already have some knowledge about the process from which the underlying data are generated, so it is possible to restrict our analysis to a specific class, where only the parameters need to be determined. For example, one may assume a normal distribution and be interested in estimates of the mean and/or the variance. Thus, whenever we limit the space of all possible processes to a smaller set, we also speak of a *model*, and, if such a set is defined up to a parametrization, the model is said to be of *parametric* kind. It is important to bear in mind that all subsequent analysis is conditional on this model and the explanatory power may suffer greatly if this is not valid. Any conclusion we draw is based on this very choice, hence, we will assume that it either matches the true process or is reasonably close.

As suggested by the title, the thesis is centered around the normal (or equivalently Gaussian) distribution, which constitutes an important class in statistics due to special properties and its frequent appearances in nature. When we restrict ourselves to the normal distribution, our model is parametric and we seek to identify mean and/or variance, or in the higher-dimensional case, mean and/or covariance *matrix*. In short, this is the most concise and compact formulation of our goal.

Although there exist estimators for the parameters of Gaussian models when all data are available in full, some interesting special cases arise if the information carried by the sample is limited. As an example, consider kernel or distance matrices, which are functions or *statistics* of the original data, but leave out vital properties: due to pairwise evaluation, a part of what characterizes each individual entity is inevitably lost. The problem becomes more evident when we think of three points in a Euclidean space: Having access to the individual properties or *features* of the points gives a more informative representation than only their relative distances. Thus, if distances are observed, there is an infinite number of geometric configurations from which the outcome *could* have originated. Unfortunately, this issue often precludes a direct application of Gaussian models in case it requires the full representation of the data. If we settle for one possible choice, it potentially determines how, say, the covariance matrix is inferred eventually. Hence, we find ourselves in a position, where some specific information is required by the model, but its reconstruction may interfere with the outcome.

In order to solve the above problem, we will apply certain transformations to the model, such that it only depends on the limited information at our disposal. Speaking in more formal terms, the idea is closely tied to the principle of *sufficiency*, stating that for a suitable statistic of the data, the *probability density* factors into two parts: one depends on the parametrization and the other is a function solely of the data. The latter term can then safely be discarded since it does not affect how the parameters are inferred. In general, we wish to remove as much irrelevant information as possible while still being able to fully and correctly distinguish two *hypotheses*. If a function satisfies this property, it is said to be *minimal sufficient* with regards to the parameters.

One may also look at this from a different point of view, where the transformation partitions the space of all data into *equivalent sets*, meaning the conclusion drawn for the parameters will be the same across each set, independent of which individual representative is selected. When a model exhibits such a property, we say it is *invariant* against certain characteristics. This concept shall be explored in more detail in the following chapters.

Apart from the removal of irrelevant information to support the inference process, there is a second justification for this concept: Assume the model has multiple parameters and we are interested only in *one* of them, then the remainder does not carry any particular meaning for us. This is a very common phenomenon in statistical analysis which garnered a lot of attention, the reason being the *separability* between *interest* and *nuisance*. Ideally, we would like to ignore all uninformative degrees of freedom, which is possible when the probability density factorizes a part which only depends on the nuisance. In such instances, these terms can safely be discarded since they do not impair our judgment of the parameter of interest. In all other cases, however, they persist and may have a critical impact on how the parameter of interest is perceived. A suitable treatment of this issue will again be based on the principle of sufficiency, where we now distinguish between interest and nuisance. All these techniques are introduced in the following section about invariances along with a more formal background.

At its core, the thesis is centered around three topics:

- A Gaussian mixture model for distances
  (Adametz and Roth, 2011; Vogt et al., 2010)
- A Gaussian graphical model for distances
  (Adametz and Roth, 2014; Prabhakaran et al., 2013)
- A Gaussian copula model for mixed data
  (Adametz et al., 2014; Hoff, 2007; Rey and Roth, 2012)

Although all methods share the normal distribution as a common foundation, we will introduce specific invariances, hereby making the models suitable to a larger class of problems. Effectively, this can be seen as a generalization, which arises from information loss in certain domains.

The first two topics are related, since they both operate on pairwise distances as opposed to full vectorial data, yet they infer different parameters. Most importantly, the fact that inference only relies on the input of a distance matrix enables us to make use of the vast set of kernel functions for seemingly any data type or domain, be it protein sequences, semantic texts, images, chemical structures or graphs. In similar vein to the *kernel trick*, e.g. (Rasmussen and Williams, 2006, p. 12), we exploit the property that objects are not required to be vectorial as long as they permit evaluating pairwise (dis)similarities.

The third topic is concerned with estimating the correlation between rvs, but tackles a conceptually different problem: here, individual data are accessible, but obscured by discrete distribution functions, which reduces the original values to a limited set of levels. This condition effectively precludes standard approaches acting on Gaussian data, making it impossible to perform inference by conventional means. By integrating suitable invariances into the model, the correlation matrix can be estimated in spite of information loss. As a result, this enables new possibilities in the context of *biological pathways*, which are demonstrated in the later course of the thesis.

This concludes the general description of the applications and the tools we will employ for inference in Gaussian models. We begin by providing the theoretical foundation of which all subsequent analysis is based upon.

1.1 Likelihood

Consider a rv X with an unknown *probability density function* (or simply *density*) $f(x)$[1]. Then, given a set of realizations

$$\{x_1, \dots, x_n\}, \tag{1.1}$$

a statistical analysis is concerned with drawing conclusions about the underlying distribution. This knowledge will give us further insights into the process and help us understand, how the data were generated.

We may already have some intuition about the family $\mathcal{F}$ of density f, hence the space of all possibilities, or *hypotheses*, reduces to the parameters θ that specify this distribution. The dependence is made explicit by writing

$$f(x\,;\theta), \tag{1.2}$$

which interprets the density as a function of θ. Since θ is not known, however, it represents a whole class of plausible densities, also called *model*. More importantly, we implicitly assume that the true $f(x)$ is a member of the class, thus all the following will be conditioned on this very choice of a model.

From a mathematical point of view, the density is a function with a fixed θ which is evaluated at different x. For our purposes, however, the roles of inputs are reversed: the data are assumed to be fixed and the interest lies in finding a hypothesis θ which best explains the observations. After all, our

[1] For discrete rvs, $f(x)$ is the *probability mass function*.

goal is to evaluate different hypotheses on the same sample. Therefore, we define the *likelihood* as

$$L(\theta) \equiv L(\theta\,;x) \propto f(x\,;\theta), \tag{1.3}$$

which tells us how *likely* hypothesis θ is given data x. It is important to note the proportionality sign, showing that all constant factors are absorbed. Due to this, the likelihood is, technically speaking, not a statistical distribution anymore and the area under the curve does not carry a meaning. Instead, its sole purpose is to distinguish different hypotheses, which is found by the *likelihood ratio*

$$\frac{L(\theta_1\,;x)}{L(\theta_2\,;x)} = \frac{L(\theta_1)}{L(\theta_2)}. \tag{1.4}$$

If the ratio is greater than 1, then, given the observations, we shall prefer θ_1 over θ_2. In the literature, this is also referred to as the *law of likelihood* (Edwards, 1992), which is based on the premise that the likelihood contains all information that is needed to fully evaluate a hypothesis (*likelihood principle*). At this point, it is clear the likelihood can only be interpreted in a *relative* fashion rather than on an absolute scale. When depicted graphically, the convention is to fix its maximum at 1.

For mathematical convenience, the likelihood is often written in its (natural) logarithmic form

$$\ell(\theta) \equiv \log L(\theta), \tag{1.5}$$

meaning products are transformed to summations. Eq. (1.4) now becomes

$$\ell(\theta_1) - \ell(\theta_2), \tag{1.6}$$

being greater than zero if hypothesis θ_1 is better supported than θ_2.

1.1.1 A Note on the Historical Developments in Statistics

According to Young and Smith (2005, p. 2f) and Efron (1998), statistics can be classified into three schools: *Bayesian*, *Fisherian* and *frequentist*. The distinction between them is not clear and sometimes even under strong dispute, which is due to the historical developments. For our purposes, however, it suffices to highlight some properties and their implications for inference.

In the *Bayesian*[2] paradigm, parameter θ is a rv itself, hence, it requires the specification of a prior belief *before* observing any datum. Using Bayes' rule, the prior is transformed by the likelihood into the *posterior*, on which inference is based. Importantly, the Bayesian concept treats a probability as the belief in a hypothesis.

In contrast to the above, the *Fisherian*[3] school assumes that θ is unknown, but fixed, thereby avoiding any prior *distribution*. Still, for inference to be most expressive, the likelihood must be conditional on everything that is already known about θ. In the same spirit, it is desirable to remove all irrelevant information contained in the data as long as judgment about θ is not impaired. This need for efficiency is more formally expressed in the principle of (minimal) sufficient statistics. The likelihood principle—perhaps the most central aspect of the Fisherian concept—naturally lead to the *maximum likelihood estimate* as an optimization task to identify the best supported parameter given the sample. In order to highlight the overlap between paradigms, one can state that the Bayesian school also obeys the likelihood principle—meaning the likelihood contains all required information to infer θ, even though θ is treated as a rv itself.

Finally, the *frequentist* approach carries over the sufficiency principle, but interprets a probability as the number of successful trials relative to their total number. In more detail, inference is treated as a decision problem which occurs *before* seeing any datum. J. Neyman and E. Pearson are often referenced as main contributors to the frequentist theory.

The above is intended as a general, non-exhaustive overview of the developments in statistics. In the course of the next section, we will introduce techniques that mainly fall into the Bayesian and Fisherian category.

[2] The term Bayesian is in honor of Thomas Bayes (1071–1761).

[3] Named after the influential statistician and biologist Sir Ronald A. Fisher (1890–1962).

1.2 Invariances

So far, we considered inference for a single parameter θ and as we learned, the likelihood enables the evaluation of hypotheses such that they can be assessed relative to each other. Using the likelihood ratio, it can be tested which value of $\theta \in \Theta$ (out of two) is better supported given the sample. Thus, if we continue this line of thought, the best θ is found at the maximum of the likelihood (assuming, of course, that the likelihood principle applies). Inference consequently becomes an optimization problem.

Many parametric models—including the Gaussian—have multiple parameters, but often, we are interested only in some of them. To make matters clear, we write $\theta = (\psi, \lambda)$, where ψ is the *parameter of interest* and λ refers to *nuisance*. In the trivial case when the likelihood factors into two independent terms, the likelihood ratio for two hypotheses ψ_1 and ψ_2 becomes

$$\frac{L(\psi_1, \lambda)}{L(\psi_2, \lambda)} = \frac{c(\lambda) \cdot L(\psi_1)}{c(\lambda) \cdot L(\psi_2)} = \frac{L(\psi_1)}{L(\psi_2)}, \tag{1.7}$$

which does not involve λ anymore. Equivalently, one can also treat $c(\lambda)$ as an unknown, but constant factor and absorb it into the proportionality constant of the likelihood. The more common situation is, however, that there exists a functional relationship between ψ and λ.

As a consequence, we may *not* use the likelihood ratio anymore to compare two hypotheses ψ_1 and ψ_2, for then the outcome depends on the unknown true λ_0. If λ was fixed at an incorrect value, we may inadvertently favor the wrong hypothesis. The appearance of nuisance parameters constitutes a fundamental problem in statistical analysis and unfortunately there is no universal solution to it. Generally speaking, the goal is to encode *invariances* into the model, such that we can perform inference as in the single-parameter case.

The following will discuss different approaches of how to remove nuisance parameters from the likelihood, each with different requirements and implications. There might be situations in which we arrive at multiple solutions, but each is valid in its own right. In other instances, the solutions of two different methods may even coincide. Yet, what ultimately matters is the ability to make robust inference about the parameter of interest. As we will learn, invariances may come at the price of information loss and require that some aspect of the data is discarded. Therefore, it is essential to strike a good balance between the sacrifice of information and gain of statistical power. The tools at our disposal can roughly be divided into three categories which are detailed next.

1.2.1 Conditional and Marginal Likelihood

When we seek to infer a parameter θ using the likelihood, in many cases not all the properties of the sample are actually needed. Instead, a lower-dimensional function $T(X)$ may suffice to arrive at the same conclusion for θ. $T(X)$ is also called a *statistic* and can be as simple as the sum of two observations or the maximum value of the sample.

For the present setting, the only incentive behind statistics is the reduction of information, meaning we may restrict ourselves to a certain aspect of our sample that is fully *sufficient* for the task at hand. Indeed, we say that a statistic is *sufficient* for θ if there is no benefit in knowing the data (in addition to the statistic). There may exist many sufficient statistics for a parameter, therefore, a *minimal sufficient* statistic is the largest possible data reduction provided that any two hypotheses are still correctly distinguished. As mentioned earlier, this is one of the principles of the Fisherian paradigm. Mathematically, a statistic T is sufficient if and only if there are functions $g(\bullet)$ and $h(\bullet)$ such that the *Fisher-Neyman* factorization holds (Davison, 2008, p. 104):

$$f(x\,;\theta) = g(t\,;\theta)\cdot h(x) \tag{1.8}$$

In particular, we see that the statistic separates the *relevant* from the *irrelevant*, where (constant) $h(x)$ is not needed for inference about θ. By the definition of the conditional density, we have

$$f(x\,|\,t\,;\theta) = \frac{f(x,t\,;\theta)}{f(t\,;\theta)}. \tag{1.9}$$

Since T is sufficient for θ (that is, T contains all necessary information about θ), the conditional density of X given T is independent of θ. Also, as $f(x, t\,;\theta) = 0$ except for $t = t(x)$, we can state $f(x, t\,;\theta) = f(x\,;\theta)$. Thus, Eq. (1.9) becomes (Davison, 2008, p. 104)

$$f(x \,|\, t) = \frac{f(x\,;\theta)}{f(t\,;\theta)} \tag{1.10}$$

or equivalently when terms are rearranged

$$f(x\,;\theta) = \underbrace{f(t\,;\theta)}_{g(t\,;\theta)} \cdot \underbrace{f(x \,|\, t)}_{h(x)}. \tag{1.11}$$

Now, we may use the likelihood $L(\theta\,; t) \propto f(t\,;\theta)$ for inference about θ instead of $f(x\,;\theta)$. As noted above, the latter contains irrelevant information, in particular $f(x \,|\, t)$, which can safely be ignored—its only application may be for internal model checking (Reid, 1995), since it does not depend on θ.

For a better understanding of sufficiency and its implication for inference, let us give a simple example: Assume we are interested in the variance of normal rv X with zero mean. Then $T(X) = |X|$ is sufficient, because the sign does not carry any meaning for the parameter of interest. Further, the sample space is partitioned into groups that are equivalent under the statistic, e.g., $t(+2) = t(-2) = 2$. We also refer to these groups as *orbits* (Young and Smith, 2005, p. 86). Hereby, the statistic is a *surjective* transformation, because each realization x maps exactly to one orbit $t(x)$ and each orbit consists of two x. The remaining question is "What is the coarsest possible set of orbits?" which is equivalent to "What is the largest possible data reduction?".

Up until this point, we only studied the sufficiency of a statistic as a means to extract relevant aspects of the sample, but the same argument holds for nuisance parameters. Again, suppose $\theta = (\psi, \lambda)$, where ψ is the parameter of interest and λ is the nuisance. Since we need to distinguish between the two, let us partition the (minimal) sufficient statistic $T = (U, V)$ for θ in such a way that the density factors as

$$f(x\,;\psi, \lambda) \propto f(u, v\,;\psi, \lambda) = f(u \,|\, v\,;\psi) \cdot f(v\,;\psi, \lambda). \tag{1.12}$$

Here, the proportional constant absorbs the Jacobian determinant due to the transformation from x to (u, v). For inference about ψ, we can now resort to

$$L(\psi\,; u \,|\, v) \propto f(u \,|\, v\,; \psi), \tag{1.13}$$

which is called *conditional likelihood* (see Davison (2008, p. 645ff), Boos and Stefanski (2003, p. 57), Severini (2001, p. 278ff), Young and Smith (2005, p. 146)). Note that the second term in Eq. (1.12) is intentionally discarded, even though it contains *some* information about ψ. The reason is, it may be too complicated to obtain, thereby outweighing the benefits (Davison, 2008, p. 656), or the loss is small (Garthwaite et al., 2002, p. 56). Clearly, this is a potential drawback of Eq. (1.12), but one could additionally require that the density of V does not depend on ψ (Reid, 1995), that is,

$$f(u, v\,; \psi, \lambda) = f(u \,|\, v\,; \psi) \cdot f(v\,; \lambda). \tag{1.14}$$

In this special case, V is called an *ancillary* for ψ in the sense that it does not contain any information about the parameter of interest (Garthwaite et al., 2002, p. 57). Hence, Eq. (1.14) solves the problem of information loss, but it may not be possible to find a suitable U and V after all.

As an alternative, we may reverse the conditioning of U and V to receive

$$f(u, v\,; \psi, \lambda) = f(u\,; \psi) \cdot f(v \,|\, u\,; \psi, \lambda), \tag{1.15}$$

for then, the first factor can be used as *marginal likelihood*, see (Severini, 2001, p. 278ff) and (Davison, 2008, p. 645ff),

$$L(\psi\,; u) \propto f(u\,; \psi), \tag{1.16}$$

thereby again ignoring the information loss due to the discarded term. In certain instances, it may be possible to arrive at a similar form as Eq. (1.14):

$$f(u, v\,; \psi, \lambda) = f(u\,; \psi) \cdot f(v \,|\, u\,; \lambda). \tag{1.17}$$

This, however, requires further assumptions about U and V as shown above. Note that the knowledge of U suffices to perform marginal inference, whereas V is often not explicitly specified unless we investigate the potential information loss.

The main idea of the above is to isolate the parameter of interest by conditioning, such that we can base the corresponding likelihood on a single factor of the density. This is a powerful approach to incorporate invariances into the model, but it may fall short due to information loss or when we simply cannot find suitable statistics. Therefore, the following explores two complementary methods for the treatment of nuisance parameters.

1.2.2 Profile Likelihood

The *profile likelihood* (Severini (2001, p. 126ff),Young and Smith (2005, p. 135ff)) is a more recent development and it borrows from the idea of maximum likelihood estimation. Essentially, it aims to replace the unknown nuisance parameter λ by a point estimate, that is best supported under the likelihood. For this, all remaining parameters are assumed to be fixed, including the parameter of interest ψ. In mathematical terms, we solve

$$\widehat{\lambda}_\psi = \underset{\lambda}{\operatorname{argmax}}\, L(\psi, \lambda), \tag{1.18}$$

where subscript $_\psi$ denotes that ψ was fixed. Inserting the estimate back into the likelihood, we arrive at

$$L_P(\psi) \equiv L(\psi, \widehat{\lambda}_\psi), \tag{1.19}$$

which is a function of ψ only. In the general likelihood $L(\psi, \lambda)$ both parameters are allowed to vary freely in the space $\Psi \times \Lambda$. By replacing the nuisance parameter with its maximum likelihood estimate, we reduce this space to $\Psi \times \{\widehat{\lambda}_\psi\}$, which can be thought of as cutting out a *profile*, hence the name.

Intuitively, the approach seems very reasonable, since it relies on the value that is best supported by the likelihood. However, if $\widehat{\lambda}_\psi$ differs from the true λ_0, our judgment of ψ can be seriously biased. Therefore, the profile likelihood works best for a large sample size, such that $\widehat{\lambda}_\psi \approx \lambda_0$. The same holds true if there is more than one nuisance parameter, but then the sample size must grow *in relation* to their number, informally speaking.

To gain further insights into the maximum likelihood estimate, let us recall that the data are an incomplete set of observations $\{x_1, \dots, x_n\}$ for rv X. Consequently, also the likelihood is random to some degree, since it is a function of the sample. The same applies to $\widehat{\theta} = \operatorname{argmax}_\theta L(\theta\,;x)$, which may change its value once more observations are available. Thus, how can we be sure that $\widehat{\theta}$ is close to the true value θ_0?

One solution is to investigate the likelihood in the parameter space surrounding $\widehat{\theta}$. This is the general idea behind the *Fisher information* (Young and Smith (2005, p. 123), Cox (2006, p. 97)), being defined as

$$i(\theta) = -\mathrm{E}\left[\frac{\partial^2 \ell(\theta\,;x)}{\partial\theta^2}\right]. \tag{1.20}$$

Here, the expectation is over the Hessian matrix of the log-likelihood, that is, the second order partial derivatives. In other words, this measure tells us about sensitivity or curvature of the log-likelihood (Cox, 2006, p. 97). If the curvature is sharp at $\widehat{\theta}$, i.e., when the likelihood strongly peaks at this value, we can be fairly certain to be close to θ_0. Correspondingly, if the log-likelihood is rather flat at $\widehat{\theta}$, it conveys only little information to discriminate different θ. For the partition $\theta = (\psi, \lambda)$, the Fisher information becomes a matrix (Young and Smith, 2005, p. 135f)

$$i(\theta) = \begin{bmatrix} i_{\psi\psi}(\psi, \lambda) & i_{\psi\lambda}(\psi, \lambda) \\ i_{\lambda\psi}(\psi, \lambda) & i_{\lambda\lambda}(\psi, \lambda) \end{bmatrix}, \tag{1.21}$$

where the off-diagonals allude to an interesting special case. In more detail, if $i_{\psi\lambda} = i_{\lambda\psi} = 0$ for some/every ψ and λ, then the parameters are said to be locally/globally *orthogonal*, see (Young and Smith, 2005, p. 143) and (Cox and Reid, 1987). This implies $\widehat{\psi}$ and $\widehat{\lambda}$ are asymptotically independent (Young and Smith (2005, p. 145), Cox (2006, p. 112)), such that $\widehat{\lambda}_\psi$ varies little in the neighborhood of $\widehat{\psi} = \operatorname{argmax}_\psi L_P(\psi)$. In practice, this only gives an approximative indication if the profile likelihood approach enables inference, yet we will encounter both working and failing examples.

On a technical note, the profile likelihood is not a *genuine* likelihood in a strict sense, because it is not based on the density of a rv (Severini, 2001, p. 323). Still, for certain models, it may coincide with the marginal (or conditional) likelihood, which *are* considered genuine.

1.2.3 Integrated Likelihood

The third approach is generally applicable, although it may not be possible to calculate the result analytically. If there is no knowledge available about λ, we can always remove its dependence by integration over its support Λ:

$$L(\psi) \propto \int_{\Lambda} L(\psi, \lambda) \, \mathrm{d}\lambda \tag{1.22}$$

The resulting likelihood is sometimes also called *(Bayesian) marginal likelihood*, or, for a better differentiation from the previous case: *integrated likelihood* (Severini, 2001, p. 306ff). In fact, Eq. (1.22) strongly resembles a Bayesian scheme with an uninformative *prior*, where the complete lack of knowledge is expressed by a uniform distribution. Following this idea, we achieve a more general formulation via

$$L(\psi \mid \beta) \propto \int_{\Lambda} L(\psi, \lambda) \cdot f(\lambda \mid \beta) \, \mathrm{d}\lambda, \tag{1.23}$$

where λ is distributed according to prior density $f(\lambda \mid \beta)$ with *hyperparameter* β. The Bayesian regime treats the likelihood as a function which transforms the prior belief into the posterior. Hence, due to integration over all parameter values, we transform the full prior in its entirety instead of only a single point.

When seen from a different perspective, Eq. (1.23) computes a weighted sum of the likelihood, where the importance of each value of the nuisance parameter is specified by the prior. This means, if the prior is highly concentrated with a peak, the integrated likelihood (the posterior) will be conditional on this choice. At the same time, a flat prior that assigns a non-zero weight to all nuisance parameters results in a very balanced, but also vague likelihood with regards to λ—the extreme is found with a uniform prior. In any case, the choice of the prior affects the maximum of the posterior, as seen by hyperparameter β in Eq. (1.23), which carries over to the integrated likelihood.

For specific combinations of likelihood and prior, the integral can be solved analytically. Further, if prior and posterior are members of the same distribution family, the prior is said to be *conjugate* to the likelihood. Similar to before, the constant normalizing factor can safely be discarded.

In review of the integrated likelihood, we can highlight its general applicability for all nuisance parameters, provided that a suitable and meaningful prior is defined. Also, under certain conditions, the solution to the integral is found analytically. On the negative side, followers of the Fisherian paradigm criticize the burden to specify a prior, meaning sometimes the expressiveness of the posterior is sacrificed in favor of computability, thereby forcing a certain interpretation into the likelihood. It goes to show that some questions do not allow a universally accepted answer and heavily depend on the point of view. To this end, we shall judge all the above approaches without bias and evaluate their performance for the situation at hand.

1.3 Gaussian Models

As suggested by the title, the thesis is centered around the normal distribution[4], which constitutes an important distribution in statistics, if not *the* most important one. The reason is that many phenomena in nature appear to be governed by normally distributed rvs. Areas of application are widespread and include for example physics, astronomy, but also psychology, social sciences and many others. While it is usually not possible to assert exact normality of an observed rv, one often resorts to the expression *approximate normal*. This is a common assumption when nothing is known about a rv, for example regarding noise terms.

A mathematically more satisfying justification for normal assumptions can be found by the *central limit theorem* (see van der Vaart (2000, p. 6f), Severini (2001, p. 28f)). Given rv X and a sequence of *independent and identically distributed* (*i.i.d.*) observations $\{x_1, \dots, x_n\}$, the following holds

$$\frac{1}{\sqrt{n}}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) \overset{\mathcal{D}}{\to} \mathcal{N}(\mu, \sigma^2) \qquad \text{as } n \to \infty, \tag{1.24}$$

[4] The normal distribution is frequently called *Gaussian* distribution due to the contributions of Carl Friedrich Gauss (1777–1855).

where μ and σ^2 are mean and variance, respectively. The theorem states that under mild conditions[5], the sum of an i.i.d. sample follows a normal distribution in the limit of infinite observations. Thus, the larger the sample size, the better the approximation becomes. Perhaps surprisingly, the result is true *regardless* of the underlying distribution of X.

The above statement partly explains the ubiquity of the normal distribution and its practical appeal, but it also alludes to a much broader role: many statistical formulae involve the sum of rvs and are therefore closely related to the concept of normality. In fact, the normal distribution is the limit of other distributions, for example the chi-squared distribution with a large degree of freedom. From a technical perspective, it also has a number of convenient properties: It is fully defined by mean and variance (with all higher moments being zero), it is symmetric, has infinite support and is infinitely differentiable. The sum or the difference of two normally distributed rvs is, again, normal and this property also applies to linear combinations of normals.

In the course of the thesis, we will consider different forms of the normal distribution, namely the univariate, multivariate and the matrix-valued case. The following introduces each individually, starting with univariate rv

$$X \sim \mathcal{N}(\mu, \sigma^2). \tag{1.25}$$

Here, μ and σ^2 are both scalar and refer to the mean and variance. The *density* of the univariate normal distribution is written as

$$f(x\,; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{1.26}$$

Due to its importance in the literature, the symbol $\phi(x) \equiv f(x\,; 0, 1)$ is often reserved for the *standard normal density* with $\mu = 0$ and $\sigma^2 = 1$.

In the multivariate case, we assume a *vector* of p rvs

$$\boldsymbol{X} \equiv [X_1, \ldots, X_p]^\top \sim \mathcal{N}_p(\boldsymbol{\mu}, \Sigma) \tag{1.27}$$

[5] Mean and variance must be finite.

having a vectorial instance $\boldsymbol{x} \in \mathbb{R}^p$. Here, the density reads

$$f(\boldsymbol{x}\,; \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} \exp\Big(-\tfrac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\Big), \tag{1.28}$$

which uses *mean vector* $\boldsymbol{\mu} \in \mathbb{R}^p$ and $p \times p$ positive-definite *covariance matrix* Σ. Finally, the most general form is achieved by a *random matrix*

$$X \sim \mathcal{N}_{p,n}(M, \Sigma \otimes \Psi), \tag{1.29}$$

where all parameters are matrices, including *mean matrix* M of size $n \times p$, *row covariance matrix* Σ of size $p \times p$ and *column covariance matrix* Ψ of size $n \times n$. In this configuration, an instance is denoted by $X \in \mathbb{R}^{p \times n}$ and the density becomes (Gupta and Nagar, 1999)

$$f(X\,; M, \Sigma, \Psi) = \\ \frac{1}{(2\pi)^{\frac{np}{2}} |\Sigma|^{\frac{n}{2}} |\Psi|^{\frac{p}{2}}} \exp\Big(-\tfrac{1}{2}\operatorname{tr}\{\Psi^{-1}(X - M)^\top \Sigma^{-1} (X - M)\}\Big). \tag{1.30}$$

We can think of matrix X as being composed column-wise of p-variate realizations, where additionally the n realizations are governed by covariance matrix Ψ. To better distinguish Σ and Ψ, we refer to them as *row* and *column* covariance matrix, respectively. It is easy to see that Eq. (1.30) coincides with Eq. (1.28) when $n = 1$ and $\Psi = 1$, which further reduces to Eq. (1.26) for $p = 1$ and $\Sigma = \sigma^2$. To this end, Table 1.1 graphically compares all the three variants of the normal distribution.

When speaking about *Gaussian models*, we assume the data originate from a source that follows one of the above classes, however, its parameters are typically unknown. The process of *inference* is then concerned with estimating these parameters from a set of observations. Having access to them enables us to better understand the generating process.

In the following chapters, we will look into the applications of clustering, graphical models and information-theoretic compression—all based on the Gaussian foundation, but under the constraints of information loss. This also involves a tailored treatment of nuisance parameters, such that the final model adheres to all required invariances.

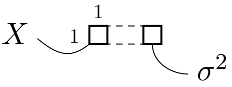
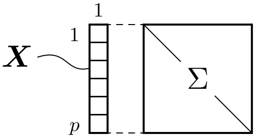
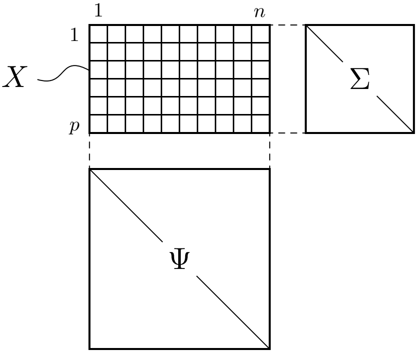
Type	**Graphical Interpretation**
univariate $X \sim \mathcal{N}(\mu, \sigma^2)$	
multivariate $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \Sigma)$	
matrix-variate $X \sim \mathcal{N}_{p,n}(M, \Sigma \otimes \Psi)$	

Table 1.1: Graphical comparison of the univariate, multivariate and matrix-variate normal distribution. A diagonal line represents symmetry. Since the mean always has the same format as its corresponding rv(s), it is omitted on the right-hand side.

1.4 Distances

The current section introduces an important topic that is used frequently throughout the thesis: squared Euclidean distances from Gaussian data. As we will learn in the following, distances naturally arise in many applications where the individual objects do not live in a Euclidean space. Take, for example, *phylogenetics* as a field in evolutionary biology, which studies the relationships of species by analyzing their genetic information. It may be computationally and algorithmically difficult to work with abstract objects like DNA sequences, but it is straight forward to compute their pairwise distances. Due to this property, distance matrices are one of the main forms of representation for constructing phylogenetic trees, reminiscent of Darwin's *tree of life* (Darwin, 1859, p. 108f).

The idea of working with abstract objects via pairwise comparison is also at the heart of kernel theory in machine learning. As of today, a myriad of kernels have been developed for virtually any domain, including graphs, strings, semantic texts, probability distributions, images and many more. When such a kernel function is evaluated for all pairwise combinations of objects, the outcome is stored in a so-called *kernel matrix*, which is symmetric and positive (semi-)definite.

There is, however, a conceptual difference between a kernel matrix and the above-mentioned distance matrix, in that kernels measure *similarity* and distances express *dissimilarity*. This may appear superficial at first, but there is a deeper connection between the two: Every kernel or similarity matrix corresponds to *exactly* one squared Euclidean distance matrix, but additionally it carries information about the point of origin of the underlying feature space. As a consequence, there can be two unique kernel matrices which map to the same distance matrix.

At this point, we make an important design decision: we henceforth assume the point of origin is *irrelevant* and argue that this is in fact a reasonable choice for many applications. As an example, take a graph kernel to compute the similarity between two chemical compounds; since the kernel implicitly operates in a potentially infinite-dimensional feature space, we lack the evidence to decide if the origin is informative or fixed arbitrarily. Due to this reason, our interest is *exclusively* confined to the part defining the distances.

Clearly, one can construct a kernel function where the point of origin is indeed meaningful, but in the context of our applications, we *intentionally* discard this information. Speaking in terms of the Fisherian concept of data reduction, the distance represents the parameter of interest and the point of origin is treated as a nuisance.

The next section will lay the technical foundation for distance-based inference and it starts by developing a geometric interpretation.

1.4.1 A Geometric Interpretation of Distances

For a better understanding of distances in the framework of kernel matrices, let us assume that all objects and their individual feature values are known. Further, suppose X is the $p \times n$ matrix containing full information about p objects living in an n-dimensional Euclidean space. As a measure of similarity, we use the $p \times p$ symmetric inner product matrix

$$S \equiv XX^\top, \tag{1.31}$$

which is positive definite if X has $n \geq p$ linearly independent columns, else it is only positive semi-definite. S inherently depends on the coordinate system of X, because for a single pair of objects $X_{i\bullet}$ and $X_{j\bullet}$, both being row vectors in $\mathbb{R}^n$, the inner product (or scalar product) corresponds to

$$S_{ij} = X_{i\bullet} X_{j\bullet}^\top = \|X_{i\bullet}\|_2 \, \|X_{j\bullet}\|_2 \cos(\alpha). \tag{1.32}$$

From this definition, it follows that the measure involves the length of both vectors as well as their angle α. As a result, the similarity of two objects depends on their position relative to the point of origin and, consequently, S changes when objects are jointly shifted in space. S is, however, invariant against rotation or reflection, which can be seen from

$$XO(XO)^\top = XOO^\top X^\top = XX^\top, \tag{1.33}$$

where O is an arbitrary $n \times n$ orthogonal transformation matrix.

When we speak of the term *distance* in the context of this work, we should correctly refer to it as *squared Euclidean distance*, meaning the underlying data are always assumed to be vectorial, even if they live in an unknown, possibly infinite-dimensional feature space (in accordance with the kernel trick). The reason for using the *squared* distance is due to its connection to the inner product, which we will see shortly, but first, the formal definition is

$$D_{ij} \equiv \|X_{i\bullet} - X_{j\bullet}\|_2^2 = (X_{i\bullet} - X_{j\bullet})(X_{i\bullet} - X_{j\bullet})^\top. \tag{1.34}$$

This leads to a $p \times p$ symmetric matrix D which has exactly one positive eigenvalue and is negative semi-definite on a $(p-1)$-dimensional subspace (Schoenberg, 1937; Gower, 1985). Fig. 1.1 depicts X, S and D.

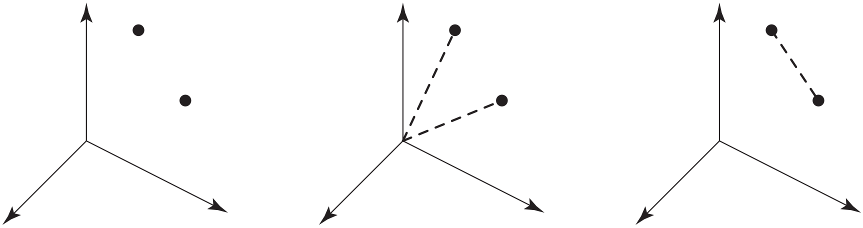


Figure 1.1: The differences between X, S and D. Left: $p = 2$ objects i and j live in a $n = 3$ dimensional space, which is the full information captured by X. Center: The scalar product S_{ij} measures the similarity of objects, which is relative to the point of origin. Right: The pairwise distance D_{ij} is independent of the point of origin (the plot shows $\sqrt{D_{ij}}$).

Finally, distance and inner product are related via

$$\begin{aligned} D_{ij} &= X_{i\bullet}X_{i\bullet}^\top + X_{j\bullet}X_{j\bullet}^\top - 2X_{i\bullet}X_{j\bullet}^\top && (1.35)\\ &= S_{ii} + S_{jj} - 2S_{ij} && (1.36)\end{aligned}$$

or written in matrix notation:

$$D = \operatorname{diag}(S)\mathbf{1}_p^\top + \mathbf{1}_p \operatorname{diag}(S)^\top - 2S. \tag{1.37}$$

The fact that D does not depend on the translation of X can readily be seen from Eq. (1.34), but what is the impact on S? Let all objects in X be jointly shifted in space, such that

$$\widetilde{X} = X + \mathbf{1}_p \boldsymbol{w}^\top \tag{1.38}$$

with $\boldsymbol{w} \in \mathbb{R}^n$. Then, for the corresponding inner product matrix, we have

$$\begin{aligned} \widetilde{S} &= \widetilde{X}\widetilde{X}^\top & (1.39)\\ &= (X + \mathbf{1}_p \boldsymbol{w}^\top)(X + \mathbf{1}_p \boldsymbol{w}^\top)^\top & (1.40)\\ &= \underbrace{XX^\top}_{S} + (X\boldsymbol{w})\mathbf{1}_p^\top + \mathbf{1}_p(X\boldsymbol{w})^\top + \boldsymbol{w}^\top\boldsymbol{w}\mathbf{1}_p\mathbf{1}_p^\top . & (1.41) \end{aligned}$$

Here, the last three terms only occur due to translation. Therefore, by varying vector $\boldsymbol{w}$, we can construct a whole set of matrices $\widetilde{S}$, which all map to the same D. For a more compact representation, notice that Eq. (1.41) can be rearranged as

$$\widetilde{S} = S + \underbrace{(X\boldsymbol{w} + \tfrac{1}{2}\boldsymbol{w}^\top\boldsymbol{w}\mathbf{1}_p)}_{\boldsymbol{u}}\mathbf{1}_p^\top + \mathbf{1}_p\underbrace{(X\boldsymbol{w} + \tfrac{1}{2}\boldsymbol{w}^\top\boldsymbol{w}\mathbf{1}_p)^\top}_{\boldsymbol{u}^\top}, \tag{1.42}$$

such that for any $\boldsymbol{w} \in \mathbb{R}^n$ there is a corresponding $\boldsymbol{u} \in \mathbb{R}^p$ without loss of generality. Hence, we can now formally define the set as

$$\mathcal{S}(D) = \left\{ \widetilde{S} \;\middle|\; \widetilde{S} = S + \mathbf{1}_p\boldsymbol{u}^\top + \boldsymbol{u}\mathbf{1}_p^\top,\ \widetilde{S} \succeq 0,\ \boldsymbol{u} \in \mathbb{R}^p \right\}. \tag{1.43}$$

In this definition, S does not have any particular meaning other than to serve as a member from which the set is spanned. Fig. 1.2 depicts an example, where two matrices X and $\widetilde{X}$ lead to the same distance matrix D.

1.4.2 Further Operations on Distances

So far, we focused on a specific type of transformation $X + \mathbf{1}_p\boldsymbol{w}^\top$ which does not enter the distances. There are, however, many other operations that *do* have an impact. As an example, assume $X + \boldsymbol{v}\mathbf{1}_n^\top$ with $\boldsymbol{v} \in \mathbb{R}^p$: Hereby, it is possible to move single objects in feature space such that their pairwise

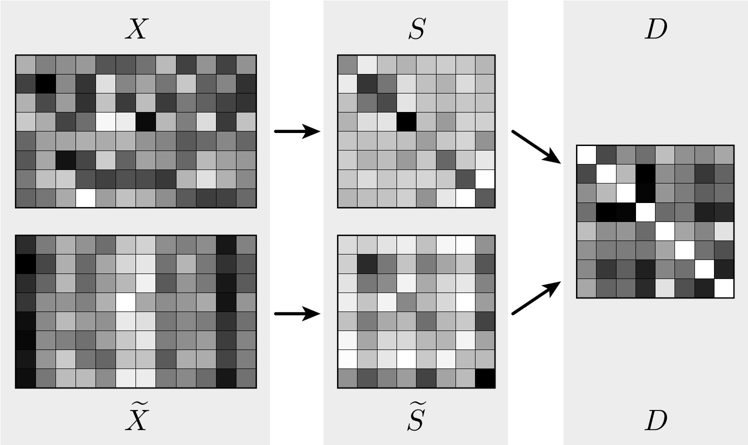


Figure 1.2: The mapping from X to S to D is surjective and involves a loss of information.

distances are altered completely. Obviously, we consider this information a vital part of the structure of a distance matrix and therefore do *not* allow modifications of this kind. Fig. 1.3 depicts both variants of the mean.

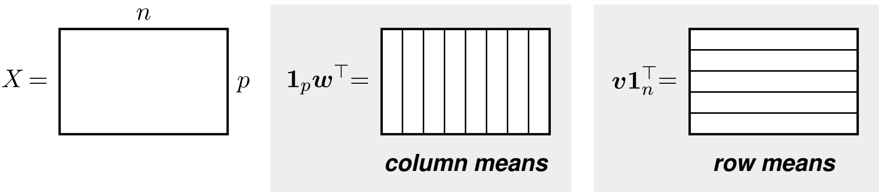

n
$X =$
p
$\mathbf{1}_p \boldsymbol{w}^\top =$
column means
$\boldsymbol{v}\mathbf{1}_n^\top =$
row means

Figure 1.3: Matrix X and two complementary mean models. Only column means are canceled in distance matrix D; row means persist.

A second transformation concerns the scaling c, which appears as cX and changes the distance matrix as c^2D. For the models being developed in the following, we decide that this parameter is uninformative, analog to our assumption concerning the point of origin. Therefore, scaling c will be treated as a nuisance.

In conclusion, all thoughts and considerations about distance matrices give a first impression as to what is required by a statistical model. To that extent, the next two chapters formulate these ideas in a more concise manner and develop the necessary modifications for Gaussian models. Also, we shall discuss their implications with regards to inference.

Chapter 2

A Gaussian Mixture Model for Distances

2.1 Introduction

Cluster analysis can best be described as finding unique and distinct groups[1] within a population, such that their resulting composition is homogeneous. In our case, we explain the data by a mixture of k normal distributions (McCullagh and Yang, 2008), where the density of a single object $\boldsymbol{x} \in \mathbb{R}^n$ is

$$f(\boldsymbol{x}) \propto \sum_{j=1}^{k} \pi_j \cdot f(\boldsymbol{x}\,; \boldsymbol{m}_j, \Psi). \tag{2.1}$$

Hereby, component j is parametrized by mixture weight π_j, mean vector $\boldsymbol{m}_j \in \mathbb{R}^n$ and covariance matrix $\Psi \in \mathbb{R}^{n \times n}$. Fig. 2.1 demonstrates an example of a Gaussian mixture in $n = 1$ dimension, where the solid line depicts the density. For a fixed variance, the contribution of each component (dashed line) is fully defined by mean m_j and mixture weight π_j. Finally, the black dots on the x axis represent $p = 50$ objects drawn from the mixture distribution. Inference reverses this generative process and aims to identify the components from the sample. This means, if a set of objects is well explained by one component, they form a cluster and will consequently be assigned the same label.

For reasons of simplicity, all clusters are assumed to have the same spherical shape, which implies $\Psi = I_n$. Since the normal density has infinite sup-

[1] The terms *group*, *cluster* and *class* are used interchangeably.

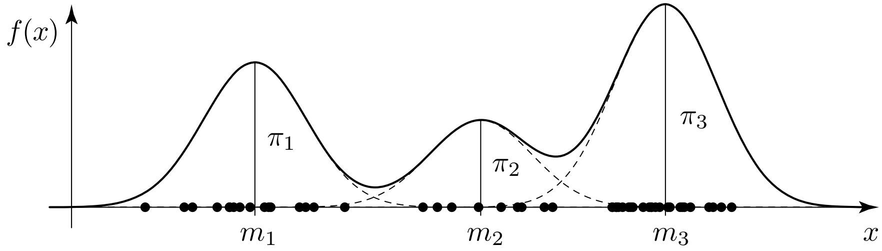


Figure 2.1: Density for a mixture of $k = 3$ Gaussians in $n = 1$ dimension.

port, each object is jointly explained by *all* k components (hence, mixture), however, the contribution of a component quickly declines with increasing distance from its mean in accordance with its bell shape.

The challenge in our setting is to harmonize the Gaussian mixture model with a distance matrix, which prevents us from directly identifying the component means, since the n-dimensional feature space is latent. Thus, we have to find an alternative way to express the clusters and their geometric configuration.

2.2 Related Work

Clustering has historically attracted a lot of attention and today constitutes a large field in machine learning. A popular representative is the *k-means* algorithm (Steinhaus, 1956; MacQueen, 1967; Jain, 2010), which minimizes a functional

$$J \equiv \sum_{j=1}^{k} \sum_{i=1}^{p} \|\boldsymbol{x}_i - \boldsymbol{m}_j\|^2, \tag{2.2}$$

that is, the sum of squared errors for p objects over all k clusters.

The method works in the following way: Randomly initialize the cluster centers $\boldsymbol{m}_\bullet$ of a fixed and predefined number k, then assign all objects to the closest cluster. Next, relocate the centers to the current mean of each cluster. These two steps are alternated until the assignments finally converge to a local optimum. Although the shortcomings of this approach are apparent, namely the dependence on the initialization and the need to fix the number of clusters beforehand, k-means can be formulated such that it only depends on the inner product of objects (*kernel k-means*, see (Schölkopf et al., 1998)) or their pairwise distances (Roth et al., 2003). Hereby, it is not required to compute the mean of a cluster explicitly.

In contrast to the above, our approach will employ a Dirichlet process mixture model, which relieves us from specifying the number of clusters beforehand. Also, we will formulate the problem in a way that it only depends on distance information.

2.3 Model

As a starting point, let us assume the simplest possible case, where the observations are generated from the matrix-variate normal distribution,

$$X \sim \mathcal{N}_{p,n}(M, I_p \otimes I_n). \tag{2.3}$$

We interpret X as a collection of p objects, each being a row vector $X_{i\bullet} \in \mathbb{R}^n$. An important observation is that both rows (= objects) and columns (= features) are independent; the clusters are solely defined by mean matrix M, which is composed of k distinct rows, see Fig. 2.2. Here, k refers to the number of clusters and $\boldsymbol{m}_j \in \mathbb{R}^n$, $j \in \{1, \dots, k\}$, are the distinct component means. Note that matrix M groups objects by cluster, but this is not a requirement. Also, the numbering of the clusters, i.e., their individual label, is arbitrary and carries no information.

When data are generated from Eq. (2.3) with M as defined in Fig. 2.2, there will be k spherical clusters, each with the same diameter. In particular, the spherical shape comes from $\Psi = I_n$ and the identical diameters are on account of $\Sigma = I_p$. It is well-known that when X is distributed as Eq. (2.3), its inner product follows a *non-central Wishart distribution*, that is,

$$S \equiv XX^\top \sim \mathcal{W}_p(n, I_p, \Theta) \tag{2.4}$$

with n degrees of freedom, covariance matrix I_p and $p \times p$ non-centrality

$$M = \begin{bmatrix} \boldsymbol{m}_1^\top \\ \vdots \\ \boldsymbol{m}_1^\top \\ \boldsymbol{m}_2^\top \\ \vdots \\ \boldsymbol{m}_{k-1}^\top \\ \boldsymbol{m}_k^\top \\ \vdots \\ \boldsymbol{m}_k^\top \end{bmatrix}$$

Figure 2.2: Objects (= rows) associated with the same cluster j are assumed to have a common mean $\boldsymbol{m}_j$. Hereby, mean matrix M is composed of individual mean vectors as shown above.

matrix $\Theta \equiv MM^\top$. The latter gives rise to a hypergeometric function in the density of S, see (Diaz-Garcia et al., 1997) and (Gupta and Nagar, 1999, p. 114), which vanishes for $M = 0_{p \times n}$, thereby leading to the simpler *central Wishart distribution*.

Unfortunately, the practical use of the non-central Wishart is severely hampered by its complicated form, and even more so, estimating the unknown Θ based on a single realization S is impossible. For this reason, the following is of particular interest: It is possible to define a central-Wishart distribution which approximates a non-central Wishart; their first moments are identical and the second moments differ by order $\mathcal{O}(n^{-1})$ (Tan and Gupta, 1982; Kollo and von Rosen, 1995). This yields

$$\mathcal{W}_p(n, I_p, \Theta) \approx \mathcal{W}_p(n, \tfrac{1}{n}MM^\top + I_p), \qquad (2.5)$$

which is a remarkable connection, because it implies that matrix S could have either originated from Eq. (2.3),

$$X \sim \mathcal{N}_{p,n}(M, I_p \otimes I_n),$$

or from the zero-mean

$$X \sim \mathcal{N}_{p,n}\Big(0_{p\times n}, (\tfrac{1}{n}MM^\top + I_p) \otimes I_n\Big). \tag{2.6}$$

Thus, in summary, the cluster-defining means were transformed into the covariance matrix of an equivalent distribution. On closer inspection of $\frac{1}{n}MM^\top + I_p$, we see that it is a $p \times p$ symmetric, block diagonal matrix with full rank p. Therefore, using a different parametrization, it can be written as

$$ZAZ^\top + I_p, \tag{2.7}$$

where $A \in \mathbb{R}^{k\times k}$ corresponds to the inner product of the k distinct mean vectors $\boldsymbol{m}_\bullet$ and $Z \in \{0,1\}^{p\times k}$ is an *indicator matrix*, such that $Z_{ij} = 1$ represents object i being a member of cluster j. Since every object can only be assigned to one cluster at a time, matrix Z has one 1 per row, leading to a total of p non-zero elements.

At this point, a legitimate question is: What is the benefit of this parametrization compared to the one using M? Recall that the observations are received in the form of a distance matrix, which permits neither explicit statements about the feature space nor the number of features n. Although we can find one possible Euclidean embedding X which corresponds to a given distance matrix, it makes a choice concerning the latent feature space and therefore potentially introduces a bias. Our decision is to avoid any reconstruction altogether; the parametrization in (Z, A) relieves us from explicitly specifying the means in terms of the underlying feature space.

Combining all statements and properties we derived so far, recall that our model assumption was to regard scaling c as uninformative. Further, the fact that we observe a distance matrix implies loss of knowledge about potential column shifts. Therefore, we arrive at the distribution

$$X \sim \mathcal{N}_{p,n}(\mathbf{1}_p\boldsymbol{w}^\top, c^2\Sigma \otimes I_n) \tag{2.8}$$

with corresponding log-likelihood

$$\begin{aligned} \ell(\Sigma, c, \boldsymbol{w}) = & -\tfrac{n}{2}\log|c^2\Sigma| \\ & -\tfrac{1}{2}\operatorname{tr}\{c^{-2}\Sigma^{-1}(X - \mathbf{1}_p\boldsymbol{w}^\top)(X - \mathbf{1}_p\boldsymbol{w}^\top)^\top\}. \end{aligned} \tag{2.9}$$

Here, $\Sigma \equiv ZAZ^\top + I_p$ are the parameters of interest and $(c, \boldsymbol{w})$ corresponds to nuisance. The following will remove the dependence on the nuisance terms in two separate steps, and we begin with translation vector $\boldsymbol{w}$. For a compact notation during the transformations of the likelihood, we will temporarily retain the symbol Σ.

2.4 Invariance against Translation

2.4.1 Marginal Likelihood Approach

For the removal of nuisance parameter $\boldsymbol{w}$, assume the statistic

$$u(X) = LX, \tag{2.10}$$

where L can be any projection matrix of size $(p-1) \times p$ that satisfies

$$L\mathbf{1}_p = \mathbf{0}_{(p-1)}. \tag{2.11}$$

Fig. 2.3 explains this mapping graphically for $p = 2$, which is loosely based on (Lay, 2011, p. 204). Notice how multiples of $\mathbf{1}_2$ (that is, all points on the dashed line) are mapped to 0. The plot used $L = [-1 \quad 1]$, however, any mapping with kernel $\mathbf{1}_2$ suffices for the purpose of removing nuisance $\boldsymbol{w}$. Other valid examples are $L = [1 \quad -1]$ and $L = [2 \quad -2]$.

The idea behind projection L is

$$LX = L(c\widetilde{X} + \mathbf{1}_p\boldsymbol{w}^\top) = L(c\widetilde{X}), \tag{2.12}$$

such that X and $c\widetilde{X}$ become equivalent, i.e., they are assigned to the same group orbit. Since L removes information from X, we could theoretically

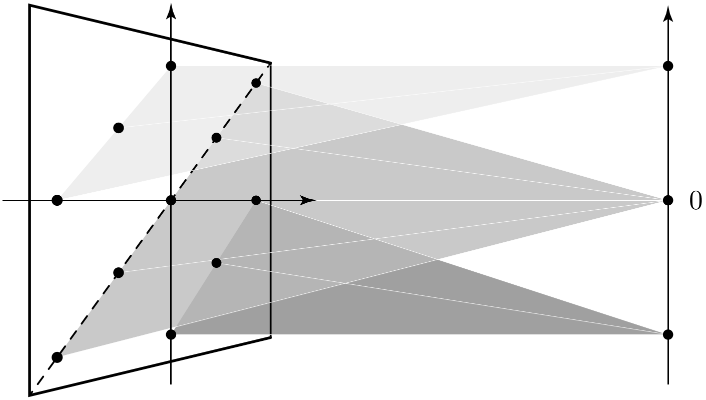


Figure 2.3: Projection L for $p = 2$ is a function $\mathbb{R}^2 \mapsto \mathbb{R}^1$.

define a statistic V to capture this very loss, say

$$v(X) = \tfrac{1}{p} X^\top \mathbf{1}_p, \tag{2.13}$$

which is in fact an estimator for column means $\boldsymbol{w}$:

$$\begin{aligned} v(X) &= \tfrac{1}{p}(c\widetilde{X} + \mathbf{1}_p \boldsymbol{w}^\top)^\top \mathbf{1}_p & (2.14) \\ &= \tfrac{1}{p} c\widetilde{X}^\top \mathbf{1}_p + \boldsymbol{w} & (2.15) \\ &= \widehat{\boldsymbol{w}}. & (2.16) \end{aligned}$$

Both statistics U and V are linear transformations of the matrix-variate normal distribution, hence they are distributed as

$$LX \sim \mathcal{N}_{(p-1),n}\Big(0_{(p-1)\times n}, (c^2 L \Sigma L^\top) \otimes I_n\Big) \tag{2.17}$$

and

$$\tfrac{1}{p} X^\top \mathbf{1}_p \sim \mathcal{N}_n\Big(\boldsymbol{w}, (\tfrac{c^2}{p} \mathbf{1}_p^\top \Sigma \mathbf{1}_p) \cdot I_n\Big), \tag{2.18}$$

respectively. Notice how the distribution of LX does not depend $\boldsymbol{w}$ anymore, which was the purpose of the transformation. When analyzed jointly, (U, V) is sufficient for Σ and c, because it captures all information about X.

Restating the definition of the marginal likelihood, Eq. (1.15), we have

$$f(u, v\,; \psi, \lambda) = \underbrace{f(u\,; \psi)}_{L(\psi\,; u)} \cdot f(v \,|\, u\,; \psi, \lambda)$$

with parameter of interest $\psi \equiv (\Sigma, c)^2$ and nuisance term $\lambda \equiv \boldsymbol{w}$. The fact that the dependence on $\boldsymbol{w}$ is removed can be seen from the reduced dimensionality of $LX \in \mathbb{R}^{(p-1)\times n}$ compared to $X \in \mathbb{R}^{p\times n}$. Due to the above factorization, the marginal log-likelihood based on Eq. (2.9) becomes

$$\begin{aligned}\ell(\Sigma, c\,; LX) = &- \tfrac{n}{2}\log|c^2 L\Sigma L^\top| \\ &- \tfrac{1}{2}\operatorname{tr}\{c^{-2}L^\top(L\Sigma L^\top)^{-1}LXX^\top\}.\end{aligned} \tag{2.19}$$

Next, we redefine a part of the trace as

$$WQ \equiv L^\top(L\Sigma L^\top)^{-1}L, \tag{2.20}$$

where $W \equiv \Sigma^{-1}$ and $Q \equiv \Sigma L^\top(L\Sigma L^\top)^{-1}L$. Further, we have

$$\begin{aligned} Q^\top WQ &= L^\top(L\Sigma L^\top)^{-1}L & (2.21)\\ &= W\Sigma L^\top(L\Sigma L^\top)^{-1}L & (2.22)\\ &= WQ. & (2.23)\end{aligned}$$

This identity will become more important in the later course, since the trace allows cyclic permutations of a product. Notice how $p\times p$ matrix Q depends on L, which results in the property $Q\mathbf{1}_p = \mathbf{0}_p$. However, we did not specify L other than to comply with $L\mathbf{1}_p = \mathbf{0}_{(p-1)}$, therefore, let us express Q solely in terms of kernel $\mathbf{1}_p$ (McCullagh, 2009), which yields

$$Q = I_p - (\mathbf{1}_p^\top W\mathbf{1}_p)^{-1}\mathbf{1}_p\mathbf{1}_p^\top W. \tag{2.24}$$

[2] Scaling c is temporarily treated as an interest parameter, because the current transformation is concerned with the removal of $\boldsymbol{w}$.

Now, Q is only a function of W, however its rank is $p-1$. Due to this, the determinant in Eq. (2.19) (after factoring out c) reads

$$
\begin{aligned}
|L\Sigma L^\top|^{-1} &= |(L\Sigma L^\top)^{-1}| && (2.25)\\
&= \det(L^\top (L\Sigma L^\top)^{-1} L) \cdot |L^\top L|^{-1} && (2.26)\\
&= \det(WQ) \cdot |L^\top L|^{-1} && (2.27)\\
&= p \cdot (\mathbf{1}_p^\top W \mathbf{1}_p)^{-1} \cdot |W| \cdot |L^\top L|^{-1}, && (2.28)
\end{aligned}
$$

see (McCullagh, 2009), where $\det(\bullet)$ represents the *generalized determinant* as product of *non-zero* eigenvalues, because WQ is rank deficient. Both p and $|L^\top L|$ in Eq. (2.28) are absorbed into the proportionality constant of the likelihood, thereby removing all remaining occurrences of L. In summary, the *translation-invariant* log-likelihood in $W = \Sigma^{-1}$ is

$$
\begin{aligned}
\ell(W,c) = \tfrac{n}{2}\log|W| - \tfrac{n}{2}\log(\mathbf{1}_p^\top W \mathbf{1}_p)&\\
- \tfrac{(p-1)n}{2}\log(c^2) - \tfrac{1}{2}c^{-2}\operatorname{tr}\{WQXX^\top\}.&
\end{aligned}
\qquad (2.29)
$$

2.4.2 Profile Likelihood Approach

For the derivation of the marginal likelihood, we resorted to a statistic that removes the dependence on $\boldsymbol{w}$. As an alternative approach, it is also possible to find the maximum likelihood estimate for $\boldsymbol{w}$ in Eq. (2.9) by

$$
\frac{\partial}{\partial \boldsymbol{w}} \ell(W,c,\boldsymbol{w}) \overset{!}{=} \mathbf{0}_n \quad \Leftrightarrow \quad \widehat{\boldsymbol{w}} = (\mathbf{1}_p^\top W \mathbf{1}_p)^{-1} X^\top W \mathbf{1}_p, \qquad (2.30)
$$

which already has strong resemblance with the previous result. Using $\widehat{\boldsymbol{w}}$, the trace of Eq. (2.9) becomes (factoring out c)

$$
\begin{aligned}
&\Big(X - (\mathbf{1}_p^\top W \mathbf{1}_p)^{-1}\mathbf{1}_p\mathbf{1}_p^\top W X\Big)^\top W \Big(X - (\mathbf{1}_p^\top W \mathbf{1}_p)^{-1}\mathbf{1}_p\mathbf{1}_p^\top W X\Big)\\
&= X^\top \Big(I_p - (\mathbf{1}_p^\top W \mathbf{1}_p)^{-1}\mathbf{1}_p\mathbf{1}_p^\top W\Big)^\top W \Big(I_p - (\mathbf{1}_p^\top W \mathbf{1}_p)^{-1}\mathbf{1}_p\mathbf{1}_p^\top W\Big) X\\
&= X^\top Q^\top W Q X\\
&= X^\top W Q X.
\end{aligned}
$$

The cyclic property of the trace allows us to rewrite the last line as $WQXX^\top$. Since $\widehat{\boldsymbol{w}}$ depends on X, the normalization term (Harville, 1974) changes to

$$p \cdot (\mathbf{1}_p^\top W \mathbf{1}_p)^{-1} \cdot |W|, \tag{2.31}$$

which is equivalent to $\det(WQ)$. Inserting $\widehat{\boldsymbol{w}}$ back into the log-likelihood and using the above identities, the profile log-likelihood coincides with the marginal log-likelihood in Eq. (2.29), however, it adds a different perspective to the previous result.

2.5 Invariance against Scaling

2.5.1 Profile Likelihood Approach

Using the translation-invariant log-likelihood from Eq. (2.29), we now aim to remove scaling factor c by calculating its maximum likelihood estimate:

$$\frac{\partial}{\partial c}\ell(W, c) = 0 \quad \Leftrightarrow \quad \widehat{c}^2 = \tfrac{1}{(p-1)n} \operatorname{tr}\{WQXX^\top\}. \tag{2.32}$$

As a result, we receive the profile log-likelihood as

$$\ell_P(W) = \tfrac{n}{2} \log|W| - \tfrac{n}{2} \log(\mathbf{1}_p^\top W \mathbf{1}_p) - \tfrac{(p-1)n}{2} \log \operatorname{tr}\{WQXX^\top\}, \tag{2.33}$$

which is invariant against scalar multiples of X.

2.5.2 Marginal and Integrated Likelihood Approach

Alternative to the above, it is also possible to apply the statistics

$$u(X) = LX/\|\mathrm{vec}(LX)\| \quad \text{and} \quad v(X) = \|\mathrm{vec}(LX)\| \tag{2.34}$$

in order to normalize the data to a fixed scale (McCullagh, 2003; Cruddas et al., 1989; Tunnicliffe-Wilson, 1989). In particular, u implies that $\|\mathrm{vec}(LX)\| \neq 0$. Following the marginal likelihood scheme with interest parameter $\psi \equiv W$ and nuisance $\lambda \equiv c$, we have

$$f(u, v\,; W, c) = \underbrace{f(u\,; W)}_{L(W\,;\,u)} \cdot f(v \mid u\,; c). \tag{2.35}$$

Hereby, the likelihood is based on the marginal density of u, which does not depend on scaling factor c. After the transformation of rvs $X \mapsto (U, V)$, we receive a marginal log-likelihood which is identical to the profile log-likelihood in Eq. (2.33).

Interestingly, Harville (1977) showed that the same scale-invariant likelihood also arises when an improper uniform prior over the real line is imposed on c. Integrating it out yields

$$L(W) \propto \int_{\mathbb{R}} L(W, c) \cdot f(c)\, \mathrm{d}c. \tag{2.36}$$

Hence, all discussed approaches lead to the same result.

2.6 The Formulation in Distances

Now that the likelihood is invariant against translation $\mathbf{1}_p \boldsymbol{w}^\top$ and scaling c, we are left with an explicit dependence on X. To remedy this shortcoming, recall the identity $WQ = Q^\top WQ$ and the cyclic property of the trace, such that we can write

$$\mathrm{tr}\{WQXX^\top\} = \mathrm{tr}\{WQXX^\top Q^\top\}. \tag{2.37}$$

Further, take the connection between $S = XX^\top$ and D in Eq. (1.37) and multiply it from left and right with Q and $Q^\top$, respectively. This yields

$$\begin{aligned} QDQ^\top &= Q\,\mathrm{diag}(S)\mathbf{1}_p^\top Q^\top + Q\mathbf{1}_p\,\mathrm{diag}(S)^\top Q^\top - 2QSQ^\top && (2.38) \\ &= -\,2QSQ^\top, && (2.39) \end{aligned}$$

where the diagonal terms of S cancel due to $Q\mathbf{1}_p = \mathbf{0}_p$. Combining all above identities, the translation- and scale-invariant log-likelihood from Eq. (2.33) finally becomes a function in D:

$$\ell(W) = \tfrac{n}{2}\log|W| - \tfrac{n}{2}\log(\mathbf{1}_p^\top W \mathbf{1}_p) - \tfrac{(p-1)n}{2}\log\operatorname{tr}\{-\tfrac{1}{2}WQD\}. \quad (2.40)$$

2.7 Inference

Summarizing our current efforts, we observe a $p \times p$ distance matrix D and want to infer the underlying covariance matrix Σ of a Gaussian mixture model, which is parametrized by $ZAZ^\top + I_p$ for obtaining a block-wise approximation. These two parameters A and Z will be estimated in Bayesian fashion, hence the following develops suitable priors.

2.7.1 A Prior for the Inner Product of Cluster Means

So far, we did not specify A other than to be the inner product of the cluster-defining means. In fact, however, there are three possible choices for A, each permitting different degrees of freedom (compare Fig. 2.4 and Fig. 2.5):

- $A \in \mathbb{R}$: $ZAZ^\top$ is block diagonal with scalar value A in all k blocks. This implies, the cluster centers have an orthogonal basis and the same distance from the point of origin.

- $A \in \mathrm{diag}$: $ZAZ^\top$ is block diagonal, however with different values in each block. Similarly, the cluster centers are still spanned by an orthogonal basis, but with arbitrary distances to the point of origin.

- $A \in \mathbb{S}_+$: A symmetric, positive-definite $k \times k$ matrix leads to a full block matrix $ZAZ^\top$. As a result, the cluster centers may have arbitrary geometry without the need for an orthogonal basis. This enables us, for example, to place three clusters on a line—something which cannot be captured with the previous two models.

For a perhaps more intuitive explanation of Fig. 2.4, recall that the block structure originates from $\frac{1}{n}MM^\top$, meaning zeros in Σ arise from orthogonal mean vectors $\boldsymbol{m}_\bullet$. The translation invariance enables us to cluster the data *as if they were centered in the latent feature space*. In other words, the

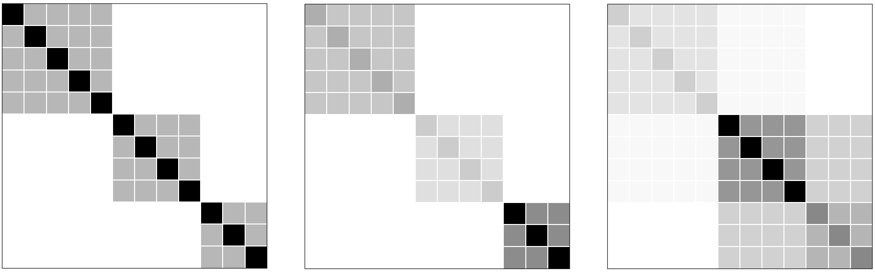


Figure 2.4: A graphical representation of covariance matrix $\Sigma = ZAZ^{\top} + I_p$ for three models of A: $\mathbb{R}$ (left), diag (center) and $\mathbb{S}_+$ (right).

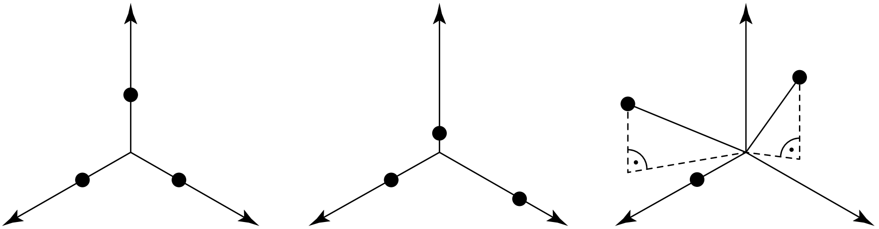


Figure 2.5: Geometry of the cluster centers corresponding to Fig. 2.4. Here, we look at only one out of many possible $(k = 3)$-dimensional subspaces, which A lives in. From left to right: All centers must be placed on an orthogonal basis and be equidistant ($A \in \mathbb{R}$); centers are placed on an orthogonal basis, but may have different distances to each other ($A \in \mathrm{diag}$); centers are allowed to have arbitrary geometry and distances to each other ($A \in \mathbb{S}_+$). Note that due to the model's invariances, every translation, rotation/reflection and scaling of these geometries is equivalent. The plots only represent *one single choice* of a coordinate system.

model was constructed in such a way that the clusters are independent of absolute location, rotation, reflection or scaling in feature space. All required information is captured by the inner product of (relative) means, A.

Since the likelihood explicitly requires precision matrix W, it is meaningful to analyze the parametrization of inverse.

Theorem 1. *$\Sigma = ZAZ^\top + I_p$ and $W = \Sigma^{-1}$ have the same block structure. Further, the cluster-defining matrix B of W is a function of A and Z.*

Proof. Assume

$$\Sigma = ZAZ^\top + I_p \tag{2.41}$$

$$W = ZBZ^\top + I_p \tag{2.42}$$

and require $\Sigma W \overset{!}{=} I_p$. This leads to

$$ZAZ^\top ZBZ^\top + ZAZ^\top + ZBZ^\top + I_p \overset{!}{=} I_p \tag{2.43}$$

$$Z(AZ^\top ZB + A + B)Z^\top = 0_{p\times p} \tag{2.44}$$

$$AZ^\top ZB + A + B = 0_{k\times k} \tag{2.45}$$

$$B = -(AZ^\top Z + I_k)^{-1}A. \tag{2.46}$$

□

From Eq. (2.46), we can deduce the following: Due to A and $Z^\top Z$ being positive definite, B is a negative-definite matrix, caused by the minus sign. Also, the Z parametrization of W implies that blocks of Σ persist in its inverse. Interestingly, a scalar A results in a $k \times k$ diagonal matrix B. This is due to the occurrence of $Z^\top Z$ which counts the number of objects per cluster on its diagonal. If the clusters are *not* of equal size, B has different values on the diagonal in spite of $A \in \mathbb{R}$. In the case of $A \in \mathrm{diag}$, B is again diagonal. This means, from the standpoint of B, there is no computational benefit of choosing $A \in \mathbb{R}$ instead of $A \in \mathrm{diag}$; generating and updating A, however, involves different costs. Finally, for the maximum degree of freedom, $A \in \mathbb{S}_+$, we have a corresponding negative-definite $B \in \mathbb{S}_-$.

In all cases, matrix B is never chosen directly, but rather computed from Z and A. Why is this? Assume there is a fixed A and a single object changes its assignment from cluster i to j. Then, $\Sigma = ZAZ^\top + I_p$ only changes by *one row* and *one column*, however, this very operation affects *multiple blocks* in W (at least 2 blocks, depending on the model for A). The reason for this behavior is due to B being a function of the individual cluster sizes, as seen in Eq. (2.46).

For a scalar A, any prior for positive reals is suitable, for example, a gamma distribution with density

$$f(A\,; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} A^{\alpha-1} \exp(-\beta A), \tag{2.47}$$

where shape $\alpha > 0$ and rate $\beta > 0$. The same choice can be carried over to the diagonal

$$A = \operatorname{diag}(A_1, \dots, A_k), \tag{2.48}$$

such that each element A_j, is i.i.d. gamma distributed.

For a positive-definite matrix A, one possibility is the Wishart distribution, having density

$$f(A\,; A_0, q) = \frac{1}{(2\pi)^{\frac{kq}{2}} |A_0|^{\frac{q}{2}} \, \Gamma_k\left(\frac{q}{2}\right)} |A|^{\frac{q-k-1}{2}} \exp(-\tfrac{1}{2} \operatorname{tr}\{A_0^{-1} A\}) \tag{2.49}$$

with $q \geq k$ degrees of freedom, $k \times k$ positive-definite matrix A_0, and $\Gamma_k(\bullet)$ as the multivariate gamma function.

Note that a prior for A is required despite scale invariance of the likelihood, even for $A \in \mathbb{R}$. Hereby, we can fix the noise scale at 1 without loss of generality, as done in $\Sigma = ZAZ^\top + I_p$, such that the scale of A carries only a *relative* meaning. To name an example, small values of A (relative to I_p) imply a large noise level.

2.7.2 A Prior for Mixture Weights and Cluster Assignments

The underlying assumption of our model was the standard mixture of Gaussians in Eq. (2.1), where the k components and its parameters

$$\{(\pi_1, \boldsymbol{m}_1), \ldots, (\pi_k, \boldsymbol{m}_k)\}, \tag{2.50}$$

are supposed to be inferred from data. It is important to stress that the above parametrization is not unique, since any pairwise shuffling produces the same outcome. As a consequence, the order is *non-identifiable*, or, speaking in previous terms, a nuisance parameter. In Bayesian inference, such a characteristic is problematic, as the ambiguity in labeling introduces modes of symmetry in the posterior, which severely interfere with its estimation. We also refer to this phenomenon as *label switching* (Redner and Walker, 1984; Stephens, 2000). The solution is to either make the clusters artificially identifiable (for example, by an ordering constraint on the component means $\|\boldsymbol{m}_1\|_2 < \cdots < \|\boldsymbol{m}_k\|_2$) or to break the symmetry by removing the dependence on the labels altogether (McCullagh and Yang, 2008). In comparison, the latter choice is more elegant, because it excludes unnecessary information and reduces the problem to what is relevant. To briefly explain their scheme, let the mixture weights follow a symmetric[3] Dirichlet distribution

$$\boldsymbol{\pi} \equiv [\, \pi_1, \ldots, \pi_k \,]^\top \sim \mathcal{D}(\xi/k), \tag{2.51}$$

where $k \geq 2$, $\xi > 0$ and $\sum \pi_j = 1$. Its density reads

$$f(\boldsymbol{\pi}\,; \xi, k) = \frac{\Gamma(\xi)}{(\Gamma(\xi/k))^k} \prod_{j=1}^{k} \pi_j^{\xi/k-1}. \tag{2.52}$$

The Dirichlet distribution is a common choice in clustering, as it can explain an infinite number of mixtures (Neal, 2000). Given weights $\boldsymbol{\pi}$, the objects

[3] Symmetry means that every mixture component is equally likely, which is a plausible choice prior to seeing the data. Therefore, a symmetric Dirichlet distribution is parametrized by a single *concentration parameter* instead of k different ones.

are then labeled with

$$\boldsymbol{z} \equiv [\, z_1, \ldots, z_p \,]^\top \tag{2.53}$$

from a multinomial distribution with density

$$f(\boldsymbol{z}\,; \boldsymbol{\pi}, k) = \frac{p!}{p_1! \cdot \ldots \cdot p_k!} \prod_{j=1}^{k} \pi_j^{p_j} \tag{2.54}$$

with p_j describing how many times label $j \in \{1, \ldots, k\}$ was observed among p objects, which also implies $\sum p_j = p$. This particular combination of distributions is mathematically convenient, since the Dirichlet is conjugate to the multinomial. Consequently, we can analytically integrate out the mixture weights from the product to receive

$$f(\boldsymbol{z}\,; \xi, k) = \int f(\boldsymbol{z}\,; \boldsymbol{\pi}, k) \cdot f(\boldsymbol{\pi}\,; \xi, k)\, \mathrm{d}\boldsymbol{\pi} \tag{2.55}$$

$$= \frac{\Gamma(\xi)}{(\Gamma(\xi/k))^k} \cdot \frac{\prod_{j=1}^{k} \Gamma(p_j + \xi/k)}{\Gamma(p+\xi)}. \tag{2.56}$$

Finally, in order to eliminate the label-switching problem, McCullagh and Yang (2008) express Eq. (2.56) in terms of blocks in $ZZ^\top$, leading to

$$f(Z\,; \xi, k_{\mathrm{mix}}) = \frac{k_{\mathrm{mix}}!}{(k_{\mathrm{mix}} - k)!} \cdot \frac{\Gamma(\xi)}{(\Gamma(\xi/k_{\mathrm{mix}}))^k} \cdot \frac{\prod_{j=1}^{k} \Gamma(p_j + \xi/k_{\mathrm{mix}})}{\Gamma(p+\xi)}, \tag{2.57}$$

where $k_{\mathrm{mix}} > k$ is the *number of mixtures*. Also, the first factor has been added to account for symmetric modes, such that unidentifiable permutations with the same outcome are combined into one. If $k_{\mathrm{mix}} \leftarrow \infty$, one arrives at the *Ewens process*, also known as *Chinese Restaurant process*, see (Aldous, 1985, p. 91) and (Pitman, 1995). Conversely, if $k_{\mathrm{mix}} \leftarrow N \in \mathbb{N}_+$, we speak of a *truncated* Ewens process which enforces an upper bound $k < N$. Note that $k \leq p$, meaning p objects can form at most $k = p$ singleton clusters.

For practical applications using an MCMC sampler, we always compare two partitions against each other via the ratio of densities. Say, we have partition G and H, then

$$r = \frac{f(G\,;\xi, k_{\text{mix}})}{f(H\,;\xi, k_{\text{mix}})}. \tag{2.58}$$

Hence, if it is possible to split off constant factors from Eq. (2.57), it suffices to evaluate only terms that actually differ. As a simple example, let partition G have p_4 objects in block 4 and p_7 objects in block 7. Further, the total number of blocks in G is k. If H supersedes G in that it switches an object from block 4 to block 7, then Eq. (2.58) simplifies to

$$r_{4\to 7} = \frac{(p_4 - 1) + \xi/k_{\text{mix}}}{(p_7 + 1) + \xi/k_{\text{mix}}}. \tag{2.59}$$

If H instead assigns the same object to a *new* block, the ratio becomes

$$r_{4\to *} = \frac{(p_4 - 1) + \xi/k_{\text{mix}}}{\xi(1 - k/k_{\text{mix}})}. \tag{2.60}$$

Therefore, the densities never need to be evaluated in full as long as we only compare incremental changes.

Finally, when using the truncated Ewens process with finite $k_{\text{mix}} \leftarrow N \leq p$, the ratio for a new cluster in Eq. (2.60) becomes exactly zero if $k = N$. This implies there can be at most $k = N - 1$ blocks, which was, in fact, the purpose of truncation.

2.7.3 Markov Chain Monte Carlo Sampling

Combining the translation- and scale-invariant likelihood from Eq. (2.40) and the above priors for A and Z, we receive the posterior

$$f(Z, A \,|\, D, \bullet\,) \propto f(D \,|\, W, n) \cdot f(Z \,|\, \xi, k_{\text{mix}}) \cdot f(A \,|\, \bullet\,), \tag{2.61}$$

where $W \equiv \Sigma^{-1} = (ZAZ^\top + I_p)^{-1}$. There is one remaining parameter in the likelihood, which has not been addressed yet: the dimensionality of the latent feature space, n. If there are fewer linear independent dimensions than objects, it is possible to identify n by the rank of D. In the typical case however, we have $n \geq p$, which means the parameter is not observable anymore and possibly infinite. The solution is to interpret n as a temperature as used in *simulated annealing*, since it appears in the exponent of every term in the likelihood. Hereby, we can effectively control the variance of the posterior samples from the MCMC process: small (large) values of n lead to large (small) temperature (= variation). We start with a small value, for example, $n \leftarrow p$, and increase it slowly until the MCMC samples finally converge, say, 1% of all objects changed their assignment in the last 1000 samples. In conjunction with a logarithmic cooling schedule, this scheme is guaranteed to converge to the global optimum *in the limit of infinite time* (Nourani and Andresen, 1998).

In order to generate an MCMC sample from the posterior, we propose a simple Metropolis-within-Gibbs scheme that is explained in Algorithm 1 and 2. Further, we also report the complexity in big $\mathcal{O}$ notation for each step: first for $A \in \mathbb{R}$ and $A \in \text{diag}$ (since they both lead to $B \in \text{diag}$), second for $A \in \mathbb{S}_+$. The sampler is initialized with the following parameters:

$$k \leftarrow 1, \quad Z \leftarrow \mathbf{1}_p, \quad A \leftarrow 1 \quad \text{and} \quad n \leftarrow p.$$

2.7.4 Complexity Analysis

The runtime of Algorithm 1 is mainly governed by the innermost loop over every object and every cluster, therefore, this is the primary target for optimization. In fact, when the posterior is naively recomputed in every iteration, the worst-case complexity for the standard Ewens process ($k_{\text{mix}} \leftarrow \infty$) adds up to $\mathcal{O}(p^5)$, which is clearly not suited for any practical application. To remedy this, observe that each operation makes only incremental changes to the previous state. In addition, an important technique is to exploit the block structure, such that actions involving $p \times p$ matrices can be reduced to k calculations. For example, we can write

$$|W| = |ZBZ^\top + I_p| = |Z^\top ZB + I_k|, \tag{2.62}$$

Algorithm 1 Gibbs sampler

```
for $i = 1$ to $p$ do
    Precompute fixed terms for the loop in $c$                $\rightsquigarrow \mathcal{O}(p)/\mathcal{O}(p)$
    for $j = 1$ to $k$ do
        Assign object $i$ to cluster $j$, compute posterior    $\rightsquigarrow \mathcal{O}(k)/\mathcal{O}(k^2)$
    end for
    Assign object $i$ to new cluster, compute posterior       $\rightsquigarrow \mathcal{O}(k)/\mathcal{O}(k^3)$
    Assign object $i$ permanently                             $\rightsquigarrow \mathcal{O}(1)/\mathcal{O}(1)$
    if object $i$ is assigned to new cluster then
        Sample new $A$, see Algorithm 2                       $\rightsquigarrow \mathcal{O}(k^2)/\mathcal{O}(k^3)$
    end if
end for
Sample new $A$, see Algorithm 2                               $\rightsquigarrow \mathcal{O}(k^2)/\mathcal{O}(k^3)$
```

Algorithm 2 Metropolis sampler for A

```
Generate proposal $A^*$ and compute posterior                 $\rightsquigarrow \mathcal{O}(k^2)/\mathcal{O}(k^3)$
if acceptance ratio $> u \sim \mathcal{U}(0,1)$ then
    $A \leftarrow A^*$                                        $\rightsquigarrow \mathcal{O}(1)/\mathcal{O}(1)$
end if
```

which is the determinant of a smaller $k \times k$ matrix, hence $\mathcal{O}(p^3)$ is reduced to $\mathcal{O}(k^3)$. Further, it is even possible to avoid full recomputation of $|Z^\top ZB + I_k|$, when operations can be written in terms of rank-1 updates. In this case, an existing QR decomposition (Lay, 2011, p. 356f) can be updated in $\mathcal{O}(k^2)$ to reflect these changes. The determinant is then obtained in $\mathcal{O}(k)$ by

$$|\det(Z^\top ZB + I_k)| = |\det(Q)| \cdot |\det(R)| = |\prod R_{ii}|. \tag{2.63}$$

Here, we used $\det(\bullet)$ for a better distinction from the absolute value. Note that the absolute value does not pose a problem, since W is positive definite by definition.

If B is a diagonal matrix, then the QR decomposition is not needed. Instead, the determinant of W breaks into the product of block-wise determinants, thus $\mathcal{O}(k)$. In addition, incremental updates are found in $\mathcal{O}(1)$. Other optimization techniques include the order of calculation, that is,

$$\mathbf{1}_p^\top W \mathbf{1}_p = \mathbf{1}_p^\top (ZBZ^\top + I_p)\mathbf{1}_p = (Z^\top \mathbf{1}_p)^\top B (Z^\top \mathbf{1}_p) + p \tag{2.64}$$

only requires "small" matrix-vector products of order k instead of costly matrix-matrix operations. A second technique employs cyclic permutation to split the trace into smaller parts, where we have

$$\begin{aligned}
&\operatorname{tr}\{-\tfrac{1}{2}WQD\} && (2.65)\\
&= \operatorname{tr}\Big\{-\tfrac{1}{2}W\Big(I_p - (\mathbf{1}_p^\top W \mathbf{1}_p)^{-1}\mathbf{1}_p\mathbf{1}_p^\top W\Big)D\Big\} && (2.66)\\
&= -\tfrac{1}{2}\operatorname{tr}\{BZ^\top DZ\} - \tfrac{1}{2}\operatorname{tr}\{D\} - \tfrac{1}{2}(\mathbf{1}_p^\top ZBZ^\top \mathbf{1}_p + p)^{-1} \\
&\quad \cdot \Big(\mathbf{1}_p^\top ZBZ^\top DZBZ^\top \mathbf{1}_p + 2 \cdot \mathbf{1}_p^\top DZBZ^\top \mathbf{1}_p + \operatorname{tr}\{D\}\Big). && (2.67)
\end{aligned}$$

From Eq. (2.67), we see that some terms occur multiple times and can be pulled out of the innermost loop: The contribution of object i in $(D_{i\bullet}Z)$ is found in $\mathcal{O}(p)$ and $\operatorname{tr}\{D\}$ can be precomputed once. Again, for a fast runtime, the evaluation order must be considered to use matrix-vector multiplications where possible.

In summary, there is a large potential for optimization due to the block structure and many reappearing terms. Table 2.1 reports the overall complexity when these properties are successfully exploited. Hereby, the results simply aggregate the complexity for the individual steps that were already stated in Algorithm 1 and 2.

Due to $A \in \mathbb{R}$ and $A \in \operatorname{diag}$ leading to a diagonal matrix B[4], both models have the same complexity in big $\mathcal{O}$ notation. When the objects can be partitioned into only few clusters, we can expect reasonable performance of the truncated Ewens process with approximately quadratic cost. Also, the flexibility of $A \in \mathbb{S}_+$ comes at the price of $\mathcal{O}(p^4)$ for the standard Ewens process, thereby preventing an application to large-scale data. Only for the truncated variant with $k_{\text{mix}} \ll p$, the model regains its practical relevance.

[4] Recall that B is the cluster-defining part of the inverse covariance matrix W.

		Ewens process	
A	B	truncated	standard
$\mathbb{R}$	diag	$\mathcal{O}(p^2 + pN^2)$	$\mathcal{O}(p^3)$
diag	diag	$\mathcal{O}(p^2 + pN^2)$	$\mathcal{O}(p^3)$
$\mathbb{S}_+$	$\mathbb{S}_-$	$\mathcal{O}(p^2 + pN^3)$	$\mathcal{O}(p^4)$

Table 2.1: Overall complexity for Algorithm 1 and 2 using the various models for A. Regarding truncation, it is guaranteed that $k < N \leq p$, which is a fixed and predetermined number. For the standard Ewens process, we have $k \leq p$, that is, p objects can form at most p singleton clusters.

The next section will explore an extension to the cluster model aimed specifically at lowering the worst-case complexity. In more detail, there is an interesting analogy between a covariance matrix and a binary rooted tree. This gives a different perspective on the centering problem, which was previously solved by making the likelihood invariant against translation.

2.8 Extension: Centering by Trees

When we incorporated translation invariance into the likelihood, the motivation was to find a projection of the data that removes any column means of the form $M = \mathbf{1}_p \boldsymbol{w}^\top$, where $\boldsymbol{w} \in \mathbb{R}^n$. The important detail is the resulting projection matrix

$$Q = I_p - (\mathbf{1}_p^\top W \mathbf{1}_p)^{-1} \mathbf{1}_p \mathbf{1}_p^\top W,$$

which has the property $Q\mathbf{1}_p = \mathbf{0}_p$. In particular, note that Q is a function of $W \equiv \Sigma^{-1}$. Let us now restate the scale- and translation-invariant log-likelihood from Eq. (2.40) with identity $WQ = Q^\top WQ$:

$$\ell(W) = \tfrac{n}{2} \log\det(Q^\top WQ) - \tfrac{(p-1)n}{2} \log\operatorname{tr}\{-\tfrac{1}{2}Q^\top WQD\}. \tag{2.68}$$

At this point, the cyclic property of the trace gives rise to two different interpretations: The log-likelihood can either be seen as a function in *transformed inverse covariance matrix* $Q^\top WQ$ given D, written as

$$\ell(Q^\top WQ\,; D), \tag{2.69}$$

or as a function in W given the *transformed distances* $QDQ^\top$, that is,

$$\ell(W\,; QDQ^\top). \tag{2.70}$$

The first option was used to derive the translation-invariant likelihood, but the second is in fact equivalent, thereby raising the question whether there is any benefit to it. To better understand its meaning, notice that

$$-\tfrac{1}{2}QDQ^\top = QSQ^\top \equiv S_* \tag{2.71}$$

is essentially a *centering operation* of the data to identify a single matrix S_* out of the equivalence set $\mathcal{S}(D)$, although S_* is strictly speaking rank-deficient. At first glance, we cannot make proper use of projection Q, because it is a function of the yet to be inferred W. However, if we had an intuition about W, hypothetically, we could approximate S_* in a preprocessing step and then fall back to the simpler translation-*variant* (but still scale-*invariant*) log-likelihood (see Section A.1 for its derivation):

$$\ell(W) = \tfrac{n}{2} \log|W| - \tfrac{np}{2} \log\operatorname{tr}\{WS_*\}. \tag{2.72}$$

In conjunction with a diagonal cluster model, $A \in \operatorname{diag}$, this leads to an algorithm with worst-case complexity of only $\mathcal{O}(p^2)$ instead of $\mathcal{O}(p^3)$. Details are found in Section A.1.

Now, assume we are given an estimate $\widehat{W} \approx W$, then the projection becomes $\widehat{Q} \approx Q$ and leads to $\widehat{S} \equiv -\frac{1}{2}\widehat{Q}D\widehat{Q}^\top$. Although $\widehat{S}$ is an approximation of the centered S_*, we are guaranteed to stay inside the set $\mathcal{S}(D)$ and leave the pairwise distances unchanged, even for poor choices $\widehat{W} \not\approx W$. Our hope is that even a rough estimate already leads to a reasonably well-centered $\widehat{S}$. In that case, the expected gain in computational performance by far outweighs the loss due to approximation. Still, there are also special instances when W can be estimated accurately, as will be shown in the following. For now it suffices to see that if we *had* access to the true W, the translation-*variant* model using S_* would be equivalent to the translation-*invariant* model based on D. This is a substantial property and speaks in strong favor for the soundness of the approach.

2.8.1 Constructing a Tree from Distances

At the current stage, we do not know W (or Q for that matter), however, there exist methods to construct a tree from a distance matrix—many of these were developed in biology to identify phylogenetic relationships between taxa (= species). More specifically, the data are mapped to a tree with p leaves, where each leaf represents one object; the distance between two leaf nodes is given by summing up edge lengths along the shortest connecting path in the tree. The goal is now to find a topology, where the distance d_{ij}, as measured between leaf i and j, matches what is given by the distance matrix, D_{ij}. In other words, we wish to minimize

$$\sum_{i=1}^{p} \sum_{j=1}^{p} (D_{ij} - d_{ij})^2. \tag{2.73}$$

From a combinatorial perspective, a binary rooted tree with p leaves has $p-1$ internal nodes and $2(p-1)$ edges, thereby leading to a total number of

$$(2p-3)!! = (2p-3)(2p-5)\cdots(c+2)\,c, \quad c = \begin{cases} 1, & \text{odd } p \\ 2, & \text{even } p \end{cases} \tag{2.74}$$

possible tree topologies (Page and Edward, 1998). This heavily outweighs the degrees of freedom in a distance matrix, which has at most $\mathcal{O}(p^2)$ distinct values, that is, all elements on the upper or lower triangular submatrix. For example, a distance matrix with $p = 12$ contains a maximum of 66 unique values, but there are already $21!! \approx 1.37 \cdot 10^{10}$ potential tree topologies. Therefore, finding a tree is a strongly underdetermined problem unless we impose further constraints on the data. One special case is the ultrametric space (Page and Edward, 1998; Felsenstein, 2003), which requires that the distances between any three points A, B and C satisfy

$$D_{AC} \leq \max\{D_{AB}, D_{BC}\}. \tag{2.75}$$

This can be seen as a stronger version of the triangle inequality, such that ABC is restricted to an isosceles triangle, where the sides of same length are longer than the third side. The triangle ABC is also allowed to be equilateral, thus having three sides of same length. If a distance matrix follows this property, there is exactly one corresponding binary rooted tree which can be recovered using the *UPGMA* algorithm (unweighted pair-group method with arithmetic mean) (Sokal and Michener, 1958). This method operates directly on the distance matrix and iteratively merges two nodes with the smallest distance into a new parent node. For every merge, the rows and columns of the children are removed from the distance matrix and replaced by new distances of the parent node to all remaining nodes. The process continues until there is only one node left—the root. Given pairs of nodes to be merged and their distances, we can draw a graphical representation of the procedure to receive the tree. Note that if the distance matrix does *not* conform to the ultrametric property—as often encountered in practice—, UPGMA produces the *closest matching tree* under the infinity norm. To demonstrate this, Fig. 2.6 shows a *non*-ultrametric D for $p = 12$ objects and the resulting tree.

The ultrametric assumption can easily be seen from the tree structure: First, it requires a root node, which implies that all species (the leaves) stem from one common ancestor. Second, all leaf nodes have the same distance to the root. In the context of phylogenetic trees, this is also known as a *clock-like topology* (Felsenstein, 2003), because it assumes a common molecular clock which governs the rate of mutation across all species. This also implies that the vertical direction serves as a time axis, where the leaf nodes live on a joint temporal front. In general, there is no formal justification for a molecular clock or even a root node, but it is a common and plausible

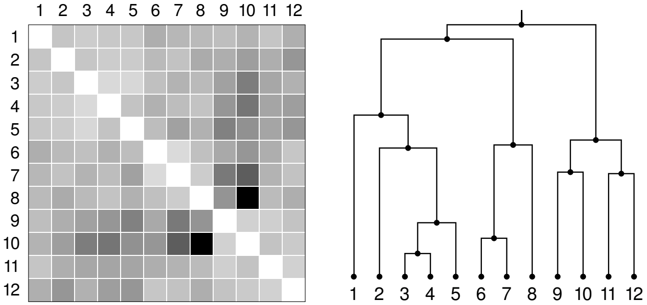


Figure 2.6: Constructing a tree from distances. Left: *Non*-ultrametric distance matrix D. Right: Closest matching binary rooted tree.

assumption for the analysis of (roughly) related species.

2.8.2 Covariance Matrix from a Tree

The purpose of constructing a tree from a distance matrix is to eventually receive an estimate of the inverse covariance matrix W. To this end, Fig. 2.7 demonstrates how every edge in the previous tree represents a binary partition of the leaf nodes. Yet, the tree—being a hierarchical clustering of leaves—contains even more information: the length of an edge indicates the importance or weight of its associated binary partition (see λ_i in Fig. 2.7).

Thus, if we combine these two sources of information and compute the weighted sum over all possible partition matrices, we receive the underlying covariance matrix (McCullagh, 2009). We can also think of this as finding a *consensus clustering* from many simple, binary partitions, where the individual partitions are highly overlapping due to the nature of a tree topology. For a mathematical description of the 2-block matrix in Fig. 2.7 corresponding to the cut of edge i, we write

$$\mathbb{1}_i \mathbb{1}_i^\top + \bar{\mathbb{1}}_i \bar{\mathbb{1}}_i^\top, \tag{2.76}$$

where $\mathbb{1}_i \in \{0,1\}^p$ and $\bar{\mathbb{1}}_i \equiv \mathbf{1}_p - \mathbb{1}_i$ are the *indicator vector* and its binary complement, respectively; their purpose is to denote which leaf node belongs to which group. In total, there are $2(p-1)$ edges in a binary rooted tree

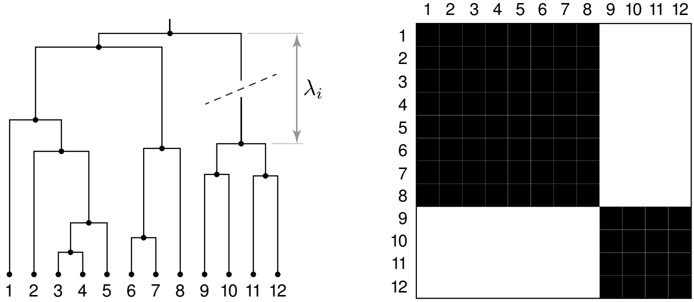


Figure 2.7: Cutting the tree from Fig. 2.6 at the indicated edge splits the leaf nodes into two groups 1–8 and 9–12 (left), which is equivalently captured by a binary partition matrix (right).

(excluding the edge above the root), thus we have

$$\Sigma_{\text{tree}} \equiv \sum_{i=1}^{2(p-1)} \lambda_i \cdot (\mathbb{1}_i \mathbb{1}_i^\top + \bar{\mathbb{1}}_i \bar{\mathbb{1}}_i^\top) \tag{2.77}$$

with λ_i being the length of edge i. Note that we call the resulting covariance matrix Σ_{tree} to distinguish it from the block-structured $\Sigma = ZAZ^\top + I_p$. Finally, Fig. 2.8 demonstrates how Eq. (2.77) is applied to the running example of the previous figures to obtain Σ_{tree}. Although the distances did not satisfy the ultrametric property, the estimate is surprisingly close to the true Σ. For a better judgment of the result, note how the data were generated: Sample $X \sim \mathcal{N}_{p,n}(0_{p\times n}, \Sigma \otimes I_n)$ with $p = 12$ and $n = 100$, then compute the corresponding distances by $D_{ij} = (X_{i\bullet} - X_{j\bullet})(X_{i\bullet} - X_{j\bullet})^\top$.

As a final remark, the time complexity for the UPGMA algorithm and the subsequent decomposition is $\mathcal{O}(p^2)$ (Murtagh, 1984; Adametz and Roth, 2011). A binary rooted tree can also be found via other distance-based methods like the single- or complete-linkage algorithm; both have efficient implementations in $\mathcal{O}(p^2)$, see (Gower and Ross, 1969; Sibson, 1973) and Defays (1977).

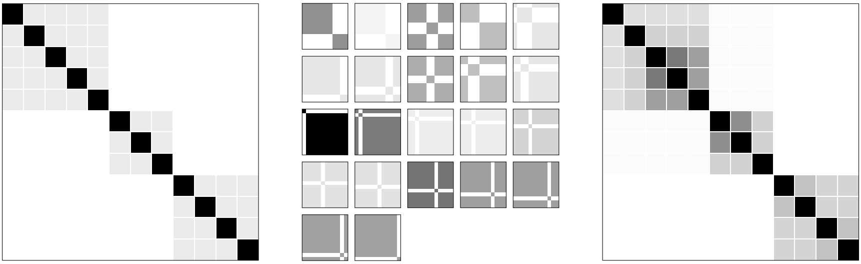


Figure 2.8: Inferring the covariance matrix from the tree in Fig. 2.6. Left: True covariance matrix underlying the distances in D. Center: All binary partition matrices multiplied by their edge length. Right: Estimated covariance matrix as the sum of weighted partition matrices. Note: The grayscale palette was rescaled in the right plot to match the range of values.

2.8.3 Computational Advantages by using the Tree

With estimated covariance matrix Σ_{tree} in place, we can now calculate $W_{\text{tree}} \equiv \Sigma_{\text{tree}}^{-1}$ and proceed to the centering operation

$$\widehat{S} = -\tfrac{1}{2}\widehat{Q}D\widehat{Q}^{\top},$$

where

$$\widehat{Q} = I_p - (\mathbf{1}_p^{\top} W_{\text{tree}} \mathbf{1}_p)^{-1} \mathbf{1}_p \mathbf{1}_p^{\top} W_{\text{tree}}. \tag{2.78}$$

From a computational point of view however, this particular step would require $\mathcal{O}(p^3)$ due to matrix inversion and a matrix-matrix product, thereby making it the most expensive calculation in the pipeline. To this end, notice how the inverse W_{tree} only enters $\widehat{Q}$ by vector

$$\boldsymbol{y} \equiv (W_{\text{tree}} \mathbf{1}_p) \in \mathbb{R}^p, \tag{2.79}$$

which gives rise to a more efficient formulation in conjunction with the tree.

Remark

The remaining steps are only given in condensed form; for a detailed proof, see (Adametz and Roth, 2011).

In short, the idea is to first find $\boldsymbol{y}$ by

$$\min_{\boldsymbol{y}} \|\Sigma_{\text{tree}}\boldsymbol{y} - \mathbf{1}_p\|^2 \tag{2.80}$$

in $\mathcal{O}(p^2)$, which exploits the tree topology to compute $\Sigma_{\text{tree}}\boldsymbol{y}$. Next, identity

$$\widehat{Q} = I_p - (\mathbf{1}_p^\top \boldsymbol{y})^{-1}\mathbf{1}_p\boldsymbol{y}^\top \tag{2.81}$$

enables us to calculate

$$\begin{aligned} \widehat{S} &= -\tfrac{1}{2}\widehat{Q}D\widehat{Q}^\top && (2.82)\\ &= -\tfrac{1}{2}D + \tfrac{1}{2}(\mathbf{1}_p^\top \boldsymbol{y})^{-1}\,\mathbf{1}_p\boldsymbol{y}^\top D + \tfrac{1}{2}(\mathbf{1}_p^\top \boldsymbol{y})^{-1}\,D\,\boldsymbol{y}\mathbf{1}_p^\top \\ &\quad - \tfrac{1}{2}(\mathbf{1}_p^\top \boldsymbol{y})^{-2}\,\mathbf{1}_p\boldsymbol{y}^\top D\,\boldsymbol{y}\mathbf{1}_p^\top && (2.83) \end{aligned}$$

in $\mathcal{O}(p^2)$. Hence, the time complexity to (i) construct the tree, (ii) decompose it into a covariance matrix and then (iii) find a centered $\widehat{S}$ is $\mathcal{O}(p^2)$. From an overall standpoint, the appeal of the approach is primarily due the significant speed up in clustering, but it also shows an alternative treatment of translation as a nuisance parameter: the extensive work on phylogenetic trees can now be used as a tool for centering, to make one choice among the set $\mathcal{S}(D)$.

2.9 Experiments

The goal of this section is to give a better picture about the cluster models, the tree extension as well as related methods. We will refer to the standard model as *TiWD* (*Translation-invariant Wishart Dirichlet* process) (Vogt et al., 2010) and the tree-based counterpart as *fastTiWD* (Adametz and Roth, 2011).

2.9.1 Synthetic Data

In the first experiment, we generate data as follows: Z is initialized as a zero matrix with $p = 500$ rows and $k = 5$ columns. Each object is then uniform randomly assigned to one of the five clusters, which approximately comprise an equal number of objects. To define the inner product of the cluster centers, $A \in \mathbb{S}_+$ is drawn from $\mathcal{W}_k(k+3, I_k)$. Regarding the noise, $ZAZ^\top$ is complemented with a scaled identity matrix that defines the noise level. Since we aim for a challenging cluster structure, the noise factor is set to 40. Note that due to scale invariance, the likelihood treats $\Sigma_1 = ZAZ^\top + 40I_p$ and $\Sigma_2 = Z(\frac{1}{40}A)Z^\top + I_p$ as equivalent, hence the *relative* noise level is sufficient. From Σ_1, we sample $X \sim \mathcal{N}(0_{p\times n}, \Sigma_1 \otimes I_n)$[5] with $n = 600$, which is finally used to compute the pairwise distances D. The process is repeated to produce a total number of 100 datasets, each of which is evaluated by TiWD, fastTiWD (both using $A \in \mathbb{S}_+$) and competing methods.

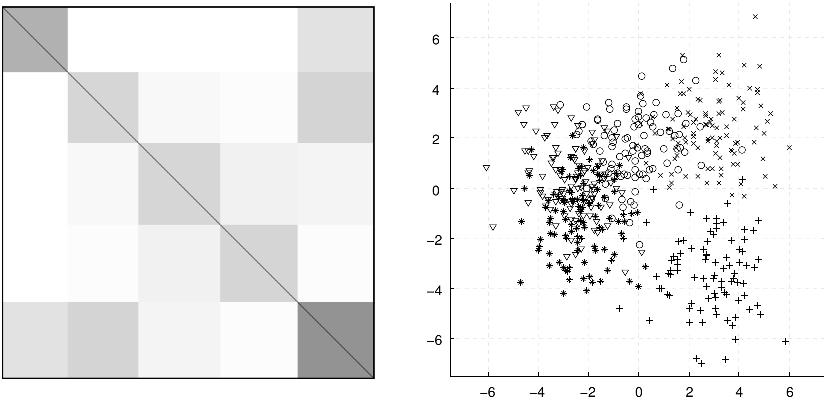


Figure 2.9: Clustering of synthetic data ($k = 5$, $p = 500$, $n = 600$). Left: True covariance matrix Σ, which leads to highly overlapping clusters. Right: Given Σ, we can sample X; the plot shows its 2D-PCA projection with true cluster labels.

Fig. 2.9 illustrates a single covariance matrix Σ and its corresponding

[5] Adding column shifts to X is not necessary here as they automatically cancel in D.

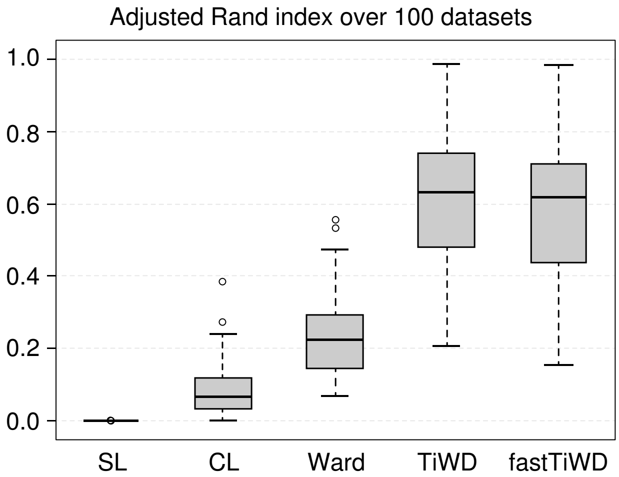


Figure 2.10: The adjusted Rand index as boxplot over 100 datasets. Methods are (from left to right): single linkage, complete linkage, Ward's method, TiWD and fastTiWD.

data matrix X in a 2D-PCA projection. The resulting clusters are highly overlapping and pose a challenging task as can be seen in Fig. 2.10. The performance of each method is measured by the adjusted Rand index between true and inferred labels. While single-linkage clustering performs worst, complete-linkage clustering seems to capture minimal structural information. Ward's method is able to exceed both, however only by a small margin. The highest results are achieved by TiWD ($A \in \mathbb{S}_+$) with a median adjusted Rand index of about 0.63, where each run consisted of 5000 iterations of the MCMC sampler. In order to put this into perspective, we also applied the centering operation using a tree construction from D. The algorithm relies on $\widehat{S}$, an estimate of the centered $S_* = -\frac{1}{2}QDQ^\top$, to use the simpler translation-variant likelihood. Although the computations are slightly simplified, the algorithm does not gain any significant benefits for $A \in \mathbb{S}_+$ as the complexity is maintained. Still, judging from its accuracy (median adjusted Rand index ≈ 0.61), fastTiWD performs virtually identical compared to its fully translation-invariant counterpart. The slight drop can be explained by the tree construction, which requires ultrametric distances and therefore identifies the closest matching tree. Aside from the difference in likelihood, all parameters were kept identical.

2.9.2 Real-World Data: Semi-Supervised Clustering of Protein Sequences

Remark

This experiment largely follows (Adametz and Roth, 2011), but adds further details and explanations.

Since fastTiWD has only quadratic complexity in combination with a scalar or diagonal model A, it is possible to apply it to fairly large datasets in a reasonable amount of time. For a demonstration of this, we select an application from biology where the task is to classify protein sequences based on a few examples with an already known label. While it is always possible to infer structure from the data without any prior knowledge, the semi-supervised setting is presumably the most interesting in practice. Hence, the question is which protein sequences fall into already existing categories and which are different enough to form groups of their own.

Due to the non-vectorial nature of proteins, it is difficult to cluster the objects when they are represented as amino-acid chains, for example, AST-$KGPSVF...$. Our solution is to abandon the original domain of the sequences and instead evaluate it in terms of pairwise alignment scores, which are straightforward to compute. We can think of this a kernel function that operates in the feature space of amino acids, hereby measuring the similarity between two inputs. The implication of a kernel is, however, that it fixes the point of origin in the latent feature space, which we assumed to be uninformative. Therefore, only the distances are actually relevant to us.

In the current problem, the semi-supervised type of clustering naturally arises due to two different databases: *SwissProt* contains protein sequences that are manually annotated with a stringent review process; *TrEMBL* is larger, not reviewed and consists only of automatic annotations. From a clustering perspective, we assume that the high-quality labels given by SwissProt are true (the supervised examples), whereas the classification from TrEMBL is untrusted and therefore completely discarded. In essence, this means assignment matrix Z contains a set of fixed rows, which are not altered during the sampling process.

Vogt et al. (2010) also analyzed this particular semi-supervised task, but the cubic complexity limited the size of the dataset: In total, $p = 3771$ globin-like protein sequences were used, out of which 1168 sequences belonged to 114 classes as given by the SwissProt database. Now, due to the significant speedup via the tree construction, it is feasible to go beyond the scope of globins and handle the superset of *oxygen binding* and *transport*, adding up to $p = 12\,290$ sequences. This set contains a much richer and diverse class of sequences among which we find *hemocyanins*, *hemerythrins*, *chlorocruorins* and *erythrocruorins*. SwissProt lists 1731 sequences with 356 known classes.

We conducted 5000 Gibbs iterations (the cluster structure stabilized after around 1100), and the total runtime on a standard computer was approximately 6 hours. In this instance, we used the scalar model of A, which requires the least amount of computations, although it formally has the same complexity in big $\mathcal{O}$ notation as its diagonal counterpart. The most interesting detail is how it compares to TiWD (while using the identical model for A), but unfortunately its runtime is beyond being feasible for clustering. Hence, we ran both methods alongside each other for 100 iterations and found an improvement of factor 192, which would result in an estimated runtime of 1152 hours (or 48 days) for TiWD. This clearly demonstrates the significant performance gain, but it also explains why a complexity reduction is undoubtedly needed for larger datasets.

In our experiment with fastTiWD, we were able to identify 23 unique clusters among the TrEMBL sequences, which are dissimilar to any class given by SwissProt. Such a result is particularly interesting from a biological standpoint, because it integrates new data into existing knowledge and highlights potential candidates for further research. The result is pictured in Fig. 2.11.

Most of the newly identified clusters contain sequences with rare, but specific structural properties. Similar to the results in (Vogt et al., 2010), we find a large cluster containing *flavohemoglobins* from a particular species of funghi and bacteria. These proteins have a certain domain architecture

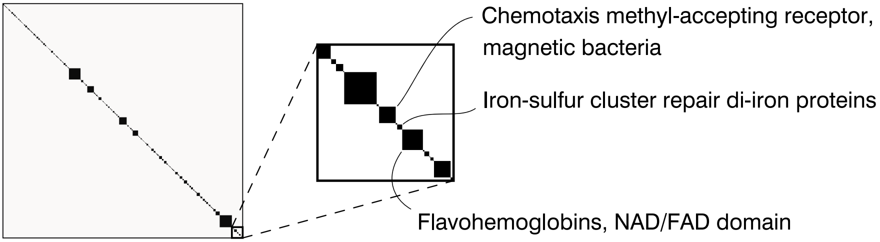


Figure 2.11: Inferred block structure for all $p = 12\,290$ protein sequences involved in oxygen binding and transport. The supervised labels are derived from the SwissProt database, which assigns 1731 sequences to one of 356 classes. Interestingly, we could find 23 new clusters among the sequences listed in the TrEMBL database, which are different from any existing class in SwissProt (see enlarged area).

in common, which is composed of a globin domain fused with ferredoxin reductase-like FAD- and NAD-binding modules. A second example is a new cluster containing proteins with a *chemotaxis methyl-accepting receptor* domain from a special class of *magnetic* bacteria, which are able to align themselves to earth's magnetic field. One potential advantage of this orientation property (known as *magnetotaxis*) is that by keeping the bacteria aligned against Brownian motion, they might be more efficient at sensing chemical signals for directing their movements (*chemotaxis*), see (Dusenbery, 2009, p. 164–167). The domain architecture of these proteins (see Fig. 2.12) is unique among all sequences in the dataset.

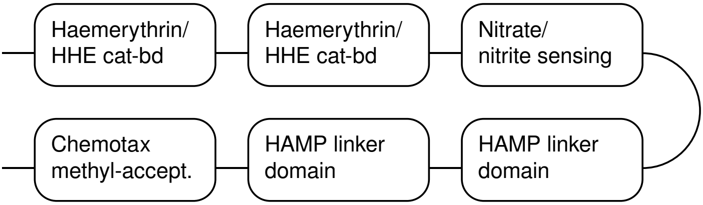


Figure 2.12: Domain architecture of the magnetic bacteria.

A last example is a cluster of *iron-sulfur cluster repair di-iron proteins*. These proteins contain a polymetallic system, the *di-iron center*, constituted by two iron ions bridged by two sulfide ions. In our dataset, such di-iron centers occur only in this particular cluster.

Overall, the purpose of this experiment is to demonstrate how fastTiWD enables analysis that was previously infeasible because of high computational requirements. Like TiWD, the Gibbs sampler can easily be altered to allow semi-supervised clustering, such that new data can be conveniently integrated into existing knowledge. Hence, fastTiWD offers all advantages and properties of TiWD at much smaller cost. Could we have arrived at the same results without clustering? Theoretically yes, considering we are able to dig into numerous protein databases and fuse their information. It would, however, also involve a large portion of domain knowledge and probably human interaction to a high degree—something that grows out of proportion for numbers of 12 000 proteins. This is in fact one of the main reasons why SwissProt is much smaller than TrEMBL.

2.10 Conclusion

In this chapter, we presented a method for clustering pairwise distances, which is based on the mixture of Gaussians. While the setting is well-defined when the full design matrix X is observed, the transfer to distances is not straightforward. The immediate challenge comes from the fact that information loss occurs, which affects a range of parameters that are otherwise essential for statistical inference. In particular, due to the nature of pairwise (dis)similarities, certain properties that characterized the original feature space become inaccessible—the individual features, the dimensionality of the feature space (if D has full rank) and the point of origin. Not only does this increase the difficulty, but the lack of knowledge also introduces a high degree of freedom of potential matrices X from which the distances *could* have originated from. Since the likelihood explicitly depends on this information, we referred to these parameters as nuisance. In the subsequent sections, it was shown how different techniques can be applied to incorporate invariances into likelihood, such that the unknown nuisance parameters are either (i) removed by marginalization via suitable statistics or (ii) substituted with their best supported estimate, hereby collapsing the parameter space into a profile. Fig. 2.13 demonstrates the invariances that were developed throughout this chapter: all variations of the data and any combinations thereof are equivalent. Hence, what uniquely defines a partition are the object-to-cluster assignments and the inner product of cluster means relative to the noise level. The plots were generated with a diagonal matrix A, which implies that the geometry of the clusters possesses an orthogonal basis (possibly coinciding with the point of origin).

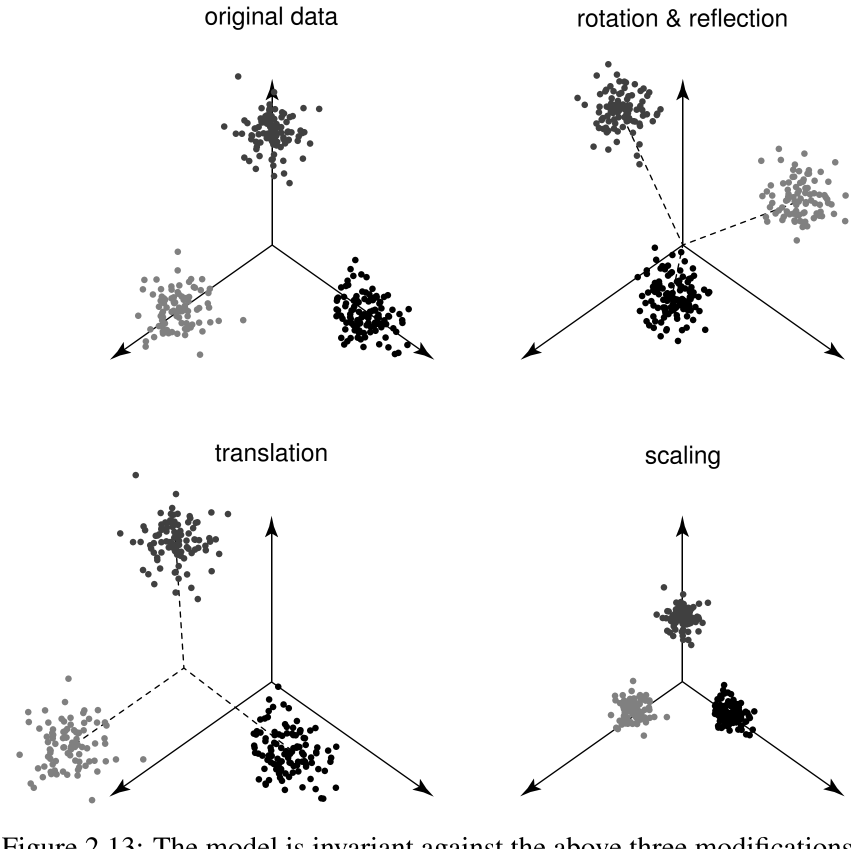


Figure 2.13: The model is invariant against the above three modifications.

Throughout the domain of clustering, we found two complementary approaches for dealing with information loss regarding the point of origin: The first was to avoid making any choice about the unknown parameter. Instead, we applied a transformation, such that the term completely cancels from the likelihood. In an extension arising from the cyclic property of the trace, the second option was to pick one candidate explicitly by means of a tree. This alternative approach does not render translation invariance obsolete as it uses an approximative centering of the data which depends on external knowledge. Yet, we gain a considerable benefit in runtime and its accuracy is in many instances very close to that of the full model, even though the data does not strictly meet all requirements, say, the ultrametric property. Maybe most importantly, it adds an alternative interpretation to the problem and further shows a potentially deeper relationship to consensus clustering.

At the core of a Gaussian mixture model, the underlying assumption is that objects form a cluster whenever they can be explained by a common mean, hence, the essential part of inference concerns the identification of cluster structure. Since we are unable to characterize these means in terms of their latent feature space, we took the intermediate step to approximate the non-central Wishart distribution by a central one, hereby expressing the cluster geometry conveniently via their inner product as captured by matrix A. The choice for its model—scalar, diagonal or symmetric positive definite—heavily depends on requirements of the user and the problem at hand, but it is important to keep in mind that the added flexibility comes at the price of substantially increased computational cost. It is relatively easy to artificially construct cluster geometries which cannot be expressed by a scalar or diagonal A, for example when three clusters are positioned on a line. In these instances, the model is likely to compensate by introducing additional clusters such that the mismatch is reduced. This clearly speaks for the most flexible variant, that is, a full matrix A, however it might not be necessary for a large number of clusters: Since A lives in the space of cluster centers, an orthogonal basis in high dimensions (= many clusters) might already be a *very good* fit, albeit not *perfect*.

In typical situations, the aspect of runtime is a much more decisive factor than modeling the cluster geometry *exactly*. Also, the data we work with might not be Gaussian after all, thereby possibly violating our assumption of spherical clusters. Hence, the sacrifice in flexibility is often subordinate and should more suitably be treated in an approximative manner. As long as the data exhibits a reasonably clear cluster separation, the correct partition can most certainly be found—in spite of a model mismatch.

Remark

Further thoughts about the model are given in Appendix A.

Chapter 3

A Gaussian Graphical Model for Distances

In this chapter we introduce a method for estimating *Gaussian graphical models* (*GGMs*) from pairwise distance data. Since it relies on the same input as the model for clustering, there is a certain amount of overlap, however, it also has considerable differences that justify a dedicated chapter.

3.1 Introduction

Let us begin with a description of the classic GGM: its main building block is a $p \times n$ matrix $\widetilde{X}$, which contains information about p objects (= rows) living in an n-dimensional space (= columns). The important assumption is that this matrix follows a matrix normal distribution, that is

$$\widetilde{X} \sim \mathcal{N}_{p,n}(M, \Sigma \otimes I_n). \tag{3.1}$$

The goal in GGMs is to identify precision matrix $W \equiv \Sigma^{-1}$, since it encodes the *conditional independences* between the p objects. More specifically, a zero element in W, say $W_{ij} = W_{ji} = 0$, represents that *object i and j are conditionally independent given all other objects*. This fundamental property *only* holds for the normal distribution and is yet another reason for its importance in statistical analyses. To this end, W captures characteristics about $\widetilde{X}$, which give deeper insights into the data. Conditional independences are explained in more detail by the following example.

Conditional Independences for Gaussian Data

Assume a vector of normal rvs distributed as

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \sim \mathcal{N}_3\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} & \Sigma_{XZ} \\ \Sigma_{YX} & \Sigma_{YY} & \Sigma_{YZ} \\ \Sigma_{ZX} & \Sigma_{ZY} & \Sigma_{ZZ} \end{bmatrix} \right). \tag{3.2}$$

Then, the conditional distribution of X and Y given Z reads

$$X, Y \mid Z \sim \mathcal{N}_2(\mathbf{0}_2, \widetilde{\Sigma}), \tag{3.3}$$

where

$$\widetilde{\Sigma} = \begin{bmatrix} \Sigma_{XX} - \Sigma_{XZ}\Sigma_{ZZ}^{-1}\Sigma_{ZX} & \Sigma_{XY} - \Sigma_{XZ}\Sigma_{ZZ}^{-1}\Sigma_{ZY} \\ \Sigma_{YX} - \Sigma_{YZ}\Sigma_{ZZ}^{-1}\Sigma_{ZX} & \Sigma_{YY} - \Sigma_{YZ}\Sigma_{ZZ}^{-1}\Sigma_{ZY} \end{bmatrix}. \tag{3.4}$$

X and Y are conditionally independent given Z, that is, the off-diagonal elements in $\widetilde{\Sigma}$ are zero, if and only if the precision matrix has the property

$$\Sigma^{-1} = W = \begin{bmatrix} W_{XX} & 0 & W_{XZ} \\ 0 & W_{YY} & W_{YZ} \\ W_{ZX} & W_{ZY} & W_{ZZ} \end{bmatrix}, \tag{3.5}$$

meaning $W_{XY} = W_{YX} = 0$.

It is common to draw a graphical representation for W in terms of an undirected graph, where the objects are vertices and the individual values in W become edges. In more detail, the edge color describes the sign (e.g., $+/-$ as black/gray) and edge thickness matches the absolute value (up to a scalar factor). The most interesting information, however, concerns the conditional independences, which correspond to the absence of an edge. Fig. 3.1 demonstrates this for an exemplary precision matrix and its associated graph.

Clearly, the estimated precision matrix is required to be *sparse* (up to a user-defined level), otherwise the graph is fully connected and its topology

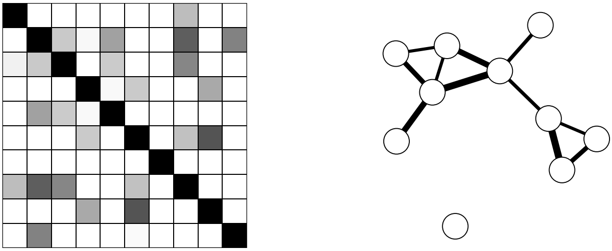


Figure 3.1: Precision matrix W (left) and its interpretation as an undirected graph (right). Matrix elements with white color contain zeros. As for the graph, self-loops due to the diagonal elements are typically omitted.

is meaningless for analyzing the conditional independences. At this point, a valid question is whether network inference is equivalent to clustering and if its block parametrization of W can be reused. From Fig. 3.2 we see that the definition of a cluster—a set of similar objects which is distinct from other sets—does not translate properly to GGMs. Instead, both approaches are rather complementary: if we had coherent subgroups among the objects, we could first cluster them and then either (i) analyze their conditional independences individually (micro level) or (ii) infer a network between clusters

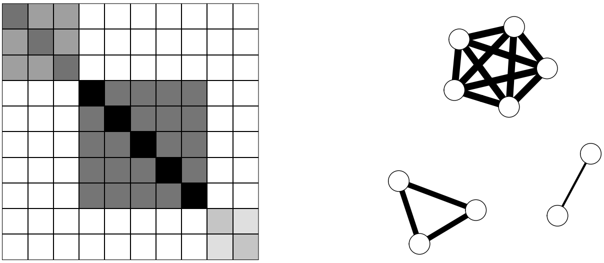


Figure 3.2: Clustering versus network inference. Block-diagonal precision matrix (left) as derived from clustering, but interpreted as an undirected graph (right). Blocks correspond to mutually independent, but internally fully-connected subgraphs (cliques).

(macro level). These are, in fact, two possibilities to effectively break down a large number of objects into smaller sets, the latter being referred to as *module networks* by Prabhakaran et al. (2013).

Due to our focus on distance-based methods, any domain is suitable for analysis as long as a kernel- or distance matrix can be computed. Consequently, we can infer a network of *probability distributions*, of *strings* or *protein sequences*, of *semantic texts* or *images*, of *chemical structures* or even *graphs* themselves. This is made possible by the 'zoo' of kernel functions which have been developed for virtually any type of input data. In that regard, the scope of potential applications is much larger than that of classical GGMs, which is limited to vectorial representations of the data.

For a proper understanding of the basic idea, let us begin with a related model: Prabhakaran et al. (2013) proposed the *Translation-invariant Wishart network* (*TiWnet*), which builds on the matrix normal distribution described in Eq. 3.1 and assumes squared Euclidean distances between object i and j,

$$D_{ij} \equiv (\widetilde{X}_{i\bullet} - \widetilde{X}_{j\bullet})(\widetilde{X}_{i\bullet} - \widetilde{X}_{j\bullet})^\top. \tag{3.6}$$

Euclidean distances also formed the basis of the cluster model from the preceding chapter, however, we will now focus on a variation of Eq. 3.1:

$$X \equiv \widetilde{X}\Psi^{\frac{1}{2}} \sim \mathcal{N}_{p,n}(M, \Sigma \otimes \Psi). \tag{3.7}$$

Both X and $\widetilde{X}$ are $p \times n$ matrices, but their difference lies in the additional feature correlation as given by $n \times n$ matrix Ψ. Although this may initially appear as a minor modification, its importance becomes more evident when we compute the corresponding distances: Suppose the existence of a feature correlation was known, then we should correctly evaluate the squared *Mahalanobis distance*

$$(D_{\text{MH}})_{ij} \equiv (X_{i\bullet} - X_{j\bullet})\Psi^{-1}(X_{i\bullet} - X_{j\bullet})^\top, \tag{3.8}$$

such that the bias from correlation is accounted for. Under this treatment, $D_{\text{MH}} = g(X)$ and $D = h(\widetilde{X})$ are equivalent.

The difficulty arises when a distance matrix is given, which by definition allows no access to its underlying features. In that situation, it is not clear which of the two distance measures was applied to generate the observations; the (now) latent features may or may not have been correlated. Hence, there are two options:

1. We decide to enforce strict feature independence as done in (Prabhakaran et al., 2013) to make the problem tractable first and foremost. This implies that the distances are assumed to be Euclidean and, to this end, Fig. 3.3 demonstrates how distances are perceived differently under feature correlation. In the worst case scenario, the data might be misinterpreted.

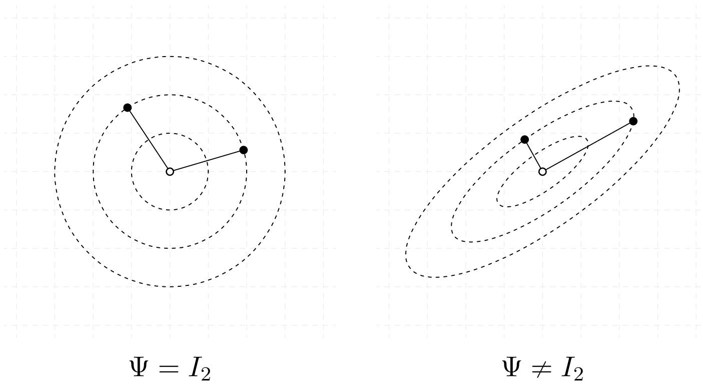
2. We choose the Mahalanobis distance, because it is more general and also includes feature independence as a special case. Unfortunately, the features are not accessible, therefore, their correlation cannot simply be removed by transformation, for example, as in

   $$X\Psi^{-\frac{1}{2}} = \widetilde{X},$$

   although it may have a big impact on the distances. To make matters worse, not even the number of features is known if it exceeds the number of objects, as we learned in Chapter 1.

In summary, the introduction of feature correlation complicates inference from distances, since it alters the data in its underlying latent space. The challenge comes from the fact that distances depend on four individual parameters: number of features n, mean matrix M, row covariance Σ and column covariance/correlation Ψ. This means, finding a network is the task of isolating Σ from the remainder, or in other words, we are required to interpret every single pairwise distance in such a way, that structural properties are distinguished from nuisance parameters, which only act on the feature space.

For a better intuition about the problem, it is always helpful to analyze the balance between number of variables and observations: We are given a $p \times p$ distance matrix D to infer a $p \times p$ precision matrix W, while an unknown portion of D might be caused by $n \times n$ feature correlation matrix Ψ with $n \gg p$. Hence, the task is significantly underdetermined. To this end, imagine a fictitious experiment in which D is subject to strong feature



$\Psi = I_2$ $\Psi \neq I_2$

Figure 3.3: Euclidean distances between $p = 3$ points when $n = 2$ features are uncorrelated (left) and correlated (right). The concentric circles on the left-hand side represent "true" equidistant points relative to the white point, while the correlation on the right-hand side transforms the circles into ellipses.

correlation, but without any conditional dependence structure, i.e., $W = I_p$. In that particular case, enforcing feature independence would falsely attribute all the information to network structure where in fact there is none. This goes to show why latent feature correlation matters and why it demands a well-considered solution.

The following describes an application, which is potentially subject to feature correlation while sharing all the aforementioned properties.

3.1.1 Example: A Network of Biological Pathways

Suppose we are in a clinical domain where the task is to study a specific disease (say colon cancer) based on a patient cohort. Since the disease influences the human body on a wide range of biological functions, a tissue sample from the affected region is taken for each patient and subsequently analyzed with a *DNA microarray*. As a result from this procedure, we simultaneously receive expression values for around 24 000 known genes in humans, which in simplistic terms could be described as "gene activity". Unfortunately, these measurements are highly prone to noise and only weakly informative when analyzed on their own. Therefore, in an attempt to alleviate these shortcomings, we instead focus on groups of functionally related genes that jointly contribute to higher-level functions. In more precise terms, such a set of genes is also called *pathway* (Curtis et al., 2005). Our hope is that the weak signals underlying a single gene are amplified in a pathway, such that visible and stable patterns can be detected.

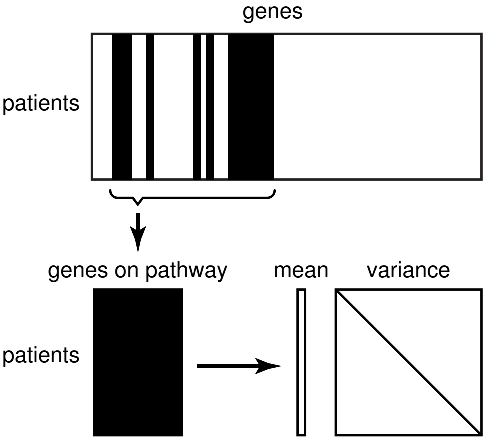


Figure 3.4: A pathway interpreted as the distribution of genes across patients.

Given the transition from single genes to high-level information, we could ask how different pathways interact with each other and if some of them are conditionally independent given others. To this end, note that a pathway contains a characteristic collection of gene expression values, meaning the object of interest is a *probability distribution*, see Fig. 3.4. It is not clear how to operate in this abstract domain directly, however, we can easily compute their pairwise dissimilarities using the *Bhattacharyya distance* (Bhattacharyya, 1943; Jebara and Kondor, 2003). This defines the foundation on which we wish to construct a network of pathway distributions. The question is: Are the patients (i.e., the features) truly independent realizations or could their common sex, age and treatment have "skewed" the distances? We will find an insightful answer in the experimental section.

3.2 Related Work

On the most basic level, all following methods rely on the matrix normal distribution

$$X \sim \mathcal{N}_{p,n}(M, \Sigma \times \Psi), \tag{3.9}$$

which is the building block for GGMs. The goal is always to infer a sparse $W \equiv \Sigma^{-1}$, however, the approaches differ in their individual assumptions about M and Ψ. An overview for all discussed variants is given in Fig. 3.5.

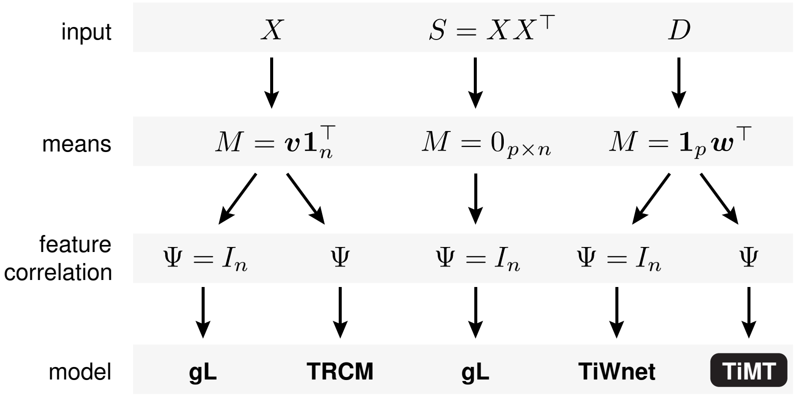


Figure 3.5: The big picture (Adametz and Roth, 2014). Different assumptions about M and Ψ lead to different models.

We begin with the distance-based approaches whose mean model is $M = \mathbf{1}_p\boldsymbol{w}^\top$, $\boldsymbol{w} \in \mathbb{R}^n$, because this information vanishes in D. Every other form of the mean, as for example row means $\boldsymbol{v}\mathbf{1}_n^\top$ with $\boldsymbol{v} \in \mathbb{R}^p$, alters D and is therefore regarded as informative. Next, if we assume independent features, we receive TiWnet (Prabhakaran et al., 2013) as introduced before. An arbitrary feature correlation falls into our current setup, leading to the *Translation-invariant Matrix-T* process (*TiMT*) (Adametz and Roth, 2014).

If matrix X is known, which contains full feature information, the mean is assumed to be $M = \boldsymbol{v}\mathbf{1}_n^\top$, that is, row means. Compared to the distance methods, this is neither better nor worse, but simply a different choice for a different application. Analog to the above, feature correlation divides this branch into the *graphical LASSO* (*gL*) (Friedman et al., 2008; Yuan and Lin, 2007) and the *Transposable Regularized Covariance Model* (*TRCM*) (Allen and Tibshirani, 2010). Both approaches are similar in that they optimize the likelihood under an $L1$ penalty, but TRCM estimates both Σ^{-1} *and* Ψ^{-1} by alternation.

The last branch uses similarity (or kernel) matrix $S = XX^\top$ with a strict zero mean assumption to avoid the centering problem. Further, when combined with independent features, we also receive gL, however, this should rather be treated as a special case due to its highly selective requirements.

3.3 Model

The starting point is the matrix normal log-likelihood:

$$\begin{aligned}\ell(W, \Psi, M\,; X) = \tfrac{n}{2}\log|W| - \tfrac{p}{2}\log|\Psi| \\ - \tfrac{1}{2}\operatorname{tr}\{W(X-M)\Psi^{-1}(X-M)^\top\}.\end{aligned} \tag{3.10}$$

To recapitulate the parameters, our goal is to infer precision matrix W, while everything related to the features is considered a nuisance:

- number of latent features n
- feature correlation Ψ
- mean matrix $M = \mathbf{1}_p \boldsymbol{w}^\top$

The reader should notice that, from the very outset, this is quite a difficult problem, because *none* of the parameters are known, not even data matrix X (or $\widetilde{X}$). Still, we can intuitively say that there must be *some* property of D that permits a statement about W, although it is not yet clear how to extract this limited information in mathematical terms. From a technical perspective, it is always possible to incorporate invariances into a model *somehow*, yet, this does not imply any guarantees about inference, as we will see. Therefore, the actual challenge is to maximize the statistical power of the model under information loss, which is particularly true for the current situation. The next section will discuss transformations of Eq. (3.10), such that it only depends on pairwise distances, and we begin by the correlation of features.

3.3.1 Invariance against Feature Correlation

Profile Likelihood Approach

Given our previous experience with the profile likelihood, it seems promising to base invariance on the maximum likelihood estimate of Ψ (McCullagh, 2008). Straight forward calculations lead to

$$\frac{\partial}{\partial \Psi}\ell(W, M\,; X, \Psi) \overset{!}{=} 0_{p\times p} \;\;\Leftrightarrow\;\; \widehat{\Psi} = \tfrac{1}{p}(X-M)^\top W(X-M), \quad (3.11)$$

meaning the best-supported feature correlation is a function of W and M. By inserting $\widehat{\Psi}$ back into the log-likelihood, we arrive at

$$\begin{aligned}&\ell_P(W, M\,; X, \widehat{\Psi})\\ &\quad= \tfrac{n}{2}\log|W| - \tfrac{p}{2}\log|\widehat{\Psi}| - \tfrac{1}{2}\operatorname{tr}\{W(X-M)\widehat{\Psi}^{-1}(X-M)^\top\} &(3.12)\\ &\quad= \tfrac{n}{2}\log|W| - \tfrac{p}{2}\log|W(X-M)(X-M)^\top|. &(3.13)\end{aligned}$$

It appears that this is a valid model, but Eq. (3.12) reveals an important limitation on closer inspection: $\widehat{\Psi}$ is an $n \times n$ matrix which can only be inverted if $n \leq p$. For the boundary condition $n = p$, we have an interesting special case:

$$\begin{aligned}&\ell_P(W, M\,; X, \widehat{\Psi})\\ &\quad= \tfrac{p}{2}\log|W| - \tfrac{p}{2}\log|W(X-M)(X-M)^\top| &(3.14)\\ &\quad= \tfrac{p}{2}\log|W| - \tfrac{p}{2}\log|W| - \tfrac{p}{2}\log|(X-M)(X-M)^\top| &(3.15)\\ &\quad= \;-\tfrac{p}{2}\log|(X-M)(X-M)^\top|. &(3.16)\end{aligned}$$

Here, the profile log-likelihood becomes degenerate as W cancels completely. As a result, the approach is only valid for $n < p$.

McCullagh (2008) identified an anomaly in the behavior of the resulting model with regards to n: Starting with a small $n \ll p$, the Fisher information grows when increasing n. However, from $n = \frac{p}{2}$ on, it declines monotonically and at $n = p$, it becomes zero. This is perhaps an unsurprising result, taking into account that the maximum likelihood estimate is only accurate if the sample size (here: p) is large relative to the number of variables (here: n). When approaching $n = p$, the performance deteriorates and the estimate becomes uninformative, e.g., if we were to compute the mean based on a single observation. Hence, this behavior is in accordance with our intuition.

In conclusion, the current profile likelihood has very limited practical appeal due to the aforementioned restrictions, and is therefore not suitable for inference in general distance matrices.

Integrated Likelihood Approach

The removal of nuisance parameters by integrating out a Bayesian prior has the advantage of being generally applicable, however, it requires a sensibly chosen distribution and the calculations may be difficult unless the prior is conjugate to the likelihood function. In our situation, a candidate distribution is required to have symmetric, positive-definite support and, to this end, Iranmanesh et al. (2010) proposed the use of an *inverse matrix gamma* prior

$$\Psi \sim \mathcal{G}_n^{-1}(\alpha, \beta, \Omega) \tag{3.17}$$

with density

$$f(\alpha, \beta, \Omega\,;\Psi) = \frac{|\Omega|^{\alpha}\,|\Psi|^{-\alpha-(n+1)/2}}{\beta^{\alpha n}\,\Gamma_n(\alpha)} \exp(-\beta^{-1}\,\mathrm{tr}\{\Omega\,\Psi^{-1}\}), \tag{3.18}$$

which is parametrized by $n \times n$ matrix $\Omega \succ 0$, $\alpha > \frac{1}{2}(n-1)$ and $\beta > 0$. Further, $\Gamma_n(\bullet)$ in the normalization constant refers to the multivariate gamma function. Note that this distribution is in fact a generalization of the *inverse Wishart distribution* in q degrees of freedom,

$$\Psi \sim \mathcal{W}_n^{-1}(q, \Omega), \tag{3.19}$$

which is received by setting $\alpha = \frac{q}{2}$ and $\beta = 2$.

As shown in (Iranmanesh et al., 2010), the inverse matrix gamma prior can be integrated out analytically in

$$f(X \mid \alpha, \beta, M, W, \Omega) = \int_{\Psi \succ 0} f(X \mid M, W, \Psi) \cdot f(\Psi \mid \alpha, \beta, \Omega) \mathrm{d}\Psi, \quad (3.20)$$

thereby leading to the *(generalized) matrix T* distribution

$$X \sim \mathcal{T}_{p,n}(\alpha, \beta, M, W, \Omega) \quad (3.21)$$

and its corresponding log-likelihood function

$$\begin{aligned} \ell(W, M\,; \alpha, \beta, X, \Omega) &= \tfrac{n}{2} \log|W| \\ &\quad - (\alpha + \tfrac{p}{2}) \log|I_p + \tfrac{\beta}{2} W (X - M) \Omega^{-1} (X - M)^\top|. \end{aligned} \quad (3.22)$$

Upon closer inspection, we see that Eq. (3.22) shares many similarities with the profile log-likelihood in Eq. (3.13). Aside from the hyperparameters of the prior, the determinant now has an additional identity matrix, which not only ensures full rank, but also serves as regularization in conjunction with $\beta > 0$, see Fig. 3.6.

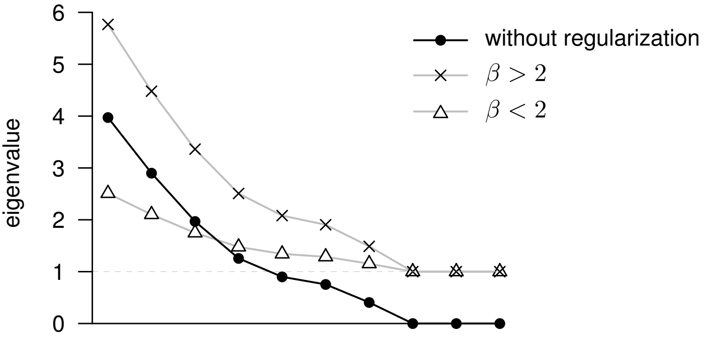


Figure 3.6: Eigenvalues of the determinant in Eq. (3.22) for synthetic data. Although the second part of the determinant can be rank-deficient (as pictured by black dots), the identity matrix ensures that all eigenvalues are ≥ 1. Additionally, the eigenvalues can be scaled with hyperparameter β.

Therefore, any $n \geq 1$ is valid, independent of p. Concerning β, there are two boundary conditions: for $\beta \to 0$, we have maximal regularization and the likelihood becomes independent of the data (i.e., a determinant equal to 1); for $\beta \to \infty$, the identity matrix "vanishes" and therefore no regularization is in effect, reminiscent of the profile log-likelihood. In fact, the integrated likelihood can often be reduced to the profile likelihood when using an improper prior (de Vos and Francke, 2008). To better understand this behavior, it is helpful to analyze the (co)variance of matrix X following the matrix T distribution:

$$\operatorname{cov}(X) = \frac{2(W^{-1} \otimes \Omega)}{\beta(2\alpha - p - 1)}. \tag{3.23}$$

Here, α and β jointly control the variance, which means that for a fixed α and $\beta \to 0$, the density becomes flat. In case of $\beta \to \infty$, however, it peaks infinitely sharp and collapses to a single element of support.

As can be seen by the variance, the hyperparameters give us a certain amount of flexibility to alter the model's behavior. Hence, we use this fact to fix $\Omega \equiv I_n$, meaning, on expectation, the latent features are assumed to be independent, which is a sensible choice prior to seeing any datum. The skeptical reader may wonder if this is a serious limitation of the model, however, we argue that the flexibility comes from the remaining hyperparameters to define a flat prior and, as a result, every possible Ψ contributes with a non-zero weight. Further, $\Omega \equiv I_n$ is an important prerequisite for the transition to distances in the next section.

3.3.2 Invariance against Column Means and Formulation in Distances

Incorporating translation invariance into the matrix T distribution (Adametz and Roth, 2014) is analog to the procedure in the matrix normal case. In essence, we formulate the likelihood not in X, but in statistic LX, where L is a $(p-1) \times p$ linear projection matrix. L satisfies the property $L\mathbf{1}_p = \mathbf{0}_{(p-1)}$, such that the column means $M = \mathbf{1}_p \boldsymbol{w}^\top$ are mapped to $0_{(p-1)\times n}$. In particular, Iranmanesh et al. (2010) proved that the matrix T distribution under affine transformations behaves in the following way:

$$AXB \sim \mathcal{T}_{p,n}(\alpha, \beta, AMB, AW^{-1}A^\top, B^\top \Omega B). \tag{3.24}$$

Currently, we have $A = L$, $B = I_n$ and $\Omega = I_n$, which yields the marginal log-likelihood

$$\begin{aligned} &\ell(W\,; \alpha, \beta, LX) \\ &\quad = -\tfrac{n}{2}\log|LW^{-1}L^\top| \\ &\qquad - (\alpha + \tfrac{p-1}{2})\log\left|I_{(p-1)} + \tfrac{\beta}{2}(LW^{-1}L^\top)^{-1}LXX^\top L^\top\right| \qquad (3.25) \\ &\quad = -\tfrac{n}{2}\log|LW^{-1}L^\top| \\ &\qquad - (\alpha + \tfrac{p-1}{2})\log\left|I_p + \tfrac{\beta}{2}L^\top(LW^{-1}L^\top)^{-1}LXX^\top\right|. \qquad (3.26) \end{aligned}$$

Note that LX is of size $(p-1) \times n$, which is reflected in Eq. (3.25). It is possible to cyclically permute the argument of the second determinant to arrive at Eq. (3.26), because the determinant is the product of eigenvalues and the pth eigenvalue is 1. In what follows, we have the previously introduced transition to Q

$$L^\top(LW^{-1}L^\top)^{-1}L = Q^\top WQ = WQ,$$

and further,

$$Q = I_p - (\mathbf{1}_p^\top W\mathbf{1}_p)^{-1}\mathbf{1}_p\mathbf{1}_p^\top W,$$

because L is arbitrary other than to satisfy $L\mathbf{1}_p = \mathbf{0}_{(p-1)}$. Again, identity

$$QDQ^\top = -2QSQ^\top$$

applies with $S = XX^\top$, which gives the translation-invariant log-likelihood in terms of D,

$$\begin{aligned} \ell(W\,; \alpha, \beta) = \tfrac{n}{2}\log|W| &- \tfrac{n}{2}\log(\mathbf{1}_p^\top W\mathbf{1}_p) \\ &- (\alpha + \tfrac{p-1}{2})\log|I_p - \tfrac{\beta}{4}WQD|. \end{aligned} \tag{3.27}$$

This formulation finally adheres to our initial model requirements in that (i) it only depends on pairwise distances D, meaning it is constant across column shifts, but more importantly, (ii) it achieves independence from any Ψ for a sensible choice of (α, β). Note that in comparison to enforcing $\Psi \equiv I_n$, our model only assumes independent features *on expectation* and it is able to account for arbitrary feature correlation due to the Bayesian prior and its hyperparameters. Hence, Adametz and Roth (2014) describe this as a *Bayesian relaxation* of feature independence.

One might ask whether Eq. (3.27) requires further transformations to account for unknown scaling. The answer is a qualified "no", because we may attribute a scalar c to feature correlation Ψ, as in

$$X = c\widetilde{X} \sim \mathcal{N}_{p,n}\Big(\mathbf{1}_p \boldsymbol{w}^\top, W^{-1} \otimes \{c^2\Psi\}\Big), \tag{3.28}$$

which is then integrated out in the matrix T posterior. Here, $\boldsymbol{w}$ absorbs the scaling without loss of generality. Equivalent to the above, Eq. (3.24) also permits $B \equiv c$, such that factor c can be merged into the hyperparameters. This intuitively makes sense, because a regularization parameter like β must always be specified *relative* to the scale of the input.

As a last remark, a critical reader may point out that $W \equiv \Sigma^{-1}$ in the matrix T distribution does not necessarily encode conditional independences as it did in the matrix normal distribution. This statement is correct, however, we regard invariances merely as technical means to isolate the parameter of interest—had we known the nuisance parameters explicitly, these operations would have not been required. Therefore, we eventually interpret W in terms of the original matrix normal distribution, *regardless* of all intermediate transformations applied to the likelihood.

3.4 Inference

Similar to Bayesian inference in the clustering application, we will now construct an MCMC sampler, which requires suitable priors for W and the remaining parameters.

3.4.1 Hyperparameters

The likelihood in its current form, see Eq. (3.27),

$$\begin{aligned}\ell(W\,;\alpha,\beta) = \tfrac{n}{2}\log|W| - \tfrac{n}{2}\log(\mathbf{1}_p^\top W\mathbf{1}_p) \\ - (\alpha + \tfrac{p-1}{2})\log|I_p - \tfrac{\beta}{4}WQD|,\end{aligned}$$

still depends on the number of latent features n, which does not appear in all terms and therefore prevents its use as an annealing parameter. A solution is found by the free hyperparameter α to express the factor as a multiple of $n/2$, such as in

$$\alpha + (p-1)/2 \overset{!}{=} vn/2. \tag{3.29}$$

Here, v is any scalar satisfying $v > 1 + (p-2)/n$ due to $\alpha > (n-1)/2$ and consequently, this yields:

$$\ell(W\,;v,\beta) = \tfrac{n}{2}\log|W| - \tfrac{n}{2}\log(\mathbf{1}_p^\top W\mathbf{1}_p) - v\tfrac{n}{2}\log|I_p - \tfrac{\beta}{4}WQD|. \tag{3.30}$$

Now, n can again be interpreted as an annealing parameter to control the variance of the distribution on a global level. Due to the introduction of v, the translation-invariant matrix T distribution has the property

$$\mathrm{cov}(LX) = \frac{(2LW^{-1}L^\top)\otimes\Omega}{\beta(vn - 2(p-1)+1)}, \tag{3.31}$$

meaning a small (large) value of vn leads to a large (small) variance. It must be stressed that v and β play an important role in inference, because they distribute the probability mass in the space of the prior, which effectively determines the scope of plausible Ψ (Adametz and Roth, 2014).

Interestingly, the matrix T model behaves exactly like TiWnet if v is set to a large value, for then the prior peaks sharply at independent features, that is, $\Omega \equiv I_n$. More details concerning model behavior are given in the experimental section.

For the practical application of our model, it is not meaningful to adjust β and v simultaneously, therefore, we propose to fix v at the smallest possible value. Further, β can be stochastically integrated out, because it depends on the particular scale of D. For this purpose, a standard gamma prior with density

$$f(\beta\,; k, \theta) \propto \theta^{-k} \beta^{k-1} \exp(-\theta^{-1}\beta), \tag{3.32}$$

is one possible choice. Here, $k > 0$ and $\theta > 0$ are shape and scale parameter, respectively.

3.4.2 A Prior for Network Analysis

Contrary to the block-structured parametrization of W in clustering, network analysis requires fine-grained control over each individual element. Moreover, a qualified prior must also be flexible and enforce sparsity. For these reasons, we adopt the construction of (Prabhakaran et al., 2013; Adametz and Roth, 2014), which consists of the two following components[1]:

- $f_1(W)$ places a 3-level uniform prior on each element W_{ij} with levels

  $$\{-1, 0, +1\},$$

  which correspond to negative/zero/positive edge weight, respectively. Also, the diagonal elements are chosen in a way that W maintains positive-definiteness at all times.

  Although the scheme can be criticized for its limitation to three levels, we argue that (i) it proved to be sufficiently flexible in practice and (ii) it can easily be extended to accommodate for more levels, say, $\{-1.0, -0.5, 0.0, +0.5, +1.0\}$. However, note that the range of values is intentional, because the MCMC sampler should not explore scaling due to the model's scale invariance.

- $f_2(W)$ enforces sparsity by penalizing the total number of non-zero W_{ij}, say, N. Possible choices include a *Laplacian* prior

  $$f_2(W\,; \lambda) \propto \exp(-\lambda N), \quad \lambda > 0 \tag{3.33}$$

[1] Alternative choices for a prior include *spike and slab* (Mitchell and Beauchamp, 1988) and *partial correlation* (Daniels and Pourahmadi, 2009).

for keeping the number of edges at a low level, or a gamma prior

$$f_2(W\,;\kappa,\eta) \propto \eta^{-\kappa} N^{\kappa-1} \exp(-\eta^{-1} N), \quad \kappa > 0,\ \eta > 0, \tag{3.34}$$

if it is desirable to cover a specific range of edge numbers, for example between 30 and 50.

3.4.3 Posterior and Algorithm

The likelihood in conjunction with all priors finally leads to the following posterior for network analysis:

$$f(W,\beta \,|\, D,\bullet\,) \propto f(D \,|\, W,\beta) \cdot f_1(W) \cdot f_2(W \,|\, \lambda) \cdot f(\beta \,|\, k,\theta). \tag{3.35}$$

Let us now give some exemplary parameter values for MCMC sampling: For a full-rank $p \times p$ distance matrix D with $p = 100$, we initialize $n = 100$ and fix v at $2 > 1 + (p-2)/n$. During sampling, n is gradually increased until W does not change anymore, say, 1 accepted proposal among the last 100 samples. At this point, n is frozen and we average over the following 1000 samples to receive the final W. This enables us to draw a graph with variable edge widths in spite of only having three levels.

The above model was introduced by Adametz and Roth (2014) as *TiMT*, the *Translation-invariant Matrix T* process. Its corresponding MCMC sampler is explained in detail in Algorithm 3. Note that superscript $*$ refers to a proposal. From a high-level perspective, the approach essentially performs a symmetric random walk in W (Prabhakaran et al., 2013), where every other W can be reached in at most $\frac{1}{2}p(p-1)$ edge flips, that is, the number of elements in the upper or lower triangular submatrix. Therefore, the prior assigns a non-zero weight to every possible graph configuration.

3.4.4 Complexity

One loop in Algorithm 3 comprises p edge flips, where each flip requires the evaluation of the posterior. Thus, a naive implementation would recompute the determinant from scratch in $\mathcal{O}(p^3)$, hence leading to complexity $\mathcal{O}(p^4)$ for the full loop, which is fairly expensive.

Algorithm 3 One loop of the MCMC sampler (Adametz and Roth, 2014)

```
Input: distance matrix $D$, temperature $n$, fixed $v > 1 + (p-2)/n$
for $i = 1$ to $p$ do
    Set $W^* \leftarrow W$
    Uniform randomly select $j \neq i$ and sample $W^*_{ij}$ from $\{-1, 0, +1\}$
    Set $W^*_{ji} \leftarrow W^*_{ij}$ and update $W^*_{ii}$ and $W^*_{jj}$ accordingly
    Compute posterior of $W^*$ in Eq. (3.35)
    if acceptance ratio $> u \sim \mathcal{U}(0,1)$ then
        $W \leftarrow W^*$
    end if
end for
Generate proposal $\beta^*$ and compute its posterior in Eq. (3.35)
if acceptance ratio $> u \sim \mathcal{U}(0,1)$ then
    $\beta \leftarrow \beta^*$
end if
```

The key observation is that a single flip in W only contributes as rank-1 matrix, which gives rise to an efficient update scheme involving the QR decomposition. In more detail, the prior with $k = 3$ levels $\{-1, 0, +1\}$ permits a total of $k(k-1) = 6$ possible flips:

$$W_{ij} \to W^*_{ij}: \qquad \left\{ \begin{array}{lll} -1 \to \;\;\, 0 & 0 \to -1 & +1 \to -1 \\ -1 \to +1 & 0 \to +1 & +1 \to \;\;\, 0 \end{array} \right\} \tag{3.36}$$

If an element W_{ij} is flipped at $(i,j) = (3,1)$, then there are 6 options[2] for

$$(W^* - W) \in \left\{ \begin{bmatrix} -1 & 0 & +1 \\ 0 & 0 & 0 \\ +1 & 0 & -1 \end{bmatrix}, \begin{bmatrix} 0 & 0 & +2 \\ 0 & 0 & 0 \\ +2 & 0 & 0 \end{bmatrix}, \dots \right\}. \tag{3.37}$$

All these matrices can be expressed as either $\boldsymbol{u}\boldsymbol{v}^\top$ or $\boldsymbol{u}\boldsymbol{v}^\top + \boldsymbol{a}\boldsymbol{b}^\top$, such that $|W^*|$ is computed in $\mathcal{O}(p^2)$ for a known QR decomposition of W.

[2] The example uses $p = 3$ for simplicity and without loss of generality.

The second determinant of the likelihood, Eq. (3.30), also decomposes into rank-1 terms, although being marginally more complex, which is why we point to (Adametz and Roth, 2014) for details.

In conclusion, one full loop in TiMT is calculated in only $\mathcal{O}(p^3)$, which is even on par with the complexity of TiWnet (Prabhakaran et al., 2013). This is a surprising result, given that the model based on the matrix T distribution is much more flexible than its Gaussian predecessor.

3.5 Experiments

> Remark
>
> This section largely follows the results by Adametz and Roth (2014), but adds more details and background information.

The derivation of the model faced a number of challenges, which were overcome by suitable invariances. What remains to be seen is how it performs in practice, in particular with and without known feature correlation.

3.5.1 Synthetic Data

Independent Features

In the first experimental setup, the goal is to confirm that TiMT—as a generalization of TiWnet—is also applicable to the standard case of fully independent features. To do this, we generate data in the following way: matrix X is sampled from a matrix normal distribution with $p = 30$, $n = 300$, $M = 0_{p \times n}$ and $\Psi = I_n$. Further, the true matrix W consists of two parts: (i) a sparse structure and (ii) real-valued weights. We obtain a challenging structure by element-wise sampling from a binomial with 1 trial and a Pareto-distributed probability of success, which results in a upper-triangular matrix of 0s and 1s. The sign of the 1 elements is then determined by a uniform distribution, that is, $u \sim \mathcal{U}(0, 1)$, where $u \geq 0.5$ yields -1. Next, the weights of the non-zero elements are independently sampled from a gamma distribution with shape and scale 2.0. As a final step, we mirror the upper to the lower triangular matrix and adjust the diagonal entries as the sum of their row plus $\epsilon = 0.5$ to ensure positive definiteness. This concludes all parameters needed to generate a single matrix X and its corresponding Euclidean distance matrix D with $D_{ij} = (X_{i\bullet} - X_{j\bullet})(X_{i\bullet} - X_{j\bullet})^\top$. Fig. 3.7 depicts an example network and the reconstruction by TiMT and TiWnet under identical parameters. Here, the MCMC sampler ran for 20 000 loops each.

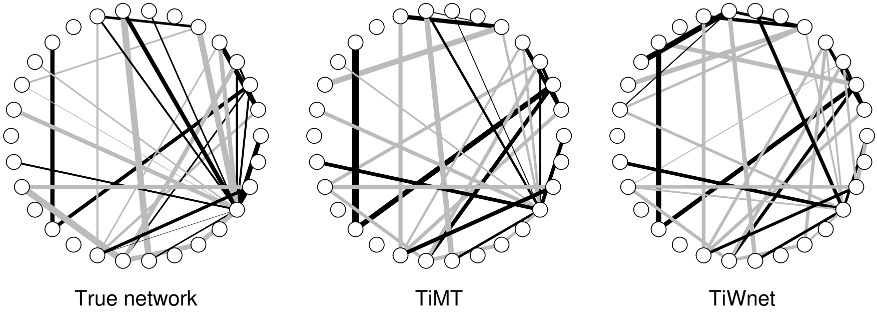


Figure 3.7: An example for synthetic data with independent features. The true network (left) has a challenging structure, but it can be recovered fairly well by both TiMT (center) and TiWnet (right). Black/gray edges refer to positive/negative sign.

For a comprehensive benchmark, the above procedure is repeated to produce a set of 100 distance matrices. Since the ground truth is known, we can calculate the accuracy of the inferred networks via F-score[3] and false positive rate, leading to an overall performance as shown by the boxplots in Fig. 3.8. The list of contestants also includes TRCM and gL, which operate on X, however. Note that a distance matrix implies unknown column means and therefore, X is translated by $M = \mathbf{1}_p \boldsymbol{w}^\top$, where each element in vector $\boldsymbol{w}$ is an independent draw from a gamma distribution with shape 0.5 and scale 10^5. As expected, the accuracy of TRCM and gL suffers heavily, because

[3] Here, we only discriminate between positive, negative and zero elements. The individual values themselves are discarded.

column means conflict with their model assumptions. This is why we also analyze TRCM in conjunction with the original zero-mean X, thus TRCM.*u* (*unshifted*). The resulting modification yields a comparable performance to the distance-based approaches.

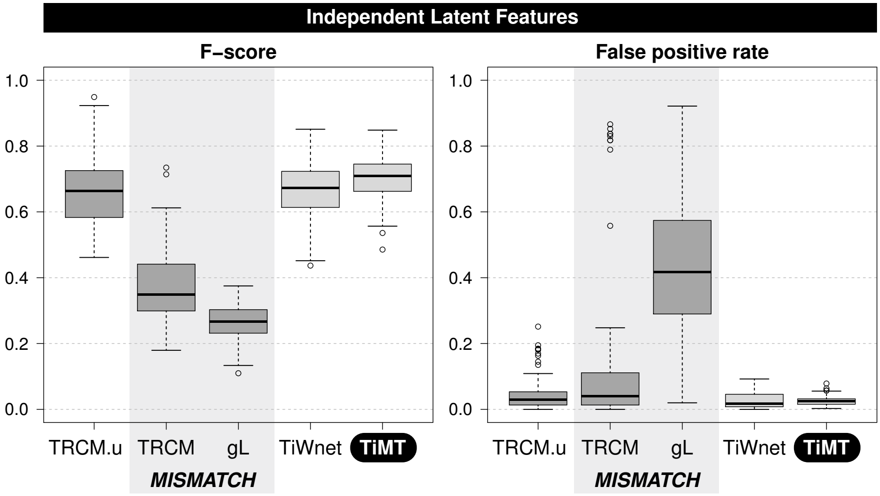


Figure 3.8: Accumulated accuracy over 100 synthetic datasets as measured by F-score and false positive rate. TRCM.*u* (using X with $M = 0_{p \times n}$), TiWnet and TiMT perform almost on par, however, TRCM and gL (both using X with $M = \mathbf{1}_p \boldsymbol{w}^\top$) fall behind, because column means do not match their model assumption.

All methods in this experiment require a sparsity parameter which determines the number of edges in the final graph and therefore has a crucial impact on the accuracy. In detail, TiMT and TiWnet use a single *fixed* value for all 100 networks, while TRCM, TRCM.u and gL were *individually optimized* for each network. This is clearly in favor of the competing methods.

Correlated Features

For the second experiment, we adopt the previous setup with $p = 30$ and $n = 300$, but introduce feature correlation Ψ that has a strong and visible impact on the data. One way of generating such a matrix is the following: We sample a $n \times 5n$ matrix $G \sim \mathcal{N}_{n,5n}(0_{n\times 5n}, I_n \otimes I_{5n})$ and add vector $\boldsymbol{a} \in \mathbb{R}^{5n}$ with gamma-distributed elements to randomly selected rows in G. This gives a full-rank feature correlation matrix $\Psi \equiv GG^\top/(5n)$. Finally, we repeat this step to arrive at a total number of 100 different matrices Ψ, each leading to one matrix X and its associated Euclidean distance D.

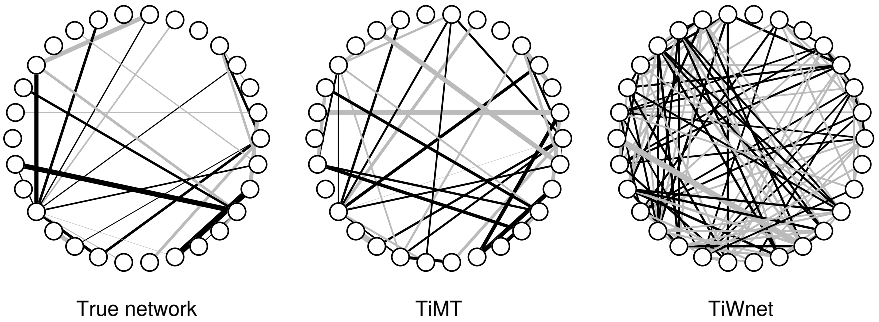


Figure 3.9: An example for synthetic data with correlated features. Black/ gray edges refer to positive/negative sign. TiMT is relatively close to the true network, but TiWnet compensates the model mismatch with additional, yet unnecessary structure. As a result, the accuracy of TiWnet suffers heavily.

Fig. 3.9 picks one example to demonstrate the behavior of TiMT and TiWnet. Note that the addition of feature correlation dramatically increases the difficulty for inference, hence, the accuracy will drop for all methods, yet, even under these conditions, TiMT recovers a network that is quite close to ground truth. This is an interesting result given that TiWnet completely overestimates the network structure. After all, it is forced to explain every observation solely by W, while TiMT can shift conflicting aspects into the prior. By looking at the overly dense topology obtained from TiWnet, a critical observer might try to tune the sparsity parameter in order to achieve a more appropriate number of edges. This, however, will not succeed, because it would remove true *and* false positive edges alike; W and Ψ are inherently “combined” in D and therefore, we cannot hope to recover the true W if the model already decided upon a separation *for us* (unless if $\Psi \equiv I_n$).

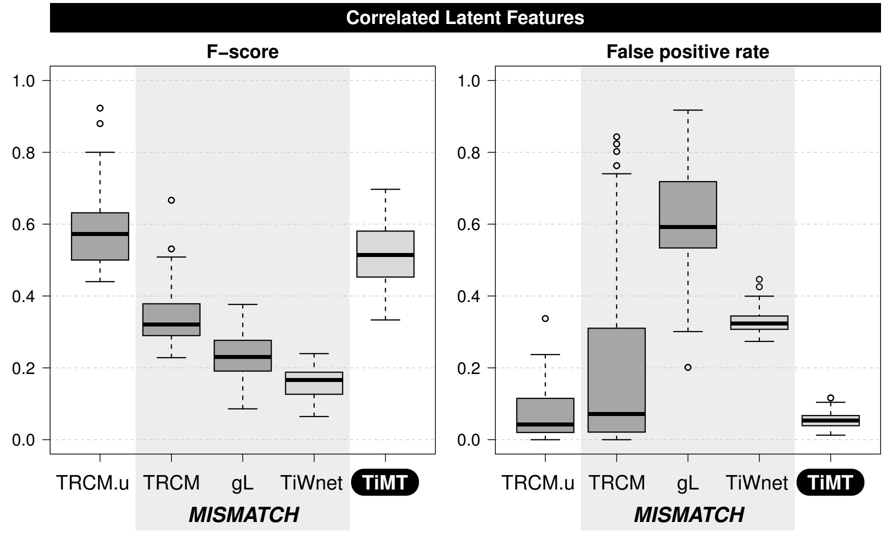


Figure 3.10: Boxplots for F-scores and false positive rates after analyzing 100 datasets with correlated features. Among the introduced methods, only the models of TRCM.u (using zero-mean X) and TiMT (using D) satisfy all requirements.

In this respect, Fig. 3.10 paints a similar picture for all contestants: the median F-score deteriorates on a global scale and every model mismatch adds another penalty, especially for TiWnet which has the lowest F-score, but also a very consistent false positive rate. Further, the model assumptions of gL are violated twice, thereby leading to the highest false positive rate and the second lowest median F-score. Given the difficulty of this task, TiMT performs remarkably well and is almost on par with TRCM.u in terms of F-score. Notice that this experiment is in strong favor of TRCM.u, because it operates on the latent, zero-mean 30×300 data matrix X with an optimized sparsity parameter, whereas TiMT arrives at the same conclusion from only a symmetric 30×30 distance matrix D without *any* knowledge about n. In that regard, TRCM.u serves as an upper bound of what could be achieved hypothetically under optimal conditions—the small remaining gap is perhaps the most impressive demonstration of invariances in the scope of the thesis.

A Note on the Ratio p/n

Both synthetic experiments used $p = 30$ and $n = 300$ to generate data, but is it necessary to repeat them for different ratios? In short, the answer is “no”; for a thorough explanation, notice that n controls the variance of the matrix normal distribution, such that small n allow samples D to differ completely from each other, which in turn renders correct inference virtually impossible. Analogously, samples D generated with large n show almost no variation and, hence, the problem becomes too easy. With this in mind, the ratio $p/n = 1/10$ was chosen as a middle ground for challenging datasets (median F-score of 0.5) while making differences between methods visible.

3.5.2 Real-World Data

Given that TiMT successfully recovers the underlying networks in the synthetic examples, we can now advance to practical domains, where a distance or kernel matrix is the only available input. On the one hand, this precludes TRCM and gL, but on the other hand it implies that there is no ground truth to compare the network against. Hence, we resort to expert knowledge in the form of side information to check individual conditional independences.

A Network of Cancer Drugs

In the first application, we are interested in finding a network of all known cancer drugs based on the similarity between chemical structures. For this purpose, we obtain a list of $p = 84$ non-experimental cancer drugs from the publicly available CancerDR[4] (Cancer Drug Resistance) database and look up their chemical structures on the NCBI pubchem website[5]. Based on the atomic coordinates in *SMILES* notation (Simplified Molecular-Input

[4] http://crdd.osdd.net/raghava/cancerdr/

[5] http://pubchem.ncbi.nlm.nih.gov

Line-Entry System), similarity matrix S is constructed by the sum over three types of graph kernels: *marginalized* Kashima et al. (2003, 2004), *lambda-K* and *pharmacophore* Mahé et al. (2006). The idea behind this is to arrive at a comprehensive representation of the compounds. In summary, matrix S is of size 84×84, from which we extract the corresponding distance matrix $D = \operatorname{diag}(S)\mathbf{1}_p^\top + \mathbf{1}_p \operatorname{diag}(S)^\top - 2S$.

Both TiMT and TiWnet use identical parameters, but produce drastically different network topologies after 15 000 iterations of the MCMC sampler, see Fig. 3.11. Note that the node size either stands for "in clinical trial" (small) or "approved" (large). In particular, the network inferred by TiMT shows some interesting properties: One of the highly-connected hubs is *Rapamycin*—a key element for cell signaling, which binds to the *MTOR* protein (Mammalian Target of Rapamycin) (Wullschleger et al., 2006). In turn, this protein is responsible for a cascade of signal-transduction pathways among which the MTOR pathway is known to be strongly deregulated in cancer. The other hub node, *Temsirolimus*, is highly related because of its

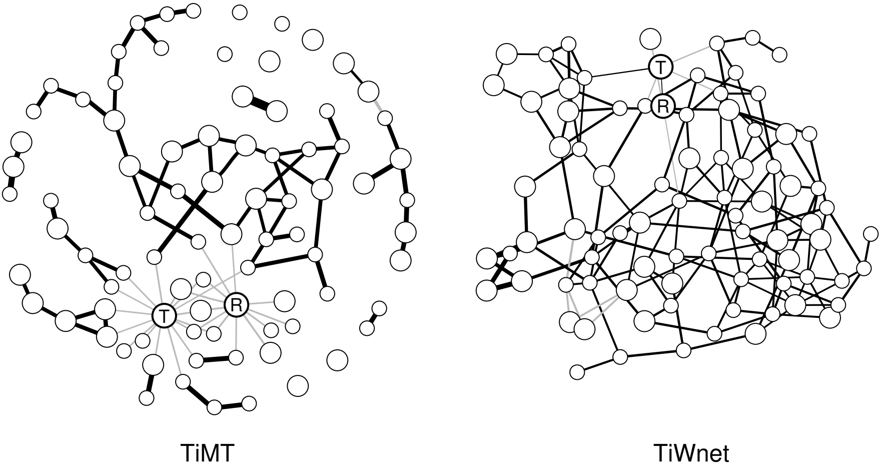


Figure 3.11: A network of $p = 84$ chemical compounds as inferred by TiMT and TiWnet with identical parameters. Due to their importance in cancer medication, *Rapamycin* (R) and *Temsirolimus* (T) are labeled in each graph.

inhibitive function regarding the MTOR protein. It appears that the graph is able to capture these deep relationships.

As we saw in the synthetic experiments, TiWnet has a tendency to overestimate the structure, because it explains any observation solely by W. Intuitively, this issue also seems to apply to the current situation, but without knowledge of X we lack the foundation to prefer either result. Still, we can explore subsets of each network and check for plausibility via experiments under artificial conditions.

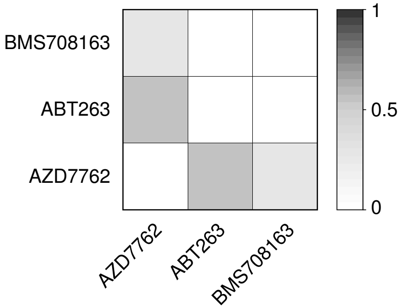


Figure 3.12: Inverse covariance of chemical compounds using cell lines.

One descriptive example is given in Fig. 3.12, where cell lines are evaluated in terms of the inverse covariance between three compounds. To arrive at this result, we extracted all available information from the COSMIC database[6] for the compounds *BMS708163*, *ABT263* and *AZD7762*. Since these data have missing values and their empirical margins are non-Gaussian, we use a Gaussian copula model to estimate the inverse covariance matrix, together with a Bayesian inference method that is capable of dealing with missing values (see the next chapter for details). According to the cell line experiments, BMS708163 and ABT263 are conditionally independent given AZD7762, and therefore, this property should also be reflected in the networks of Fig. 3.11. Thus, we remove AZD7762 from both networks and check for remaining paths between ABT263 and BMS708163, which is illustrated in Fig. 3.13.

[6] http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/

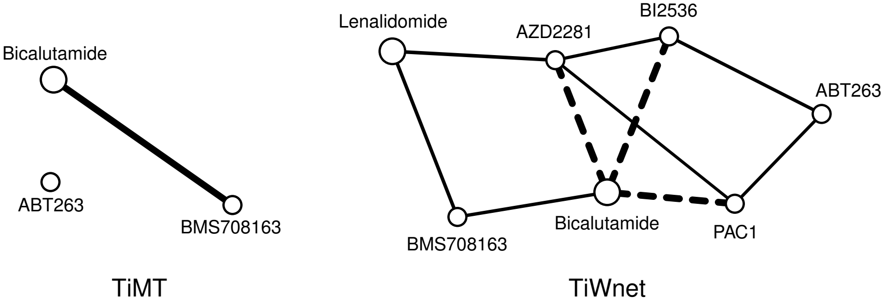


Figure 3.13: Subgraphs from networks in Fig. 3.11 after removing compound AZD7762. Dashes represent connections across several other nodes.

While the network obtained from TiMT correctly supports conditional independence between BMS708163 and ABT263 (no remaining path between them after removing AZD7762), the network found by TiWnet produces a more complex and fairly interconnected graph. If we wanted to arrive at the correct independence pattern, we would have to block at least two additional nodes, for example, *BI2536* and *PAC1*. The above subnetwork is only one example among many, where the reconstruction of TiMT agrees with cell line experiments.

A Network of Biological Pathways

The second application revisits the introductory example of Section 3.1.1 as a means to derive a network of biological pathways in cancer patients. We start with the publicly available dataset of Sheffer et al. (2009) containing the expression values for 13 437 genes across 182 colon cancer patients (category *primary tumor*). Our motivation is the departure from single-gene analysis, which (i) heavily suffers from the imbalance between unknowns (genes) and measurements (patients) and (ii) is very prone to noise. Thus, we turn to the *KEGG* database[7] (*Kyoto Encyclopedia of Genes and Genomes*) and extract information for all $p = 276$ pathways, each defined as a specific

[7] http://www.genome.jp/kegg/

collection of genes. Using this, the gene expression matrix decomposes into 276 submatrices, where overlaps are explicitly allowed due to the pivotal role of many genes. Every submatrix has a characteristic distribution of genes, thus we calculate the mean and variance for each patient to arrive at 182-dimensional mean vector $\boldsymbol{m}$ and 182×182 covariance matrix C for each of the pathways. In other words, the data we base inference on has the form

$$\{(\boldsymbol{m}_1, C_1),\ (\boldsymbol{m}_2, C_2),\ \ldots,\ (\boldsymbol{m}_{276}, C_{276})\}. \tag{3.38}$$

As stated in the introduction, it is straight-forward to work with vectorial objects, but currently we have a set of distributions, whose integration into a GGM is non-trivial. To this end, the Bhattacharyya distance (Bhattacharyya, 1943; Jebara and Kondor, 2003) is a pairwise distance measure for distributions, which finally gives rise to a 276×276 matrix D.

Due to the nature of a distance matrix, there is no access to vectorial features, but it is clear this information resides in mean and covariance of the patients. Consequently, we expect that the *assumed* correlation of patients influences the distances to some degree, simply because all patients belong to the *primary tumor* category, which exhibit similar symptoms, receive similar treatment and possibly share many clinical properties. Therefore, we should not treat them as independent realizations, intuitively speaking.

Using the Bhattacharyya distance matrix, we run TiMT and TiWnet with identical parameters for 20 000 iterations of the MCMC sampler each. Since this is a moderately large dataset for GGMs, we measure the runtime separately and arrive at the following duration: TiWnet finishes after $3{:}00$ hours, TiMT requires $3{:}10$ hours, and a naive $\mathcal{O}(p^4)$ implementation of TiMT takes around 20 hours to complete. TiWnet has the same cubic complexity as TiMT, but does not use hyperparameter β, thus it is slightly faster.

The final results are reported in Fig. 3.14, which again shows a sparse network for TiMT and a densely connected graph for TiWnet. In spite of the lack of ground truth, this is a clear indicator for latent feature correlation, since both methods have the same Gaussian foundation other than the invariance against Ψ. As alluded in the synthetic experiments, the assumption of independent features in TiWnet can not be compensated by inducing more sparsity; this would cancel both true and false positive edges.

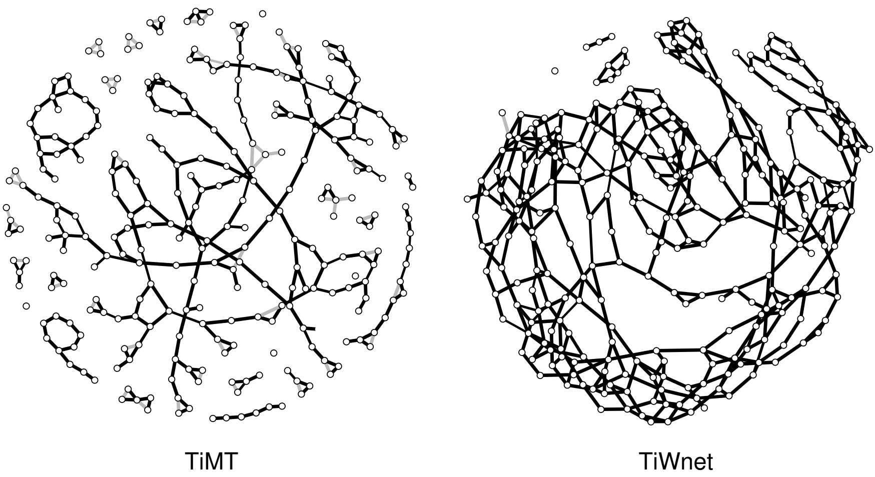


Figure 3.14: A network of $p = 276$ pathways in colon cancer.

Although TiMT produced a sparser graph, there is no guarantee that it actually captures meaningful conditional independences, hence, we consult the *BioGRID* database[8] for protein-protein interactions and compute the functional overlap between pathways relative to the union of their parts. This information is completely detached from the colon cancer dataset, as it relies only on "raw" pathway definitions and aggregated expert knowledge.

Given that the medical study targets colon cancer, we analyze three pathways which indicate elevated susceptibility (Peltomäki, 2001; Fortini et al., 2003): *base excision repair* (token: 96), *mismatch repair* (token: 98) and *cell cycle* (token: 114). Extracting the corresponding subnetworks of these pathways including all neighboring pathways in range of two edges, we arrive at Fig. 3.15. Interestingly, the protein-protein interactions hint at a particular structure, where pathway 98 links pathway 96 and 114, but 96 and 114 share no common basis. This information is in exact support of the subgraph by TiMT, which places 96 and 114 on diverging branches. In contrast, TiWnet infers a densely connected structure.

All in all, we do not know whether a GGM is truly a good fit for this

[8] http://thebiogrid.org, version 3.2

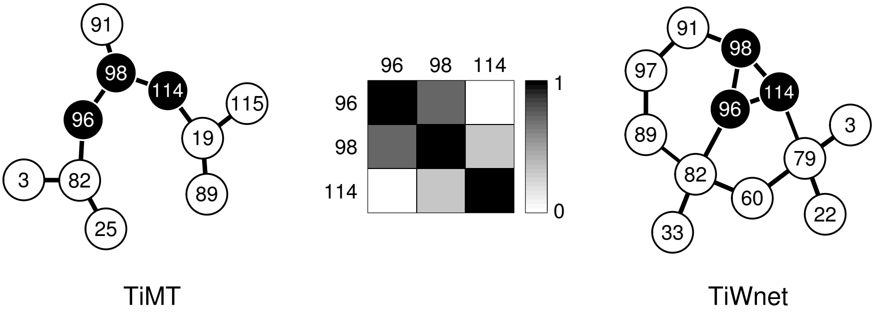


Figure 3.15: Conditional independences in detail. From both networks in Fig. 3.14, we extract a subgraph of three pathways (solid black nodes) including all neighbors in reach of two edges (white nodes). The matrix in the center shows external information on pathway similarity based on their relative number of protein-protein interactions from the BioGRID database.

domain, although there are numerous subnetworks with confirmed independence patterns. Still, this type of analysis was only made possible by distance-based approaches and, in its current state, it represents an interesting step towards a more thorough understanding of pathways. Note that pathway interactions constitute a relatively young field, which is typically analyzed in terms of static gene/protein set intersections and raw count information, meaning it primarily caters to data mining aspects. A case-driven analysis, as shown for this colon cancer dataset, is unexplored for the most part. After all, our inferred network does not explain pathway interactions under regular conditions, but rather what happens in colon cancer *specifically*.

3.6 Conclusion

In this chapter, we presented how to incorporate the framework of Gaussian Graphical Models (GGMs) into the scope of distance-based methods. As we learned, distances not only obscure translational information, but also the underlying feature space, including the correlation of features. This poses a serious challenge for any statistical model, because row precision matrix W (the parameter of interest) and Ψ (the nuisance) jointly impact the perceived distances D and the task is to recover their correct separation.

We may, of course, simplify the problem by assuming strict feature independence, however, this consequently forces the model to explain the data solely by adding structure, which can produce misleading results, as shown in the synthetic experiments. In fact, in many applications it is plausible to assume feature correlation, yet,—assuming it can be detected—we are unable to remove it without access to the underlying features. Due to this reason, TiMT expects a non-trivial Ψ, but never tries to reconstruct it explicitly. By the definition of the integrated likelihood, we account for *every possible* Ψ and transform its prior belief into the posterior using the likelihood. In contrast, the profile likelihood makes an attempt in finding an explicit reconstruction at the price of requiring $n \ll p$, which disqualifies its application to the large majority of distance matrices.

The combination of integrated and marginal likelihood for the invariance against feature parameters may appear arbitrary at first, however, note that translation affects merely one direction in an p-dimensional space (illustrated by $(p-1) \times p$ projection L), while feature correlation concerns a much larger scope with potentially $n \gg p$, which is more suitably expressed by a Bayesian prior. Without question, it is technically possible to formulate a likelihood that is invariant against any right-multiplication of X (i.e., the profile likelihood approach), as for example $X\Psi^{\frac{1}{2}}$. However, the subsequent information loss significantly impairs the statistical power of the model—especially in the domain of distances. This goes to show that the Bayesian relaxation of strict feature independence (Adametz and Roth, 2014) is a superior, but also much more careful approach. After all, we are finally able to use all latent parameters of the matrix normal while only requiring pairwise distances on input.

On the computational side, it was shown that the algorithm of TiMT can be elegantly reformulated in terms of $\mathcal{O}(p^3)$, which brings it on par with TiWnet despite the greatly increased flexibility. Most importantly, TiMT does not suffer from the additional degree of freedom as we confirmed experimentally: if the true features are independent, TiMT and TiWnet recover the same conditional independences. Moreover, when the hyperparameters of TiMT are chosen in a way that the underlying prior assigns all its mass to feature independence, both methods behave identically. Therefore, TiMT fulfills all requirements of a proper generalization.

Remark

Additional thoughts about the model are given in Appendix B.

Chapter 4

A Gaussian Copula Model for Mixed Data

The third and final pillar of the thesis describes a model for estimating the dependence structure of multivariate distributions, in particular *meta-Gaussian* distributions. The approach due to Hoff (2007) is complementary to the distance-based methods, but can be used for an interesting generalization of the information bottleneck (Tishby et al., 1999; Chechik et al., 2007; Rey and Roth, 2012). Hereby, it is possible to find an information-theoretic compression of continuous *and* discrete random variables (rvs) with possibly missing values—all while maintaining the benefits of the Gaussian foundation.

4.1 Introduction

The entry point of this chapter is the univariate normal distribution,

$$X \sim \mathcal{N}(\mu_X, \sigma_X^2), \tag{4.1}$$

where X is a rv with mean μ_X, variance σ_X^2 and a realization x. Further, X has density f_X and *(cumulative) distribution function*

$$F_X(x) \equiv \int_{-\infty}^{x} f_X(t)\mathrm{d}t, \tag{4.2}$$

which is non-decreasing and satisfies $F_X(-\infty) = 0$ and $F_X(\infty) = 1$ (Nelsen, 2007, p. 17). A fundamental property is given by the *probability integral transform* (Genest and Rivest (2001),Davison (2008, p. 39)): If

X follows a continuous distribution with distribution function F_X, then

$$U \equiv F_X(X) \sim \mathcal{U}(0,1), \tag{4.3}$$

which is the uniform distribution on interval $[0,1]$. Fig. 4.1 provides a graphical interpretation. Note that this relationship is also a constructive way of generating samples from F_X (Nelsen, 2007, p. 40f): first, draw from a uniform distribution to receive u, then transform it via $x = F_X^{-1}(u)$; the inverse F_X^{-1} is also known as *quantile function*. If F_X is strictly increasing, the transformation is bijective, meaning x and u have a one-to-one correspondence.

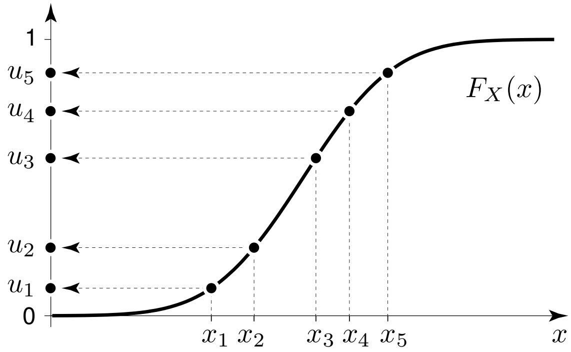


Figure 4.1: The continuous distribution function $F_X(x)$ transforms realizations x_i onto the uniform interval $[0,1]$.

4.2 Dependence and Correlation

In the case of multiple normal rvs, say, X and Y, we are interested in their dependence, because it tells us, for example, that a large x implies a large y. If the rvs are jointly Gaussian, they follow a multivariate normal distribution, denoted by

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathcal{N}_2\left(\boldsymbol{\mu} = \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \, \Sigma = \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{bmatrix}\right) \tag{4.4}$$

with mean vector $\boldsymbol{\mu}$, covariance matrix Σ and distribution function F_{XY}. The covariance matrix, in particular, determines the dependence between rvs. It is composed of

$$\Sigma = \begin{bmatrix} \operatorname{cov}(X,X) & \operatorname{cov}(X,Y) \\ \operatorname{cov}(Y,X) & \operatorname{cov}(Y,Y) \end{bmatrix}, \tag{4.5}$$

which can be rewritten using the identities $\operatorname{cov}(X,X) = \operatorname{var}(X) \equiv \sigma_X^2$ and $\operatorname{cov}(X,Y) = \operatorname{cov}(Y,X) = \rho\,\sigma_X\sigma_Y$, such that

$$\Sigma = \begin{bmatrix} \sigma_X^2 & \rho\,\sigma_X\sigma_Y \\ \rho\,\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix}. \tag{4.6}$$

Further, by the definition of the *linear correlation* (Pearson, 1895),

$$\operatorname{corr}(X,Y) \equiv \frac{\operatorname{cov}(X,Y)}{\sigma_X\sigma_Y}, \tag{4.7}$$

we receive the corresponding *correlation matrix* as

$$R = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}. \tag{4.8}$$

In matrix notation, the connection is found by

$$R = \operatorname{diag}(\Sigma)^{-\frac{1}{2}}\,\Sigma\,\operatorname{diag}(\Sigma)^{-\frac{1}{2}}, \tag{4.9}$$

hence, due to affine transformations of the multivariate normal, we have

$$\begin{bmatrix} \sigma_X^{-1} & 0 \\ 0 & \sigma_Y^{-1} \end{bmatrix} \left(\begin{bmatrix} X \\ Y \end{bmatrix} - \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix} \right) \equiv \begin{bmatrix} \bar{X} \\ \bar{Y} \end{bmatrix} \sim \mathcal{N}_2(\mathbf{0}_2, R), \tag{4.10}$$

where $\bar{X}$ and $\bar{Y}$ are *standard normal rvs*, each with a marginal distribution function Φ of mean zero and unit variance.

For $\rho \equiv 0$ (that is, $R \equiv I_2$), we say that X and Y are *uncorrelated*, which, in the Gaussian case, is equivalent to being independent (Schmidt, 2007). As a result, the joint distribution function factors into the product of univariate marginals:

$$F_{XY}(x, y) = F_X(x) \cdot F_Y(y). \tag{4.11}$$

Note that for general distributions, uncorrelated rvs are still dependent.

4.3 Copula

The rationale behind introducing correlation is to find the most "efficient" description of dependence between X and Y. To this end, Sklar (1959) showed that every multivariate distribution decomposes into two distinct parts: the marginals, which are specific to each rv, and the *copula* (Nelsen (2007, p. 10ff); Schmidt (2007); Aas (2004)), which captures their dependence. Therefore, we write

$$F_{XY}(x, y) = C(F_X(x), F_Y(y)), \tag{4.12}$$

where copula C is a mapping of the unit square to $[0, 1]$, which is exemplified by Fig. 4.2. An important property of Eq. (4.12) is that if F_X and F_Y are continuous, C is unique; else it is only unique on the range of F_X and F_Y (Schmidt, 2007). The above decomposition can also be formulated in terms

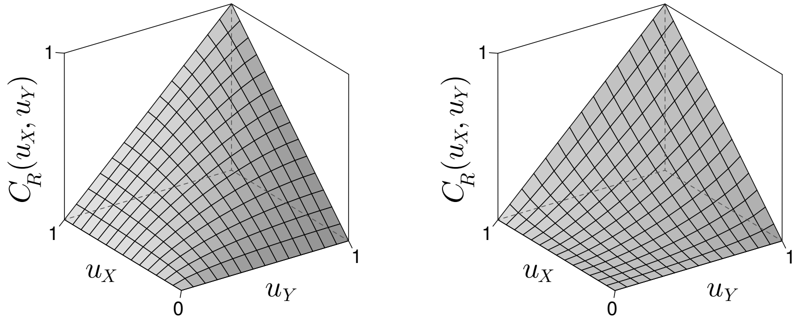


Figure 4.2: Gaussian copula for $\rho = +0.5$ (left) and $\rho = -0.5$ (right).

of densities, that is

$$f_{XY}(x,y) = c(F_X(x), F_Y(y)) \cdot f_X(x) \cdot f_Y(y), \tag{4.13}$$

where c refers to the *copula density*

$$c(u_X, u_Y) \equiv \frac{\partial^2}{\partial u_X \partial u_Y} C(u_X, u_Y) \tag{4.14}$$

with $(u_X, u_Y) \in [0,1]^2$.

At this point, we have already implicitly used the Gaussian copula C_R, which is inherent to every multivariate normal distribution and fully defined by correlation matrix R (as denoted by the subscript). In many applications acting on Gaussian data, the covariance matrix is the object of interest, say, for the Gaussian graphical model based on distance matrices, yet, what we actually seek can often be reduced to correlation matrix R, which is devoid of scaling of the marginals, that is, the nuisance.

To estimate R in the Gaussian case, we can do the following: Assume there are n realizations $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ from

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \sigma_X^2 & \rho\,\sigma_X\sigma_Y \\ \rho\,\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix} \right), \tag{4.15}$$

where the marginals F_X and F_Y are known. Then, it is possible to apply the probability integral transform twice to receive

$$\bar{X} \equiv \Phi^{-1}(F_X(X)) \quad \text{and} \quad \bar{Y} \equiv \Phi^{-1}(F_Y(Y)), \tag{4.16}$$

where $\bar{X}$ and $\bar{Y}$ are standard normal rvs with marginal distribution function Φ. This means, for any realization x_i, we can calculate the associated *normal score* $\bar{x}_i = \Phi^{-1}(F_X(x_i))$, which applies analogously to y_i and $\bar{y}_i$. A visual interpretation is given in Fig. 4.3. Finally, the correlation matrix is estimated by

$$\widehat{R} = \begin{bmatrix} 1 & \widehat{\rho} \\ \widehat{\rho} & 1 \end{bmatrix} \qquad \text{with} \qquad \widehat{\rho} = \frac{1}{n-1} \sum_{i=1}^{n} \bar{x}_i \bar{y}_i. \tag{4.17}$$

Although a mapping from one Gaussian rv to another is not meaningful in practice[1], the key observation is the removal of nuisance parameters mean and scale, which are both attributes of the marginal. Hereby, we can treat the observations *as if they were generated from a standard multivariate normal distribution*, which enables us to estimate the correlation matrix as above.

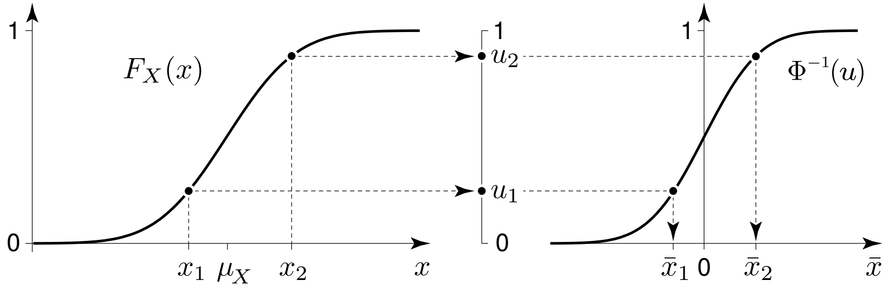


Figure 4.3: Calculating the normal score $\bar{x}_i$ of a realization x_i involves the distribution function $F_X(x)$ and the quantile function of the standard normal distribution, $\Phi^{-1}(u)$.

Given that Sklar's theorem decouples marginals from dependence structure, the next step is to replace the Gaussian marginals with other distribution functions. This leads to the class of *meta-Gaussian* distributions (Rey and Roth, 2012), which all have the Gaussian copula in common. Note that estimating the correlation matrix of a meta-Gaussian distribution is identical to Eq. (4.17), since it only depends on the normal scores.

In case the distribution functions F_X and F_Y are unknown, we can instead use the *empirical marginals* (Rey and Roth, 2012),

$$\widehat{F}_X(x) \equiv \frac{\operatorname{rank}(x)}{n+1}, \tag{4.18}$$

which evaluates the *rank* of x among n realizations to define a step function between $1/(n+1)$ and $n/(n+1)$. Fig. 4.4 illustrates the construction of $\widehat{F}_X$ using the example of actual gene expression values, where every observation

[1] Instead of the probability integral transform, we could have also applied linear transformations to arrive at the same result.

contributes as a step of height $1/(n+1)$. With an increasing number of expression values, $\widehat{F}_X$ approaches the true underlying distribution function. The best result is achieved for all $n = 313$ measurements (right plot).

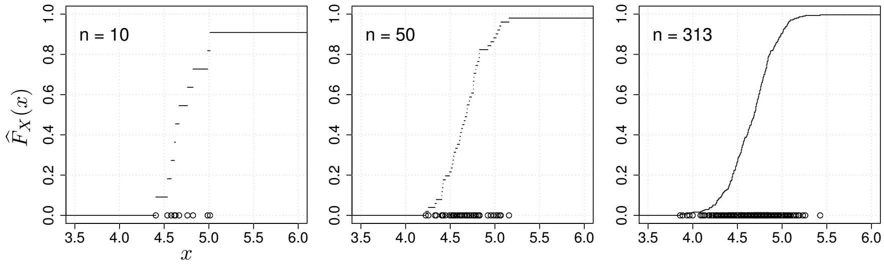


Figure 4.4: Empirical distribution function for expression values of gene *TNF* (*Tumor Necrosis Factor*), data due to (Sheffer et al., 2009).

The following example further helps to understand the notion.

Example

Assume we have $n = 6$ observations with unknown F_X,

$$\boldsymbol{x} = \{1.05, 0.60, 0.21, 1.01, 0.96, 1.51\},$$

and corresponding ranks

$$\text{rank}(\boldsymbol{x}) = \{3, 2, 5, 4, 1, 6\}.$$

Then the empirical distribution function in Eq. (4.18) yields

$$\widehat{F}_X(\boldsymbol{x}) = \{0.43, 0.29, 0.71, 0.57, 0.14, 0.86\},$$

leading to the normal scores

$$\bar{\boldsymbol{x}} = \Phi^{-1}(\widehat{F}_X(\boldsymbol{x})) = \{-0.18, -0.57, 0.57, 0.18, -1.07, 1.07\}.$$

Note that the normal scores do not change under scaling, translation or any transformation that leaves the ranks intact, hence, this describes a very general class of invariances.

Suppose rvs X and Y follow a joint meta-Gaussian distribution, where F_X and F_Y are unknown. Computing the normal scores using their empirical marginals

$$\bar{x}_i = \Phi^{-1}\left(\frac{\operatorname{rank}(x_i)}{n+1}\right) \quad \text{and} \quad \bar{y}_i = \Phi^{-1}\left(\frac{\operatorname{rank}(y_i)}{n+1}\right) \tag{4.19}$$

gives rise to the *Gaussian rank correlation* (Boudt et al., 2012)

$$\widehat{R}_G = \begin{bmatrix} 1 & \widehat{\rho} \\ \widehat{\rho} & 1 \end{bmatrix} \quad \text{with} \quad \widehat{\rho} = \sum_{i=1}^{n} \frac{\Phi^{-1}\left(\frac{\operatorname{rank}(x_i)}{n+1}\right) \Phi^{-1}\left(\frac{\operatorname{rank}(y_i)}{n+1}\right)}{\Phi^{-1}\left(\frac{i}{n+1}\right)^2}. \tag{4.20}$$

Notice how the observations are evaluated only in terms of their ranks; the marginals F_X and F_Y hereby become irrelevant. An important property of rank-based estimators is robustness against outliers, because extreme values either have rank 1 or n and therefore their contribution does not skew the results as much as their value would. This distinguishes rank correlation from linear correlation (Schmidt, 2007).

4.4 Discrete Random Variables and Missing Values

As we learned in the previous section, calculating the normal scores enables us to robustly estimate the underlying correlation matrix, which in turn fully defines the dependence structure of a meta-Gaussian distribution. The setup, however, is only valid for continuous rvs, because their distribution functions are non-decreasing. By contrast, discrete rvs have *ties* among realizations, such that also their corresponding normal scores collapse into a finite number of levels. As a result, the mapping

$$\bar{x}_i = \Phi^{-1}(F_X(x_i)) \tag{4.21}$$

does *not* produce unique normal scores anymore (Fig. 4.5), but the converse

$$x_i = F_X^{-1}(\Phi(\bar{x}_i)) \tag{4.22}$$

still holds (Fig. 4.6). Here, F_X^{-1} is the *generalized quantile function*.

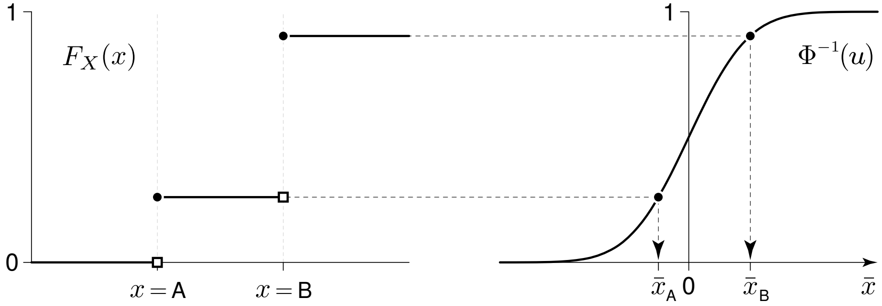


Figure 4.5: A discrete rv does not have unique realizations, which leads to ties among normal scores.

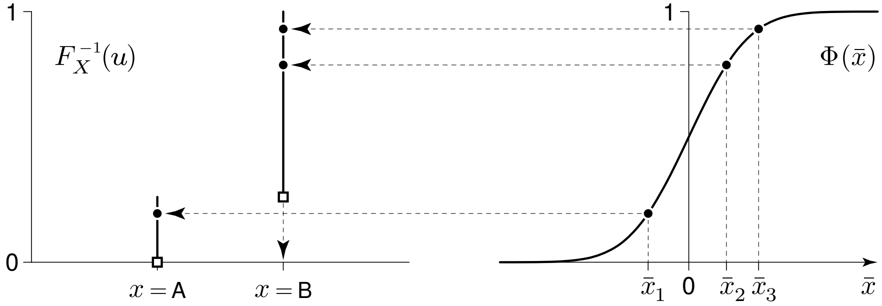


Figure 4.6: Continuous $\bar{x}_i$ can easily be mapped to their discrete counterparts using the generalized quantile function (pictured left).

The above implies a loss of information due to the discrete distribution function, which prevents us from estimating the underlying correlation matrix R. Unfortunately, there are many applications using discrete rvs, for example, a medical study may involve age, sex or the cancer stage of a patient. In the current situation, we are not able to estimate R if at least one rv is discrete. To this end, Hoff (2007) defined the *extended rank likelihood*, which assumes a unique normal score for every observation, but without modeling any distribution function. Instead, the approach relies on the fundamental fact that if two observed values have the relation

$$x_1 < x_2, \tag{4.23}$$

their associated normal scores must comply:

$$\bar{x}_1 < \bar{x}_2. \tag{4.24}$$

We can easily confirm this property for a continuous rv in Fig. 4.3 and a discrete rv in Fig. 4.5.

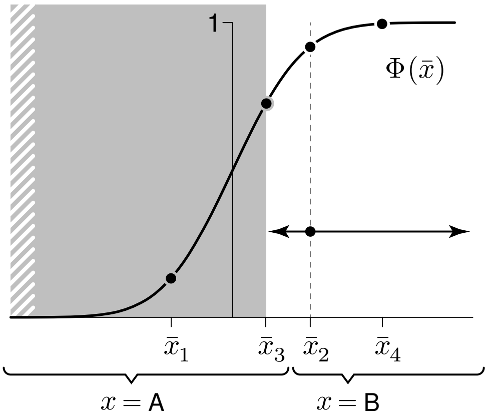


Figure 4.7: Normal scores must only obey to the (incomplete) ordering constraints given by their discrete observations.

As a consequence, a normal score $\bar{x}_i$ is valid if it obeys to all ordering constraints which are defined by their observed values x. Fig. 4.7 gives a graphical representation of the constraints for the binary rv of Fig. 4.5. In this example, $\bar{x}_1$ and $\bar{x}_3$ are associated with level A, while $\bar{x}_2$ and $\bar{x}_4$ correspond to level B. Here, $\bar{x}_2$ must satisfy $\bar{x}_2 > \max(\bar{x}_1, \bar{x}_3)$, which excludes the highlighted region in gray (it continues to minus infinity). If discrete realization x_2 was missing, its normal score $\bar{x}_2$ would not be constrained.

In the case of k levels, normal scores may be confined from left *and* right. Note that the ordering constraints also generalize to continuous data, which can be seen as an n-level discrete variable without ties. In this case, a normal score $\bar{x}_i$ is limited to one of n intervals; if $n \to \infty$, it becomes equivalent to the mapping $\bar{x}_i = \Phi^{-1}(F_X(x_i))$.

4.5 The Extended Rank Likelihood

Remark

This section primarily follows the model given in (Hoff, 2007).

How do these properties fit into our concept of invariances? To answer this question, let us begin by defining a vector of p standard normal rvs

$$\bar{\boldsymbol{X}} \equiv [\bar{X}_1, \ldots, \bar{X}_p]^\top \tag{4.25}$$

distributed as

$$\bar{\boldsymbol{X}} \sim \mathcal{N}_p(\mathbf{0}_p, R) \tag{4.26}$$

with density $f(\bar{\boldsymbol{X}}\,; R)$. For the observed continuous/discrete/mixed rvs, we write (without bar notation)

$$\boldsymbol{X} \equiv [X_1, \ldots, X_p]^\top, \tag{4.27}$$

where the connection is formally given by $X_i = F_i^{-1}(\Phi(\bar{X}_i))$ for $i \in \{1, \ldots, p\}$. Since we only want honor the order relations tying together $\boldsymbol{X}$ and $\bar{\boldsymbol{X}}$, we introduce the notation $\bar{\boldsymbol{X}} \in Z$ as used in (Hoff, 2007), where Z is the set of all possible $\bar{\boldsymbol{X}}$ satisfying the order constraints of n realizations.

Notice that Z is fully defined by the discrete rvs $\boldsymbol{X}$. Fig. 4.8 gives a better intuition about the setup.

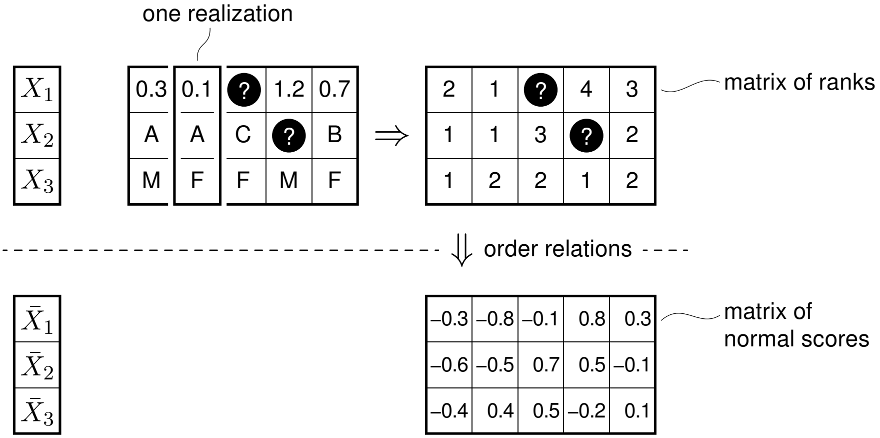


Figure 4.8: $p = 3$ mixed rvs $\boldsymbol{X}$ have $n = 5$ realizations with ranks as shown on the top right. Question marks refer to missing values. The bottom row shows *one* example for corresponding realizations of $\bar{\boldsymbol{X}} \in Z$, which conforms to the order relations.

Now, recall the derivation of the marginal likelihood in Eq. (1.15) from the introduction of invariances (Section 1.2),

$$f(u, v\,; \psi, \lambda) = f(u\,; \psi) \cdot f(v \,|\, u\,; \psi, \lambda),$$

where (U, V) are sufficient statistics, ψ is the parameter of interest and λ represents the nuisance terms. Due to the factorization, ψ becomes isolated and the marginal likelihood is given by $L(\psi\,; u) \propto f(u\,; \psi)$. If applied to the current situation, the density factors as given in (Hoff, 2007),

$$\begin{aligned} f(&\bar{\boldsymbol{X}} \in Z, \boldsymbol{X}\,; R, F_1, \dots, F_p) \\ &= f(\bar{\boldsymbol{X}} \in Z\,; R) \cdot f(\boldsymbol{X} \,|\, \bar{\boldsymbol{X}} \in Z\,; R, F_1, \dots, F_p), \end{aligned} \tag{4.28}$$

and we are now in a position to define the likelihood

$$L(R\,;\bar{\boldsymbol{X}} \in Z) \propto f(\bar{\boldsymbol{X}} \in Z\,;R) \tag{4.29}$$

which is devoid of the marginal distribution functions and only relies on the ordering constraints given by Z.

Our task is to infer R given mixed data, which can now be solved via Bayesian inference: complementing the likelihood with an inverse Wishart prior (Hoff, 2007)

$$R \sim \mathcal{W}_p^{-1}(\nu, R_0) \tag{4.30}$$

with ν degrees of freedom and scale matrix R_0, we receive the posterior

$$f(R \,|\, \bar{\boldsymbol{X}} \in Z, \nu, R_0) \propto f(\bar{\boldsymbol{X}} \in Z \,|\, R) \cdot f(R \,|\, \nu, R_0). \tag{4.31}$$

Unfortunately, however, samples from the posterior would not be valid correlation matrices satisfying $\operatorname{diag}(R) = \mathbf{1}_p$, therefore, Hoff (2007) modifies Eq. (4.31),

$$f(B \,|\, \bar{\boldsymbol{X}} \in Z, \nu, B_0) \propto f(\bar{\boldsymbol{X}} \in Z \,|\, B) \cdot f(B \,|\, \nu, B_0), \tag{4.32}$$

where B is a $p \times p$ covariance matrix, which is scaled to

$$R \leftarrow \operatorname{diag}(B)^{-\frac{1}{2}}\, B\, \operatorname{diag}(B)^{-\frac{1}{2}}. \tag{4.33}$$

Averaging over the posterior samples then yields the estimate $\widehat{R}$. Note that although the standard normal rvs $\bar{\boldsymbol{X}}$ are currently only a byproduct to calculate R, we can also estimate the normal scores by averaging, which requires an adjustment of scaling:

$$\bar{X}_i \leftarrow \bar{X}_i / \sqrt{B_{ii}}, \qquad i \in \{1, \ldots, p\}. \tag{4.34}$$

These normal scores play an important role in the information bottleneck, which is introduced in the next section.

Algorithm 4 describes a simple Gibbs sampler to calculate the posterior in Eq. (4.32) as used in (Adametz et al., 2014), which in turn is an extended version of (Hoff, 2007). The major differences are the estimation of normal scores and the use of a Wishart prior (instead of an *inverse* Wishart prior) to avoid matrix inversion in the innermost loop. The notation $V_{-i,j}$ stands for column vector j of matrix V without element i.

Algorithm 4 Sampling correlation matrix R and matrix $\bar{X}$ of normal scores

```
Input: $p \times n$ matrix $X$ containing observations from $p$ rvs
Set $V \leftarrow I_p,\ V_0 \leftarrow \epsilon\, I_p,\ \epsilon > 0,\ \nu \leftarrow p + 1$
Initialize $\bar{X}$ by $\bar{X}_{i\bullet} \leftarrow \Phi^{-1}\left(\frac{\text{rank}(X_{i\bullet})}{\#\text{levels}+1}\right)$ for $i \in \{1, \ldots, p\}$
for $k = 1 \ldots N_{\text{samples}}$ do
   for rv $i = 1 \ldots p$ do
      for level $r$ in $X_{i\bullet}$ do
         Find lower bound $a \leftarrow \max(\bar{X}_{i\bullet} \mid X_{i\bullet} < r)$
         Find upper bound $b \leftarrow \min(\bar{X}_{i\bullet} \mid X_{i\bullet} > r)$
         for every $j \in \{1, \ldots, n\}$ where $X_{ij} = r$ do
            Set $\mu \leftarrow -V_{ii}^{-1}\, V_{i,-i}\, \bar{X}_{-i,j}$
            Set $\sigma^2 \leftarrow V_{ii}^{-1}$
            Sample $\bar{X}_{ij} \sim \mathcal{N}(\mu, \sigma^2, a, b)$
         end for
      end for
   end for
   Sample $V \sim \mathcal{W}_p(\nu + n, (V_0 + \bar{X}\bar{X}^\top)^{-1})$
   Compute $B \leftarrow V^{-1}$
   Compute $R \leftarrow \text{diag}(B)^{-\frac{1}{2}}\, B\ \text{diag}(B)^{-\frac{1}{2}}$
   Compute $\bar{X}_{i\bullet} \leftarrow \bar{X}_{i\bullet} / \sqrt{B_{ii}}$ for $i \in \{1, \ldots, p\}$
end for
```

Notice how the normal scores are sampled from a *truncated normal distribution* $\mathcal{N}(\mu, \sigma^2, a, b)$. In case of missing values, there is no upper or lower bound, which leads to the unconstrained normal distribution (Hoff, 2007). As a final remark, the complexity of Algorithm 4 is $\mathcal{O}(p^3 + np^2)$ per full loop of the sampler, where it is assumed that $n \gg p$.

4.6 Application: The Information Bottleneck

Given our ability to estimate correlation matrix R for mixed rvs, we now turn to an application that benefits from this: The *information bottleneck* (Tishby et al., 1999) assumes two rvs X and Y, and the goal is to find a *compression* T of X while preserving information about Y. Its elegance is due to the fact, that choosing Y determines which aspect of X is *relevant*. In information-theoretic terms, the idea is to force the information of Y contained in X through a limited set of *code words* T (the *bottleneck*), where the mapping is stochastic: $f(t \,|\, x)$. This leads to a variational problem of minimizing a functional

$$J \equiv I(X\,;Y) - \beta\, I(T\,;Y) \tag{4.35}$$

with regards to code-word mapping $f(t \,|\, x)$. Here, $I(\bullet)$ measures *mutual information* between two rvs and $\beta > 0$ is a Lagrange parameter, which controls the trade-off between the best compression ($\beta \to 0$) or being most informative about Y ($\beta \to \infty$). The above is a very broad problem formulation, which only requires access to the joint distribution $F_{XY}(x, y)$, yet it makes no statement about the distribution *families*. For this reason, there exists no general closed-form solution, which motivated Chechik et al. (2007) to analyze the special case of Gaussian rvs. The *Gaussian information bottleneck* assumes random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ of size p and q, respectively, having the joint multivariate normal distribution

$$\begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{bmatrix} \sim \mathcal{N}_{p+q}\left(\begin{bmatrix} \mathbf{0}_p \\ \mathbf{0}_q \end{bmatrix}, \begin{bmatrix} \Sigma_{\boldsymbol{X}} & \Sigma_{\boldsymbol{XY}} \\ \Sigma_{\boldsymbol{YX}} & \Sigma_{\boldsymbol{Y}} \end{bmatrix} \right). \tag{4.36}$$

Interestingly, this implies that also $\boldsymbol{T}$ is Gaussian (Chechik et al., 2007),

$$\boldsymbol{T} \equiv A\boldsymbol{X} + \boldsymbol{\xi}, \tag{4.37}$$

which depends on $p \times p$ projection matrix A and noise component $\boldsymbol{\xi} \sim \mathcal{N}_p(\mathbf{0}_p, \Sigma_{\boldsymbol{\xi}})$. Calculating A involves the left eigenvectors $\boldsymbol{v}_i$ and eigenvalues

$\lambda_i \leq 1$ of matrix

$$\begin{aligned} B &\equiv \Sigma_{\boldsymbol{X}|\boldsymbol{Y}} \Sigma_{\boldsymbol{X}}^{-1} && (4.38) \\ &= I_p - \Sigma_{\boldsymbol{XY}} \Sigma_{\boldsymbol{Y}}^{-1} \Sigma_{\boldsymbol{YX}} \Sigma_{\boldsymbol{X}}^{-1} && (4.39) \end{aligned}$$

together with scaling

$$\alpha_i = \begin{cases} \sqrt{\dfrac{\beta(1-\lambda_i)-1}{\lambda_i \boldsymbol{v}_i^\top \Sigma_{\boldsymbol{X}} \boldsymbol{v}_i}} & \text{if } \beta > (1-\lambda_i)^{-1} \\ 0 & \text{else,} \end{cases} \tag{4.40}$$

where $\lambda_1 \leq \ldots \leq \lambda_p$. Finally, this yields the projection matrix

$$A = \begin{bmatrix} \alpha_1 \boldsymbol{v}_1^\top \\ \vdots \\ \alpha_p \boldsymbol{v}_p^\top \end{bmatrix}. \tag{4.41}$$

Notice that due to the definition of matrix B in Eq. (4.38), there can be at most $N \equiv \min(p, q)$ eigenvalues $\lambda_i < 1$, which implies N non-zero rows in matrix A if β is suitably large. To this extent, medium to small values of β introduce even further rows of zero in A, thereby losing information about $\boldsymbol{Y}$ in favor of a better compression of $\boldsymbol{X}$. As described above, the most informative compression is achieved when A contains all N non-zero rows for large β; further increasing β only affects the scaling.

Although the Gaussian information bottleneck has a convenient closed-form solution, it is by definition limited to normal rvs. To this end, Rey and Roth (2012) reformulated functional J of Eq. (4.35) in terms of copula densities to receive

$$H(c_{\boldsymbol{X}}) + H(c_{\boldsymbol{T}}) - H(c_{\boldsymbol{XT}}) - \beta\Big(H(c_{\boldsymbol{Y}}) + H(c_{\boldsymbol{T}}) - H(c_{\boldsymbol{YT}})\Big), \tag{4.42}$$

where $H(\bullet)$ denotes *entropy*. This reveals two fundamental properties:

- The *general* information bottleneck problem is independent of the marginals $F_1, \ldots, F_{p+q}$.

- The *Gaussian* information bottleneck is optimal for the family of meta-Gaussian distributions, because a copula is invariant against strictly increasing transformations, i.e., the marginals.

This generalization is also known as the *meta-Gaussian information bottleneck* (*MGIB*) (Rey and Roth, 2012). From a practical standpoint, the difference in comparison to the Gaussian information bottleneck only concerns Eq. (4.38), which now becomes

$$B \equiv R_{\boldsymbol{X}|\boldsymbol{Y}} R_{\boldsymbol{X}}^{-1}. \tag{4.43}$$

While the approach of Rey and Roth (2012) requires *continuous* meta-Gaussian distributions, we can now apply the correlation estimate for mixed data from Algorithm 4, thereby further generalizing the information bottleneck, see Fig. 4.9.

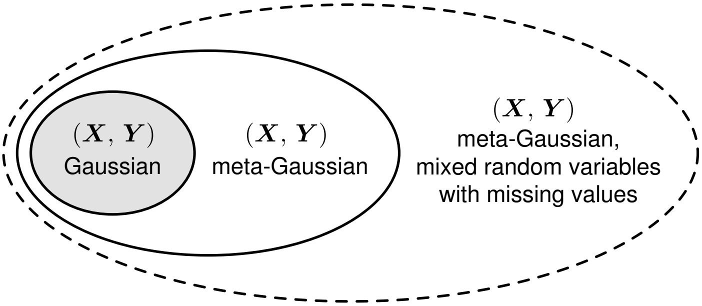

$(\boldsymbol{X}, \boldsymbol{Y})$
Gaussian
$(\boldsymbol{X}, \boldsymbol{Y})$
meta-Gaussian
$(\boldsymbol{X}, \boldsymbol{Y})$
meta-Gaussian,
mixed random variables
with missing values

Figure 4.9: The big picture comparing the scope of methods.

Let us now briefly summarize the necessary steps for compression of mixed data with missing values:

1. For p- and q-dimensional random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$, estimate $(p+q) \times (p+q)$ correlation matrix R and p-dimensional standard normal random vector $\bar{\boldsymbol{X}}$ using the extended-rank-likelihood sampler of Algorithm 4.

2. Compute $p \times p$ matrix $B = R_{\boldsymbol{X}|\boldsymbol{Y}} R_{\boldsymbol{X}}^{-1}$ including its left eigenvectors $\boldsymbol{v}_i$ and eigenvalues λ_i.

3. Calculate $p \times p$ projection matrix A for a given β, see Eq. (4.41).

4. Find p-dimensional compression $\boldsymbol{T} = A\bar{\boldsymbol{X}} + \boldsymbol{\xi}$.

4.6.1 Removal of Information

Complementary to highlighting a relevance variable $\boldsymbol{Y}$ in compression $\boldsymbol{T}$, one might also be interested in the opposite, that is, a compression free of nuisance information. In medical applications for example, age and sex are often known to have unwanted effects, which require a normalization of the data. The conventional approach is to introduce a dummy coding for binary or discrete rvs and then perform multiple regression to cancel their interaction (Lay, 2011, p. 372). This quickly becomes cumbersome in the presence of multiple discrete rvs with many levels.

The more convenient alternative is to apply the meta-Gaussian information bottleneck with a minor modification (Adametz et al., 2014): Recall that projection matrix A in compression $\boldsymbol{T} = A\boldsymbol{X} + \boldsymbol{\xi}$ specifically captures every aspect of $\boldsymbol{X}$ that is *relevant* to $\boldsymbol{Y}$. If A is replaced by Q, whose columns $Q_{\bullet j}$ span the *nullspace* of A, satisfying

$$A \cdot (Q_{\bullet j}) = \mathbf{0}_p, \tag{4.44}$$

then compression

$$\boldsymbol{T} = Q^\top \boldsymbol{X} + \boldsymbol{\xi} \tag{4.45}$$

is explicitly *devoid* of $\boldsymbol{Y}$. Due to this notion, we also refer to $\boldsymbol{Y}$ as *irrelevance* variables (Adametz et al., 2014). Fig. 4.10 and Fig. 4.11 give examples for A and Q including their associated compression.

Analog to the case of compression, we can control to which extent information about $\boldsymbol{Y}$ is removed. By choosing a value for trade-off parameter $\beta > 0$, the resulting compression $\boldsymbol{T}$ may cancel only main effects of $\boldsymbol{Y}$ (small β) or a wide range of its characteristics (large β). Table 4.1 exemplary describes different levels of β along with their interpretation regarding $\boldsymbol{T}$. Note that projection A can only have as many as $N \equiv \min(p, q)$ non-zero rows, where p and q refer to the length of random vector $\boldsymbol{X}$ and $\boldsymbol{Y}$,

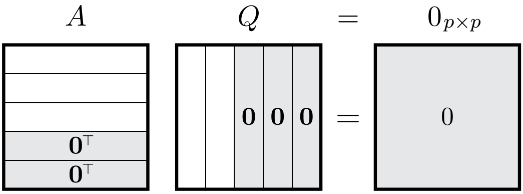


Figure 4.10: A schematic example of matrix A and Q for $p = 5$.

respectively. Therefore, in the above example of

$$\boldsymbol{X} = [\,\text{gene }1, \ldots, \text{gene }p\,]^\top \qquad \text{and} \qquad \boldsymbol{Y} = [\,\text{age}, \text{sex}\,]^\top,$$

projection matrix Q removes a maximum of two directions for a sufficiently large β, whose value depends on the left eigenvalues of $B = R_{\boldsymbol{X}|\boldsymbol{Y}} R_{\boldsymbol{X}}^{-1}$, see Eq. (4.40) and Eq. (4.43). Intuitively speaking, by applying Q, we transform the space $\mathbb{R}^p$ to subspace $\mathbb{R}^{(p-2)}$, where the effects of age and sex are mapped to zero.

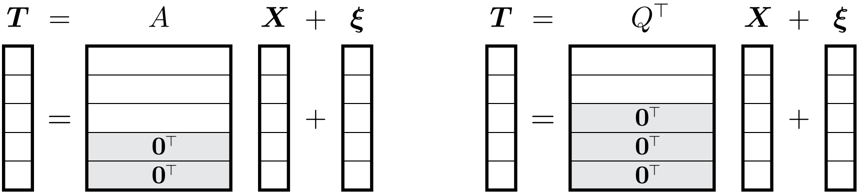

$\boldsymbol{T}$ $=$ A $\boldsymbol{X}$ $+$ $\boldsymbol{\xi}$
$=$
$\mathbf{0}^\top$
$\mathbf{0}^\top$
$+$
$\boldsymbol{T}$ $=$ $Q^\top$ $\boldsymbol{X}$ $+$ $\boldsymbol{\xi}$
$=$
$\mathbf{0}^\top$
$\mathbf{0}^\top$
$\mathbf{0}^\top$
$+$

Figure 4.11: Using the matrices from Fig. 4.10, a compression either highlights (left) or discards (right) side information $\boldsymbol{Y}$.

The information conveyed by Y_1 and Y_2 is *jointly* removed from the compression, meaning there is no *exclusive* correspondence between T_1 and Y_1 (or T_2 and Y_2). This is analog to PCA, which computes linear combinations of the underlying variables in order to identify directions of largest variance.

A	$Q^\top$	**Interpretation**
$\begin{array}{\|c\|}\hline \mathbf{0}^\top \\ \hline \mathbf{0}^\top \\ \hline \mathbf{0}^\top \\ \hline \mathbf{0}^\top \\ \hline \mathbf{0}^\top \\ \hline \end{array}$	$\begin{array}{\|ccccc\|}\hline 1 & & & 0 & \\ & 1 & & & \\ & & 1 & & \\ 0 & & & 1 & \\ & & & & 1 \\ \hline \end{array}$	No characteristic about $\boldsymbol{Y}$ is captured in A, therefore, applying $Q^\top = I_p$ does not alter the data: compression $\boldsymbol{T}$ is equal to $\boldsymbol{X}$ (ignoring noise component $\boldsymbol{\xi}$).
$\begin{array}{\|c\|}\hline \\ \hline \mathbf{0}^\top \\ \hline \mathbf{0}^\top \\ \hline \mathbf{0}^\top \\ \hline \mathbf{0}^\top \\ \hline \end{array}$	$\begin{array}{\|c\|}\hline \\ \hline \\ \hline \\ \hline \\ \hline \mathbf{0}^\top \\ \hline \end{array}$	Projection A captures the dominant/main effect of $\boldsymbol{Y}$ on $\boldsymbol{X}$ and therefore, $Q^\top$ specifically removes a single direction.
$\begin{array}{\|c\|}\hline \\ \hline \\ \hline \\ \hline \mathbf{0}^\top \\ \hline \mathbf{0}^\top \\ \hline \end{array}$	$\begin{array}{\|c\|}\hline \\ \hline \\ \hline \mathbf{0}^\top \\ \hline \mathbf{0}^\top \\ \hline \mathbf{0}^\top \\ \hline \end{array}$	Projection A captures several effects of $\boldsymbol{Y}$ on $\boldsymbol{X}$ and $Q^\top$ keeps everything pertaining to the remaining subspace.
$\begin{array}{\|c\|}\hline \\ \hline \\ \hline \\ \hline \\ \hline \mathbf{0}^\top \\ \hline \end{array}$	$\begin{array}{\|c\|}\hline \\ \hline \mathbf{0}^\top \\ \hline \mathbf{0}^\top \\ \hline \mathbf{0}^\top \\ \hline \mathbf{0}^\top \\ \hline \end{array}$	Matrix $Q^\top$ extracts a single direction from the space of $\boldsymbol{X}$; the remainder is canceled completely.
$\begin{array}{\|c\|}\hline \\ \hline \\ \hline \\ \hline \\ \hline \\ \hline \end{array}$	$\begin{array}{\|c\|}\hline \mathbf{0}^\top \\ \hline \mathbf{0}^\top \\ \hline \mathbf{0}^\top \\ \hline \mathbf{0}^\top \\ \hline \mathbf{0}^\top \\ \hline \end{array}$	All facets of $\boldsymbol{Y}$ are deemed important and shall be removed from $\boldsymbol{X}$, hence, no information is left after applying $Q^\top$.

Table 4.1: Parameter β determines to which extent information about $\boldsymbol{Y}$ is removed from compression $\boldsymbol{T} = Q^\top \boldsymbol{X} + \boldsymbol{\xi}$. The plots range from small β (top row) to large β (bottom row). It is assumed that A can have up to $N \equiv \min(p, q) = p$ non-zero rows.

4.7 Experiments

4.7.1 Synthetic Data

Since the estimation of a correlation matrix from mixed data is the core component of this chapter, we begin with an analysis of Algorithm 4. Let us therefore generate a covariance matrix Σ with visible dependencies between $p = 10$ rvs, constructed by the same scheme as in the synthetic experiment for Gaussian graphical models, p. 84ff. By rescaling, we receive the correlation matrix as $R = \mathrm{diag}(\Sigma)^{-\frac{1}{2}}\, \Sigma \; \mathrm{diag}(\Sigma)^{-\frac{1}{2}}$, which is used in

$$\bar{\boldsymbol{X}} \sim \mathcal{N}_p(\mathbf{0}_p, R) \tag{4.46}$$

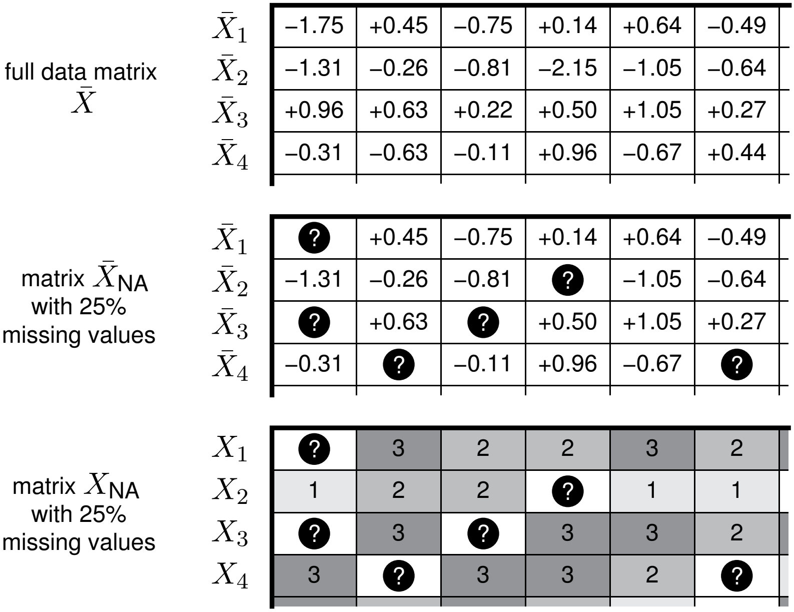
for $n = 100$ independent draws. All observations are stored as columns in $p \times n$ matrix $\bar{X}$, from which we compute the $p \times p$ Gaussian rank correlation $\widehat{R}_G$ (subscript G), see Eq. (4.20).

In a first setup, we use matrix $\bar{X}$ containing the normal scores and run the extended-rank-likelihood sampler (subscript E) to obtain $\widehat{R}_E$ by averaging over 1000 sweeps. It is important to keep in mind that both estimators do *not* use the normal scores in $\bar{X}$ directly, but rather only their *ranks*. Hence, we could have theoretically applied, say, beta or gamma marginals, however, their purpose is defeated by the fact that continuous marginal transformations preserve the rank. In our previous terminology, all such operations remain inside the invariance class (that is, the group orbit of the rank statistic) and therefore, they do not bias inference.

In order to make the task more challenging, we introduce 5, 10, 15, 20 and 25% missing values at randomly selected elements of $\bar{X}$ to receive matrix $\bar{X}_{\text{NA}}$, which is *not* covered by the Gaussian rank correlation. The extended rank likelihood handles missing values by assuming an *unconstrained* normal distribution; all observed values follow a *truncated* normal distribution which satisfies the given ranks. For a better understanding of the data and its composition, Fig. 4.12 depicts one dataset.

The obtained correlations $\widehat{R}_E$ and $\widehat{R}_G$ are compared by

$$\ell(\widehat{R}_E\,; \bar{X}) - \ell(\widehat{R}_G\,; \bar{X}), \tag{4.47}$$



full data matrix $\bar{X}$

$\bar{X}_1$	−1.75	+0.45	−0.75	+0.14	+0.64	−0.49
$\bar{X}_2$	−1.31	−0.26	−0.81	−2.15	−1.05	−0.64
$\bar{X}_3$	+0.96	+0.63	+0.22	+0.50	+1.05	+0.27
$\bar{X}_4$	−0.31	−0.63	−0.11	+0.96	−0.67	+0.44

matrix $\bar{X}_{\text{NA}}$ with 25% missing values

$\bar{X}_1$	?	+0.45	−0.75	+0.14	+0.64	−0.49
$\bar{X}_2$	−1.31	−0.26	−0.81	?	−1.05	−0.64
$\bar{X}_3$	?	+0.63	?	+0.50	+1.05	+0.27
$\bar{X}_4$	−0.31	?	−0.11	+0.96	−0.67	?

matrix X_{NA} with 25% missing values

X_1	?	3	2	2	3	2
X_2	1	2	2	?	1	1
X_3	?	3	?	3	3	2
X_4	3	?	3	3	2	?

Figure 4.12: An excerpt from one dataset as used in the experiment. Top: Normal scores, where rvs are correlated according to R. Center: Same matrix as above, but 25% of all values are removed. Bottom: The remaining observations are transformed into discrete values with only 3 levels; hereby, every rv has a different discrete marginal, but the order of levels is supported by the order of normal scores. Each matrix has size 10×100.

where $\widehat{R}_E$ used $\bar{X}_{\text{NA}}$ and $\widehat{R}_G$ resorted to $\bar{X}$. For a proper evaluation, we repeat the above procedure to generate a total of 1000 correlation matrices R and matrix-realizations $\bar{X}$. The results can be seen in Fig. 4.13 (left).

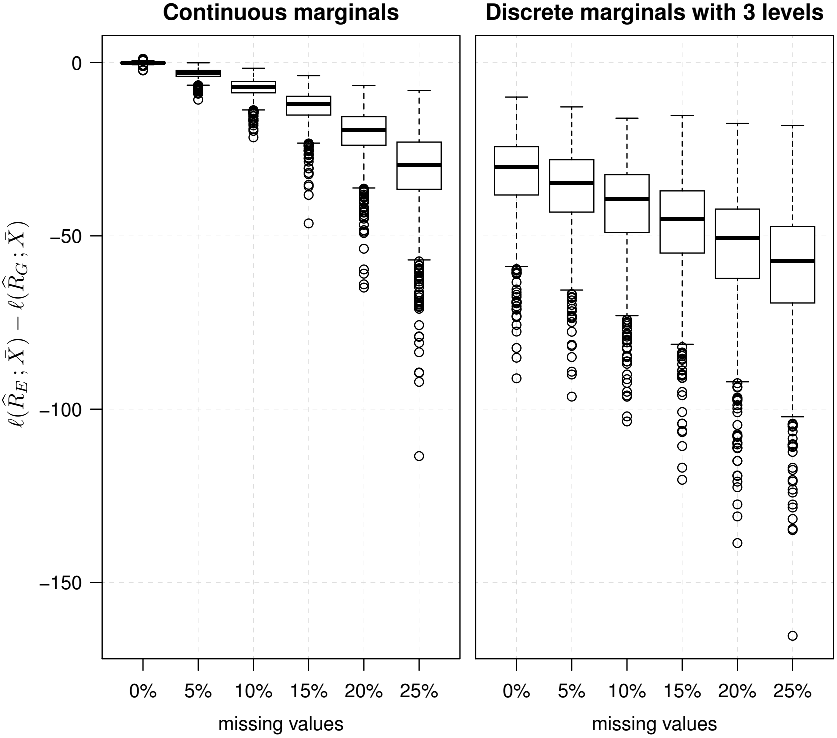


Figure 4.13: Boxplots over the difference in log-likelihood between $\widehat{R}_E$ and $\widehat{R}_G$ using 1000 synthetic datasets $\bar{X}$ with $p = 10$ and $n = 100$.

As expected, $\widehat{R}_E$ and $\widehat{R}_G$ are virtually identical for continuous marginals and no missing values, while the performance only slightly decreases when few data are missing. The removal of more than 5% has a visible impact (notice the use of the *log* form).

In a second experiment, we transform all 1000 matrices $\bar{X}$ row-wise into discrete observations of only 3 levels, where the order of levels is in accordance with the ranks of the normal scores. This means, $\widehat{R}_E$ is now inferred from matrix X_{NA}, which refers to discrete observations with missing values. As a result, this transformation considerably increases the difficulty due to ties among ranks, see again Fig. 4.12. For a proper comparison, we again remove 5, 10, 15, 20 and 25% of the data at the exact same locations as before. To this end, Fig. 4.13 (right) reports the outcome for discrete marginals over all 1000 datasets. The performance without missing values is roughly equivalent to that of continuous marginals with 25% missing values. Further, introducing discrete marginals increases uncertainty, which is to be expected.

The reader should keep in mind that the second experiment demonstrates the extended-rank-likelihood sampler in situations with significant information loss and rather large numbers of missing values. Real-world data with mixed marginals is likely situated somewhere in between both experiments.

4.7.2 Real-World Data – Pathway-Based Cancer Analysis

For a practical application of the meta-Gaussian information bottleneck to mixed data, let us revisit the pathway domain, as introduced in Section 3.1.1. Recall that a pathway is a distinct group of genes, which contributes to a specific biological function. For our purposes, we again adapt the pathway definitions from the KEGG database, which enables us to abandon single-gene analysis. Hereby, we solve two conceptual issues:

- Identifying single factors among 24 000 genes on the basis of only hundreds of patients is a strongly underdetermined problem. As a result, this often leads to the false discovery of genes, which perfectly explain a target, but cannot be reproduced independently (Ein-Dor et al., 2006). The solution is either to dramatically increase the sample size, or to turn to a pathway-centric analysis, such that the number of candidate genes is limited. Hereby, we can make practical use of rich biological expert knowledge.

- Pairing functionally related genes amplifies signals which are weak, but consistent. Such patterns may not be discoverable by other means, especially given the high noise level of a single gene.

Due to these reasons, the objective is to fuse all information of a gene set and extract its characteristic properties. One suitable tool is PCA, which identifies the directions of largest variance in the genes. Note however, that variance merely represents a single choice among a host of relevant biological properties. For instance, we could ask "What information does a pathway convey about cancer stage?". This is in fact a prime example of the meta-Gaussian information bottleneck for mixed data, which can now be handled with the extended rank likelihood. Fig. 4.14 gives a schematic outline.

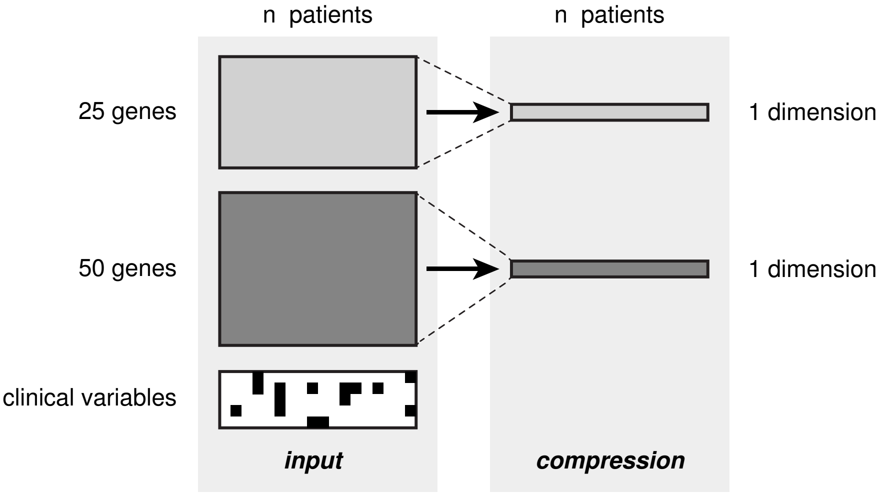


Figure 4.14: Overview of pathway compression. The above shows two pathways (gene sets) comprising gene expression values for n patients. The goal is to fuse and compress their information into a single dimension while highlighting clinical properties. Black boxes symbolize missing values in clinical data.

Clinical information is often subject to missing values due to various reasons: in some cases, information is redacted to protect a patient's privacy, in other situations, health concerns or the need for additional surgery prevent complete measurements. For a better intuition about the data, Fig. 4.15 reports a set of exemplary measurements.

Our analysis is based on the colon cancer dataset of Sheffer et al. (2009) as in the previous experiments for Gaussian graphical models, however, different from before we now use all $n = 313$ patients, falling into the categories

$$
\begin{array}{r|ccccc|}
 & \text{patient } 1 & \text{patient } 2 & \text{patient } 3 & \dots & \text{patient } n \\
\text{gene } 1 & 1.48 & 0.59 & -0.24 & \dots & -1.57 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
\text{gene } p & -0.81 & -1.01 & 0.32 & \dots & 2.15 \\
\hline
\text{age} & 68 & \boxed{?} & 43 & \dots & 39 \\
\text{sex} & F & M & F & \dots & \boxed{?} \\
T & 3 & 2 & \boxed{?} & \dots & 4 \\
N & 1 & 0 & \boxed{?} & \dots & 0 \\
M & 1 & 0 & \boxed{?} & \dots & 1
\end{array}
$$

Figure 4.15: A fictitious sample of gene expression data accompanied by clinical features. The matrix shows measurements for p genes and $q = 5$ clinical traits across n patients.

healthy (53 patients), *primary tumor* (184 patients), *polyp* (46 patients) and *metastasis* (30 patients). Further, the dataset comprises a total of 13 437 genes. In addition to gene expression values, a separate table lists clinical features age, sex and *TNM* tumor staging, where $T = \{T0, T1, T2, T3, T4\}$ is the size of the tumor, $N = \{N0, N1, N2, N3\}$ represents the spread to adjacent/distant lymph nodes and $M = \{M0, M1\}$ stands for absence/presence of metastasis in distant body regions. Note that for some patients, lymph nodes are not evaluated thus leading to missing values. In summary, the clinical side information consists of binary or discrete observations with various numbers of levels.

The data are analyzed in the following way: First, select a pathway and look up its p genes from the KEGG database. Then, proceed in two steps:

1. Normalize gene expression to remove bias from age and sex:

    a) For random vectors $\boldsymbol{X}$ (p genes) and $\boldsymbol{Y}$ (age, sex), estimate the joint correlation matrix R and standard normal random vector $\bar{\boldsymbol{X}}$ using the extended-rank-likelihood sampler of Algorithm 4.

    b) Calculate $B = R_{\boldsymbol{X}|\boldsymbol{Y}} R_{\boldsymbol{X}}^{-1}$, projection A for a sufficiently large

β and its nullspace matrix Q.

c) Compute compression $\boldsymbol{T}_{\text{norm}} = Q^\top \bar{\boldsymbol{X}} + \boldsymbol{\xi}$.

2. Compress gene expression values with regards to cancer stage:

   a) For random vectors $\boldsymbol{X} \equiv \boldsymbol{T}_{\text{norm}}$ (p genes after normalization) and $\boldsymbol{Y}$ (T, N, M), estimate R and $\bar{\boldsymbol{X}}$ using Algorithm 4.

   b) Calculate $B = R_{\boldsymbol{X}|\boldsymbol{Y}} R_{\boldsymbol{X}}^{-1}$ and projection A for a small β.

   c) Compute compression $\boldsymbol{T}_{TNM} = A\bar{\boldsymbol{X}} + \boldsymbol{\xi}$.

The resulting rv $\boldsymbol{T}_{TNM}$ captures all information of the gene expression values pertaining to cancer stage while being free from the bias of age and sex. As it is difficult to visually interpret the compression of every pathway, we evaluate the outcome using software package *Pathifier* (Drier et al., 2013). This method measures the *deregulation score* for each patient as the distance from the average healthy patient, see Fig. 4.16. In addition, the patient with the largest distance defines score of 1. Since every pathway captures different aspects of the patients, the distances are computed for each pathway separately.

By default, Pathifier first projects the gene expression values onto eigenvectors corresponding to largest variance (PCA), which we now replace by the MGIB compression. The comparison can be found in Fig. 4.17 along with TNM cancer staging and a selection of pathways that are known to be affected in colon cancer. Here, rows are pathways, columns are patients and their gray value corresponds to the distance from the average healthy patient along the principal curve. Patients are categorized into the classes healthy (H), metastasis (M), polyp (P) or primary tumor (T). The bottom rows show cancer staging TNM, which contains missing values (white bars) for (M) and (P) patients. Note that the order of patients and pathways is identical in both plots.

The compression identifies interesting new patterns in the data, which divide the primary tumor patients (T) into two very distinct subgroups. It appears the presence of metastasis leads to a characteristic activation of the pathways VEGF, MTOR and ErbB signaling and the more general "Pathways in Cancer". In particular, patients with cancer stage $M = 1$ (metastasis in

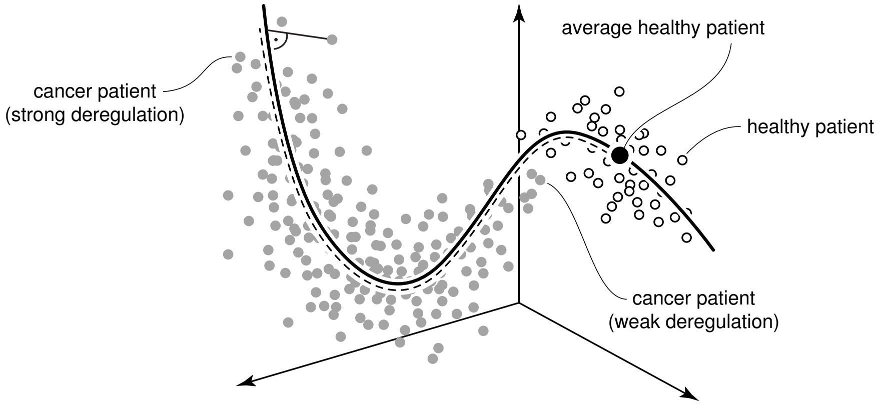


Figure 4.16: Pathifier assumes that—for each pathway—the patients form point clouds with meaningful structure. After identifying the skeleton line (the *principal curve*), the algorithm measures the distance of a cancer patient to the average healthy patient *along the curve*. The greater the distance, the stronger a patient is deregulated.

distant body regions) and $N = \{2, 3\}$ (spread to distant lymph nodes) behave similarly to metastasis patients (patient category M).

The deregulation scores based on principal components do not exhibit any of these patterns, which goes to show that variance does not necessarily capture all meaningful facets. In that regard, the information bottleneck is a method of feature extraction akin to PCA, but with the added flexibility to determine what is relevant to the user. The experiment demonstrates how previously unused side information can now be incorporated to gain a more expressive and problem-specific representation of genes.

Lastly, the (meta-)Gaussian information bottleneck has another interesting property for pathway analysis: Recall that the optimal compression is a linear projection of $\boldsymbol{X}$, which means it is possible to identify the contribution of each rv (= gene). This information can be leveraged to trace back single factors within the confines of a pathway. For more details and an in-depth discussion of the results, we point the reader to (Adametz et al., 2014).

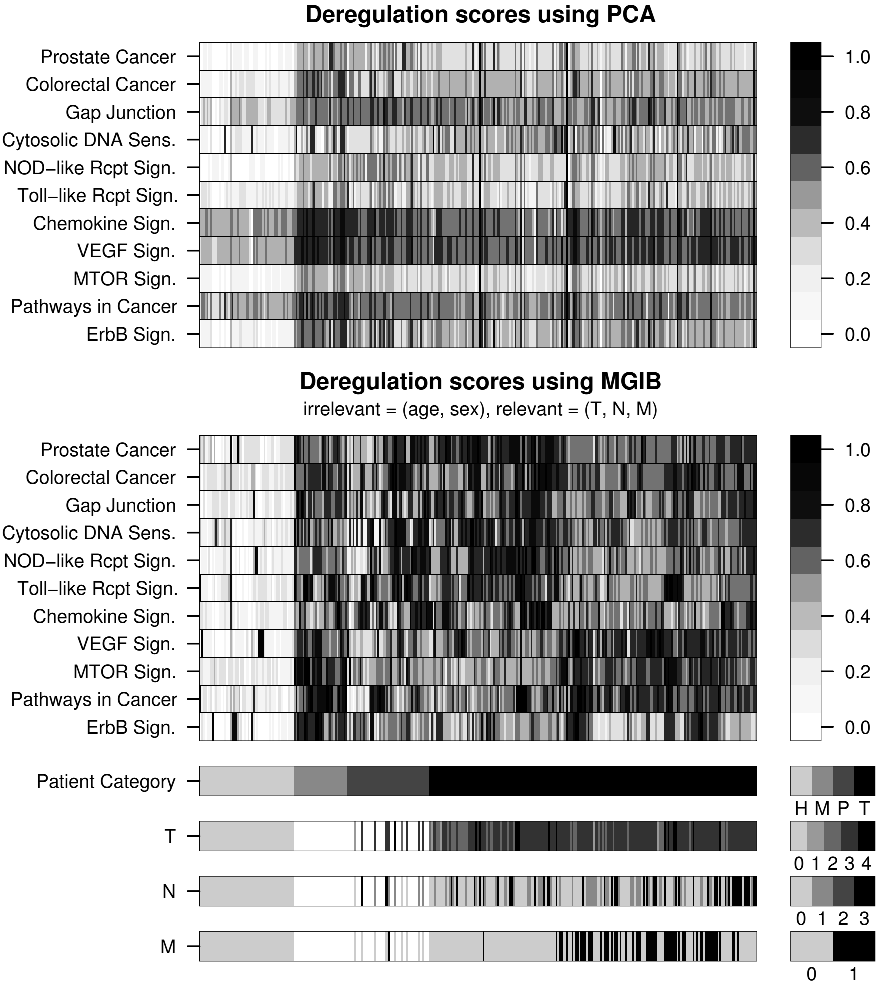


Figure 4.17: Pathifier computes the deregulation scores based on PC projection (top) and MGIB compression (bottom) for the colon cancer dataset of Sheffer et al. (2009).

4.8 Conclusion

At the heart of the current chapter, we focused on the dependence of rvs when i.i.d. realizations are directly accessible. To this end, Sklar's theorem (Sklar, 1959) provided the theoretical foundation to decompose any multivariate distribution into univariate marginals and the so-called copula, which induces dependencies. In the Gaussian case, the copula is fully specified by correlation matrix R, which makes it the parameter of interest. Moreover, disentangling marginal properties and dependencies forms the basis of meta-Gaussian distributions (Rey and Roth, 2012), which maintain the Gaussian copula but permit arbitrary marginals.

As a consequence of the probability integral transform, it is possible to convert continuous realizations back and forth into normal scores (realizations from a standard normal rv) *if the distribution function is known.* In turn, this gives rise to a straight-forward estimation of the correlation matrix. On the contrary, *if the distribution function is not known*, we can resort to empirical marginals, leading to an estimate that is only a function of the ranks (Boudt et al., 2012).

Since a host of applications solely require the dependence of Gaussian rvs, the generalization from Gaussian to meta-Gaussian distributions vastly extends the range of practical applications (Rey and Roth, 2012). Still, this setup excludes discrete rvs due to ties among the ranks, which would otherwise violate the uniqueness property of the copula. In accordance with the overarching subject of the thesis, the issue is rooted in information loss and the need to identify unique ranks, while the evidence does not permit any such statement. Intuitively speaking, discrete observations *do* carry information with regards to correlation and treating them as an incomplete set of order relations is the fundamental principle behind the extended rank likelihood (Hoff, 2007).

Finally, the ability to derive the correlation between continuous and discrete rvs opens new possibilities, as was demonstrated by the information bottleneck and its application to pathways. Here, we are interested in compressing a gene set in such a way that it only maintains characteristics related to side information, hence, the information bottleneck describes the notion of relevance in the data. While clinical features are typically only used at later stages of gene expression analysis, it is now possible to integrate this information very early on to extract subtle aspects of the data. In addition, we can not only highlight specific features, but also remove them in a unified fashion. Notice that the applied invariance is indispensable for this type of analysis—it represents an important step towards generalizing information-theoretic feature extraction.

Remark

Appendix C discusses links to distance-based methods.

Chapter 5

Discussion & Outlook

5.1 A Global Perspective on Invariances

In review of the three models introduced in this work, the concept of invariances played a major role in statistical inference. On the most fundamental level, all discussed applications have in common that a sample offers limited information with regards to the initial model. This gives rise to what was called nuisance parameter λ, which prevents us from evaluating hypotheses for an interest parameter ψ. We lack the means to distinguish

$$\frac{L(\psi_1, \lambda\,; x)}{L(\psi_2, \lambda\,; x)} \overset{?}{\gtrless} 1, \tag{5.1}$$

because the choice of λ determines if the likelihood favors ψ_1 over ψ_2 or vice versa. If we were to fix the unknown λ at an incorrect value, the conclusion drawn for ψ could potentially be wrong. Hence, it goes to show that problems of this kind require a cautious and well-considered treatment. Unfortunately, however, there exists no universal solution and the requirements often vary depending on the application at hand.

To this end, the thesis discussed an assortment of methods established in the literature: marginal and conditional likelihood, profile likelihood and integrated likelihood. These approaches all have their specific strengths and weaknesses, but most importantly, they do not guarantee a robust estimation of ψ or that inference is possible in the first place.

Let us demonstrate this schematically by Fig. 5.1 using the example of the marginal likelihood. Here, the box represents all available information contained in sample x and by applying statistic $u = U(x)$, only a certain aspect of it is used. Among the contents of x, we find a portion related to interest parameter ψ (white area), while the remainder is considered irrelevant. The dashed line symbolizes that U is unable to capture all necessary properties in their entirety, hence, some information is inadvertently lost. For our purposes, a statistic U is required to be strictly free of λ, therefore a loss can often not be avoided and is minimal at best.

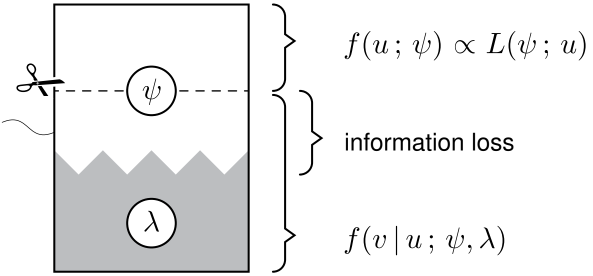

x
ψ
λ
$f(u\,;\,\psi) \propto L(\psi\,;\,u)$
information loss
$f(v\,|\,u\,;\,\psi,\lambda)$

Figure 5.1: Abstract interpretation of the marginal likelihood.

The above alludes to the fact that many problems do not allow an accurate isolation of the subset concerning ψ. In such instances, the loss is assumed to be negligible and therefore, certain characteristics are intentionally discarded in favor of the complete marginalization of nuisance λ. There are certainly situations, where ψ and λ can be separated perfectly, however, these should rather be treated as special cases. In general, the model always loses robustness, especially when λ comprises multiple parameters, as for example column means, scaling and feature correlation. This also implies that with every additional statistic, the available support is potentially cut back further and further. Therefore, incorporating invariances is, generally speaking, at the expense of statistical power—the question is whether the remainder still permits inference.

It is important to stress that depending on the nuisance parameter and the specific model, there may not necessarily be a statistic which meets our requirements. In some instances, suitable statistics *do* exist, but the remaining information about X is insufficient for inference. To name one example, the removal of column means via projection L with kernel $\mathbf{1}_p$ succeeded, because it affects a one-dimensional subspace in a $p \times p$ distance matrix. Yet, the profile likelihood using an $n \times n$ maximum-likelihood estimate of feature correlation Ψ failed, because the invariance against right-multiplication of arbitrary matrices renders the model virtually insensitive to detecting conditional independences. Therefore, the mere existence of a statistic does not necessarily imply viability.

Interestingly, different techniques sometimes lead to the same likelihood, as could be observed for translation or scale invariance. On a purely technical level, this does not offer any benefit, but it adds a new interpretation to the result and enhances our understanding of the model. A conceptually different invariance was applied in the third domain to estimate the normal scores based on non-unique rank information. Considering that the extended rank likelihood builds on classical marginalization, there is an enormous diversity of potential invariances.

5.2 Invariances as a Means of Generalization

If a model depends only on, say, pairwise distances instead of the full normal-distributed data matrix, we regard it as being more efficient, because it uses less information to infer the same outcome. The more important implication is however, that we can now apply it to *any* domain provided that it permits the computation of distances—regardless if the underlying data were vectorial or not. To this extent, the vast portfolio of kernel functions opens a huge field of potential applications, which, on a technical level, can all be treated as if they originated from a normal distribution. Its appeal is due to the fact that we can operate in a possibly infinite-dimensional feature space without the need to specify it explicitly.

The underlying Gaussian assumption is of course not appropriate for every application, but the normal distribution appears in many natural phenomena and is therefore a common choice when many rvs contribute additively or when nothing is known about the inner workings of a process. Further, if a distance matrix is the sole input to infer, say, a Gaussian graphical model, then the discussed methods TiMT and TiWnet are, to our knowledge, the only choice. In that regard, the contributions of this work cover...

- a Gaussian mixture model for clustering in distance matrices,
- a Gaussian graphical model to infer conditional independences in distance matrices and
- a Gaussian copula model for the compression of mixed data with missing values.

All three approaches inherit the core properties from their preceding models, but they now enter new ground for applications that could *not* be analyzed before. Hence, removing the dependence on nuisance parameters always implies a generalization.

It should be noted, however, that the standard models have an advantage over distance-based approaches in case of vectorial data, since they are not subject to potential information loss. Although the invariances were constructed in a way to maintain statistical power, say, to counter the effects of latent feature correlation in Gaussian graphical models, a certain loss is unavoidable and can only be minimized. This could be seen from the synthetic experiment in Section 3.5.1, where the distance-based approach (TiMT) closely follows the model with access to the underlying data matrix (TRCM), but it can never *exceed* it. Therefore, having full information serves as an upper bound of what is achievable under ideal conditions. The same applies to the Gaussian copula model if we compare the performance for discrete data and its true continuous representation.

Finally, for a side-by-side comparison of all models, see Table 5.1.

5.3 A Note on the Mean Model

Throughout the development of invariances in Gaussian models, we often referred to *column means*, because these are motivated by the definition of a distance matrix; in terms of the Gaussian copula model, we are limited to *row means* exclusively. Hence, it is justified to ask if there are alternatives beyond these examples. First of all, recall that data in our context are distributed as

$$X \sim \mathcal{N}_{p,n}(M, \Sigma \otimes \Psi),$$

	(fast)TiWD	**TiMT**	**Copula Model**
Application			
	clustering	network inference	compression
Distribution			
		$X \sim \mathcal{N}_{p,n}(M, \Sigma \otimes \Psi)$	
Parameter			
input	D (or S)	D (or S)	X
mean M	$\mathbf{1}_p \boldsymbol{w}^\top$	$\mathbf{1}_p \boldsymbol{w}^\top$	$\boldsymbol{v}\mathbf{1}_n^\top$
row covariance Σ	$\mathbb{S}_+^{p \times p}$	$\mathbb{S}_+^{p \times p}$	$\mathbb{S}_+^{p \times p}$
column covariance Ψ	I_n	$\mathbb{S}_+^{n \times n}$	I_n
parameter of interest	Σ	Σ^{-1}	R from Σ
Model infers...			
	cluster assignments of rows	conditional independences of rows	correlation of rows
Invariance against...			
number of columns n	✓	✓	—
row means $\boldsymbol{v}\mathbf{1}_n^\top$	—	—	✓
column means $\mathbf{1}_p \boldsymbol{w}^\top$	✓	✓	—
scaling c	✓	✓	✓
column covariance Ψ	—	✓	—
rank-preserv. operations	—	—	✓

Table 5.1: Comparison between the three introduced models, which all build on the matrix normal distribution. $\mathbb{S}_+$ refers to the set of symmetric positive-definite matrices.

where mean matrix M has the same format $p \times n$ as a realization X. The fact that we observe a single matrix X (or D) does not permit any conclusive statement for a general M, same as in the scalar case, if a single observation x shall be used to identify the mean μ. The only deduction we could make is $M \equiv X$ (or $\mu \equiv x$), therefore it is imperative to assume a direction of repeated measurements:

$$M = \mathbf{1}_p \boldsymbol{w}^\top, \ \ \boldsymbol{w} \in \mathbb{R}^n \qquad \text{or} \qquad M = \boldsymbol{v}\mathbf{1}_n^\top, \ \ \boldsymbol{v} \in \mathbb{R}^p. \tag{5.2}$$

Making a choice between the two is closely tied to the composition and interpretation of matrix X, that is, *objects* and *features* or similarly *rvs* and *realizations*. In that regard, Fig. 5.2 explains the differences between the distance-based methods and the Gaussian copula approach. Notice that in spite of their assumptions, both aim to infer row covariance matrix Σ[1].

As can be seen, distances are independent of column means, but sensitive to a modification of rows. The Gaussian copula model interchanges these assumptions, because rows correspond to rvs and columns are independent realizations. Theoretically, it is possible to center *both* rows and columns simultaneously, however, this is not meaningful for either model.

5.4 Nuisance Parameters and Embeddings

For distances, statements about the underlying feature space were avoided, however, a skeptical reader may propose an embedding X of distance matrix D, as done, for example, in *multidimensional scaling* (Gower, 1966; McCullagh, 2009). The rationale behind this is to resort to the standard normal likelihood with the intention of skipping invariances altogether. Notice the idea does not cause computational problems, but rather introduces a potential

[1] We could alter the notation of the copula model, such that the mean models become identical, however, this would also reverse the roles of Σ and Ψ. The intention behind the current notation is to unify the frameworks of distances and copulas. In doing so, we accept having two different models of the mean.

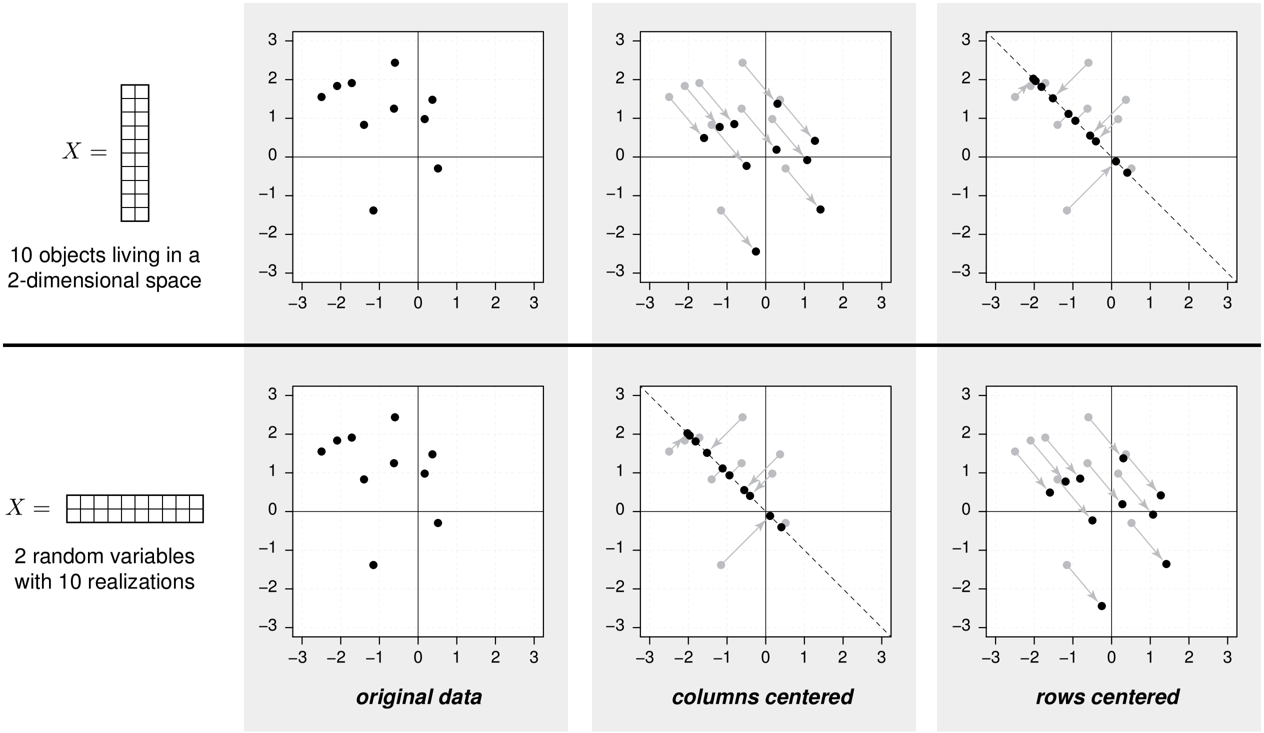


Figure 5.2: Comparing mean models for the distance-based methods (top row) and the Gaussian copula approach (bottom row). Although the difference is trivial and only due to transposition, the interpretation of X decides which centering operation is applicable.

bias, which may alter the likelihood ratio

$$\frac{L(\psi_1, \lambda)}{L(\psi_2, \lambda)} \tag{5.3}$$

if $\lambda \neq \lambda_0$. In fact, there is an infinite number of Euclidean configurations which are equivalent from the standpoint of distances. At the same time, each makes a different choice concerning the nuisance parameters.

Fig. 5.3 depicts this situation: Every box represents a matrix and a diagonal line symbolizes symmetry. A distance matrix D implies a space of potential originating matrices $S + \mathbf{1}_p \boldsymbol{u}^\top + \boldsymbol{u}\mathbf{1}_p^\top$ for any $\boldsymbol{u} \in \mathbb{R}^p$. If we decide on one candidate $S = XX^\top$, there is another space of corresponding matrices $X \in \mathbb{R}^{p \times n}$, which contains n unknown features (columns) as well as orthogonal projections thereof. Making a choice among X is the same as estimating unidentifiable nuisance parameters.

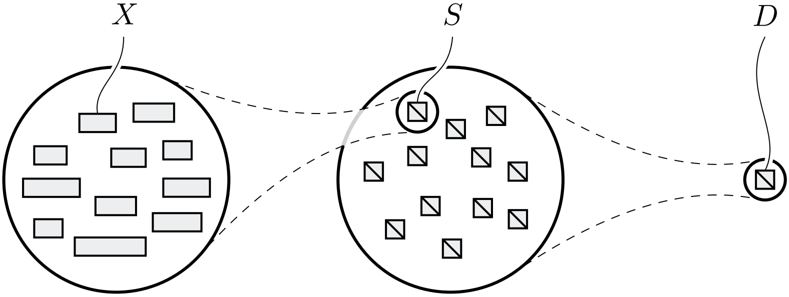


Figure 5.3: A schematic interpretation of X, S and D when seen from the perspective of distances.

Inferring the correlation matrix from mixed inputs pursues a conceptually different line of thought: By sampling the normal scores, the underlying data are explicitly reconstructed from the limited number of order constraint, however, it avoids specifying the marginals, which are attributed to the nuisance parameters. The reconstruction quality very much depends on the number of levels and observations, as well as the occurrence of missing values.

5.5 Support for Inference

The following picks up the idea of embeddings, but interprets it from a different standpoint: As mentioned above, the foundation of all models is

$$X \sim \mathcal{N}_{p,n}(M, \Sigma \otimes \Psi),$$

from which we are interested only in ψ, while the remainder is considered as nuisance λ. Suppose the application at hand defines $\psi \equiv \Sigma^{-1}$ and

$$\lambda \equiv \{M, \Psi\}, \tag{5.4}$$

then we would proceed by incorporating invariances, such that the density becomes independent of λ. Note that if X is accessible, its features are known and the number of columns n can simply be read off.

Computing the distances D is essentially a statistic that obscures information, which implies that the nuisance parameters grow in number to

$$\lambda \equiv \{M, \Psi, \mathcal{X}\}, \tag{5.5}$$

where $\mathcal{X}$ refers to the *feature space* of X and includes n. To this end, the density for $M = 0_{p\times n}$ and $\Psi = I_n$ reads

$$f(X\,; \Sigma) \propto \exp(-\tfrac{1}{2}\operatorname{tr}\{\Sigma^{-1}XX^\top\}), \tag{5.6}$$

meaning X and its feature space $\mathcal{X}$ only enter by way of the inner product. In other words, the density already partitions $\mathcal{X}$ into equivalent sets. For a general $\Psi \neq I_n$, however, all features interact with each other in

$$f(X\,; \Sigma, \Psi) \propto \exp(-\tfrac{1}{2}\operatorname{tr}\{\Sigma^{-1}X\Psi^{-1}X^\top\}). \tag{5.7}$$

This is in fact the main reason why feature correlation in distances is a highly challenging problem that requires a Bayesian approach—it now affects the possibly high-dimensional $\mathcal{X}$. Our solution was specifically tailored to the distance domain as we successfully avoided any explicit statement about Ψ *or* $\mathcal{X}$ altogether. Instead, both entered the likelihood only implicitly via the hyperparameters of the prior and therefore, it was not necessary to specify

X directly, akin to the kernel trick.

A related issue applies to the Gaussian copula model, where X is accessible in discretized form instead of its true representation. Using incomplete rank information, the task was to estimate the normal scores, which, in turn, give rise to correlation matrix R. Fig. 5.4 picks up the idea that reconstruction of the underlying normal matrix X is a task of identifying the true nuisance parameters, λ_0.

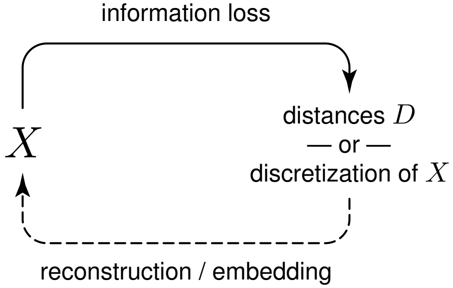


Figure 5.4: Nuisance parameter λ seen as a reconstruction. Matrix X is only accessible in reduced form, which makes it difficult to assess the parameter of interest due to information loss. Reverting this process is equivalent to identifying the true λ_0, because feature space $\mathcal{X}$ is in part a nuisance parameter.

All previous discussion was mainly focused on incorporating suitable invariances into the likelihood, which represents a gradual removal of dependence from X. However, when we think of the data reduction or efficiency principle as a recurring theme in the Fisherian likelihood concept, we could also ask which distribution our inference is *actually* based on, had we access to the true nuisance parameters λ_0 including all missing information about feature space $\mathcal{X}$. Hereby, we obtain the following interpretations:

- Clustering model (fast)TiWD assigns objects to the same cluster if they can be explained by a distribution with shared mean. We have

$$X \sim \mathcal{N}_{p,n}(M, I_p \otimes I_n), \tag{5.8}$$

where each row in M is associated with one cluster center $\boldsymbol{m}_j \in \mathbb{R}^n$, $j \in \{1 \ldots k\}$. Most properties of M pertaining to the feature space are lost when calculating the pairwise distances, which lead to a model that only requires the inner product $\frac{1}{n} M M^\top$. At the same time, this formulation introduces equivalence classes among cluster geometries.

- At its core, inferring a Gaussian graphical model via TiMT relies on

  $$X \sim \mathcal{N}_{p,n}(0_{p \times n}, W^{-1} \otimes I_n). \tag{5.9}$$

  Contrary to (fast)TiWD, it assumes the distances have been altered by feature correlation Ψ.

- The Gaussian copula model is based on the underlying matrix

  $$\bar{X} \sim \mathcal{N}_{p,n}(0_{p \times n}, R \otimes I_n), \tag{5.10}$$

  which is treated as n independent realizations of a p-dimensional standard normal random vector $\bar{\boldsymbol{X}}$. Many vital characteristics of $\bar{X}$ are lost due to discretization and missing values, such that the approach only uses the non-unique rank information.

The above is a description of the core mechanics if the data were completely devoid of nuisance terms. Its intention is to give an insight into the support for inference in terms of the underlying matrix normal distribution.

5.6 Outlook

Invariances constitute a powerful tool in statistical analyses, which offers new possibilities for existing, well-established models. As we learned, these techniques frequently involve a trade-off between removal of information and statistical explanatory power, hence, the essential question is how far we can go while maintaining a robust model for the task at hand. A boundary could be seen by the invariance against feature correlation in Gaussian graphical models (Adametz and Roth, 2014), which initially appeared as infeasible: the nuisance term affected a significant portion of the data which even exceeded the dimensionality of the parameter of interest. By using a flat prior in conjunction with the integrated likelihood approach, we could strike a careful balance to isolate what is relevant. For the first time, the model accounted for all available parameters of the matrix normal distribution when only a distance matrix is given on input.

Clustering in distances, as a generalization of the standard mixture of Gaussians, described two complementary approaches to the centering problem: either the likelihood is modified to be constant across the set of unidentifiable nuisance parameters (Vogt et al., 2010), or we intentionally make a decision and select the nuisance parameter in a preprocessing step (Adametz and Roth, 2011). The latter relied on external knowledge from phylogenetics to center the data via a simple tree construction. Although the result is only optimal if the distances satisfy the ultrametric inequality, the closest matching tree is often sufficient to produce very good clustering results at significantly lower cost. Leaving computational aspects aside, the most interesting aspect concerns theoretical links to ensemble methods like consensus clustering, which—similar to a binary tree—aggregate a large number of simple building blocks into complex decisions. We conjecture a deeper connection between tree decomposition, translation invariance and the aforementioned methods.

Not all invariances necessarily lead to meaningful models, though, be it due to excessive information loss or negligible practical benefits. As an example, we refer to the matrix T likelihood for clustering (see Appendix B), which permits elliptical clusters under the restriction of equal orientation. Unfortunately, a single-matrix distribution conflicts with arbitrary cluster alignments, thereby making it impossible to bring this assumption in line with our existing framework of distances.

Finally, the generalization from Gaussian to meta-Gaussian distributions (Rey and Roth, 2012) considerably expands the range of potential applications. The approach in (Adametz et al., 2014) enables the treatment of mixed continuous and discrete data in the information bottleneck (Tishby et al., 1999; Chechik et al., 2007), which is promising especially for the domain of pathway analysis. Here, it allows to identify subtle new patterns in gene expression data that would otherwise be lost when maximizing variance. We think of the meta-Gaussian information bottleneck as a module for feature extraction, which can be plugged into any existing pipeline whenever side information is available.

The discussion goes to show how the standard Gaussian models—which are appealing theoretically, but often limited practically—experience a substantial leap forward by invariances, thereby establishing the basis for a vast spectrum of new applications.

If we go beyond the presented work, there exist many fields of research that deeply rely on invariances, one of which is *computer vision*. As part of the greater goal to reach human levels of visual understanding, *object recognition* represents an elementary building block therein. In more detail, a task may require invariances against the camera viewing angle or lighting conditions, therefore, the literature is frequently concerned with *invariant representations* (Sohn and Lee, 2012; Monasse and Guichard, 2000) or in a more abstract sense, the *learning of invariances* (Wiskott and Sejnowski, 2002). From a human standpoint, the existence of a solution may be clear intuitively, yet, a concise mathematical formulation is often hard if not impossible. Regarding an invariance against lighting conditions, one might simply define a filtering step that only leaves outer object contours—a statistic in our terminology. Such an ad-hoc solution is valid in the sense of the nuisance parameter, but it might render the recognition task (or inference) exceedingly difficult; after all, the mentioned filter would discard a large amount of relevant information. The gist is that invariances of this kind demand a high degree of problem-specific knowledge.

The human cognitive system is already equipped with an enormous number of invariances (DiCarlo and Cox, 2007; Wallis and Rolls, 1997), which are vital for its remarkable performance (Quiroga et al., 2005) but also its stringent energy conservation. To cite one simple example related to visual perception, stimuli in the retina are only processed and forwarded if certain patterns are detected—irrelevant information is discarded at a very early stage. This has an interesting analogy to our implementation of invariances: As could be seen in the clustering domain, we sometimes have the choice to either incorporate a desired invariance directly into the core of the inference process *or* we shift it into a preprocessing step and receive a simpler model. In this regard, computational requirements are often a decisive factor, but a faster runtime may also involve approximative solutions or a loss of statistical robustness.

In a nutshell, the general concept of invariances has tremendous appeal for data analysis, because it concerns a wide array of facets: it can be used to generalize existing models, it is an essential requirement for many statistical problems and, most importantly, it deepens our understanding of the underlying probabilistic mechanisms.

Appendix A
Gaussian Mixture Model

A.1 Translation-Variant Likelihood

Let $X \sim \mathcal{N}_{p,n}(0_{p\times n}, (c^2\,\Sigma)\otimes I_n)$, then its inner product $S = XX^\top$ follows a *central* Wishart distribution $S \sim \mathcal{W}_p(n, c^2\,\Sigma)$. The log-likelihood in inverse covariance $W \equiv \Sigma^{-1}$ and scaling c reads

$$\ell(W,c) = \tfrac{n}{2}\log|c^{-2}\,W| - \tfrac{1}{2}\operatorname{tr}\{c^{-2}\,WS\}. \tag{A.1}$$

In order to remove nuisance parameter c, we first compute its maximum likelihood estimate

$$\frac{\partial}{\partial c}\ell(W,c) \overset{!}{=} 0 \qquad \Leftrightarrow \qquad \widehat{c}^2 = \tfrac{1}{np}\operatorname{tr}\{WS\} \tag{A.2}$$

and then insert it back into Eq. (A.1) to receive the scale-invariant profile log-likelihood

$$\ell_P(W) = \tfrac{n}{2}\log|W| - \tfrac{np}{2}\log\operatorname{tr}\{WS\}. \tag{A.3}$$

This coincides with the marginal log-likelihood, analog to scale invariance in the translation-invariant case.

A.1.1 Complexity

Evaluating the translation-variant likelihood requires fewer computations than its translation-invariant counterpart. Specifically for clustering, we have

$W = ZBZ^\top + I_p$ with $k \times k$ matrix B. Hereby the determinant becomes

$$|W| = |ZBZ^\top + I_p| = |BZ^\top Z + I_k| \tag{A.4}$$

and the trace transforms to

$$\mathrm{tr}\{WS\} = \mathrm{tr}\{ZBZ^\top S + S\} = \mathrm{tr}\{BZ^\top SZ\} + \mathrm{tr}\{S\}. \tag{A.5}$$

Besides precomputing term $\mathrm{tr}\{S\}$, it is possible to exploit incremental updates. This means, we can apply the same techniques analog to the translation-invariant likelihood:

- $B \in \mathrm{diag}$: Updating the determinant requires $\mathcal{O}(1)$, because it only involves diagonal terms. The trace builds on the diagonal elements of the block sums $Z^\top SZ$, which in total consumes $\mathcal{O}(p)$ for assigning one object to all k existing clusters. The assignment to a new cluster, however, is constant. Finally, generating and evaluating proposal B^* is $\mathcal{O}(k)$.
- $B \in \mathbb{S}_-$: The determinant is found by a QR decomposition for rank-1 updates in $\mathcal{O}(k^2)$; the trace term only requires the k diagonal entries of its argument, hence it suffices to update the block sums $Z^\top SZ$, such that an assignment of one object to one existing cluster consumes $\mathcal{O}(k)$. The most expensive operation is to compute B from scratch when A grows by one dimension or changes completely (as with proposal A^* and its counterpart of the inverse, B^*): this requires $\mathcal{O}(k^3)$, albeit not in the innermost loop. Evaluating the likelihood for B^* falls back to only $\mathcal{O}(k^2)$.

When all computations are combined as in Algorithm 1 and 2, the overall complexity can be summarized as shown in Table A.1. Unfortunately, the most flexible model, $A \in \mathbb{S}_+$, does not benefit from the simpler likelihood, however, the worst-case complexity for $A \in \mathrm{diag}$ is successfully reduced from cubic to square. In combination with tree construction, decomposition and centering to receive matrix $\widehat{S}$, a diagonal model for A is the middle ground between flexibility and computational cost. Hereby, the full pipeline consumes $\mathcal{O}(p^2)$.

		Ewens process	
A	B	truncated	standard
$\mathbb{R}$	diag	$\mathcal{O}(p^2)$	$\mathcal{O}(p^2)$
diag	diag	$\mathcal{O}(p^2)$	$\mathcal{O}(p^2)$
$\mathbb{S}_+$	$\mathbb{S}_-$	$\mathcal{O}(p^2 + pN^3)$	$\mathcal{O}(p^4)$

Table A.1: Overall complexity of clustering with Algorithm 1 and 2 when using the translation-variant model. diag, $\mathbb{S}_+$ and $\mathbb{S}_-$ represent the set of diagonal, symmetric positive-definite and negative-definite matrices, respectively. Truncation assumes $k_{\text{mix}} \leftarrow N \in \mathbb{N}_+$, where $k < N \leq p$.

A.2 Translation Invariance and its Relation to Kernel PCA

Translation invariance can be interpreted as a special way of centering the data and to this end, it is natural to ask how it compares to projections used in other methods. One popular example is *kernel principal component analysis* (*kernel PCA*) (Schölkopf et al., 1998), which constructs

$$Q_{\text{PCA}} \equiv I_p - \tfrac{1}{p}\mathbf{1}_p\mathbf{1}_p^\top \tag{A.6}$$

to receive the centered inner product matrix

$$S_{\text{PCA}} \equiv Q_{\text{PCA}} S Q_{\text{PCA}}^\top. \tag{A.7}$$

At first glance, this appears as a valid choice, because it satisfies $Q_{\text{PCA}}\mathbf{1}_p = \mathbf{0}_p$, such that all column means $\mathbf{1}_p\boldsymbol{w}^\top$, $\boldsymbol{w} \in \mathbb{R}^n$, are mapped to zero. However, in comparison to the projection for translation invariance,

$$Q = I_p - (\mathbf{1}_p^\top W \mathbf{1}_p)^{-1}\mathbf{1}_p\mathbf{1}_p^\top W,$$

we see that both coincide for $W \equiv I_p$. This means, kernel PCA assumes the special case of fully independent rows (= objects), something that is not meaningful in the context of clustering. Even though it can technically be applied to our domain, this particular choice of centering would introduce a serious bias.

The problem can also be seen from another point of view: each of the n columns in X is allowed to have an individual mean according to $M = \mathbf{1}_p \boldsymbol{w}^\top$. As a result, they are independently, but *not* identically distributed. In other words, the first column follows $\mathcal{N}_p(w_1 \cdot \mathbf{1}_p, \Sigma)$, while the second is distributed as $\mathcal{N}_p(w_2 \cdot \mathbf{1}_p, \Sigma)$, where $w_1 \neq w_2$. This property violates the assumptions of kernel PCA.

A.3 Translation Invariance and Cluster Geometry

The flexibility of the model $A \in \mathbb{S}_+$ leads to an insightful connection regarding translation invariance:

Theorem 2. *A full $k \times k$ positive-definite matrix $\widetilde{A}$ is flexible enough to model any cluster geometry without the need for translation invariance.*

Proof. Let mean matrix M consist of a cluster-geometry-defining part M_A and a translational part M_T, that is,

$$M = M_A + M_T, \tag{A.8}$$

where $M_T = \mathbf{1}_p \boldsymbol{w}^\top$, $\boldsymbol{w} \in \mathbb{R}^n$ and $M_A M_A^\top = Z A Z^\top$. Here, A is the inner product of all k cluster means,

$$A = \begin{bmatrix} \boldsymbol{m}_1^\top \\ \boldsymbol{m}_2^\top \\ \vdots \\ \boldsymbol{m}_k^\top \end{bmatrix} \begin{bmatrix} \boldsymbol{m}_1 & \boldsymbol{m}_2 & \dots & \boldsymbol{m}_k \end{bmatrix}, \tag{A.9}$$

with $\boldsymbol{m}_\bullet \in \mathbb{R}^n$. Then, we have

$$
\begin{aligned}
&\tfrac{1}{n}MM^\top \\
&\quad = \tfrac{1}{n}(M_A + M_T)(M_A + M_T)^\top && \text{(A.10)} \\
&\quad = \tfrac{1}{n}(M_A M_A^\top + M_A M_T^\top + M_T M_A^\top + M_T M_T^\top) && \text{(A.11)} \\
&\quad = \tfrac{1}{n}(ZAZ^\top + Z\boldsymbol{a}\mathbf{1}_p^\top + \mathbf{1}_p\boldsymbol{a}Z^\top + \boldsymbol{w}^\top\boldsymbol{w}\mathbf{1}_p\mathbf{1}_p^\top) && \text{(A.12)} \\
&\quad = \tfrac{1}{n}(ZAZ^\top + Z\boldsymbol{a}\mathbf{1}_k^\top Z^\top + Z\mathbf{1}_k\boldsymbol{a}^\top Z^\top + \boldsymbol{w}^\top\boldsymbol{w}Z\mathbf{1}_k\mathbf{1}_k^\top Z^\top) && \text{(A.13)} \\
&\quad = Z\tfrac{1}{n}(A + \boldsymbol{a}\mathbf{1}_k^\top + \mathbf{1}_k\boldsymbol{a}^\top + \boldsymbol{w}^\top\boldsymbol{w}\mathbf{1}_k\mathbf{1}_k^\top)Z^\top && \text{(A.14)} \\
&\quad = Z\widetilde{A}Z^\top. && \text{(A.15)}
\end{aligned}
$$

Eq. (A.12) used the fact that the mixed product of $M_A M_T^\top$ leads to $Z\boldsymbol{a}\mathbf{1}_p^\top$ with vector

$$
\boldsymbol{a} \equiv \begin{bmatrix} \boldsymbol{m}_1^\top \\ \boldsymbol{m}_2^\top \\ \vdots \\ \boldsymbol{m}_k^\top \end{bmatrix} \boldsymbol{w}. \tag{A.16}
$$

Next, Eq. (A.13) relied on the identity $Z\mathbf{1}_k = \mathbf{1}_p$. As can be seen above, there is always a positive-definite matrix $\widetilde{A}$ that captures *both* cluster-defining A *and* translation $\mathbf{1}_p\boldsymbol{w}^\top$ simultaneously. □

Due to this property, we can use the most flexible model for A and explicitly learn the column translation. There is, however, a practical implication in the sense that large shifts will dominate the values of A by $\frac{1}{n}\boldsymbol{w}^\top\boldsymbol{w}\mathbf{1}_k\mathbf{1}_k^\top$ in Eq. (A.14). Hence, in Bayesian inference, a prior for A should ideally peak at this scale, otherwise the exploration of the space is slow.

The question is whether to model the inner product of the cluster means on an absolute scale or only on the relative part that defines the cluster structure. Translation invariance explicitly distinguishes these two and allows us to treat the cluster means as if they were centered.

A.4 Variable Cluster Diameters

The focus on the covariance matrix was to model the inner product of the cluster means via different assumptions for matrix A, but at the same time the noise level was identical for all clusters. It is easy to imagine a dataset, which does not fit into this category and instead exhibits clusters of different diameter. Hence, the covariance model can be extended by writing

$$\Sigma = ZAZ^\top + \mathrm{diag}(Z\boldsymbol{a}), \tag{A.17}$$

where vector $\boldsymbol{a} \in \mathbb{R}^k$ defines the noise level for all clusters and $Z\boldsymbol{a}$ distributes its values according to the block structure. In the simplest case, we have $\boldsymbol{a} = \mathbf{1}_k$, thus leading to $\mathrm{diag}(Z\boldsymbol{a}) = \mathrm{diag}(\mathbf{1}_p) = I_p$.

The reader should notice that the added degree of freedom also entails further complications, since it must not conflict with scale invariance. This means, vector $\boldsymbol{a}$ should only define the noise level *relative* to the scale of matrix A. If this is ignored, it forces A to follow $\boldsymbol{a}$ in scale and consequently, the MCMC sampler would unnecessarily explore equivalent cluster configurations. A solution could be to enforce additional constraints, as for example

$$\sum_{j=1}^{k} a_j = 1. \tag{A.18}$$

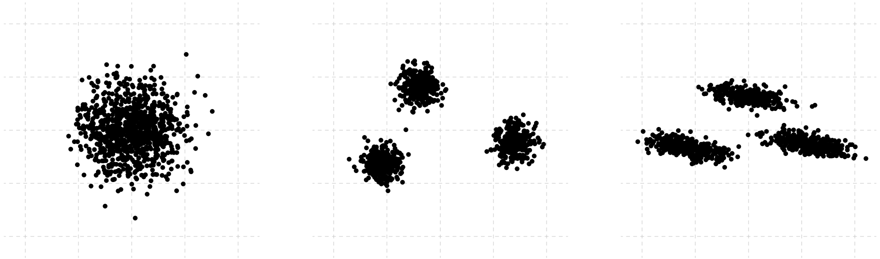
Variable cluster diameters are ultimately a valid extension, but it is important to keep in mind that (i) it requires a suitable and flexible prior, (ii) it must avoid redundancy and should *not* interfere with the scale of matrix A, and (iii) it has a practical impact on the sampling process due to the increased parameter space. For these reasons, we acknowledge the idea, yet maintain a fixed noise term I_p.

Appendix B

Gaussian Graphical Model

B.1 Clustering and Feature Correlation

Having developed the matrix T distribution as a generalization of the matrix normal, it is possible to revisit the Gaussian mixture model for distance matrices. Recall that the main difference between clustering and network inference lies in the parametrization of covariance matrix $\Sigma \equiv W^{-1}$, where clustering uses block structure $ZAZ^\top + I_p$ with A referring to the inner product of cluster centers. For a better understanding of feature correlation Ψ in the context of clustering, let us generate three configurations from the matrix normal distribution $\mathcal{N}_{p,n}(0_{p\times n}, \Sigma \otimes \Psi)$, as depicted in Fig. B.1.



$\Sigma = I_p$
$\Psi = I_n$

$\Sigma = ZAZ^\top + I_p$
$\Psi = I_n$

$\Sigma = ZAZ^\top + I_p$
$\Psi \neq I_n$

Figure B.1: A $(p = 1000) \times (n = 2)$ sample X of the matrix normal distribution is generated according to three different configurations.

If both covariances are identity matrices, the structure is trivial and all objects correspond to a single spherical cluster. The addition of a block-structured Σ partitions the objects into three groups, each being of spherical shape, and finally, $\Psi \neq I_n$ induces elliptical clusters. The important observation is that *all ellipses have the same orientation*, because Ψ affects all rows in the same fashion, regardless of their cluster assignment. Therefore, the model in Fig. B.1 (right) is indeed an improvement over strictly spherical clusters, albeit only a minor step forward.

From a clustering perspective, a more interesting dataset would comprise elliptical clusters with *different* orientation, see Fig. B.2 (left). Unfortunately, however, it is not possible to generate this from a single matrix normal distribution; here, we applied different feature correlations Ψ_1 and Ψ_2 to the clusters. If we ignore this warning and deliberately evaluate the matrix T likelihood, the outcome looks like Fig. B.2 (right). Clearly, the model cannot explain the observations properly, which is why the most pronounced elliptical cluster determines the global orientation and all remaining clusters are aligned accordingly. Moreover, there is only one way for the model to compensate the overall mismatch—introducing additional clusters.

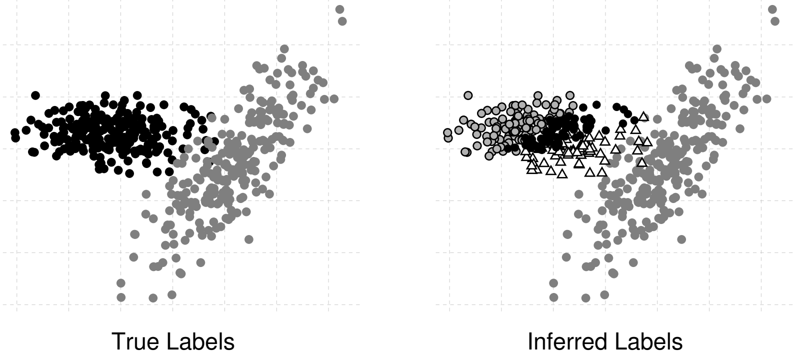


Figure B.2: Clustering $p = 500$ objects using the matrix T likelihood with block-parametrized W, where the data were generated by *two different* matrix normal distributions. This is a clear violation of the single-matrix assumption. The plots shows X (500×2).

In conclusion, feature correlation in the matrix normal distribution only has practical relevance for Gaussian graphical models; its transfer to clustering adds some flexibility, but the resulting benefit is minor.

Appendix C

Gaussian Copula Model

C.1 Distances and Meta-Gaussian Distributions

Regarding inference in pairwise distances, it first appears promising to decouple the dependence structure from everything pertaining only to the individual rvs, however, recall that this explicitly requires access to the realizations or at least their rank information. It is indeed possible to compute pairwise distances from a meta-Gaussian distribution, provided that all p marginals are continuous (that is, they permit a scalar product). Unfortunately, the resulting distance matrix lacks the foundation to separate dependence from marginals, or in other words: Sklar's theorem is not applicable.

A second argument concerns the mean model: The meta-Gaussian setup assumes that rows correspond to rvs, each following an individual marginal distribution. Hence, it accounts for *row means*

$$M = \boldsymbol{v}\mathbf{1}_n^\top, \tag{C.1}$$

where $\boldsymbol{v} \in \mathbb{R}^p$. This, however, conflicts with the definition of pairwise distances, which depends on above information as, for example, in the Gaussian mixture model. Instead, a distance matrix cancels *column means* $M = \mathbf{1}_p\boldsymbol{w}^\top$ with $\boldsymbol{w} \in \mathbb{R}^n$. Hence, the approaches are not compatible with each other, except for the trivial case when both means are zero.

C.2 Correlated and Independent Features

When developing the Gaussian graphical model for the domain of distances, a considerable portion was spent to explain the consequences of latent feature correlation, thereby leading to a model based on Mahalanobis distances rather than Euclidean distances. In that regard, we utilized the gene expression dataset of Sheffer et al. (2009) to demonstrate that a network of pathways is very likely subject to correlation among patients. However, revisiting the application using the information bottleneck, we required the patients to be independent, which is obviously a contradiction.

It is true that the meta-Gaussian extension for mixed data requires i.i.d. realizations, because it depends on the ranks to estimate the underlying correlation matrix. If this independence property were violated, the realizations would consequently change their order or become shuffled in the extreme case. Detecting the occurrence of such a condition is a non-trivial problem, especially for discrete rvs with only few levels.

In defense of the pathway experiments, note that the Gaussian graphical model was restricted to *only primary tumor patients* in conjunction with characteristic distributions across genes, while the information bottleneck used expression values from *all patient groups*. Regarding the second, we explicitly accounted for the effects of age and sex, thereby removing the main confounding factors that are known to skew our perception of gene expression data, see for example (Licastro et al., 2005; Baffert et al., 2004; Pal and Hurria, 2010; Söderlund et al., 2010). It appears plausible that these factors also contribute to latent feature correlation as observed between TiMT and TiWnet. Due to this, we argue that both experiments are sufficiently different and do not compete, but rather complement each other. The combination of both results leads to a more cohesive interpretation of pathways.

Bibliography

Aas, K. (2004). Modelling the Dependence Structure of Financial Assets: A Survey of Four Copulas. Technical report, Norsk Regnesentral, SAMBA.

Adametz, D., Rey, M., and Roth, V. (2014). Information Bottleneck for Pathway-Centric Gene Expression Analysis. In *Pattern Recognition*, Lecture Notes in Computer Science, pages 81–91. Springer.

Adametz, D. and Roth, V. (2011). Bayesian Partitioning of Large-Scale Distance Data. In *Advances in Neural Information Processing Systems 24*, pages 1368–1376. Curran Associates, Inc.

Adametz, D. and Roth, V. (2014). Distance-Based Network Recovery under Feature Correlation. In *Advances in Neural Information Processing Systems 27*, pages 775–783. Curran Associates, Inc.

Aldous, D. J. (1985). *Exchangeability and Related Topics*. Lecture Notes in Math. Springer.

Allen, G. and Tibshirani, R. (2010). Transposable Regularized Covariance Models with An Application to Missing Data Imputation. *Annals of Applied Statistics*, 4:764–790.

Baffert, F., Thurston, G., Rochon-Duck, M., Le, T., Brekken, R., and McDonald, D. M. (2004). Age-Related Changes in Vascular Endothelial Growth Factor Dependency and Angiopoietin-1-Induced Plasticity of Adult Blood Vessels. *Circulation Research*, 94(984):984–992.

Bhattacharyya, A. (1943). On a Measure of Divergence between two Statistical Populations Defined by their Probability Distributions. *Bulletin of the Calcutta Mathematical Society*, 35:99–109.

Boos, D. D. and Stefanski, L. A. (2003). *Essential Statistical Inference*. Springer.

Boudt, K., Cornelissen, J., and Croux, C. (2012). The Gaussian Rank Correlation Estimator: Robustness Properties. *Statistics and Computing*, 22(2):471–483.

Chechik, G., Globerson, A., Tishby, N., and Weiss, Y. (2007). Information Bottleneck for Gaussian Variables. *Journal of Machine Learning Research*, 6:165–188.

Cox, D. R. (2006). *Principles of Statistical Inference*. Cambridge University Press.

Cox, D. R. and Reid, N. (1987). Parameter Orthogonality and Approximate Conditional Inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, 49(1):1–39.

Cruddas, A. M., Reid, N., and Cox, D. R. (1989). A Time Series Illustration of Approximate Conditional Likelihood. *Biometrika*, 76:231–237.

Curtis, R. K., Oresic, M., and Vidal-Puig, A. (2005). Pathways to the Analysis of Microarray Data. *Trends in Biotechnology*, 23(8):429–435.

Daniels, M. and Pourahmadi, M. (2009). Modeling Covariance Matrices via Partial Autocorrelations. *Journal of Multivariate Analysis*, 100(10):2352–2363.

Darwin, C. R. (1859). *On the Origin of Species by Means of Natural Selection; or, The Preservation of Favoured Races in the Struggle for Life*. Murray, London.

Davison, A. C. (2008). *Statistical Models*. Cambridge University Press.

de Vos, A. and Francke, M. (2008). Bayesian Unit Root Tests and Marginal Likelihood. Technical report, Departement of Econometrics and Operation Researchs, VU University Amsterdam.

Defays, D. (1977). An Efficient Algorithm for a Complete Link Method. *The Computer Journal (British Computer Society)*, 20(4):364–366.

Diaz-Garcia, J., Gutierrez, J., and Mardia, K. (1997). Wishart and Pseudo-Wishart Distributions and Some Applications to Shape Theory. *Journal of Multivariate Analysis*, 63:73–87.

DiCarlo, J. J. and Cox, D. D. (2007). Untangling Invariant Object Recognition. *Trends in Cognitive Sciences*, 11(8):333–341.

Drier, Y., Sheffer, M., and Domany, E. (2013). Pathway-Based Personalized Analysis of Cancer. In *Proceedings of the National Academy of Sciences*, pages 6388–6393.

Dusenbery, D. B. (2009). *Living at Micro Scale: The Unexpected Physics of Being Small*. Harvard University Press.

Edwards, A. W. F. (1992). *Likelihood*. Johns Hopkins University Press.

Efron, B. (1998). R. A. Fisher in the 21st century. *Statistical Science*, 13(2):95–122.

Ein-Dor, L., Zuk, O., and Domany, E. (2006). Thousands of Samples are Needed to Generate a Robust Gene List for Predicting Outcome in Cancer. In *Proceedings of the National Academy of Sciences*, pages 5923–5928.

Felsenstein, J. (2003). *Inferring Phylogenies*. Sinauer Associates.

Fortini, P., Pascucci, B., Parlanti, E., D'Errico, M., Simonelli, V., and Dogliotti, E. (2003). The Base Excision Repair: Mechanisms and its Relevance for Cancer Susceptibility. *Biochimie*, 85(11):1053–1071.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse Inverse Covariance Estimation with the Graphical Lasso. *Biostatistics*, 9(3):432–441.

Garthwaite, P. H., Jolliffe, I. T., and Jones, B. (2002). *Statistical Inference*. Oxford University Press, second edition.

Genest, C. and Rivest, L. P. (2001). On the Multivariate Probability Integral Transformation. *Statistics & Probability Letters*, 53(4):391–399.

Gower, J. C. (1966). Some Distance Properties of Latent Root and Vector Methods Used in Multivariate Analysis. *Biometrika*, 53(3–4):325–338.

Gower, J. C. (1985). Properties of Euclidean and non-Euclidean Distance Matrices. *Linear Algebra and its Applications*, 67:81–97.

Gower, J. C. and Ross, G. J. S. (1969). Minimum Spanning Trees and Single Linkage Cluster Analysis. *Journal of the Royal Statistical Society*, 18(1):54–64.

Gupta, A. K. and Nagar, D. K. (1999). *Matrix Variate Distributions*. PMS Series. Addison-Wesley Longman.

Harville, D. (1974). Bayesian Inference for Variance Components using only Error Contrasts. *Biometrika*, 61:383–384.

Harville, D. (1977). Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. *Journal of the American Statistical Association*, 72(358):320–338.

Hoff, P. D. (2007). Extending the Rank Likelihood for Semiparametric Copula Estimation. *Annals of Applied Statistics*, 1(1):273.

Iranmanesh, A., Arashi, M., and Tabatabaey, S. M. M. (2010). On Conditional Applications of Matrix Variate Normal Distribution. *Iranian journal of Mathematical Sciences and Informatics*, 5:2:33–43.

Jain, A. K. (2010). Data Clustering: 50 Years beyond K-means. *Pattern Recognition Letters*, 31(8):651–666.

Jebara, T. and Kondor, R. (2003). Bhattacharyya and Expected Likelihood Kernels. In *Conference on Learning Theory*.

Kashima, H., Tsuda, K., and Inokuchi, A. (2003). Marginalized Kernels between Labeled Graphs. In *Proceedings of the Twentieth International Conference on Machine Learning*, pages 321–328. AAAI Press.

Kashima, H., Tsuda, K., and Inokuchi, A. (2004). *Kernels for Graphs*, pages 155–170. MIT Press.

Kollo, T. and von Rosen, D. (1995). Approximating by the Wishart Distribution. *Annals of the Institute of Statistical Mathematics*, 47:767–783.

Lay, D. (2011). *Linear Algebra and its Applications*. Addison Wesley, 4 edition.

Licastro, F., Candore, G., Lio, D., Porcellini, E., Colonna-Romano, G., Franceschi, C., and Caruso, C. (2005). Innate Immunity and Inflammation in Ageing: A Key for Understanding Age-Related Diseases. *Immunity & Ageing*, 2(1):8.

MacQueen, J. B. (1967). Some Methods for Classification and Analysis of Multivariate Observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297.

Mahé, P., Ralaivola, L., Stoven, V., and Vert, J.-P. (2006). The Pharmacophore Kernel for Virtual Screening with Support Vector Machines. *Journal of Chemical Information and Modeling*, 46(5):2003–2014.

McCullagh, P. (2003). A Note on Marginal Likelihood for Gaussian Models. Technical report, Department of Statistics, University of Chicago.

McCullagh, P. (2008). Marginal Likelihood for Parallel Series. *Bernoulli*, 14:593–603.

McCullagh, P. (2009). Marginal Likelihood for Distance Matrices. *Statistica Sinica*, 19:631–649.

McCullagh, P. and Yang, J. (2008). How Many Clusters? *Bayesian Analysis*, 3:101–120.

Mitchell, T. and Beauchamp, J. (1988). Bayesian Variable Selection in Linear Regression. *Journal of the American Statistical Association*, 83(404):1023–1032.

Monasse, P. and Guichard, F. (2000). Fast Computation of a Contrast-Invariant Image Representation. *IEEE Transactions on Image Processing*, 9(5):860–872.

Murtagh, F. (1984). Complexities of Hierarchical Clustering Algorithms: State of the Art. *Computational Statistics Quarterly*, 1:101–113.

Neal, R. (2000). Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics*, 9(2):249–265.

Nelsen, R. B. (2007). *An Introduction to Copulas*. Springer Series in Statistics. Springer.

Nourani, Y. and Andresen, B. (1998). A Comparison of Simulated Annealing Cooling Strategies. *Journal of Physics A: Mathematical and General*, 31(41):8373–8385.

Page, R. D. M. and Edward, H. C. (1998). *Molecular Evolution: A Phylogenetic Approach*. Blackwell Science.

Pal, S. K. and Hurria, A. (2010). Impact of Age, Sex, and Comorbidity on Cancer Therapy and Disease Progression. *Journal of Clinical Oncology*, 28(26):4086–4093.

Pearson, K. (1895). Note on Regression and Inheritance in the Case of two Parents. *Proceedings of the Royal Society of London*, 58(347–352):240–242.

Peltomäki, P. (2001). DNA Mismatch Repair and Cancer. *Mutation Research*, 488(1):77–85.

Pitman, J. (1995). Exchangeable and Partially Exchangeable Random Partitions. *Probability Theory and Related Fields*, 102(2):145–158.

Prabhakaran, S., Adametz, D., Metzner, K. J., Böhm, A., and Roth, V. (2013). Recovering Networks from Distance Data. *Journal of Machine Learning Research*, 92:251–283.

Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., and Fried, I. (2005). Invariant Visual Representation by Single Neurons in the Human Brain. *Nature*, 435(7045):1102–1107.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press.

Redner, R. and Walker, H. (1984). Mixture Densities, Maximum Likelihood, and the EM Algorithm. *SIAM Review 26*, 26:195–239.

Reid, N. (1995). The Roles of Conditioning in Inference. *Statistical Science*, 10(2):138–157.

Rey, M. and Roth, V. (2012). Meta-Gaussian Information Bottleneck. In *Advances in Neural Information Processing Systems 25*, pages 1925–1933.

Roth, V., Laub, J., Kawanabe, M., and Buhmann, J. M. (2003). Optimal Cluster Preserving Embedding of Non-Metric Proximity Data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12):1540–1551.

Schmidt, T. (2007). Coping with Copulas. *Copulas: From Theory to Application in Finance*, 108:3–34.

Schoenberg, I. J. (1937). On Certain Metric Spaces Arising from Euclidean Spaces by a Change of Metric and Their Imbedding in Hilbert Space. *Annals of Mathematics*, 38(2):787–793.

Schölkopf, B., Smola, A., and Müller, K. R. (1998). Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation*, 10(5):1299–1319.

Severini, T. A. (2001). *Likelihood Methods in Statistics*. Oxford University Press.

Sheffer, M., Bacolod, M. D., Zuk, O., Giardina, S. F., Pincas, H., Barany, F., Paty, P. B., Gerald, W. L., Notterman, D. A., and Domany, E. (2009). Association of Survival and Disease Progression with Chromosomal Instability: A Genomic Exploration of Colorectal Cancer. In *Proceedings of the National Academy of Sciences*, pages 7131–7136.

Sibson, R. (1973). SLINK: An Optimally Efficient Algorithm for the Single-Link Cluster Method. *The Computer Journal (British Computer Society)*, 16(1):30–34.

Sklar, A. (1959). *Fonctions de répartition à n dimensions et leurs marges*. Université Paris.

Söderlund, S., Granath, F., Broström, O., Karlen, P., Löfberg, R., Ekbom, A., and Askling, J. (2010). Inflammatory Bowel Disease Confers a Lower Risk of Colorectal Cancer to Females than to Males. *Gastroenterology*, 138(5):1697–1703.

Sohn, K. and Lee, H. (2012). Learning Invariant Representations with Local Transformations. In *Proceedings of the 29th International Conference on Machine Learning*, pages 1311–1318.

Sokal, R. R. and Michener, C. D. (1958). A Statistical Method for Evaluating Systematic Relationships. *University of Kansas Scientific Bulletin*, 28:1409–1438.

Steinhaus, H. (1956). Sur la Division des Corps Matériels en Parties. *Bulletin de l'Académie Polonaise des Sciences*, 4(12):801–804.

Stephens, M. (2000). Dealing with Label Switching in Mixture Models. *Journal of the Royal Statistical Society, Series B*, 62:795–809.

Tan, W. Y. and Gupta, R. P. (1982). On Approximating the Non-central Wishart Distribution by Central Wishart Distribution: A Monte Carlo Study. *Communications in Statistics-Simulation and Computation*, 11(1):47–64.

Tishby, N., Pereira, F. C., and Bialek, W. (1999). The Information Bottleneck Method. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377.

Tunnicliffe-Wilson, G. (1989). On the Use of Marginal Likelihood in Time Series Model Estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 51(1):15–27.

van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press.

Vogt, J. E., Prabhakaran, S., Fuchs, T. J., and Roth, V. (2010). The Translation-Invariant Wishart-Dirichlet Process for Clustering Distance Data. In *Proceedings of the 27th International Conference on Machine Learning*, pages 1111–1118.

Wallis, G. and Rolls, E. T. (1997). Invariant Face and Object Recognition in the Visual System. *Progress in Neurobiology*, 51(2):167–194.

Wiskott, L. and Sejnowski, T. J. (2002). Slow Feature Analysis: Unsupervised Learning of Invariances. *Neural Computation*, 14(4):715–770.

Wullschleger, S., Loewith, R., and Hall, M. N. (2006). TOR Signaling in Growth and Metabolism. *Cell*, 124(3):471–484.

Young, G. A. and Smith, R. L. (2005). *Essentials of Statistical Inference*. Cambridge University Press.

Yuan, M. and Lin, Y. (2007). Model Selection and Estimation in the Gaussian Graphical Model. *Biometrika*, 94(1):19–35.