Correlation and Dependency across the scales

Masterarbeit aus der Physik

Vorgelegt von
Wolfram Barfuß
7. Mai 2015

Institut für Theoretische Physik I
Friedrich-Alexander-Universität Erlangen-Nürnberg

BetreuerInnen: Prof. Dr. Ana-Sunčana Smith
Prof. Dr. Tomaso Aste

Abstract

We live in a time where the increasing availability of data poses novel challenges regarding the interpretation of the interdependencies inhered in the data. If we categorize a dataset along the scale of the ratio of the number of variables p over the number of observations T two problems asking for the interdependencies inherent in the datasets will be addressed. One question focuses on high dimensional (i.e. $p/T \geq 1$), the other one on low dimensional data (i.e. $p/T < 1$).

On the one side, the analysis of a high dimensional dataset requires the extraction of the underlying interaction graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Assuming that the data is describable by a stationary probability distribution we identify the multivariate Gaussian distribution as one particular result of a maximum Entropy ansatz. In theory, for these Gaussian distributions the key to the interaction graph is the inverse of the covariance matrix $\mathbf{\Sigma}^{-1}$ pointing to the area of Gaussian Markov Random Fields where zero entries in $\mathbf{\Sigma}^{-1}$ correspond to non-adjacent nodes in the interaction graph which both correspond to conditional independence relations of the variables. However, in practice, we have no access to the real covariance matrix $\mathbf{\Sigma}$ and consequently have to work with its maximum likelihood estimate, the sample covariance matrix $\mathbf{S}$. But due to noise and the finiteness of the time series $\mathbf{S}$ is inhering errors we need to address. Furthermore, from the principle of parsimony we want to introduce parameters (i.e. non-zero entries of the inverse covariance matrix $\mathbf{\Sigma}^{-1}$) to our model only if the data requires them. We summarize these challenges with the following two sub-problems: a) Choosing the set of edges $\mathcal{E}$ of the underlying graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and the zero/non-zero pattern of the estimate of the inverse covariance matrix $\mathbf{J}$, respectively. b) Assigning values to the chosen non-zero entries of $\mathbf{J}$. We propose the following approach to deal with these problems: 1) starting from filtering techniques applied to the covariance matrix, we filter out the most reliable information resulting in a set of edges $\mathcal{E}$ that form a decomposable graph; 2) we exploit the fact that for Gaussian Markov Random Fields the associated probability function factorizes in a neat and handy way and use the circumstance that marginal Gaussian distributions have covariance matrices which are sub-matrices of the global covariance matrix; 3) this allows us to invert the sample covariance matrix **lo**cally and compose these local parts to a sparse **glo**bal estimate $\mathbf{J}$. For this reason we call our approach *LoGo*. This method of inverting $\mathbf{S}$ is exact given that the underlying graph $\mathcal{G}$ is decomposable which is advantageous because it will not introduce new errors while at the same time it uses less information of $\mathbf{S}$ than the standard inversion process which makes it more stable to initial errors of $\mathbf{S}$. Based on these theoretical findings we construct improved filtering algorithms. Likelihood tests on sample and real data comparing *LoGo* we the

state-of-the-art method of estimating the inverse covariance matrix for high-dimensional data show promising benefits of our method.

On the other side a particular system of low-dimensional data has been studied. Here, we examine the packings of ellipses of different elongations into an area such that a specified area fraction of aggregated area of ellipses over the total background area is reached and present a systematic way how one can analyse these systems. After constructing the Voronoi tessellation we focus on the analysis of the correlation matrix between several morphological measures. In particular, we are able to study this system in a fruitful way by visualizing the correlation matrix along the elongation of the packed ellipses and the area fraction. From here, we are able to relate the characteristics of the phase-space correlation matrix to the actual form of the packings. Thus, we find that at high area fractions the elongation of the packed ellipses is of great importance regarding the structure of the packings whereas at low area fraction the packings are independent of the shape of the particles.

With both ways of analysing the interdependencies among the variables inside a set of data I show how a particular algorithm or a particular visualization, based on well established concepts like the covariance or the correlation coefficient, serve as powerful tools to extract the important information of the specific datasets.

Acknowledgements

I am deeply thankful to Ana Smith who convinced me of the advantages of a proper education in physics back in the spring of 2013. Surely, without this conversation this theses would not exist. Further I thank her for her engaging way of working and her useful comments, from which I could learn a lot.

I thank Sara Kaliman for our well going collaboration on the random packings including her work on the analysis code which was essential in the process. My thanks goes also to Jayant Pande who introduced me to the packing simulation and the whole group here in Erlangen where I always felt welcome. I want to thank Ira Röllinghoff who had always an open door to overcome the bureaucracy. I thank especially Jakov Lovrić with whom I collaborated as well on the random packings project very productively. In particular he focused on the assessment and fitting of the distributions of the individual morphological measures which I only broached in this work. On top of that, I am also very thankful that I had the chance to meet and work with Jakov in Zagreb. My thanks goes also to the group there which was more than hospitable. And I thank Gerd Schröder-Turk who, together with Ana Smith, set up the contact with Tomaso Aste at the University College of London (UCL) where I went to for seven month and worked out serious parts of this thesis. I thank Klaus Mecke in the name of the Elite Graduate Program of Physics for providing financial support for this journey. It was truly an experience I do not want to miss. For my time in London I want to thank Tiziana Di Matteo and Guido Massara for our collaboration on *LoGo* and the Network Filtering. I am especially thankful having had the opportunity to visit my first scientific conference with a poster of our work in Rhodes, Greece in July, 2014. Here, my thanks goes also to the Leonardo-Kolleg of the University of Erlangen-Nuremberg for financing this trip.

My very special thanks goes to Tomaso Aste from whom I learned a lot during my stay at UCL. The frequent lunch meetings we had were always a pleasure to me. I am very grateful for his supervision, his helpful comments, his way of communicating that really stimulates collaborative progress and his visit to Erlangen for my defence.

Contents

1 Introduction

1.1 Motivation

Unquestionable, *big data* is one of the buzzwords of our time. With internet companies collecting large amounts of user generated data a fundamental new situation for understanding social systems emerges. Other areas, where data becomes *big* include financial markets or biological systems, like for example gene expression data. One might come to the conclusion that this novel situation calls for novel tools to analyse the newly generated data. For physicists however, the sophisticated dealing with large amounts of data is nothing new. In fact, data - from experiments or observations - has always been the foundation of all proper physical models. Take for example the *Large Hadron Collider* at *CERN*, primary a physical experiment, it can be regarded easily as a big data analysis centre, since their production of data is so enormous that they have only the capacity to store a small proportion of the generated data for later analysis[1].

Nevertheless, the term *big data* may indicate the increasing importance of data for society, science and business. And thus proper tools to analyse and visualize data in a meaningful way become more and more relevant to gain a sophisticated understanding of the respective systems. Of special interest here are methods to extract or filter the most relevant information out of a big data set, since from its entirety nothing can be learned. On the other hand, it is often a question of the proper way of visualizing the given data that advances insight. Both approaches shall be addressed in the following.

To approach data in its most general form, let us regard them as a collection of variables with corresponding observations. Throughout this thesis the question of the mutual relationships between the variables is of special importance. From a scientific point of view exact relations between variables seem certainly preferable over probabilistic descriptions. Unquestionable, a parade example for such a deterministic theory is classical mechanics. Because of its great success in predicting the evolution of mechanical systems it had an enormous impact on other sciences like economics and hence on society and culture in general (Backhaus, 2012). On the other hand, with quantum mechanics probability theory is intertwined at natures micro scale. And although in principal chance disappears in the classical limit statistical physics has proven to be a useful description of many particle systems. Here, one can already read its statistical nature from its name. Moreover Jaynes (1957a,b, 2003) considers statistical mechanics merely as a

[1] `http://home.web.cern.ch/about/computing/processing-what-record`, accessed March 31, 2015

special form of statistical inference. So, probabilistic relations between variables have been, are and will be important descriptions along the way to understand the world a little bit better.

Now various tools, often called *statistics*, exist, to describe statistical relations, like for example the mean, median or the variance. However, in this work we will focus to a lesser extend on the description of one variable, like in the form of its probability distribution. The emphasis is rather put on the interactions between variables, where also several measures are in use to quantify the relations between two variables. One prominent example of a measure of association between two variables is the *Pearson product-moment correlation coefficient* which is still widely used and remarkably unchanged over 100 years after its invention. Lee Rodgers and Nicewander (1988) give a brief summary of the history of this correlation coefficient before they take into account thirteen ways to interpret it. So for example, as I will show below (Section 2.1) it indicates a linear relationship between two random variables. An illustrative and not too serious example is shown in Figure 1.1.

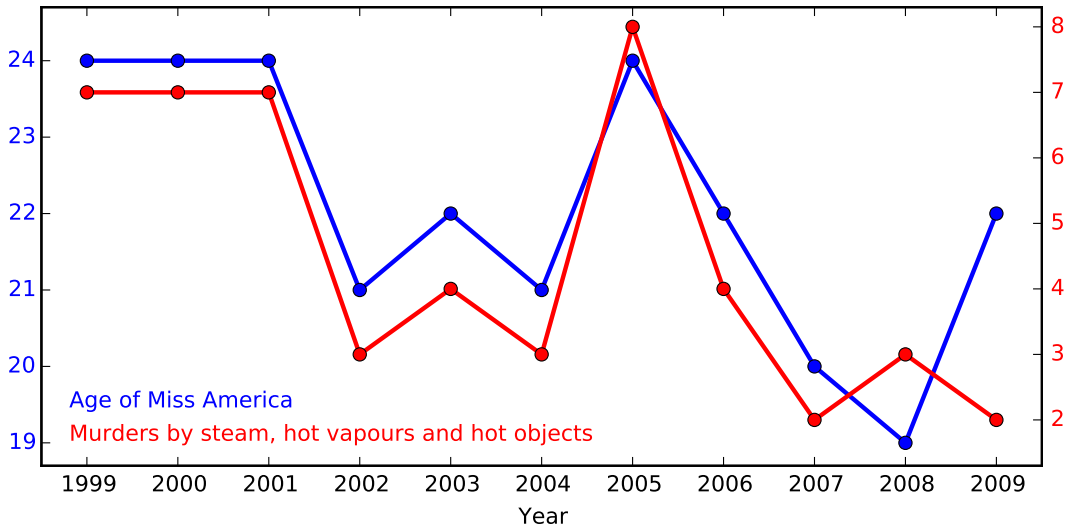

Figure 1.1: The age of Miss America (blue) is correlated to murders by steam, hot vapours and hot objects (red) with a correlation coefficient of 0.87. This is a relative high values since the correlation coefficient is bounded between -1 and 1, where 1 means perfect correlation, 0 means no correlation and -1 means perfect anti-correlation. This graph is shamelessly adapted from `http://www.tylervigen.com/`, accessed April 1, 2015

It might arise the question what the age of Miss America has to do with freaky murders. This Figure might be a good example of the prominent phrase *correlation does not imply causation*. And in fact it is much harder, if possible at all, to quantify the causal relationship by a single measure. Certainly up to now, it is the scientists task to detect causal relationships by proposing models that link the observed phenomena in a causal

way. Correlation measures may help along the way.

In general, our sets of data consist of more than just two variables. This brings us to another prominent feature of describing today's world: Networks, which are able to represent the mutual relationships of a set of variables. Originated from the mathematical field of graph theory it has been found fruitful to put an emphasis on the topology and the evolution of network structures (instead of using random graphs) to better describe many situation of the real world, such as the internet, neural and ecological systems (Albert and Barabási, 2002; Boccaletti et al., 2006). Since to my knowledge, there is no clear distinction between the notion of a *graph* and a *network*, I will not introduce one here and use these terms interchangeable, although it seems to me that *graphs* are more related to their mathematical treatment and *networks* point the light more to the real world. Throughout this work, it will be the combination of probability and network theory from which new insights arise.

In conclusion, interesting times lie ahead of us. The increasing availability of data enables us to extend the core methodology of physics to other areas than natural phenomena. At its core, it is the combination of empirical data - whether gained from experiments or observations - and sophisticated modelling. Only good models, i.e. models that can and are (at least partly) falsified against data from the real world, are able to provide causal explanations. But before one can start creating such models one has to deal with a proper analysis of the data. In this sense, I want take up the interest in novel data analysis techniques and contribute to this area with this work.

1.2 Research Question

In particular I will examine in this thesis how one can analyse sets of data in two fairly distinct situations. First of all, I now define a set of data as a data matrix $\mathbf{X} \in \mathbb{R}^{T \times p}$ where T denotes the number of observations and p the number of variables, such that $\mathbf{X}$ takes the form

		Variables					
		1	2	...	i	...	p
Observations	1	X_{11}	X_{12}	...	X_{1i}	...	X_{1p}
	2	X_{21}	X_{22}	...	X_{2i}	...	X_{2p}
	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
	t	X_{t1}	X_{t2}	...	X_{ti}	...	X_{tp}
	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
	T	X_{T1}	X_{T2}	...	X_{Ti}	...	X_{Tp}

Examples of these kinds of data include *multivariate data*, where the variables consists of p characteristics of interest, like age, sex, income, measured for T subjects and *time*

series data, where p variables are measured over T time steps to name just the two (Pourahmadi, 2013). So in this work, we are interested in general in the analysis of such data sets, in particular in the mutual relationships between the variables such that one can formulate the central research question of this thesis according to:

How can we extract, analyse and visualize the dependency structure between a set of variables, given some observations, such that we can improve our understanding about the system?

An import distinction with regard to this question concerns the relation between p and T. Usually it is assumed to have a large number of observations T compared to the number of variables p to get meaningful statistical results. Consequently the other case where T is small compared to p raises novel difficulties, that I will explain and address in the respective sections, since in this work both cases, the first *low dimensional* and the second *high dimensional* case will be addressed separately. So to summarize:

high dimensional	$p \gtrsim T$	Chapter 3,
low dimensional	$p \ll T$	Chapter 4.

This brings me to the outline of the thesis which I want to explain in the following.

1.3 Outline

This work consists of three main chapters. In Chapter 2 I will provide the basics and foundations regarding the notions of dependence, covariance and correlation (Section 2.1). Further, in Section 2.2, I will show how to derive a probability model out of just four assumptions and requirements, respectively. These would be first that a given set of data is describable by a probability distribution. Second, we assume that we have no more knowledge that the data provides and will therefore take a maximum Entropy ansatz. Third, we treat the probability distribution as stationary, i.e. we assume it will not change along the data series and fourth, we require that our model recovers the first and second moments of the data. One can fairly question these assumptions from a scientific standpoint. But since the result will be the well known and widely used multivariate Gaussian distribution this part can be seen as an illustration which assumptions flow into this model.

In the following two chapters I will concentrate on the two particular cases of data in the spectrum of number of variables and number of observations. Chapter 3 will concentrate on the high dimensional case where our data matrix consists of relatively many variables and comparable few observations. In Section 3.1 I will stress why usual statistical methods fail. Consequently, we will develop a novel method in Section 3.3 to extract the dependency relation between the variables after having introduced the relevant theoretical background of the merger of probability and network theory in Section

3.2. In particular, I show that this merger, taking seriously, gives us guidance how to find parameters of our maximum Entropy model of Section 2.2 by combining the relevant aspects of both theories. Section 3.4 will test and discuss our new method with artificial and real-world data before we summarize this chapter in Section 3.5.

Subsequently, in Chapter 4 I will provide an example how one can analyse the dependency structure of a low dimensional data set. In particular, I focus on the morphological properties of two dimensional random packings of ellipses with respect to different elongations of the ellipses and area fractions of area covered by the ellipses over the total background area. In Section 4.1 I will motivate our interest in these kind of systems from a biological perspective. Section 4.2 gives details about the data generation and analyses methods before I outline the results in Section 4.3. In particular, we are able to relate specific characteristics of our technical dependency analysis to the actual packed systems, which confirms the validity of our method. Here, I summarize our findings of this chapter in Section 4.3 before I give a general conclusion in Chapter 5.

2 Foundations

2.1 (In)Dependency & Correlation

Let us begin with some core notions of this work.

Joint & marginal probability. Let x and y be two proper random variables. For the probability to observe x and y in conjunction, one writes the *joint probability* as

$$f(x, y). \tag{2.1}$$

On the other hand, the *marginal probability* denotes the probability to observe x independently from y,

$$f_x(x) \tag{2.2}$$

and the probability to observe y independently from x,

$$f_y(y), \tag{2.3}$$

respectively.

(In)Dependence. One says that x is *independent* from y and similarly that y is independent from x (abbreviated with $x \perp y$) if and only if one can write equivalently

- $f(x, y) = f_x(x) f_y(y)$,
- $f(x|y) = f_x(x)$,
- $f(y|x) = f_y(y)$.

Note that the equivalence of the three statements above is in fact the well known *Bayes theorem*,

$$f(x|y) = \frac{f(x, y)}{f_y(y)} = f_x(x) \tag{2.4}$$

In the contrary, if the equal signs do not hold, we speak of *dependent* variables.

The same statements of independence hold true for the cumulative distributions. x and y are independent if and only if

$$F_{x,y}(x, y) = F_x(x) F_y(y) \tag{2.5}$$

since,

$$F_{x,y}(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(x,y)dxdy \tag{2.6}$$

$$= \int_{-\infty}^{x}\int_{-\infty}^{y} f_x(x)f_y(y)dxdy \tag{2.7}$$

$$= \int_{-\infty}^{x} f_x(x)dx \int_{-\infty}^{y} f_y(y)dy = F_x(x)F_y(y) \tag{2.8}$$

Covariance. Now, to measure the dependence between the two random variables, let us take the ansatz

$$F_{x,y}(x,y) - F_x(x)F_y(y). \tag{2.9}$$

So our dependency measure would indeed vanish, if x and y are independent. In fact Höffding (1940) proved, that Equation 2.9, integrated over the entire support gives the well known *covariance*

$$\mathrm{Cov}(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \big(F_{x,y}(x,y) - F_x(x)F_y(y)\big)dxdy \tag{2.10}$$

which is more commonly defined as

$$\mathrm{Cov}(x,y) = \mathbb{E}[(x - \mathbb{E}[x])(y - \mathbb{E}[y])] \tag{2.11}$$

$$= \mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y]. \tag{2.12}$$

Note, that $\mathrm{Cov}(x,x)$ equals the variance of variable x. As we already observed, if the two variables are independent then $\mathrm{Cov}(x,y) = 0$. However, the converse is not true in general. There may be true dependencies despite a vanishing covariance. Therefore Rényi (1959) formulated seven postulates which should be fulfilled by a suitable measure of dependence and consequently various other measures exist. Nevertheless, the covariance plays an important role (as we will also see below) and thus, we proceed with the closely related correlation.

Correlation. The *Person product-moment correlation coefficient* $\mathrm{Corr}(x,y)$ is defined as

$$\mathrm{Corr}(x,y) = \frac{\mathrm{Cov}(x,y)}{\sigma(x)\sigma(y)} \tag{2.13}$$

where $\sigma(x) = \sqrt{\mathrm{Cov}(x,x)}$ denotes the standard deviation. So the correlation coefficient can be regarded as a normalized covariance which is obtained by transforming the random variables x, y to be measured in units of their inverse standard deviations $\sigma(x)^{-1}, \sigma(y)^{-1}$. Therefore the correlation is a bounded measure of dependency $-1 \leq \mathrm{Corr}(x,y) \leq 1$, which makes it often very useful in practice.

But what kind of dependence does the correlation cover? In the following I show how the correlation and the covariance are related to *linear* dependence. In order to achieve

this goal let me briefly introduce the concept of *regression*. Here we assume that we can model the dependence of variable y from variable x by a function

$$g(x) = y. \tag{2.14}$$

The quality of this assumption gets quantified by the loss function

$$L(y, g(x)) = \mathbb{E}\big[(y - g(x))^2)\big], \tag{2.15}$$

which shall be minimized by the best fitting function

$$\hat{g}(x) = \underset{g}{\operatorname{argmin}} \left(\mathbb{E}\big[L\big(y, g(x)\big)\big]\right). \tag{2.16}$$

If we now further assume g to be a *linear* function $y = g(x) = a + bx$ we obtain $\hat{g}$ by minimising

$$\begin{aligned} L(y, g(x)) &= \mathbb{E}\big[\big(y - g(x)\big)^2\big] \\ &= \mathbb{E}\big[\big(y - a - bx\big)^2\big] \\ &= \mathbb{E}[y^2 + a^2 + b^2x^2 - 2ay - 2bxy + 2abx] \\ &= \mathbb{E}[y^2] + a^2 + b^2\mathbb{E}[x^2] - 2a\mathbb{E}[y] - 2b\mathbb{E}[xy] + 2ab\mathbb{E}[x]. \end{aligned} \tag{2.17}$$

So,

$$\frac{dg}{da} = 2a - 2\mathbb{E}[y] + 2b\mathbb{E}[x] \stackrel{!}{=} 0 \tag{2.18}$$

$$\Rightarrow \quad a = \mathbb{E}[y] - b\mathbb{E}[x] \tag{2.19}$$

And

$$\begin{aligned} \frac{dg}{db} &= 2b\mathbb{E}[x^2] - 2\mathbb{E}[xy] + 2a\mathbb{E}[x] \\ &= 2b\mathbb{E}[x^2] - 2\mathbb{E}[xy] + 2(\mathbb{E}[y] - b\mathbb{E}[x])\mathbb{E}[x] \stackrel{!}{=} 0 \end{aligned} \tag{2.20}$$

$$\begin{aligned} \Rightarrow \quad b &= \frac{\mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y]}{\mathbb{E}[x^2] - \mathbb{E}[x]\mathbb{E}[x]} \\ &= \frac{\mathrm{Cov}(x, y)}{\sigma_x^2} \end{aligned} \tag{2.21}$$

Thus,

$$b \cdot \frac{\sigma_x}{\sigma_y} = \mathrm{Corr}(x, y) \tag{2.22}$$

From Equation 2.22 we can state that the correlation coefficient equals the slope of the regression line of the normalized variables. Thus, the correlation is a measure of linear dependence.

To summarize our findings about the correlation coefficient we can state that

- $\text{Corr}(x, y) = 0$ if x and y are independent
- $\text{Corr}(x, y) = 1$ if x and y are perfectly positively linearly dependent
- $\text{Corr}(x, y) = -1$ if x and y are perfectly negatively linearly dependent

However, if $\text{Corr}(x, y) = 0$ the two random variables x and y still may be non-linearly dependent, e.g. $y = x^2$. For eleven additional ways on how to interpret the correlation coefficient the reader is referred to Lee Rodgers and Nicewander (1988).

Multivariate Gaussians. So far we have obtained the relation

$$\text{if } \; x_i \perp x_j \quad \Rightarrow \quad \text{Cov}(x_i, x_j) = 0 \tag{2.23}$$

and noticed that in general the converse does not hold true for a set of random variables $\{x_1, \ldots, x_p\}$. However, in the case where these random variables follow a multivariate Gaussian distribution

$$f(x_1, \ldots, x_p) = \frac{1}{\sqrt{(2\pi)^p \det \boldsymbol{\Sigma}}} \exp\Big(-\frac{1}{2} \sum_{i,j=1}^{p} (x_i - \mu_i) \Sigma^{-1}_{ij} (x_j - \mu_j) \Big) \tag{2.24}$$

where $\boldsymbol{\mu} \in \mathbb{R}^p$ denotes the mean and $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$ denotes the covariance matrix, it holds that

$$x_i \perp x_j \quad \Leftrightarrow \quad \text{Cov}(x_i, x_j) = 0. \tag{2.25}$$

To see this consider the following property of the multivariate Gaussian distribution (Rue and Held, 2005). If we divide $\mathbf{x} = (x_1, \ldots, x_p)$ into two parts, $\mathbf{x} = (\mathbf{x}_A^{\mathrm{T}}, \mathbf{x}_B^{\mathrm{T}})^{\mathrm{T}}$, and split $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ accordingly,

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_A \\ \boldsymbol{\mu}_B \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{AA} & \boldsymbol{\Sigma}_{AB} \\ \boldsymbol{\Sigma}_{BA} & \boldsymbol{\Sigma}_{BB} \end{pmatrix}, \tag{2.26}$$

we obtain that $\mathbf{x}_A$ is distributed according to a multivariate Gaussian distribution as well, but with covariance matrix $\boldsymbol{\Sigma}_{AA}$ being the respective sub-matrix of $\boldsymbol{\Sigma}$, thus

$$\mathbf{x}_A \sim \mathcal{N}(\boldsymbol{\mu}_A, \boldsymbol{\Sigma}_{AA}). \tag{2.27}$$

This is true for any subdivision. Now consider the joint distribution $f(x_i, x_j)$. Consequently it has the covariance matrix

$$\boldsymbol{\Sigma}_{\{i,j\}} = \begin{pmatrix} \Sigma_{ii} & \Sigma_{ij} \\ \Sigma_{ji} & \Sigma_{jj} \end{pmatrix}. \tag{2.28}$$

So if $\text{Cov}(x_i, x_j) = \Sigma_{ij} = 0$ than the off-diagonal entries of the corresponding inverse matrix $\boldsymbol{\Sigma}^{-1}_{\{i,j\}}$ must be zero as well and the probability distribution factorizes according

to

$$f(x_i, x_j) = \frac{1}{\sqrt{(2\pi)^2 \det \mathbf{\Sigma}_{\{i,j\}}}} \exp\Big(-\frac{1}{2}\big(\Sigma_{ii}^{-1}(x_i - \mu_i)^2 + \Sigma_{jj}^{-1}(x_j - \mu_j)^2\big)\Big) \tag{2.29}$$

$$= \frac{1}{\sqrt{2\pi\Sigma_{ii}}} \exp\Big(-\frac{1}{2}\Sigma_{ii}^{-1}(x_i - \mu_i)^2\Big) \frac{1}{\sqrt{2\pi\Sigma_{jj}}} \exp\Big(-\frac{1}{2}\Sigma_{jj}^{-1}(x_j - \mu_j)^2\Big) \tag{2.30}$$

$$= f(x_i) f(x_j), \tag{2.31}$$

which states in fact, that x_i and x_j are independent.

Example. Let us consider an example of a multivariate Gaussian distribution consisting of three variables $\mathbf{x} = (x_1, x_2, x_3)$, mean $\boldsymbol{\mu} = (0, 0, 0)$ and covariance matrix

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & 0 & 0.5 \\ 0 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{pmatrix}. \tag{2.32}$$

From that we can immediately see that $x_1 \perp x_2$, since $\Sigma_{12} = 0$ and therefore $f(x_1, x_2) = f(x_1) f(x_2)$. However, this is not the full picture. If we include x_3 into the analysis we find that the joint probability distribution $f(x_1, x_2, x_3)$ does not factorize in any way, since the inverse covariance matrix

$$\mathbf{\Sigma}^{-1} = \begin{pmatrix} 1.5 & 0.5 & -1 \\ 0.5 & 1.5 & -1 \\ -1 & -1 & 2 \end{pmatrix}. \tag{2.33}$$

contains no zero entries. So in our example system x_1 is independent from x_2 but only if we do not look at x_3. This maybe unsatisfying result points us to the notion of conditional independence.

Conditional independence. Two random variables x and y are said to be conditionally independent given a random variable z iff we can write

$$f(x, y|z) = f_x(x|z) f_y(y|z). \tag{2.34}$$

This is also denoted as $x \perp y|z$. Later we will identify the zero/non-zero structure of the inverse covariance matrix with conditional independence relations of the random variables of a multivariate Gaussian distribution, which will provide the basis for a key result of this work.

2.2 A Maximum Entropy Model

In the following, I want to show the assumptions one makes deriving the well-known multivariate Gaussian distribution. Mainly they consist of (1) that the data is describable by a probability density function, (2) that one does not know more about the data than the data and hence, uses a maximum Entropy ansatz, (3) that the probability density function is stationary, i.e. that its time derivative vanishes and (4) that one requires the probability density function to recover the statistical moments of the data.

Problem statement. As before, let us consider T observations of p variables such that our multivariate series has the form $X_{ti} \in \mathbb{R}$ with $i = 1 \dots p$ and $t = 1 \dots T$. If one assumes that this data is describable by a joint probability density function $f(\mathbf{x}|\mathbf{J})$ where $\mathbf{x} \in \mathbb{R}^p$ denotes the random vector corresponding to the p variables and $\mathbf{J}$ denotes a set of parameters, two questions arise:

- Which form does the joint probability density function have?
- What are the values of the parameters $\mathbf{J}$?

Entropy maximization. We derive such a probability density function from the principle of maximum Entropy. The Entropy of such a function is defined as

$$H(f(\mathbf{x})) = -\int dx^p f(\mathbf{x}) \ln f(\mathbf{x}), \tag{2.35}$$

where $\int dx^p$ denotes $\int_{-\infty}^{\infty} dx_1 \cdots \int_{-\infty}^{\infty} dx_p$ and $f(\mathbf{x})$ is short for $f(\mathbf{x}|\mathbf{J})$.

The idea is now to maximize the Entropy under some constraints indicating important characteristics of the data. This approach has similarities with the maximum Entropy principle of statistical mechanics (Jaynes, 1957a,b) and is widely used in many applications, such as detecting neural (Schneidman et al., 2006) and genetic interactions (Lezon et al., 2006).

Here, we maximize the Entropy under the following constraints:

0) $1 \stackrel{!}{=} \int dx^p f(\mathbf{x})$

1) $\langle x_i \rangle = \frac{1}{T} \sum_{t=1}^{T} X_{ti} \stackrel{!}{=} \int dx^p x_i f(\mathbf{x})$

2) $\langle x_i x_j \rangle = \frac{1}{T} \sum_{t=1}^{T} X_{ti} X_{tj} \stackrel{!}{=} \int dx^p x_i x_j f(\mathbf{x})$

where constraint **0)** is the normalization condition and constraints **1)** and **2)** represent the condition that the first and second sample moments of the data should be recovered by the probability density function $f(\mathbf{x})$.

With the respective Lagrange multipliers $J^{(0)}, \mathbf{J}^{(1)}, \mathbf{J}^{(2)}$ plus additional factors $\frac{1}{k!}$ for each Lagrange multiplier $J^{(k)}$ the constrained entropy $\hat{H}$ looks like

$$\hat{H} = \int dx^p \left[-f \ln f - \frac{1}{0!} J^{(0)} f - \frac{1}{1!} \sum_i J_i^{(1)} x_i f - \frac{1}{2!} \sum_{ij} J_{ij}^{(2)} x_i x_j f \right], \tag{2.36}$$

where we abbreviate $f(\mathbf{x})$ with f.

Assuming that the data is describable by a stationary probability density function, the derivation of $\hat{H}$ with respect to f gives us the stationary condition

$$\frac{d\hat{H}}{df} = \int dx^p \left[-\ln f - 1 - J^{(0)} - \sum_i J_i^{(1)} x_i - \frac{1}{2} \sum_{ij} J_{ij}^{(2)} x_i x_j \right] \stackrel{!}{=} 0, \tag{2.37}$$

which is fulfilled with

$$-\ln f - 1 - J^{(0)} - \sum_i J_i^{(1)} x_i - \frac{1}{2} \sum_{ij} J_{ij}^{(2)} x_i x_j = 0 \tag{2.38}$$

or equivalently

$$f = \exp\left(-1 - J^{(0)} - \sum_i J_i^{(1)} x_i - \frac{1}{2} \sum_{ij} J_{ij}^{(2)} x_i x_j \right). \tag{2.39}$$

With

$$\frac{d^2\hat{H}}{df^2} = -\int dx^p \frac{1}{f} < 0, \tag{2.40}$$

since $f(\mathbf{x}) > 0$ for all $\mathbf{x}$, we show that we have found a maximum.

Upon defining $Z = \exp(1 + J^{(0)})$ we get a probability density function

$$f = \frac{1}{Z} \exp\left(-\sum_i J_i^{(1)} x_i - \sum_{ij} \frac{1}{2} J_{ij}^{(2)} x_i x_j \right) \tag{2.41}$$

which has striking similarity with the Gibbs-Boltzmann distribution of the well known Ising model. Note that there are important differences between the case where the variables $X_{ti} \in \mathbb{R}$ like in our setting and the case where the variables are elements of a restricted set M, for example $M = \{-1, 1\}$ referring to the original Ising model. In the latter scenario, normalization is always guaranteed independent of $J^{(1)}, J^{(2)}$ which is not the case in the first scenario due to the integration over the whole space. We want to take this into account by referring with the term Ising model only to the second case

with restricted variables. The problem of finding the right parameters $J^{(1)}, J^{(2)}$ from the moments $\langle x_i \rangle$ and $\langle x_i x_j \rangle$ is called the Inverse Ising problem (Roudi et al., 2009; Sessak and Monasson, 2009; Ravikumar et al., 2010; Ricci-Tersenghi, 2012; Aurell and Ekeberg, 2012).

Taylor Expansion. There is no reason why one should take only the first and the second moments into account while maximizing the Entropy. Theoretically it is possible to extend the constraints up to an arbitrary Kth moment

$$\textbf{K)} \quad < \underbrace{x_i \dots x_l}_{K \text{ times}} >= \tfrac{1}{T} \sum_{t=1}^{T} \underbrace{X_{ti} \dots X_{tl}}_{K \text{ times}} \stackrel{!}{=} \int dx^p \underbrace{x_i \dots x_l}_{K \text{ times}} p(\mathbf{x})$$

which leads to the probability density function

$$f = \frac{1}{Z} \exp\left(- \sum_i J_i^{(1)} x_i - \frac{1}{2} \sum_{ij} J_{ij}^{(2)} x_i x_j \cdots - \frac{1}{K!} \sum_{\underbrace{i \dots l}_{K \text{ times}}} J^{(K)}_{\underbrace{i \dots l}_{K \text{ times}}} \underbrace{x_i \dots x_l}_{K \text{ times}} \right) \tag{2.42}$$

where $\mathbf{J}^{(k)}$ are tensors of order k, i.e. $\mathbf{J}^{(1)}$ is a vector and $\mathbf{J}^{(2)}$ is a matrix as can easily be seen in Equation 2.41.

Note further that due to the symmetry of the moments all tensors $\mathbf{J}^{(k)}$ are symmetric. This gives us the possibility to express the negative exponent

$$U^{(K)} := \sum_i J_i^{(1)} x_i + \frac{1}{2} \sum_{ij} J_{ij}^{(2)} x_i x_j \cdots + \frac{1}{K!} \sum_{\underbrace{i \dots l}_{K \text{ times}}} J^{(K)}_{\underbrace{i \dots l}_{K \text{ times}}} \underbrace{x_i \dots x_l}_{K \text{ times}} \tag{2.43}$$

as a Taylor expansion up to the Kth order with

$$J^{(k)}_{\underbrace{i \dots l}_{k \text{ times}}} := \frac{\partial^k U}{\underbrace{\partial_i \cdots \partial_l}_{k \text{ times}}} |_{\mathbf{x}=0} \tag{2.44}$$

where $U = U(x_1 \dots x_p)$ is a scalar function and $\int dx^p e^{-U(\mathbf{x})}$ converges such that one can write

$$f(\mathbf{x}) = \frac{1}{Z} e^{-U(\mathbf{x})} \tag{2.45}$$

with the partition function $Z := \int dx^p e^{-U(\mathbf{x})}$. Note that Equation 2.45 is almost written in the general form of the *exponential family*. This points already to Section 3.2 where we will discuss graphical models, that can often be viewed naturally as exponential families (Wainwright and Jordan, 2008).

3 Network Filtering

3.1 Introduction

In this chapter we want to study high dimensional data sets, i.e. where the number of observations T is comparable or even smaller than the number of variables p. As I will point out, the key to a meaningful estimate of the model's parameters lies in the filtering of the data to obtain the relevant information. In Section 2.2 we derived the well-known multivariate Gaussian distribution from the more general principle of maximum Entropy and saw that in this model, the inverse of the covariance matrix coincides with the model's parameters. However, in the practical case, that we are dealing with here, we have only some finite set of data and therefore no access to the true covariance matrix. If we had the original covariance matrix, the task of finding the parameters of the model would be a very easy one, since only a simple matrix inversion would be required. So let us look at the empirical or sample covariance matrix $\mathbf{S}$ more in detail. Generally, it is defined as

$$S_{ij} = \frac{1}{T}\sum_{t=1}^{T}\left(X_{ti} - \left(\frac{1}{T}\sum_{t=1}^{T}X_{ti}\right)\right)\left(X_{tj} - \left(\frac{1}{T}\sum_{t=1}^{T}X_{tj}\right)\right). \tag{3.1}$$

Note that $\frac{1}{T}\sum_{t=1}^{T}X_{ti}$ computes the sample mean of the ith variable. But in what way are these *sample estimates* of the true mean and the true covariance matrix distinguished over other possible estimators?

Maximum Likelihood Estimators. Since, in fact, these are so called *maximum likelihood estimators*, I want to introduce in the following the general idea behind maximum likelihood. Let us consider in total generality a probability distribution $f(\mathbf{x}|\mathbf{J})$ of a random variable $\mathbf{x}$ with some set of parameters $\mathbf{J}$. The maximum likelihood estimator asks for the most probable set of parameters $\hat{\mathbf{J}}$ given a set of observations $\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_T\}$. According to Bayes' theorem we can rearrange $f(\mathbf{x}|\mathbf{J})$ according to

$$f(\mathbf{x}|\mathbf{J}) = \frac{f(\mathbf{J}|\mathbf{x})f(\mathbf{x})}{f(\mathbf{J})}. \tag{3.2}$$

Now we interpret $f(\mathbf{J}|\mathbf{x})$ as the probability that the parameters $\mathbf{J}$ have produced the observed data $\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_T\}$, i.e. $f(\mathbf{J}|\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_T\})$. The idea is now to maximize this probability to find the most likely parameters, given the observed data. Assuming

initially a uniform distribution for the parameters $f(\mathbf{J})$ we get for the maximum likelihood estimator $\hat{\mathbf{J}}$:

$$\begin{aligned}\hat{\mathbf{J}} &= \underset{\mathbf{J}}{\operatorname{argmax}}\, f(\mathbf{J}|\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_T\}) && (3.3)\\ &= \underset{\mathbf{J}}{\operatorname{argmax}} \left(\frac{f(\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_T\}|\mathbf{J}) f(\mathbf{J})}{f(\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_T\})} \right) && (3.4)\\ &= \underset{\mathbf{J}}{\operatorname{argmax}} \left(f(\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_T\}|\mathbf{J}) f(\mathbf{J}) \right) && (3.5)\\ &= \underset{\mathbf{J}}{\operatorname{argmax}}\, f(\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_T\}|\mathbf{J}), && (3.6)\end{aligned}$$

where we used Equation 3.2 from (3.3) to (3.4). From (3.4) to (3.5) we exploit that $f(\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_T\})$ is independent from $\mathbf{J}$ and lastly we assumed from (3.5) to (3.6) that $\mathbf{J}$ is initially uniformly distributed and $f(\mathbf{J})$ therefore a constant. Now we treat $f(\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_T\}|\mathbf{J})$ as the joint probability distribution

$$f(\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_T\}|\mathbf{J}) = f(\mathbf{X}_1|\mathbf{J}) \cdot f(\mathbf{X}_2|\mathbf{J}) \cdot \ldots \cdot f(\mathbf{X}_T|\mathbf{J}), \tag{3.7}$$

which is already the *likelihood function* to be maximized,

$$\mathcal{L}(\mathbf{J};\, \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_T) = f(\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_T\}|\mathbf{J}) = \prod_{i=1}^{T} f(\mathbf{X}_i|\mathbf{J}). \tag{3.8}$$

Let us now apply Equation 3.8 to the multivariate Gaussian distribution (Equation 2.24) to obtain an expression for the likelihood function that uses only matrix calculus. Direct insertion yields

$$\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{t=1}^{T} (2\pi)^{-p/2} \left(\det \boldsymbol{\Sigma}\right)^{-1/2} \exp\left(-\frac{1}{2} (\mathbf{X}_t - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{X}_t - \boldsymbol{\mu}) \right). \tag{3.9}$$

Replacing $\boldsymbol{\mu}$ by its maximum likelihood estimate, which is the sample mean $\langle \mathbf{X} \rangle$, where $\langle X \rangle_i = \frac{1}{T} \sum_{t=1}^{T} X_{ti}$ of our original data matrix X_{ti}, the likelihood function reads

$$\mathcal{L}(\mathbf{\Sigma}) = (2\pi)^{-\frac{pT}{2}} (\det \mathbf{\Sigma})^{-\frac{T}{2}} \exp\left(-\frac{1}{2}\sum_{t=1}^{T}(\mathbf{X}_t - \langle\mathbf{X}\rangle)^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{X}_t - \langle\mathbf{X}\rangle)\right) \tag{3.10}$$

$$= (2\pi)^{-\frac{pT}{2}} (\det \mathbf{\Sigma})^{-\frac{T}{2}} \exp\left(-\frac{1}{2}\sum_{t=1}^{T}\mathrm{Tr}\left((\mathbf{X}_t - \langle\mathbf{X}\rangle)^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{X}_t - \langle\mathbf{X}\rangle)\right)\right) \tag{3.11}$$

$$= (2\pi)^{-\frac{pT}{2}} (\det \mathbf{\Sigma})^{-\frac{T}{2}} \exp\left(-\frac{1}{2}\sum_{t=1}^{T}\mathrm{Tr}\left((\mathbf{X}_t - \langle\mathbf{X}\rangle\,(\mathbf{X}_t - \langle\mathbf{X}\rangle)^{\mathrm{T}}\mathbf{\Sigma}^{-1}\right)\right) \tag{3.12}$$

$$= (2\pi)^{-\frac{pT}{2}} (\det \mathbf{\Sigma})^{-\frac{T}{2}} \exp\left(-\frac{1}{2}\mathrm{Tr}\left(\sum_{t=1}^{T}(\mathbf{X}_t - \langle\mathbf{X}\rangle)(\mathbf{X}_t - \langle\mathbf{X}\rangle)^{\mathrm{T}}\mathbf{\Sigma}^{-1}\right)\right) \tag{3.13}$$

$$= (2\pi)^{-\frac{pT}{2}} (\det \mathbf{\Sigma})^{-\frac{T}{2}} \exp\left(-\frac{1}{2}T\,\mathrm{Tr}\left(\mathbf{S}\mathbf{\Sigma}^{-1}\right)\right) \tag{3.14}$$

where $\mathbf{S}$ denotes the sample covariance matrix (Equation 3.1). Using the relation $\det \mathbf{\Sigma} = 1/\det \mathbf{\Sigma}^{-1}$ the log likelihood reads

$$\ln \mathcal{L}(\mathbf{\Sigma}) = \frac{T}{2}\left(\ln\det(\mathbf{\Sigma}^{-1}) - \mathrm{Tr}\left(\mathbf{S}\mathbf{\Sigma}^{-1}\right) - p\ln(2\pi)\right). \tag{3.15}$$

This equation can also be interpreted as a function of the inverse covariance matrix $\mathbf{\Sigma}^{-1}$ and therefore as possible starting point to find respective estimators, as we will see in Section 3.4.1.

The curse of dimensionality. As I pointed out, with the sample covariance we actually have a well founded estimator for the true covariance matrix. Now, where do the problems emerge? In fact, they come from the high dimensionality of our data matrix $\mathbf{X}$. Consider the shifted data matrix $\mathbf{Y} = \mathbf{X} - \langle\mathbf{X}\rangle$ which has mean zero. It is possible to write Equation 3.1 of the sample covariance matrix also in the form

$$\mathbf{S} = \frac{1}{T}\mathbf{Y}^{\mathrm{T}}\mathbf{Y}. \tag{3.16}$$

From here we can easily see that in the high-dimensional setting $\mathbf{S}$ becomes singular and hence not invertible, which makes it useless as an estimator for our model's parameters, since these coincide with the inverse covariance matrix.

Moreover, it is known that the variance of the inverse coefficients follows (Hotelling, 1953; Schäfer and Strimmer, 2005)

$$\mathrm{var}(S^{-1}_{ij}) = \frac{1}{T - p + 1}. \tag{3.17}$$

In other words, the variability, or noise increases for low sample size, which makes the sample covariance estimator imprecise for high-dimensional data.

For these reasons, other estimators for the inverse covariance matrix are needed. In the following we will develop a method, that inverts the covariance matrix locally, in low dimensional spaces, and combines the local inversions to a global estimate of the inverse covariance matrix. By the local inversions, we are able to avoid this *curse of dimensionality*. We continue this Chapter with Section 3.2 where I provide the essential theoretical background of the merger of probability and network theory. Here I will also introduce the idea of filtering out the significant information through Network Filtering Techniques. In Section 3.3 I than present how the local inversions are able to form a global estimate of the inverse covariance matrix, before we test and discuss our novel method in Section 3.4. Finally, in Section 3.5 we summarize our findings of this Chapter.

3.2 Graphical Models

Graphical Models combine the two fields of probability theory and graph theory in a concise way. Therefore random variables are mapped to the vertices of the graph. Their key feature now is the representation of conditional (in)dependence statements of the probability distribution by the graph structure which for itself offers the possibility of a clear visual representation. Murphy (2001) offers a slightly more comprehensive but still brief introduction to graphical models than the one that is given here. The book by Lauritzen (1996) regards this field from a very technical and mathematical perspective whereas Koller and Friedman (2009) put more emphasis on their use and application, e.g. in computer science and machine learning.

In general two types of graphical models exist. *Bayesian Networks* which use a directed graph and *Markov Random Fields* which go together with undirected graphs. After a brief introduction to the relevant basics of graph theory in Section 3.2.1 we will focus on the latter in Section 3.2.2 with a special emphasis on the Maximum Entropy Model we derived in Section 2.2. This section on Graphical Models ends with a key principal of modern science, *Occam's Razor* or the *principle of parsimony*, that gets translated practically into graph or *network filtering* approaches in Section 3.2.3.

3.2.1 Graph Theory Basics

I want to begin this section with the relevant basics of graph theory mostly following Diestel (2010).

Basics. A *graph* is defined as a tuple $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ is the set of vertices (also called nodes) in the graph, and $\mathcal{E}$ is the set of edges $\{i, j\}$, where $i, j \in \mathcal{V}$ and $i \neq j$. If $\{i, j\} \in \mathcal{E}$, there is an undirected edge from vertex i to vertex j, otherwise, there is no edge between vertex i and vertex j. In most cases, we will assume that $\mathcal{V} = \{1, 2, \ldots, p\}$,

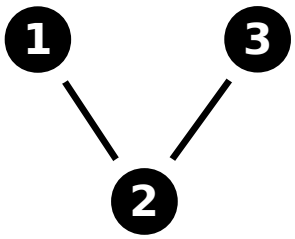

Figure 3.1: An example of an undirected labelled graph with $p = 3$ vertices, here $\mathcal{V} = \{1, 2, 3\}$ and $\mathcal{E} = \{\{1, 2\}, \{2, 3\}\}$. We also see that $\text{ne}(1) = 2, \text{ne}(2) = \{1, 2\}, \text{ne}(\{1, 2\}) = 3$, and 2 separates 1 and 3. For the subgraph $\mathcal{A} = \{1, 2\}, \mathcal{V}^{\mathcal{A}} = \{1, 2\}$ and $\mathcal{E}^{\mathcal{A}} = \{\{1, 2\}\}$.

in which case the graph is called *labelled*. A simple example of an undirected graph is shown in Figure 3.1.

The *neighbours* of vertex i are all vertices in $\mathcal{G}$ having an edge to vertex i,

$$\text{ne}(i) = \{j \in \mathcal{V} : \{i, j\} \in \mathcal{E}\}. \tag{3.18}$$

We write $i \overset{\mathcal{G}}{\sim} j$ if vertices i and j are neighbours in graph $\mathcal{G}$, or just $i \sim j$ where the graph is implicit. A direct consequence of the definition is that $i \sim j \Leftrightarrow j \sim i$. We can extend this definition to a set $\mathcal{A} \subset \mathcal{V}$, where we define the neighbours of $\mathcal{A}$ as

$$\text{ne}(\mathcal{A}) = \bigcup_{i \in \mathcal{A}} \text{ne}(i) \backslash \mathcal{A}. \tag{3.19}$$

The neighbours of $\mathcal{A}$ are all vertices not in $\mathcal{A}$, but adjacent to a vertex in $\mathcal{A}$. Figure 3.1 illustrates this definition.

In the following we want to deal with some notions of graph structures that will play an important role in the further course of this thesis.

Separators. We start by defining a *path* from i_1 to i_m as a sequence of distinct vertices in $\mathcal{V}, i_1, i_2, \ldots, i_m$, for which $\{i_j, i_{j+1}\} \in \mathcal{E}$ for $j = 1, \ldots, m - 1$.

Now, we are able to introduce the local structure of a *separator*. A subset $\mathcal{S} \subset \mathcal{V}$ *separates* two vertices $i \neq \mathcal{S}$ and $j \neq \mathcal{S}$, if every path from i to j contains at least one vertex from $\mathcal{S}$. Two disjoint sets $\mathcal{A} \subset \mathcal{V} \backslash \mathcal{S}$ and $\mathcal{B} \subset \mathcal{V} \backslash \mathcal{S}$ are separated by $\mathcal{S}$, if all $i \in \mathcal{A}$ and $j \in \mathcal{B}$ are separated by $\mathcal{S}$, i.e. we cannot walk on the graph starting somewhere in $\mathcal{A}$ ending somewhere in $\mathcal{B}$ without passing through $\mathcal{S}$. We refer to $\mathcal{S}$ with the term *separator*.

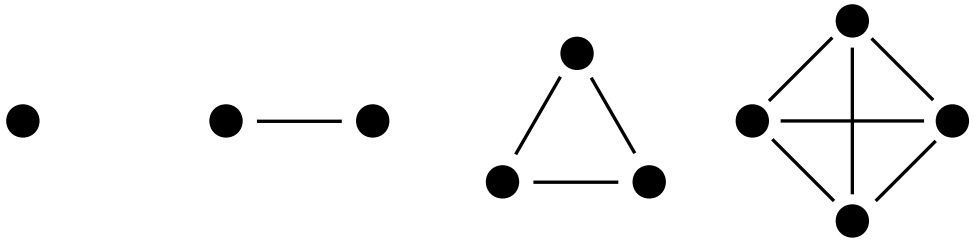

Figure 3.2: First four k-cliques, note that for example a 4-clique contains several 3-,2- and 1-cliques but it is denoted here as 4-clique because it is a maximal 4-clique.

Cliques. Further we define the notion of a *subgraph* $\mathcal{G}^{\mathcal{A}}$ of $\mathcal{G}$. Let $\mathcal{A}$ be a subset of $\mathcal{V}$. Then $\mathcal{G}^{\mathcal{A}}$ denotes the graph restricted to $\mathcal{A}$, i.e., the graph we obtain after removing all vertices not belonging to $\mathcal{A}$ and all edges where at least one vertex does not belong to $\mathcal{A}$. Precisely, $\mathcal{G}^{\mathcal{A}} = \{\mathcal{V}^{\mathcal{A}}, \mathcal{E}^{\mathcal{A}}\}$, where $\mathcal{V}^{\mathcal{A}} = \mathcal{A}$ and

$$\mathcal{E}^{\mathcal{A}} = \{\{i,j\} \in \mathcal{E} \text{ and } \{i,j\} \in \mathcal{A} \times \mathcal{A}\}. \tag{3.20}$$

See Figure 3.1 for an illustration of this definition.

A graph $\mathcal{G}$ is called *complete*, if all its vertices a pairwise adjacent. In that case the vertices of $\mathcal{G}$ are *fully connected*. Now a *clique* $\mathcal{C} \subset \mathcal{V}$ is defined as fully connected subset of vertices. Consequently, all the members of a clique are neighbours. A *maximal clique* is a clique such that there is no larger clique that contains the clique (Barber, 2013). With k-clique we denote a clique containing k vertices, therefore $k = |\mathcal{C}|$. Figure 3.2 illustrates the first four k cliques. It is to mention here, that with the term *clique* it is also common to refer to a *maximal clique*. However, the context should provide clarity in these cases.

Decomposable Graphs Finally, let us define a global graph structure, i.e. one that describes a property of the whole graph. Several names exist for this graph structure, such as *decomposable*, *triangulated* or *chordal graphs*. In the following we need the concept of a *loop* of length k which is a path containing k vertices that starts and returns to the same vertex. A *chord* is an edge that connects two non-adjacent vertices in a loop.

An undirected graph is decomposable if every loop of length 4 or more has a chord. One can show that these graphs can be divided into a unique set of cliques $\mathcal{C}_1, \ldots, \mathcal{C}_M$ such that each pair of neighbouring cliques are separated by a separator $\mathcal{S}_m$ of a unique set of $\mathcal{S}_2 \ldots, \mathcal{S}_M$. For a proof together with a more detailed, but still brief introduction to chordal graphs the reader is referred to the paper by Blair and Peyton (1992). This makes it possible to represent decomposable graphs as so called *clique trees* where the cliques of the decomposable graph become the vertices of the clique tree and the separators of the decomposable graph become the edges of the clique tree. Figure 3.3 shows an example

of a decomposable graph together with one possible clique tree. Note how the graph can be decomposed to its set of cliques.

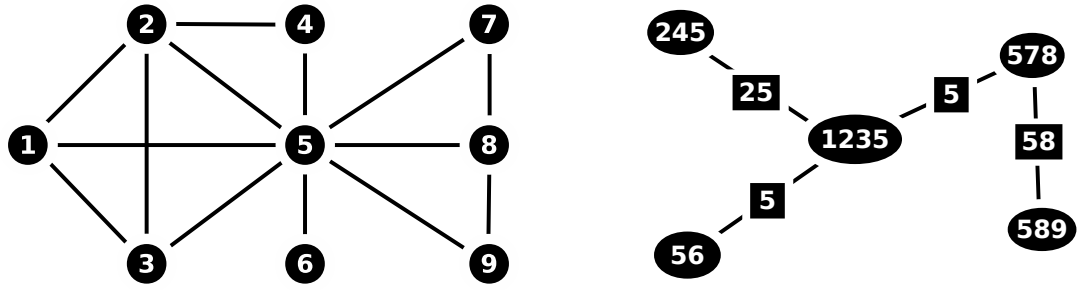

Figure 3.3: An example of a decomposable graph (LEFT) and its representation as a clique tree (RIGHT) where the edges equal the separators $\mathcal{S}$ and are represented as rectangles and the vertices equal the cliques $\mathcal{C}$ and are represented as ellipses. Note that the separators are the intersection of neighbouring cliques. Note further that the tree representation is not unique, the edge with the separator 5 leading to the vertex 578 could also lead to the vertex 589.

3.2.2 (Gaussian) Markov Random Fields

Markov Random Fields are one type of Graphical Model. As such they are probabilistic models where a graph is used to represent the structure of conditional dependence between random variables. The key element is the factorization of the multivariate dependency structure into a set of conditional independences connected through a graph linking conditionally dependent variables.

More precisely, a Markov Random Field is defined as a probability distribution $f(x_1, \ldots, x_p)$ with an associated graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and a certain degree of independence statement:

- The edges of $\mathcal{G}$ connect variables that are not conditionally independent given all other variables
  $x_i \perp x_j \mid \mathbf{x}_{\setminus ij}$ if $\{i, j\} \notin \mathcal{E}$ and $i \neq j$ *Pairwise Markov Property* **P**

- A variable is conditionally independent of all other variables given its neighbours in the network representation
  $x_i \perp \mathbf{x}_{\setminus\{i,\mathrm{ne}(i)\}} \mid \mathbf{x}_{\mathrm{ne}(i)}$ $\forall i \in \mathcal{V}$ *Local Markov Property* **L**

- Consider two disjoint, non-empty sets of variables $\mathcal{A}$ and $\mathcal{B}$, where every path from a node in $\mathcal{A}$ to a node in $\mathcal{B}$ passes through a separating subset $\mathcal{S}$. Any two subsets of variables separated by a separating subset are conditionally independent:
  $\mathbf{x}_{\mathcal{A}} \perp \mathbf{x}_{\mathcal{B}} \mid \mathbf{x}_{\mathcal{S}}$ *Global Markov Property* **G**

From Markov properties to decomposition. These statements of conditional independence are not equivalent in the general case. In fact, one can show that

$$\mathbf{G} \Rightarrow \mathbf{L} \Rightarrow \mathbf{P} \tag{3.21}$$

A proof can be found in Lauritzen (1996). However, the contrary implications, and consequently their equivalence, can be established by the requirement of a positive and continuous density function (Lauritzen, 1996).

Closely related to these independence statements is the property of factorization. The probability density function $f(x_1, \ldots, x_p)$ is said to factorize according to $\mathcal{G}$, iff

$$f(x_1, \ldots, x_p) = \frac{1}{Z} \prod_{\mathcal{C}} \phi_{\mathcal{C}}(\mathbf{x}_{\mathcal{C}}) \tag{3.22}$$

holds for non-negative functions $\phi_{\mathcal{C}}$ (also called potentials) and $\mathcal{C}$ being the cliques of $\mathcal{G}$ such that $\mathbf{x}_{\mathcal{C}}$ denotes the marginal vector constituted by the vertices of $\mathcal{C}$ (Lauritzen, 1996). For the case in which the potentials are strictly positive, this is also called a Gibbs distribution (Barber, 2013). Denoting this factorization property with $\mathbf{F}$, one can proof (Lauritzen, 1996) that

$$\mathbf{F} \Rightarrow \mathbf{G} \Rightarrow \mathbf{L} \Rightarrow \mathbf{P}. \tag{3.23}$$

Now it is the Hammersley and Clifford Theorem (Lauritzen, 1996) that ensures the contrary implications for positive and continuous probability density functions f. It follows that for these types of distributions the conditional independence relations stating that two non-adjacent variables are independent given all other variables, implies the factorization of the distribution over the cliques of the graph $\mathcal{G}$. See Barber (2013, p.61) for an illustrative example.

If additionally the graph $\mathcal{G}$ is decomposable made of M cliques $\mathcal{C}_m$ and $M-1$ separators $\mathcal{S}_m$, one can write the corresponding probability density function as

$$f(\mathbf{x}) = \frac{\prod_{m=1}^{M} f(\mathbf{x}_{\mathcal{C}_m})}{\prod_{m=2}^{M} f(\mathbf{x}_{\mathcal{S}_m})} \tag{3.24}$$

where $f(\mathbf{x}_{\mathcal{C}_m})$ ($f(\mathbf{x}_{\mathcal{S}_m})$) are the marginal distribution of the variables constituting $\mathcal{C}_m$ ($\mathcal{S}_m$) (Lauritzen, 1996).

Two examples. After this theory part, let us try to build up some intuition by considering the following example. Let $f(1, 2, 3, 4)$ be a probability distribution of four variables, where we abbreviate x_i with i, on the graph given in Figure 3.4. Note that this graph is decomposable with cliques $\mathcal{C}_1 = \{1, 2, 3\}$ and $\mathcal{C}_2 = \{2, 3, 4\}$ and the separator $\mathcal{S}_2 = \{3, 4\}$ as we can easily see.

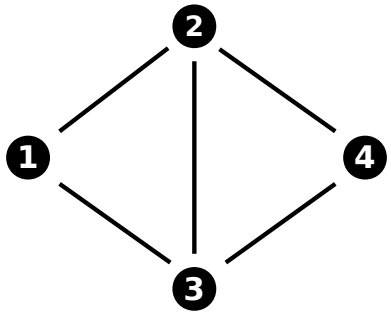

Figure 3.4: Associated graph of the probability distribution $f(1, 2, 3, 4)$

Bayes theorem tells us that we can write the joint probability distribution $f(1, 2, 3, 4)$ as

$$f(1, 2, 3, 4) = f(1|2, 3, 4)f(2, 3, 4), \tag{3.25}$$

where f is now a generic symbol for the respective probability distribution of its arguments. Now we apply the local Markov property that says 1 is independent from 4 given its neighbours $2, 3$ such that we can write

$$f(1, 2, 3, 4) = f(1|2, 3)f(2, 3, 4), \tag{3.26}$$

If we continue applying Bayes theorem we get

$$f(1, 2, 3, 4) = f(1|2, 3)f(2|3, 4)f(3|4)f(4). \tag{3.27}$$

Now, replacing $f(1|2, 3)$ with $f(1, 2, 3)/f(2, 3)$ and analogously the other conditionals results in

$$f(1, 2, 3, 4) = \frac{f(1, 2, 3)}{f(2, 3)} \frac{f(2, 3, 4)}{f(3, 4)} \frac{f(3, 4)}{f(4)} f(4) \tag{3.28}$$

$$= \frac{f(1, 2, 3)f(2, 3, 4)}{f(2, 3)}. \tag{3.29}$$

which is in fact equivalent to

$$f(\mathbf{x}) = \frac{f(\mathbf{x}_{\mathcal{C}_1})f(\mathbf{x}_{\mathcal{C}_2})}{f(\mathbf{x}_{\mathcal{S}_1})} \tag{3.30}$$

The theory provided above ensures that this process of decomposing the probability function is possible for all Markov Random Fields with a decomposable graph.

That is why we can now easily write down the decomposition of the probability function with the associated graph of Figure 3.3 as

$$f(1, \dots 9) = \frac{f(1, 2, 3, 5)f(2, 5)f(5, 6)f(5, 7, 8)f(5, 8, 9)}{f(2, 5)f(5, 8)f(5)f(5)}. \tag{3.31}$$

Note that the Seperator $\mathcal{S} = 5$ actually appears twice.

Gaussian Markov Random Fields. If we now focus on our maximum Entropy model (Equation 2.41) which is in fact equivalent with the well known and widely used multivariate Gaussian distribution

$$f(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^p \det \mathbf{\Sigma}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right), \tag{3.32}$$

we gain the additional property that the zero/non-zero structure of the inverse covariance matrix $\mathbf{\Sigma}^{-1}$ coincides with the graph structure of the associated graph and therefore encodes the conditional independence relations.

So a Gaussian Markov Random Field is defined as the probability distribution Equation 3.32 with respect to a labelled graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ such that

$$\Sigma^{-1}_{ij} \neq 0 \;\Leftrightarrow\; \{i,j\} \in \mathcal{E} \qquad \forall\, i \neq j, \tag{3.33}$$

and therefore x_i is conditionally independent of x_j iff $\Sigma^{-1}_{ij} = 0$, i.e.

$$x_i \perp x_j \mid \mathbf{x}_{\setminus ij} \;\Leftrightarrow\; \Sigma^{-1}_{ij} = 0 \qquad \forall\, i \neq j. \tag{3.34}$$

For a proof that this relationship actually holds see Rue and Held (2005).

3.2.3 The principle of parsimony

Occam's Razor, the presumption that a simpler theory is of more virtue than a more complicated one if both theories are able to make predictions at the same level, is widespread in the areas of philosophy, theology and science (Baker, 2013).

In the context of covariance estimation this is expressed by the *principle of parsimony* (Dempster, 1972). Let us consider the off-diagonal entries of the estimate of the inverse covariance matrix $\mathbf{J}$ as $p(p-1)/2$ individual parameters J_{ij}. The principle of parsimony states that in this setting of parametric model fitting the parameters should be introduced sparingly and only when the data indicate that they are required. It is of importance to make ourselves clear that parameter reduction involves a trade-off between two kinds of errors. If a substantial number of parameters can be set to null values, the amount of noise in a fitted model due to errors of estimation is substantially reduced. On the other hand, errors of misspecification are introduced because the null values are incorrect.

This idea of estimating a sparse inverse covariance matrix divides our problem into two sub-problems:

1. Choosing the set of edges $\mathcal{E}$ of the underlying graph $\mathcal{G}$ and the zero/non-zero pattern of $\mathbf{J}$, respectively.
2. Assigning values to the chosen non-zero entries of $\mathbf{J}$.

Of course, various ways exist how to deal with these two sub-problems. An integrated approach on selecting the graph structure and estimating the non-zero values simultaneously based on a regularized regression is presented in Section 3.4.1. This is done for the purpose of comparison, since a key contribution of this work follows a different approach. Therefore we keep the two sub-problems separated and first try to find a suitable network structure (see below). Afterwards we show how to assign values to the non-zero values (Section 3.3).

Information Filtering Networks. In the context of Graphical Models a sparse inverse covariance matrix $\mathbf{J}$ is equivalent to a sparse connected graph $\mathcal{G}$. So the principle of parsimony translates into the need of filtering techniques of densely connected graphs into more simpler relevant graphs still containing the relevant information. That is why we want to explore certain kind of (sparse) graph structures and techniques for network filtering.

One example of such a graph structure would be the *maximum spanning tree* which was applied for example to detect hierarchical structures in financial markets by Mantegna (1999). To be precise we define a graph as *connected* if there is at least one path from every vertex to any other vertex. A graph is *singly connected* if there is exactly one path between every pair of vertices or in other word the graph is a *tree*. Given a connected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ a *spanning tree* is a graph $\mathcal{G}' = (\mathcal{V}, \mathcal{E}')$ such that its edges become a subset of the edge of $\mathcal{G}$, $\mathcal{E}' \subseteq \mathcal{E}$, resulting in singly connected graph $\mathcal{G}'$ containing all the vertices $\mathcal{V}$. Now let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a *weighted graph*, i.e. with assigned *weights* to each edge. A *maximum spanning tree* (MST) is a spanning tree such that the sum of all weights is at least as large as the sum of any other spanning tree.

So a maximum spanning tree is the most sparse structure still connecting all vertices with each other by maximizing the sum of the weights of the edges. Consequently it contains a number of edges of the same order as the number of vertices. In practise several algorithms exist how to compute a maximum spanning tree. In particular, Kruskal's algorithm (Kruskal Jr., 1956) performs as shown in Algorithm 1: First it creates a list containing all weights in descending order. Then it inserts edges to the graph, beginning from the top of the list. Before every insertion it checks if the graph is still loop free. If an edge causes a loop it is left out. After considering the whole list of descended weights, the resulting graph is the maximum spanning tree.

In a sense, we can regard the process of computing a maximum spanning tree as filtering out the most relevant information to a densely connected structure. However the reduction to a tree is a very drastic one, hence loosing valuable information is most probable. In order to achieve a less drastic filtering process it has been found useful to characterise graphs or networks according to their possible embeddings on hyperbolic surfaces (Aste et al., 2005, 2012). Intuitively speaking, a graph can be embedded on a surface if it can be drawn on that surface without edges crossing. This idea of information filtering networks on hyperbolic surface was applied for example to study stock

Algorithm 1 Computing a Maximum Spanning Tree $\mathcal{T}$ (Kruskal Jr., 1956)

Require: $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is a connected graph
1: Create descending list of edges $\mathcal{L}$
2: **for** edge e in list $\mathcal{L}$ **do**
3: insert edge e into new graph $\mathcal{T}$
4: **if** graph $\mathcal{T}$ is not loop-free **then**
5: remove edge e from graph $\mathcal{T}$
6: **end if**
7: **end for**

market data (Tumminello et al., 2005, 2007), interest rates (Di Matteo et al., 2005) or risk diversification (Musmeci et al., 2014).

Now, every graph can be embedded on a surface with sufficiently high *genus* g, which then serves as a measure of complexity of the network. The genus of a topological surface is the largest number of non-intersecting simple closed curves that can be drawn on the surface without separating it, i.e. its number of holes. So for example a sphere has genus $g = 0$ whereas a torus (doughnut) has genus $g = 1$. Consequently, a lower complexity is assigned to a network embedded on a sphere than to a network embedded on a torus. Now, it is known that a complete graph with $p \geq 3$ vertices can be embedded on a surface with genus

$$g \geq g' = \left\lceil \frac{(p-3)(p-4)}{12} \right\rceil, \tag{3.35}$$

where $\lceil x \rceil$ denotes the ceiling function which returns the smallest integer greater or equal to x. It immediately follows that also for non complete graphs with p vertices there exists always a topological surface where they can be embedded in. Nevertheless, other difficulties arise since graphs with only a few edges may be embedded in topologically simpler surfaces (with lower genus than g') (Aste et al., 2005).

Moreover, it is known that a graph with a fixed genus g has at most $3(p-2+2g)$ edges whereas a maximum spanning tree contains only $p-1$. So the relative increase in the number of edges - and therefore also the relative increase of information gained about the system - is the largest from the spanning tree to a graph of genus $g = 0$ compared to the increase between aligning genuses g and $g+1$.

So the planar ($g = 0$) graph is not only the simplest but also the one with the most significant information increase, which qualifies it for further investigation (Tumminello et al., 2005). With $3(p-2)$ edges the planar graph can be constructed as a clique tree consisting of 4-cliques separated by 3-separators (i.e. a separator consisting of a 3-clique), such that each separator is only separating two 4-cliques. If a separator would separate more than two 4-cliques, the resulting graph would not be planar any more. Consider further the following argument how to construct a 4-clique tree separated by 3-separators: Let us begin with one 4-clique. Up to now the graph consists of 4 vertices and $6 = 3(4-2)$ edges. Now every additional vertex will be inserted with 3 edges

connecting to an open 3-clique separator which then makes the open separator an actual separator separating the old 4-clique and the one originated with the insertion of the new vertex. By connecting only to open separator it is ensured that the graph remains planar. Consequently, the number of edges of this graph increases by three with every additional vertex, so in total the number of edges is $3(p-2)$. See Figure 3.5 for an illustration for this procedure. One can also show that such a planar graph constructed with a similar algorithm then the one of the maximum spanning tree actually results in a graph that contains the maximum spanning tree and can therefore be regarded as a directed extension (Tumminello et al., 2005).

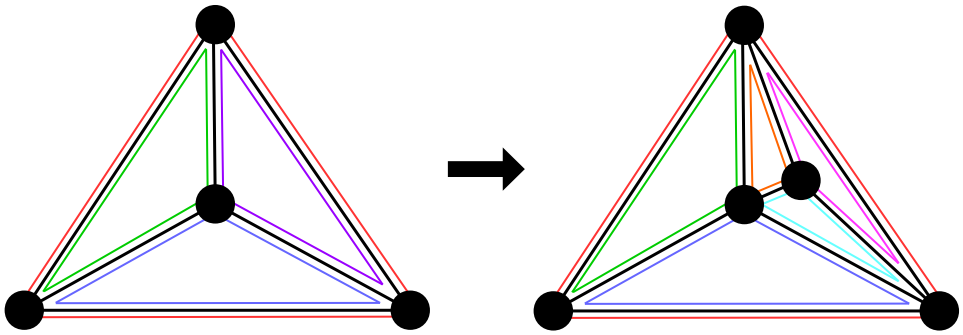

Figure 3.5: (LEFT) A 4-clique with coloured possible separators (blue, green, purple, red). A fifth vertex can be inserted into each of the possible separators, connecting to the three corresponding vertices, such that the whole graph remains planar. After the insertion (RIGHT), the possible separator is gone, but three new possible separators have emerged. By proceeding analogously one results in a maximal planar graph with $3(p-2)$ edges that is simultaneously a decomposable clique tree consisting of 4-cliques separated by 3-separators.

Later we will see the importance that the planar graph can be regarded as a clique tree, since it makes it treatable as decomposable graph for which also the decomposition of the probability density holds as we saw in the previous section.

Summary. In this section I have shown how the language of Graphical Models combines probability and network theory. After the relevant basics of graph theory with a special emphasis on decomposability and clique trees, we saw how general statements about Markov Random Fields, (a particular type of Graphical Model) can be applied to our maximum Entropy model, the multivariate Gaussian distribution. Moreover we explored several graph structures in the context of information filtering as a practical interpretation of the principal of parsimony. By that we have laid all necessary foundations to continue with the results of this Chapter. As we will see, the key contribution will be the combination of the topics treated so far.

3.3 Results

In the previous section we saw how information filtering techniques function as a reductionist tool to obtain the relevant and therefore useful information of a system. Hence, they are ideally suitable to address the first of our two sub-questions stated in Section 3.2.3: Choosing a graph structure or equivalently the zero/non-zero pattern of the estimate of the inverse covariance matrix. We are left with the second sub-question, how to assign values to the non-zero entries. This will be addressed in the following. Up to now we have collected all the relevant background knowledge, so that we are finally able to combine the bits and pieces to a convenient equation which does the assignment of values to the non-zero entries in a distinct manner. Since it exploits the fact that the associated graph structure is decomposable to use **lo**cal information of the interaction between just a few variables to obtain a **g**l**o**bal estimate of the inverse covariance matrix we named this formula and the resulting method *LoGo*. The detailed derivation of this equation is given in 3.3.1.

(Un)fortunately this is not the end of the story. In Section 3.2.3 I motivated why different graph *structures* (namely the maximum spanning tree and the 4-clique tree) qualify to filter out information of the system. For example, the maximum spanning tree keeps only the edges with the largest weights, such that the resulting graph is a tree connecting still all vertices/variables. But in our context it is not straight forward what these weights are. They could be the covariance, or the correlation, or something different at all, depending on the desired result. So in short, the theory of information filtering networks tells us something about the *structure* of the graph. We are left with obtaining a specific realization of the structure by some algorithm. We explore how the *LoGo* formula is applied to the two graph structures above in Section 3.3.2 and Section 3.3.3 by giving some explicit expressions and suggesting algorithms how to obtain a *good* realization of the specific structure.

3.3.1 The LoGo Inverse

In the following we can derive very efficiently an equation how to assign values to the non-zero entries of an estimate of an inverse covariance matrix with an underlying decomposable graph in a distinct way. Additionally some of its properties will be outlined as well.

With Equation 3.24 we found how a general probability distribution factorizes over a decomposable graph. It is basically the product of the marginal distributions of all cliques divided by the product of the marginal distributions of all separators. Now for our maximum Entropy model, the multivariate Gaussian distribution, we have obtained in Equation 2.27 that these marginal Gaussians are in fact Gaussian as well but with the respective sub covariance matrix of the whole covariance matrix. This makes it now possible to invert these marginal local covariance matrices to obtain a local estimate of

the inverse covariance matrix. And because of the factorization and the fact that in the Gaussian case the matrix algebra of the inverse covariance matrix is done in the exponent it is further possible to sum up all local inverses to a global estimate of the inverse covariance matrix.

So, given a decomposable graph with p vertices, M cliques and therefore $M-1$ separators numbered from 2 to M for reasons that will become clear later, we can express the global inverse covariance matrix $\mathbf{\Sigma}^{-1}$ according to the following summation

$$\mathbf{\Sigma}^{-1} = \sum_{m=1}^{M} \left[(\mathbf{\Sigma}_{\mathcal{C}_m})^{-1} \right] - \sum_{m=2}^{M} \left[(\mathbf{\Sigma}_{\mathcal{S}_m})^{-1} \right], \tag{3.36}$$

where $[\dots]$ denotes an $p \times p$ matrix with the values of the sub-matrix $(\mathbf{\Sigma}_{\mathcal{C}_m})^{-1}$ or $(\mathbf{\Sigma}_{\mathcal{S}_m})^{-1}$ at the indices corresponding to the labels of the vertices in $\mathcal{C}_m$ or $\mathcal{S}_m$ and zeros elsewhere. However, in the practical case we have no access to the true covariance matrix $\mathbf{\Sigma}$, which is why we use the sample covariance matrix $\mathbf{S}$ to obtain an estimate of the inverse covariance matrix $\mathbf{J}$ according to

$$\mathbf{J} = \sum_{m=1}^{M} \left[(\mathbf{S}_{\mathcal{C}_m})^{-1} \right] - \sum_{m=2}^{M} \left[(\mathbf{S}_{\mathcal{S}_m})^{-1} \right]. \tag{3.37}$$

Note, that this result

- is *exact*, given the underlying (decomposable) graph, and will therefore not introduce new errors. Of course in practice the crucial assumption is the one of an underlying decomposable graph. However, this is motivated in Section 3.2.3.
- replaces the inversion of a $p \times p$ matrix by multiple inversions of *local* matrices of lower dimensions. For example, in the case of a 4-clique tree only inversions of 4×4 and 3×3 matrices have to be performed. This means that in principle only 4 observations in our data matrix would be enough to result in a non-singular global estimate.
- is a *maximum likelihood estimate*. Consider all estimates of $\mathbf{\Sigma}^{-1}$ with the same underlying network as $\mathbf{J}$. Among all these estimates $\mathbf{J}$ is the one with the maximum likelihood. This can be seen from the fact that $\mathbf{S}$ for itself is the maximum likelihood solution of the covariance matrix. Moreover, in practise we see that $\mathbf{J}^{-1}$ coincides with $\mathbf{S}$ at the non-zero entries of $\mathbf{J}$. According to Dempster (1972), this is a characteristic of the maximum likelihood estimate among all estimates with the same underlying graph.

Since Equation 3.37 is a maximum likelihood solution for itself we proceed in the following to obtain the particular graph, such that the resulting estimate for the inverse covariance maximizes the likelihood, which we derived for the multivariate Gaussian

distribution in Equation 3.15. Before we concentrate ourselves on the particular graph structures I want to show that for the addend in the middle of Equation 3.15,

$$\mathrm{Tr}(\mathbf{SJ}) = p \tag{3.38}$$

holds, independent of the chosen graph. I begin straightforwardly by plugging in Equation 3.37 for $\mathbf{J}$ and apply simple rules for computing matrices and the trace. Note that

$$\left[(\mathbf{S}_{\mathcal{C}_m})^{-1}\right] \text{ and } \left[(\mathbf{S}_{\mathcal{S}_m})^{-1}\right]$$

are $p \times p$-matrices.

$$\mathrm{Tr}(\mathbf{SJ}) = \mathrm{Tr}\left(\mathbf{S}\Big(\sum_{m=1}^{M}\left[(\mathbf{S}_{\mathcal{C}_m})^{-1}\right] - \sum_{m=2}^{M}\left[(\mathbf{S}_{\mathcal{S}_m})^{-1}\right]\Big)\right) \tag{3.39}$$

$$= \mathrm{Tr}\Big(\sum_{m=1}^{M}\mathbf{S}\left[(\mathbf{S}_{\mathcal{C}_m})^{-1}\right] - \sum_{m=2}^{M}\mathbf{S}\left[(\mathbf{S}_{\mathcal{S}_m})^{-1}\right]\Big) \tag{3.40}$$

$$= \sum_{m=1}^{M}\mathrm{Tr}\left(\mathbf{S}\left[(\mathbf{S}_{\mathcal{C}_m})^{-1}\right]\right) - \sum_{m=2}^{M}\mathrm{Tr}\left(\mathbf{S}\left[(\mathbf{S}_{\mathcal{S}_m})^{-1}\right]\right) \tag{3.41}$$

Now we use the fact that $[(\mathbf{S}_{\mathcal{C}_m})^{-1}]$ and $[(\mathbf{S}_{\mathcal{S}_m})^{-1}]$ have zero entries at indices not in $\mathcal{C}_m$ and $\mathcal{S}_m$, respectively. Therefore $(\mathbf{S}[(\mathbf{S}_{\mathcal{C}_m})^{-1}])$ and $(\mathbf{S}[(\mathbf{S}_{\mathcal{S}_m})^{-1}])$ will have zeros everywhere on the diagonal except at elements with indices in $\mathcal{C}_m$ and $\mathcal{S}_m$, respectively, where they will have ones. Therefore

$$\mathrm{Tr}(\mathbf{SJ}) = \sum_{m=1}^{M}|\mathcal{C}_m| - \sum_{m=2}^{M}|\mathcal{S}_m|, \tag{3.42}$$

where $|...|$ denotes the cardinality of the cliques and separators, respectively. Now we argue that with the first sum we over count the number of vertices in $\mathcal{G}$. Thus, for each pair of cliques we need to subtract the number of vertices they have in common which is exactly the number of vertices of a separator. Hence, we obtain Equation 3.38.

It follows that we can safely ignore this middle addend of Equation 3.15 in our considerations how to construct a maximum likelihood network since it will always be the number of variables, independently of the network. Consequently the task of finding a maximum likelihood network translates into the task of finding a maximum determinant network.

In the following section I want to apply the LoGo Equation 3.37 to two distinct graph structures that have proven to obtain meaningful sparse filtered networks, the spanning tree and the 4-clique tree.

3.3.2 LoGo on a spanning tree

In our language a tree is a decomposable graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ that consists only of 2-cliques and 1-separators meaning that every clique of a tree is a subset of $\mathcal{V}$ consisting exactly of two elements and every separator consists exactly of one element. But unlike in the case of the planar 4-clique tree a separator can actually appear more than once in the case of a spanning tree. In order to obtain an explicit relation for the tree we need the explicit expression for the inverses of symmetric 1×1- and 2×2-matrices according to Equation 3.37. They are given by

$$\left(S_{ii}\right)^{-1} = \left(\tfrac{1}{S_{ii}}\right) \tag{3.43}$$

and

$$\begin{pmatrix} S_{ii} & S_{ij} \\ S_{ij} & S_{jj} \end{pmatrix}^{-1} = \frac{1}{S_{ii}S_{jj} - S_{ij}^2} \begin{pmatrix} S_{jj} & -S_{ij} \\ -S_{ij} & S_{ii} \end{pmatrix}. \tag{3.44}$$

To derive an explicit expression for $\mathbf{J}$ we split $\mathbf{J}$ into three parts: the non-neighbouring off-diagonal, the neighbouring off-diagonal and the diagonal elements.

For an off-diagonal element J_{ij} of two distinct and non-neighbouring vertices i and j we can immediately state that $J_{ij} = 0$. It follows directly form the definition of Gaussian Markov Random Fields. Additionally we can argue that if i and j are not neighbours there will not be a clique $\mathcal{C}_m$ containing both vertex i and j and consequently there is no contribution to J_{ij} in the sum of Equation 3.37.

Now, consider the case that the vertices i and j are neighbours. Consequently there will be a 2-clique $\mathcal{C} = \{i, j\}$ contributing in Equation 3.37 to J_{ij}. In fact, this will be the only contribution to J_{ij}. On the one hand there cannot be another clique containing i and j since a tree has only 2-cliques. On the other hand in a tree no separator can contribute to an off-diagonal element since they only consist of one element for which the respective 1×1 sub covariance matrix lies on the diagonal. Therefore $J_{ij} = -S_{ij}/(S_{ii}S_{jj} - S_{ij}^2)$ according to Equation 3.44 if i and j are neighbours.

Finally, we take into account an arbitrary diagonal element J_{ii}. We easily convince ourselves that vertex i is an element of every 2-clique that consists of i and one neighbour of i and second that $\{i\}$ is a separator exactly the number of neighbours of i minus one

times. Thus, inserting Equations 3.43 and 3.44 into Equation 3.37 we obtain

$$J_{ii} = \Big(\sum_{l \in \mathrm{ne}(i)} \frac{S_{ll}}{S_{ii}S_{ll} - S_{il}^2} \Big) - (|\,\mathrm{ne}(i)| - 1)\frac{1}{S_{ii}} \tag{3.45}$$

$$= \sum_{l \in \mathrm{ne}(i)} \frac{S_{ll}}{S_{ii}S_{ll} - S_{il}^2} - \Big(\sum_{l \in \mathrm{ne}(i)} \frac{1}{S_{ii}} \Big) + \frac{1}{S_{ii}} \tag{3.46}$$

$$= \frac{1}{S_{ii}} + \sum_{l \in \mathrm{ne}(i)} \left(\frac{S_{ll}}{S_{ii}S_{ll} - S_{il}^2} - \frac{1}{S_{ii}} \right) \tag{3.47}$$

$$= \frac{1}{S_{ii}} + \sum_{l \in \mathrm{ne}(i)} \left(\frac{S_{ll}S_{ii}}{(S_{ii}S_{ll} - S_{il}^2)S_{ii}} - \frac{S_{ii}S_{ll} - S_{il}^2}{(S_{ii}S_{ll} - S_{il}^2)S_{ii}} \right) \tag{3.48}$$

$$= \frac{1}{S_{ii}} + \sum_{l \in \mathrm{ne}(i)} \frac{1}{S_{ii}S_{ll} - S_{il}^2} \frac{S_{il}^2}{S_{ii}} = \tag{3.49}$$

$$= \frac{1}{S_{ii}} \Big(1 + \sum_{l \in \mathrm{ne}(i)} \frac{S_{il}^2}{S_{ii}S_{ll} - S_{il}^2} \Big) \tag{3.50}$$

Putting all together, we can state the explicit *LoGo* Equation for trees as follows

$$J_{ij} = \begin{cases} 0 & \text{if } i \text{ and } j \text{ are not adjacent} \\ -\frac{S_{ij}}{S_{ii}S_{jj} - S_{ij}^2} & \text{if } i \text{ is adjacent to } j \text{ and } i \neq j \\ \frac{1}{S_{ii}} \big(1 + \sum_{l \in \mathrm{ne}(i)} \frac{S_{il}^2}{S_{ii}S_{ll} - S_{il}^2} \big) & \text{if } i = j \end{cases} \tag{3.51}$$

Network Algortihm. Since it is not possible to explore every possible spanning tree because of the huge number of possibilities, we are now left with presenting an argument how to choose the set of edges $\mathcal{E}$ efficiently that maximizes the log likelihood given in Equation 3.15. We have already shown that the term $\mathrm{Tr}(\mathbf{SJ}) = p$, the number of vertices. Thus it is independent from the choice of $\mathcal{E}$.

As a small test of consistency we will verify this result using the explicit Equation 3.51. In particular we will show that each element on the diagonal of **SJ** is 1 and therefore

$\mathrm{Tr}(\mathbf{SJ}) = p$. So let us consider the diagonal element $(SJ)_{ii}$,

$$(SJ)_{ii} = \sum_{j=1}^{p} S_{ij} J_{ji} \tag{3.52}$$

$$= S_{ii} J_{ii} + \sum_{\substack{j=1 \\ j \neq i}}^{p} S_{ij} J_{ji} \tag{3.53}$$

$$= S_{ii} \frac{1}{S_{ii}} \Big(1 + \sum_{l \in \mathrm{ne}(i)} \frac{S_{il}^2}{S_{ii} S_{ll} - S_{il}^2} \Big) + \sum_{\substack{j=1 \\ j \neq i}}^{p} S_{ij} J_{ji} \tag{3.54}$$

$$= 1 + \sum_{l \in \mathrm{ne}(i)} \frac{S_{il}^2}{S_{ii} S_{ll} - S_{il}^2} + \sum_{j \in \mathrm{ne}(i)} S_{ij} \frac{-S_{ij}}{S_{ii} S_{jj} - S_{ij}^2} \tag{3.55}$$

$$= 1 \tag{3.56}$$

Hence, we are left with maximizing the first term of Equation 3.15, i.e. the determinant of $\mathbf{J}$.

The idea is now to define a new weights-matrix W_{ij} whose corresponding maximum spanning tree applied to Equation 3.51 maximizes $\det \mathbf{J}$. We note that the determinant, as well as the trace, are invariant properties of a matrix under transformations in space. As such, the determinant of a matrix can be expressed as the product of its eigenvalues like the trace can be expressed as their sum. So in other words, we are trying to find the spanning tree network whose product of eigenvalues is at its maximum. Since we do not have a direct link to the determinant nor to the eigenvalues we argue the eigenvalues which maximize their sum will also have a large product and therefore a large determinant. So in the following we will define weights with which we calculate a maximum spanning tree that maximizes the trace of $\mathbf{J}$. If there were no edges at all (i.e. all off-diagonal entries are zero) the trace of $\mathbf{J}$ would look like the following according to Equation 3.51:

$$\mathrm{Tr}(\mathbf{J}) = \frac{1}{S_{11}} + \frac{1}{S_{22}} + \cdots + \frac{1}{S_{pp}}. \tag{3.57}$$

Imagine now that we add an edge between vertex i and j so the trace will change according to

$$\mathrm{Tr}(\mathbf{J}) = \frac{1}{S_{11}} + \cdots + \frac{1}{S_{pp}} + \frac{1}{S_{ii} S_{jj} - S_{ij}^2} \frac{S_{ij}^2}{S_{ii}} + \frac{1}{S_{jj} S_{ii} - S_{ji}^2} \frac{S_{ji}^2}{S_{jj}} \tag{3.58}$$

$$= \frac{1}{S_{11}} + \cdots + \frac{1}{S_{pp}} + \frac{S_{ij}^2}{S_{ii} S_{jj} - S_{ij}^2} \left(\frac{1}{S_{ii}} + \frac{1}{S_{jj}} \right) \tag{3.59}$$

Hence, we can define a weights matrix $\mathbf{W}$ to compute the trace maximizing spanning tree for $i \neq j$ by

$$W_{ij} = \frac{S_{ij}^2}{S_{ii}S_{jj} - S_{ij}^2} \left(\frac{1}{S_{ii}} + \frac{1}{S_{jj}} \right). \tag{3.60}$$

Note, that we do not have to define the diagonal elements of $\mathbf{W}$ because they have no influence on finding a maximum spanning tree. Now we can use Algorithm 1 with weight matrix $\mathbf{W}$ to filter out a tree with respect to a large determinant of $\mathbf{J}$. If we are working with standardized variables, i.e. using the correlation matrix instead of the covariance matrix Equation 3.60 reduces to $W_{ij} = 2S_{ij}^2/(1 - S_{ij}^2)$ and hence, the resulting maximum spanning tree is equivalent to the maximum spanning tree obtained by a weight matrix with the squared entries of the correlation matrix $W'_{ij} = S_{ij}^2$ which is easier to compute.

3.3.3 LoGo on a planar 4-clique tree

In the following segment we want to apply our general *LoGo* Equation 3.37 to planar graphs which consist of 4-cliques separated by 3-separators, such that every 3-separator is part of no more than two 4-cliques, as we have seen in Section 3.2.3. Since an explicit expression of the *LoGo* equation would require the explicit inverses of 4×4 and 3×3-matrices stating the explicit result of Equation 3.37 for the planar 4-clique tree is not very instructive.

Network Algorithm. Anyhow, more interesting than an explicit expression of the *LoGo* Equation is the question how to construct the actual planar 4-clique network. Like in Section 3.3.2 we are interested in the particular graph such that its corresponding inverse covariance matrix maximizes the likelihood function (Equation 3.15). We have already seen that this translates directly into the task of finding the network such that the corresponding determinant of $\mathbf{J}$ is maximal.

Because of the decomposability of our Gaussian Markov Random Fields we can exploit rules regarding block-diagonal matrices for computing the determinant of $\mathbf{J}$ (Lauritzen, 1996) such that we can write

$$\det \mathbf{J} = \frac{1}{\det\left(\mathbf{S}_{\mathcal{C}_1}\right)} \prod_{m=2}^{M} \frac{\det\left(\mathbf{S}_{\mathcal{S}_m}\right)}{\det\left(\mathbf{S}_{\mathcal{C}_m}\right)}. \tag{3.61}$$

Equation 3.61 predetermines the logic of the greedy algorithm that we use to build the planar 4-clique network. It is shown conceptually in Algorithm 2.

Note, that in line 4, a not inserted vertex gets inserted with $\mathcal{C}_m$ as well. Further let me mention that one could also take other *gain* measures into account than the fraction

Algorithm 2 Estimating a maximum likelihood planar 4-clique tree $\mathcal{T}$

Require: $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is a connected graph associated to Gaussian Markov Random Field with sample covariance matrix $\mathbf{S}$x

1: insert $\mathcal{C}_1$ with minimal $\det(\mathbf{S}_{\mathcal{C}_1})$ into $\mathcal{T}$
2: **while** there are not-yet-inserted-vertices left **do**
3: create a list of $\det(\mathbf{S}_{\mathcal{S}_m})/\det(\mathbf{S}_{\mathcal{C}_m})$ of all possible combinations of separators of $\mathcal{T}$ and not-yet-inserted-vertices
4: insert $\mathcal{C}_m$ into $\mathcal{T}$ with maximum $\det(\mathbf{S}_{\mathcal{S}_m})/\det(\mathbf{S}_{\mathcal{C}_m})$
5: **end while**

of the determinants. For example, the square of the correlation has been considered as well as a promising alternative. Moreover I want to comment on line 1. In practise, computing every possible combination of vertices to form the initial 4-clique is very costly, especially for large p. So a shortcut of taking the four most correlated vertices as the initial clique has been proven practicable.

Summary. In this section I showed how one can write the inverse covariance matrix of a Gaussian Markov Random Field if the underlying graph is decomposable. Moreover I showed how this equation applies to two distinct decomposable graph structures, the spanning tree and the planar 4-clique network. For the spanning tree an explicit expressions for the inverse were given. Additionally, I proposed two algorithms how to find the actual network with respect to a maximum likelihood ansatz of the resulting estimate $\mathbf{J}$ for the sparse inverse covariance matrix.

3.4 Discussion

Having obtained a practical equation to assign values to the non-zero entries of the sparse inverse covariance matrix and an algorithm how to choose an actual planar 4-clique graph we are now able to use our LoGo method in practise. In order to have a strong opponent to which we can compare the testing results I briefly introduce the state-of-the-art method of finding a sparse estimate of the inverse covariance matrix in Section 3.4.1. In Section 3.4.2 I will then give some testing results based on simulated data series before presenting some real word example case in Section 3.4.3.

3.4.1 Penalized likelihood

The way of estimating sparse inverse covariance matrices which I named here *penalized likelihood* is a widely studied method (Meinshausen and Bühlmann, 2006; Banerjee et al., 2006, 2008; Yuan and Lin, 2007; Friedman et al., 2008; Ravikumar et al., 2011; Hsieh

et al., 2011; Oztoprak et al., 2012). Despite this variety of literature all these approaches follow the same basic principle I want to introduce in the following.

As we have seen before, a maximum likelihood problem refers to finding the parameters A_{ij} that maximize the likelihood function for the multivariate Gaussian distribution (Equation 3.15). The idea is to solve this maximum likelihood problem with an added $\ell 1$-norm penalty $\sum_{ij} |A_{ij}|$ term which was first introduced as the *Lasso*, the least absolute shrinkage and selection operator, by Tibshirani (1996).

He proposes this technique in the context of a usual regression situation to retain the good features of the standard techniques by combining ordinary least squares estimates, subset selection and ridge regression. Subset selection provides interpretable models but can be extremely variable because it is a discrete process, the coefficients are either retained or dropped from the model. Small changes in the data can result in very different models being selected which can reduce its prediction accuracy. Ridge regression on the other hand is a continuous process and hence more stable, however it does not set any coefficients to zero which results in models more difficult to interpret. Therefore the *Lasso* tries to combine the advantages of both subset selection and ridge regression by setting some coefficients to zero and shrinking the others. It tries to solve the two sub-problems stated in Section 3.2.3 simultaneously.

In fact, the *Lasso* consists of a standard least squares ansatz and an added $\ell 1$-norm penalty term. It exploits the convexity of the general least square ansatz and the form of the $\ell 1$-norm to produce parameters with zero value ($A_{ij} = 0$) with a higher fraction compared for example to the squared penalty $\sum_{ij} A_{ij}^2$ of the ridge regression as pointed out by Tibshirani (1996, Figure 2). Regularization parameters $\lambda_{ij} > 0$ for each parameter control the sparsity.

The *Lasso* was adapted to the negative log likelihood problem by Banerjee et al. (2006). This ansatz uses the convexity of the negative log likelihood as a function of the parameters J_{ij}, i.e. the entries of the estimate of the inverse covariance matrix, similarly to the convexity of the least square ansatz.

Taking all together the penalized likelihood estimator reads

$$\mathbf{J} = \underset{\mathbf{X} \succ 0}{\operatorname{argmin}} \left(- \ln \det \mathbf{X} + \operatorname{Tr} (\mathbf{S}\mathbf{X}) + \sum_{i,j=1}^{N} \lambda_{ij} |X_{ij}| \right). \tag{3.62}$$

In practise several algorithms for this optimisation problem exist. In fact, this is the origin for the variety of literature about this problem. For the following tests we rely on the algorithm *Glasso* by Friedman et al. (2008) for which popular implementations exist in *R* or *Python*.

3.4.2 Simulations

In the following we will compare our *LoGo* result with the state-of-the-art method *Lasso* in practise. Before we turn our attention to a real-world example we want to observe their performance with simulated data which have two important advantages.

- First of all we can be sure that the data series is true Gaussian which is highly doubtful in the real-world case. Therefore no new errors are introduced here and the data series matches our assumptions we have made along the development of *LoGo*. (*Lasso*, by the way, also assumes an underlying Gaussian distribution.)

- Further, with simulated data, we are able to cover the whole range of settings, from high-dimensional, where the number of observations T is small compared to the number of variables p up to almost the asymptotic limit where in theory $T \to \infty$.

However, there is one big open question. How to generate a random symmetric positive-definite matrix? As it turns out, in order to generate a data series according to a multivariate Gaussian distribution one needs to specify a covariance matrix as an input parameter which is required to be a symmetric positive definite matrix. Since to my knowledge there is no systematic study of random symmetric positive definite matrices we avoid answering this question and use a sample covariance matrix, obtained from the real-world data series of Section 3.4.3. By that we are sure that we use a meaningful covariance structure in that sense, that important characteristics of this covariance matrix result from the real-world, where our methods aims to contribute useful insights. But we still benefit from the advantages listed above by using a simulated data series.

Methods. We are using several methods to obtain a sparse estimate of the original inverse covariance matrix. These were on the one hand the *LoGo* (as described in Section 3.3) with a network finding algorithm based on the determinants of the sub-matrices as shown in Algorithm 2, called *LoGoDet*. Further we use *LoGo* where we retrieve the network based on the square of covariance matrix entries as the gain function in Algorithm 2, called *LoGoSqu*. Additionally, we also compute an estimate based on *LoGo* but with a spanning tree as the underlying graph structure, i.e. *LoGoST*. This spanning tree is a maximum spanning tree based on the weights defined in Equation 3.60. In contrast to the *LoGo*s two *Lasso*s were computed with the widely used *Glasso* Algorithm (Friedman et al., 2008) for comparison. One uses a cross-validation technique to obtain a best possible estimate (*LassoCV*), at the other one, the regularization parameter was set such that the resulting estimate has approximately the same fixed number of zero entries as the *LoGo* estimates on the planar 4-clique tree. These *Lasso*-type algorithms were used via the implementation of the *python* package *scikit-learn* (Pedregosa et al., 2011).

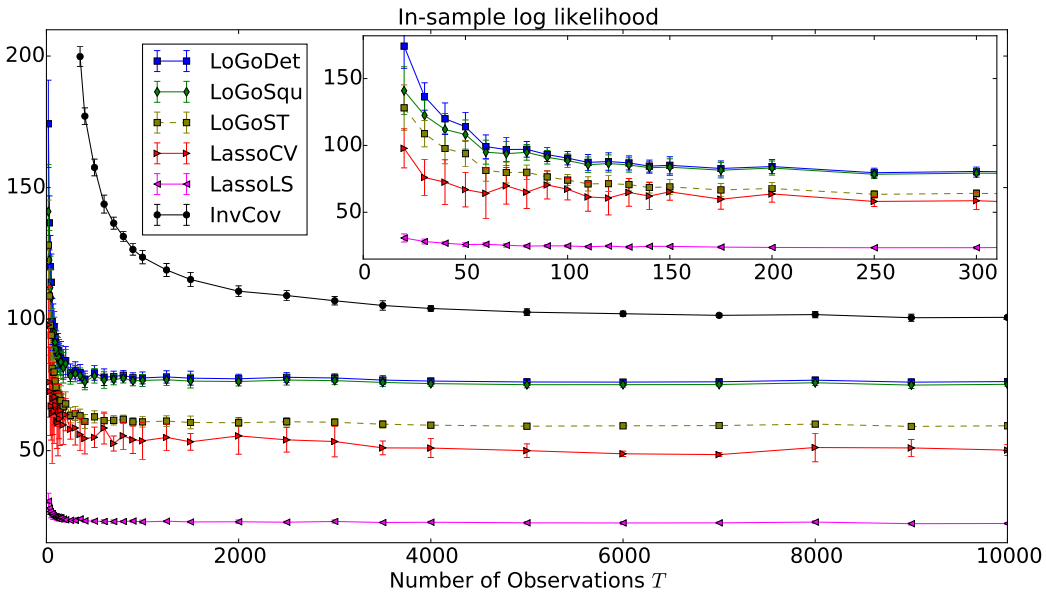

Figure 3.6: Log likelihood for various estimates of the inverse correlation matrix of simulated data for the in-sample-case, i.e. where the likelihood is computed with the set of data, that was used to compute the estimates. The in-sample log likelihood is plotted versus the length of the dataset T. The number of variables is fixed at $p = 300$. The estimates were computed with the methods described in Section 3.4.2. Additionally *InvCov* (black) denotes the inverse of the sample covariance matrix. The dots are the mean of 10 sample runs, the errorbars denote the corresponding standard deviations.

In-sample log likelihood. In Section 3.1 I motivated that the likelihood measures how well a set of parameters describe a set of observed data. In the following test I choose randomly $p = 300$ of in total $p_{\max} = 342$ variables (in order not to rely on a specific set of variables) and compute several estimates for the inverse covariance matrix for different lengths of the series T (x-axis). The estimates were computed according to the methods described above. Additionally, I invert the sample covariance matrix for $T > p$ and call this method *InvCov* (black line). Finally I compute the *log likelihood* on the set of data, that was used to compute the estimates (in-sample) with respect to the *Null* model, where the *Null* estimate is a simple p-dimensional one matrix. So with Equation 3.15 we get for the in-sample log likelihood with estimate $\mathbf{J}$,

$$\frac{1}{T}\Big(\ln \mathcal{L}(\mathbf{J}) - \ln \mathcal{L}(\mathbf{1}_p)\Big) = \frac{1}{2}\Big(\ln\det(\mathbf{J}) - \mathrm{Tr}(\mathbf{SJ}) + \mathrm{Tr}(\mathbf{S})\Big), \tag{3.63}$$

Note that for all *LoGo* estimates $\mathrm{Tr}(\mathbf{SJ}) = p$ for the in-sample case. Further when we use standardized data, i.e. correlation instead of covariance matrices $\mathrm{Tr}(\mathbf{S})$ reduces to p. Figure 3.6 shows the mean result for 10 sample runs for standardized data. The errorbars denote the standard deviation of the ten samples.

In Figure 3.6 we see clearly that in total both *LoGo* methods on a planar 4-clique tree (blue and green lines) perform distinctly better than both *Lasso* methods (red

and magenta lines). *LoGo* on a spanning tree (olive line) is doing only slightly but still better then the *Lasso*s. Concerning the underlying graph structure of *LoGo* this clearly shows that the planar 4-clique tree indeed filters the data less drastically than the spanning tree and thereby regains important information. For a large enough sample size ($T > p = 300$) it is possible to compute the inverse of the sample covariance matrix as well (black line). It is no surprise that it outperforms every other method, since it is by definition the maximum likelihood estimate. So in conclusion, we observe that *LoGo* on a planar 4-clique tree does better at the in-sample log likelihood than the cross-validated and the equal sparse *Lasso*. In fact, that is what the network retrieval algorithm based on the determinant (Algorithm 2) was intended to do. Surprising is, that *LoGoSqu* performs almost equally good. On the other hand, note, that for *LoGoST*, computing the maximum spanning tree based on Equation 3.60 and the square of the correlation matrix entries is actually equivalent. In future works one could continue exploring this relationship between the determinant, the square of the correlation matrix entries and the likelihood. This is important, since the computation time for the determinants is higher than for the squares of the covariance matrix entries, which makes the *LoGoSqu* more advantageous regarding computation speed in comparison to *LoGoDet*. However, this difference in computation time is only minor compared to the huge execution time that is needed to compute the *LassoCV* estimate.

Off-sample log likelihood. The in-sample log likelihood illustrated the principal testing set-up we want to extend now further. We have seen that it is possible with *LoGo* to beat the state-of-the-art method at the log likelihood at the in-sample case where we compute the likelihood with the same data that was used to compute the estimates. In practice, however, this quantity is not of much interest, since the data is already observed. It is not interesting how well a model describes old data but how well it is suitable for predictions. That is why we want to repeat the whole testing process of the in-sample case but instead on calculating the log likelihood on the set of data that was used to compute the estimates we are calculating the log likelihood on a new set of data from the Gaussian distribution with the same length T. One also speaks of a *training set* on which the methods calculate the estimates and a *testing set* on which the estimates are tested. Consequently the off-sample log likelihood reads

$$\frac{1}{T}\Big(\ln \mathcal{L}(\mathbf{J}) - \ln \mathcal{L}(\mathbf{1}_p)\Big) = \frac{1}{2}\Big(\ln\det(\mathbf{J}) - \mathrm{Tr}(\mathbf{S}\mathbf{J}) + \mathrm{Tr}(\mathbf{S})\Big), \tag{3.64}$$

where $\mathbf{S}$ denotes now the covariance matrix of the off-sample (testing set). Figure 3.7 shows the result for standardized data with an additional yellow line, denoted *Max*, that is no real estimate but the inverse sample covariance matrix calculated on the testing set. So it constitutes by definition the maximum reachable value.

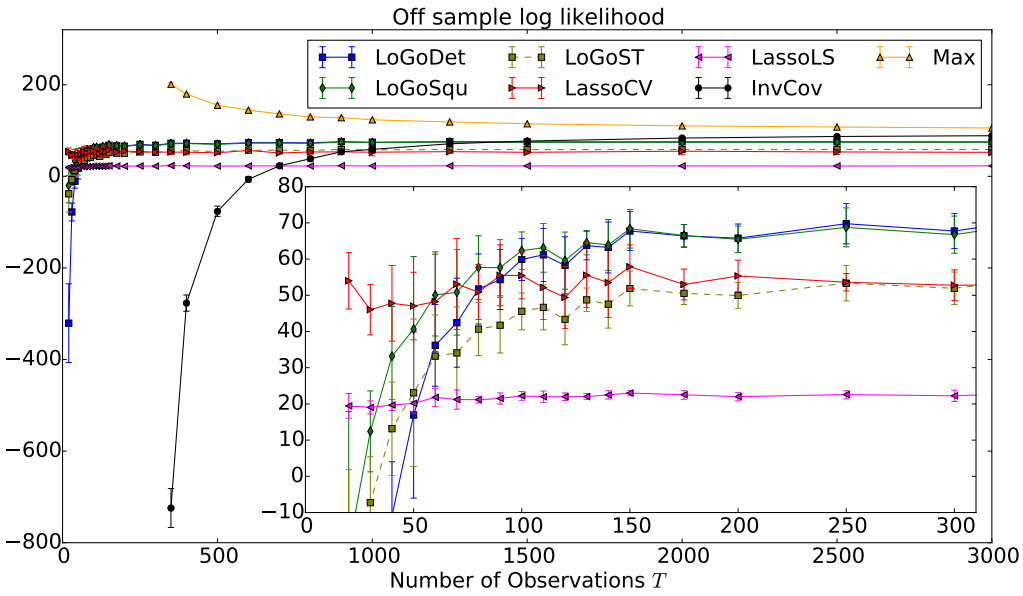

Figure 3.7: Log likelihood for various estimates of the inverse correlation matrix of simulated data for the off-sample-case, i.e. where the estimates were computed on a training set and the likelihoods on a testing set of equal length. The off-sample log likelihood is plotted versus the length of the datasets T. The number of variables is fixed at $p = 300$. The estimates were computed with the methods described in Section 3.4.2. Additionally *InvCov* (black) denotes the inverse of the sample correlation matrix and *Max* denotes the inverse of the sample correlation matrix of the testing set, i.e. the maximum value possible for the log likelihood. The dots are the mean of 10 sample runs, the errorbars denote the corresponding standard deviations.

First of all, we are able to observe that all *LoGo* and *Lasso* methods are able to estimate an inverse covariance matrix that yields a likelihood that is comparable to the regime of infinite observations, approximated at the most right of Figure 3.6, even at the high-dimensional regime where $T \lesssim p$ and the sample covariance matrix (black line) is theoretical impossible to invert or yields very bad results for the likelihood. For very few observations ($20 \lesssim T \lesssim 50$) we note that both *Lasso* algorithms perform distinctly better than all three *LoGo* methods with *LoGoST* being worse than *LoGoSqu*, but better than *LoGoDet*. The bad performance of the latter could result from the intention to maximize the likelihood on the training set that lets also the inverse covariance (black line) perform so poorly at few observations. Note, that $T = 50$ observations corresponds to a number of observations of only one sixth of the number of variables, $T = p/6$. In total, we observe that just as in Figure 3.6 *LoGoSqu* does comparable well than *LoGoDet*. So for $50 \lesssim T \lesssim 100$ both *LoGo* algorithms on a 4-clique tree perform approximatively as good as the cross-validated *Lasso*, that aims to find the regularization parameter that yields the best possible off-sample likelihood. Hence, it is not surprising that for all T *LassoCV* outperforms *LassoLS*. If one looks at the number of parameters of these models, i.e. the number of non-zero off-diagonal entries, one will find that the *LassoCV* model needs more parameters than the other sparse estimates, since for the

*LoGo*s the number of parameter is fixed by definition and *LassoLS* is calculated such that its number of parameters is equal to the one of *LoGo* on a 4-clique tree. Thus, with the principle of parsimony in mind, this would favour here rather the *LoGos* than *LassoCV* since both are doing equally well in describing the data while the *LoGo*s being more sparse. Being even more sparse *LoGoST* performs worse than *LassoCV* at this regime. Then for $T \gtrsim 150 = p/2$ both planar 4-clique *LoGo* methods outperform both *LassoCV* algorithms along the way and *LoGoST* is doing approximately equivalent or slightly better than *LassoCV*. Note that it takes around 700 observations until the sample inverse (black line) lies in the area of our sparse estimates and a total of around 1500 observation to finally beat the *LoGo*s. This is clearly out of the high-dimensional regime, where *LoGo* aims to contribute. Considering the early stage of the development of *LoGo* these are very promising results.

3.4.3 Real-world example

Let us finally take a look at some dataset from the real world. In the following we will repeat our previous tests except we will not use simulated data but instead 342 US stock prices. As before, for one sample run we choose randomly $p = 300$ out of the 342 variables in order not to rely on a specific set of variables and compute the estimates of the inverse covariance matrix as described in Section 3.4.2 and the corresponding values of the likelihood for various numbers of observations T.

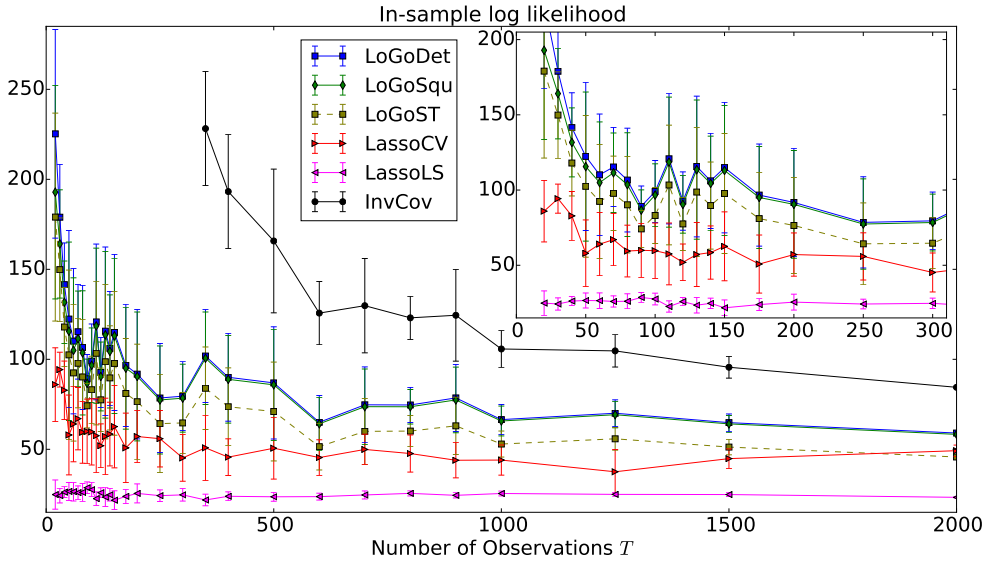

Figure 3.8: Log likelihood for various estimates of the inverse correlation matrix of US stock market data for the in-sample-case, i.e. where the likelihood is computed with the set of data, that was used to compute the estimates. The in-sample log likelihood is plotted versus the length of the dataset T. The number of variables is fixed at $p = 300$. The estimates were computed with the methods described in Section 3.4.2. Additionally *InvCov* (black) denotes the inverse of the sample covariance matrix. The dots are the mean of 10 sample runs, the errorbars denote the corresponding standard deviations.

In-sample log likelihood. The in-sample log likelihood for standardized variables is shown in Figure 3.8. First of all it strikes that Figure 3.8 is comparable to Figure 3.6 but with distinctly more noise indicated by the larger error bars, i.e. standard deviations of the sample runs. This clearly justifies the tests with simulated data. Nevertheless, we still are able to observe that especially the planar 4-cliques *LoGo*s outperform both *Lasso* methods at all T. By itself, this is an encouraging result. However, as before, the off-sample case is of more interest.

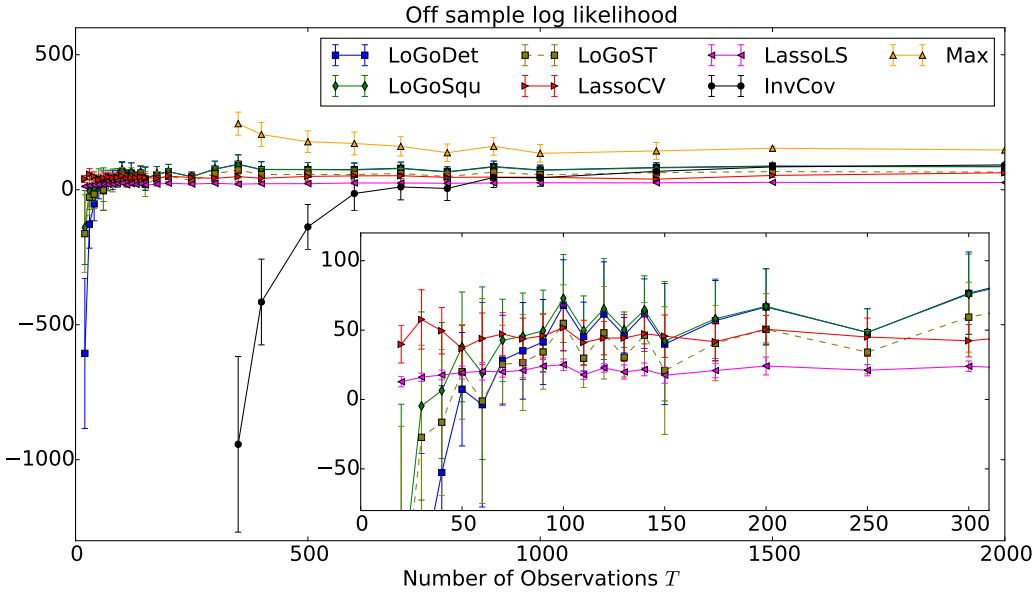

Figure 3.9: Log likelihood for various estimates of the inverse correlation matrix of US stock market data for the off-sample case, i.e. where the estimates were computed on a training set and the likelihoods on a testing set of equal length. The off-sample log likelihood is plotted versus the length of the datasets T. The number of variables is fixed at $p = 300$. The estimates were computed with the methods described in Section 3.4.2. Additionally *InvCov* (black) denotes the inverse of the sample correlation matrix and *Max* denotes the inverse of the sample correlation matrix of the testing set, i.e. the maximum value possible for the log likelihood. The dots are the mean of 10 sample runs, the errorbars denote the corresponding standard deviations.

Off-sample log likelihood. Figure 3.9 shows the off-sample log likelihood (Equation 3.64) for the methods described in Section 3.4.2 to obtain a sparse estimate of the inverse covariance matrix. Additionally the inverse of the sample covariance matrix of the training set (*InvCov*) and the inverse of the sample covariance matrix of the testing set (*Max*) is shown for $T > p$. Overall, we observe a similar behaviour than for the test with simulated data (Figure 3.7). Although here the level of noise has increased, as can be seen in the larger error bars ,which is equivalent to the in-sample case (Figure 3.8), on average it still holds that both planar 4-clique tree *LoGo*s outperform the *Lasso*s for $T \gtrsim 150$. The log likelihood of *LoGoST* fits in general between the one of *LassoCV* and the one of both planar 4-clique *LoGo*s. For $T < 150$ we recover also the facts from

Figure 3.7 of the simulated data case. At very few observations the *LoGo*s are doing distinctly worse than the *Lasso*s. But already for $T \gtrsim 50$ one enters a regime where the performance of all methods is on an equal footing. By that we are able to reproduce our results of the tests with simulated data also with data from the real world which now clearly proofs the promising state of *LoGo*.

3.5 Conclusion

In this chapter I introduced a novel approach to the problem of estimating a sparse inverse covariance matrix in the high-dimensional data setting. First I showed why the sample covariance fails to be a useful estimate at high dimensions. I also motivated the idea behind the maximum likelihood estimates and derived the explicit expression of the log likelihood equation as a function of the inverse covariance matrix for the multivariate Gaussian distribution. I proceeded with presenting the theoretical background to establish the links between network filtering techniques and Gaussian Markov Random Fields. Here I showed how these concepts together can be interpreted as an operational tool of the fundamental principal of parsimony.

Based on these theoretical findings we derived a practical equation that assigns unique values to the non-zero entries of a matrix with an underlying decomposable graph structure, such that this matrix serves as a sparse estimate of the inverse covariance matrix. The key of this ansatz, we call *LoGo*, is to invert the sample covariance matrix **lo**cally and compose the local parts together to a **glo**bal estimate of the inverse covariance matrix. It has several beneficial features. On the one hand it is exact given the underlying graph structure is decomposable, i.e. it will not introduce new errors. Further the values are assigned to the non-zero entries such that the result is a maximum likelihood estimate among all possible estimates of the inverse covariance matrix with the same underlying graph. On the other hand it uses less information to estimate the inverse covariance matrix than the inversion of the sample covariance matrix which makes it more stable to initial errors due to noise or the finiteness of the data series. I illustrated the use of *LoGo* for two graph structures, the spanning tree and the planar 4-clique tree. For each, I proposed improved filtering algorithms to obtain the actual network with the aim to maximize the likelihood on the give set of data.

Comparisons of *LoGo*, both on the spanning tree and the planar 4-clique tree, with the state-of-the-art method of finding a sparse inverse covariance matrix *Lasso* were performed on several likelihood measures, both with simulated and real-world data. Although *LoGo* is still at its infancy it managed to compete with and partly to outperform *Lasso* at the respective test scores. At the same time *LoGo* is computationally more efficient and usually more sparse than the *Lasso* estimates. More tests and practical applications have to be performed and examined in order to asses the full potential that lies in this idea.

4 Random Packings

4.1 Introduction

In the following chapter we turn away our attention from the high dimensional data sets and focus on a classical example of low dimensional data where the number of observations exceeds the number of variables by far. Nevertheless, here also the correlation matrix will help us to understand the underlying system better. But in contrary to the last chapter, the key in this case lies in a proper visualization, as we will see below.

In particular, we will deal here with the morphological properties of two-dimensional random packings of ellipses of different elongations up to different area fractions. The area fraction hereby is defined as the area covered by the packed ellipses over the total background area. Besides a general interest of packing problems (Weaire and Aste, 2008) this approach is motivated especially from biology, namely from the study of epithelial tissue, that only occurs two-dimensionally. While experimenting with these kinds of tissues, one lets the tissue cells artificially grow, where they start at an area fraction of approximately 0.2 until they reach an area fraction of approximately 0.65. Further one has found, that the cell nuclei can be very well approximated with ellipses and the actual cells with a Voronoi tessellation. Figure 4.1 serves as an illustration of these concepts.

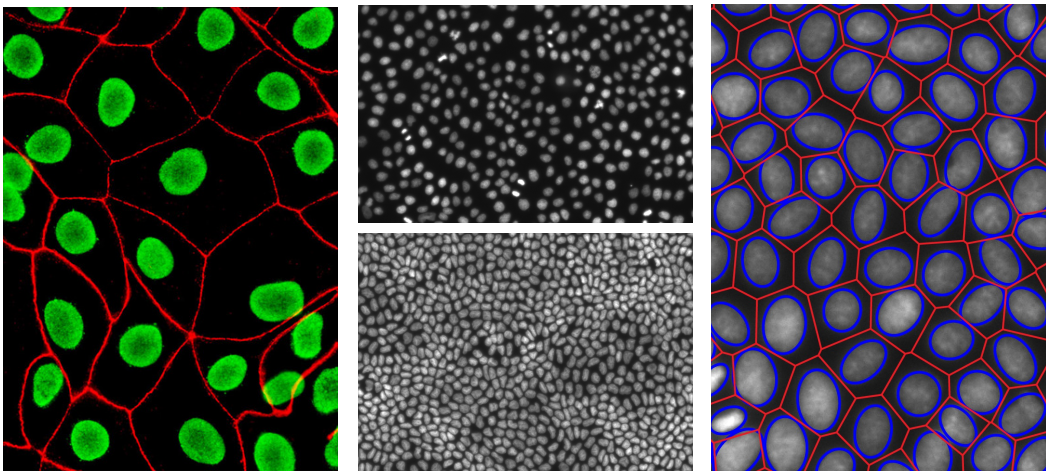

Figure 4.1: Epithelial Tissue. (LEFT) The nuclei are coloured in green, the cell boundaries are shown in red. (MID) Growth process: (MID-TOP) At low area fraction of approx. 0.2. (MID-BOTTOM) At high area fraction of approx. 0.65. (RIGHT) Voronoi tessellation of epithelial tissue. The nuclei are approximated with ellipses and the cell boundaries with the Voronoi cells.

Now from the point of view of epithelial tissue it seems evident to systematically study the characteristics of packings of ellipses of different elongations along different area fraction for a better understanding of this biological system. In fact, it is of great interest to determine the specific features of epithelial tissue, that just result from the fact that the epithelial cells are randomly packed ellipses at a specific area fraction. In the contrary, those features, one cannot provide a packing explanation for, one has to search for a true biological reason. In that manner, the following analysis contributes to a reference study of random packings.

This chapter will continue with outlining a clear way to systematically study two dimensionally random packings of ellipses for different area fractions and different elongations of the ellipses in Section 4.2. Section 4.3 will present and discuss the results by an integrated way of visualizing the global correlation matrix. Finally, in Section 4.4 I will conclude and summarize the main findings.

4.2 Methods

In the following I will present the methodology we used to systematically study the packing of ellipses of different elongations up to different area fractions.

4.2.1 Voronoi Tessellations

I start by briefly explaining the concept of a Voronoi tessellation, which is the partition of a plane with T generating points into Voronoi cells, such that each Voronoi cell contains exactly one generating point and every point in a given Voronoi cell is closer to its generating point than to any other. Figure 4.2 illustrates this definition. The black dots represent the generating points and each of the coloured fields corresponds to one Voronoi cell.

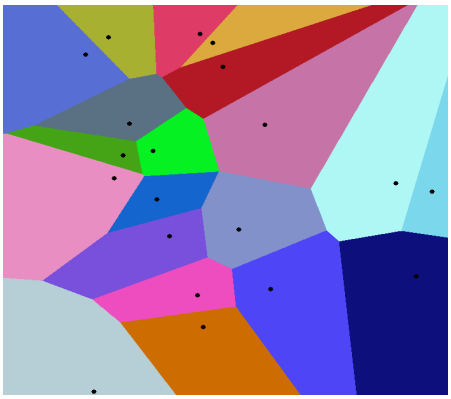

Figure 4.2: Example of a Voronoi tessellation. The black dots represent the generating points and each of the coloured fields corresponds to one Voronoi cell.

4.2.2 Morphological Measures

Now let us define the morphological measures that we compute for every Voronoi cell in the tessellation. These are

- the area of the cells,
- the number of neighbours,
- the elongation of the cells,
- the distance from the centre of mass of the packed particle to the centre of mass of the Voronoi cell,
- the perimeter of the cells,
- the mean of the contact lengths of the cells with their neighbours and
- the standard deviation of the contact lengths.

Two Voronoi cells are said to be neighbours if they share one part of their boundary. To determine the elongation of the cells we used the principle axes of the moment of inertia (assuming a homogeneous mass density) to fit an ellipse to the cell. From the semi-axes of the ellipse, we compute the elongation $e = a/b$ where a (b) denotes the longer (shorter) semi-axis.

On top of that we were also observing the behaviour of the rescaled versions of the last four of this list, which come in units of length. The rescaled measure is obtained by dividing the actual measure by the square root of the area which is equivalent to setting the area to 1.

4.2.3 Simulation

Now, having understood the concept of Voronoi tessellations, we will give a brief overview on how we actually generated the packings that we will study later on.

In order to simulate that type of two dimensional random packings we are interested in, we use a collision-driven molecular dynamics algorithm (Donev et al., 2005a,b). Conceptually the simulation works like explained in the following. First one chooses the desired elongation of the particles to be packed,

$$e^{-1} = \frac{b}{a}, \tag{4.1}$$

where b (a) denotes the shorter (longer) semi-axis of the ellipse shaped particles. The inverse of the elongation e^{-1} was chosen in order to have a bounded parameter, i.e. $0 < e^{-1} \leq 1$. Moreover, by that it servers to distinguish between the elongation of the

packed particles (denoted with e^{-1}) and the elongation of the Voronoi cells (denoted with e). The desired area fraction reads

$$\rho = \frac{\sum_{i=1}^{T} \text{area of ellipse } i}{\text{Total background area}}, \tag{4.2}$$

where T denotes the total number of particles. The algorithm will place T point particles randomly into the background area and let them grow according to the set elongation e^{-1}. If the desired area fraction ρ is reached the simulation stops with the centre of mass coordinates and orientations of the particles as an output. If two particles collide during the growth process before ρ is reached, the particles are able to rotate and move slightly in order to fit in the free space to reach higher area fractions. Figure 4.3 illustrates the algorithm conceptually.

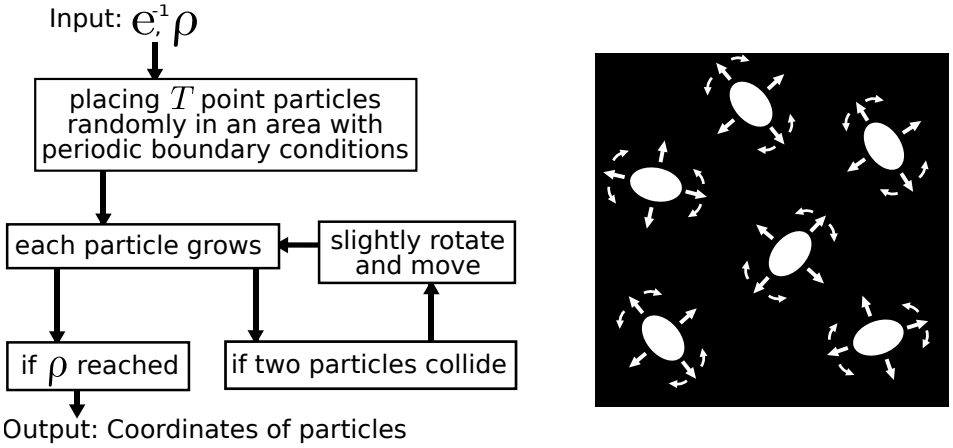

Figure 4.3: Conceptual illustration of the algorithm to produce the random packings.

Now from the final coordinates of the particles we are able to calculate the Voronoi tessellation and the morphological measures listed above. Here, two possible ways of calculating the tessellation are possible. On the one hand one could compute the Voronoi tessellation on the basis of the centre of mass of the packed particles. From these centres of mass the Voronoi cells are constructed as shown in Figure 4.2. On the other hand one can also construct the Voronoi tessellation on the basis of the shape of the packed ellipses, i.e. the generating points become generating shapes and consequently the Voronoi cells follow the form of their ellipses. Moreover, it is ensured that each Voronoi cell covers fully its particle, which is not the case for the centre of mass based tessellation. These make the shape based tessellation the more realistic one, which is also found in practice, since here the shape based tessellation coincides better with the real tissue cell boundaries than the centre of mass based ones. On the downside for shape based tessellations one has to state that they are computationally more difficult to calculate. Since we found out that the results of these two ways of constructing the Voronoi cells do not differ qualitatively I will use from now on the centre of mass based tessellations.

4.2.4 Correlation Analyses

Now we have several data matrices with the morphological measures as variables and the number of cells as observations. Each data matrix corresponds to a pair of area fraction ρ and elongation e^{-1} which are the parameters of interest of this study. Consequently, following the general theme of this thesis we will apply Equation 2.13 to compute the correlation matrices for these sets of data. However, now we have twelve correlation matrices, since we used three area fractions $\{0.2, 0.5, 0.8\}$ and four different inverse elongations $\{0.3, 0.5, 0.8, 1.0\}$ (Thus, an elongation of $e^{-1} = 1$ corresponds to packed circles). So the question arises, how one can represent these twelve correlation matrices in an integrated way such that one can observe the overall system at once?

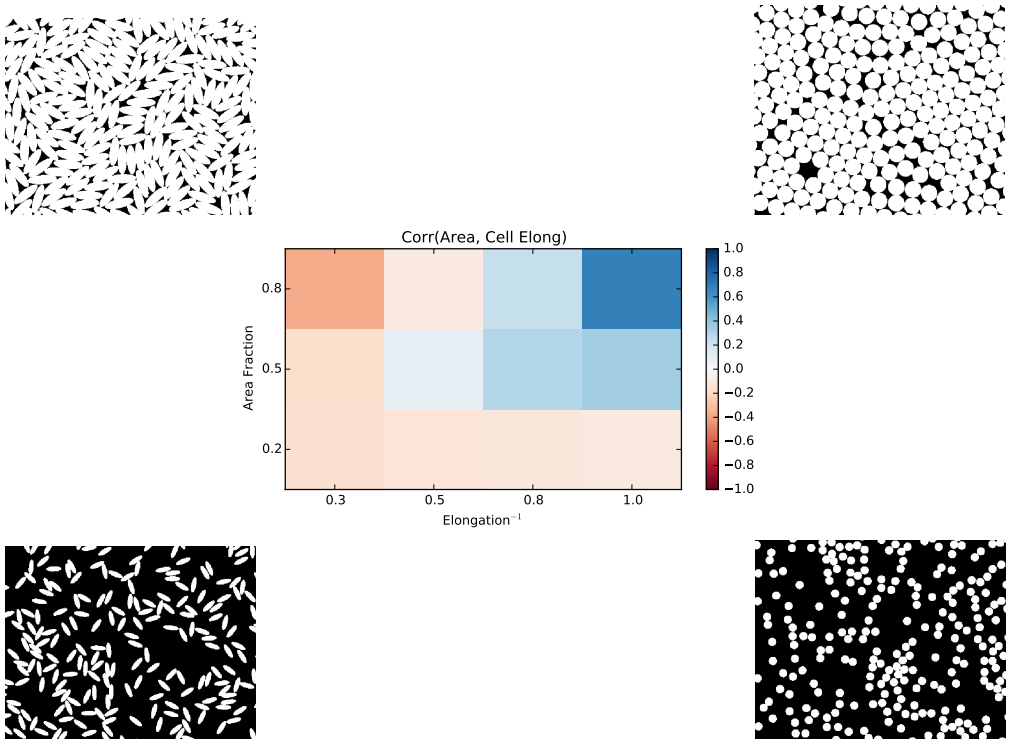

Figure 4.4: Example of the representation of correlation between two measures (here area and elongation of the cells). The elongation of the packed particles is put on the x-axis, the area fraction on the y-axis and the value of the correlation is encoded in the color (blue corresponds to positive correlation, red to negative correlation). Additionally, examples of the packings for the corner cases are shown.

Figure 4.4 illustrates the proposed approach. Here an *entry* of the global correlation matrix is shown. It consists of a plot of the correlation between two morphological measures for the whole phase space of elongation of the packed particles (x-axis) and area fraction (y-axis). The values of the correlation coefficients are encoded in the colour, where blue denotes a positive and red a negative correlation. Additionally for

the extreme cases (at the corners) images of the packed systems are shown to support this way of visualizing the correlation.

4.3 Results

Having introduced all the necessary methodology we are now able to focus on the inter-relations the morphological measures have with each other. But before we proceed with the analyses of the correlations I want to give a brief overview how the measures behave individually in the phase space.

4.3.1 Individual Measures

The study of the individual morphological measures involves the observation and comparison of the histograms of the measures for various elongations of the packed particles and area fractions of the packings as shown in Figure 4.5 for low area fraction.

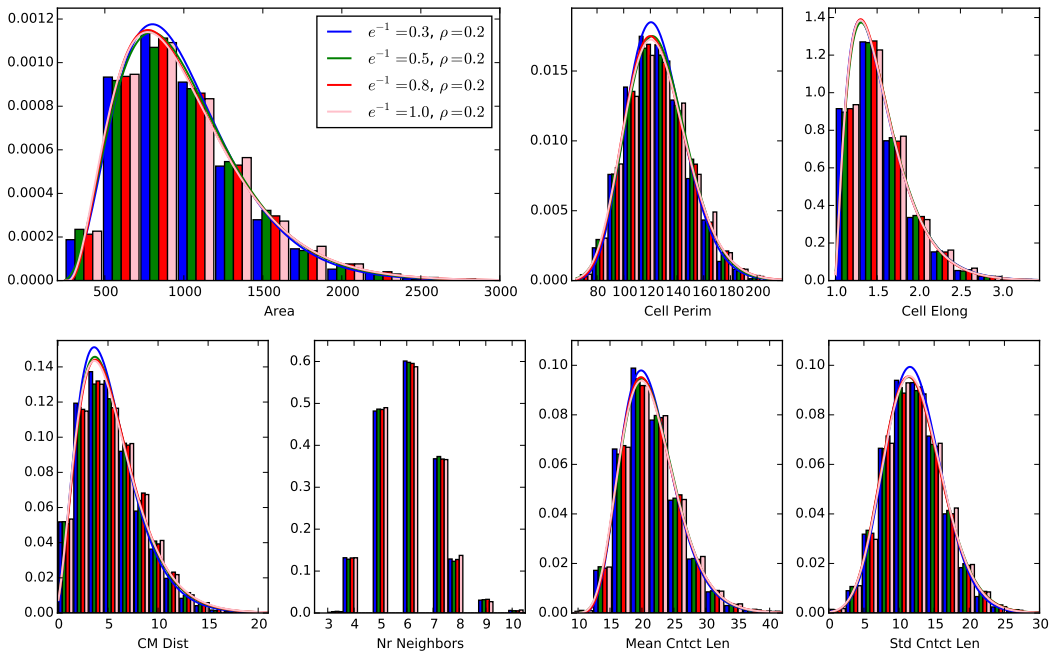

Figure 4.5: Histograms of various morphological measures at low area fraction $\rho = 0.2$ and various elongations e^{-1} of the packed particles. Additionally, except for the discrete number of neighbours, gamma distributions with maximum likelihood parameters are shown as well. Despite a relative good accordance between the data and the fits, a more detailed analyses revealed that the proposed occurrence of gamma distributions in these packings is still open to be debated. At low area fraction it strikes, that the histograms qualitatively do not differ along the elongations e^{-1}.

Additionally the estimation of specific probability density functions for the distributions are here of interest. It has been proposed by Aste and Di Matteo (2008) that gamma distributions should emerge naturally in these kinds of packing systems. Accordingly gamma distributions are shown in Figure 4.5 as well (except for the discrete number of neighbours) where the parameters of the gamma distribution were found by the maximum likelihood method (see Section 3.1). However, a more detailed analysis (that I will not present here) revealed, that this proposition is still open to be debated. Nevertheless, it strikes that the histograms do almost not differ for the various elongations of the packed particles. This can be explained, as we will also see later at the correlation analyses, by the fact that at low area fractions the morphological properties of these packings are independent of the shape of the particles.

On the other side Figure 4.6 shows the histograms of the morphological measures plus the gamma distributions with maximum likelihood parameters for high area fraction. Here, we can clearly observe that different elongations of the packed particles result in different distributions. We observe from Figure 4.6 the general trend is that for more elongated packed particles $e^{-1} \to 0.3$ the distributions get wider. This indicates on the one hand the very regular packing at high area fraction for circles ($e^{-1} = 1.0, \rho = 0.8$) and on the other hand the thereof different structure of the packing at high area fraction for very elongated particles ($e^{-1} = 0.3, \rho = 0.8$). Both packings will also be recognized and discussed through the correlation analyses below.

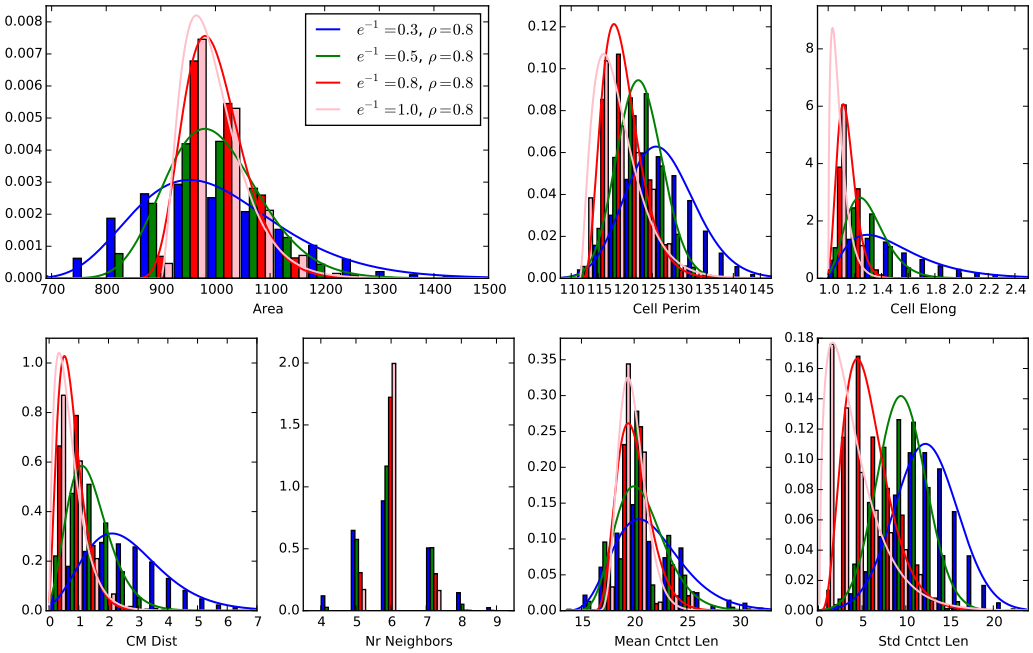

Figure 4.6: Histograms of various morphological measures at high area fraction $\rho = 0.8$ and various elongations e^{-1} of the packed particles. Additionally, except for the discrete number of neighbours, gamma distributions with maximum likelihood parameters are shown as well.

But note that even the wide distributions for very elongated packed particles and high area fraction of Figure 4.6 are still more narrow than the corresponding distributions for low area fraction at Figure 4.5.

4.3.2 Correlation Overview

We start the analyses of the inter-relations between the morphological measures with an overview how their mutual correlation coefficient varies along the phase space of elongation of the packed particles and area fraction of the packing in Figure 4.7. In the following I want to highlight and discuss certain aspects of this plot.

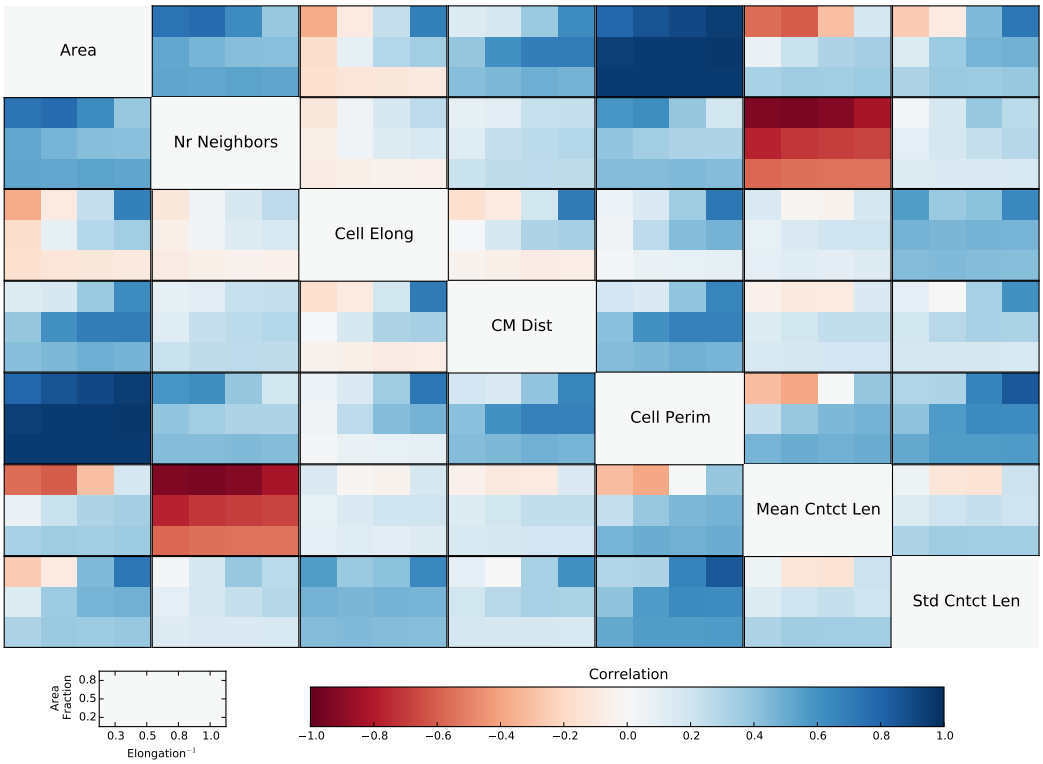

Figure 4.7: Overview of the correlation of all morphological measures. Similar to Figure 4.4 at each *entry* of the correlation matrix, the elongation of the packed particles is put on the x-axis, the area fraction on the y-axis and the value of the correlation is encoded in the color (blue corresponds to positive correlation, red to negative correlation).

4.3.3 At Low Area Fraction

Looking especially at the bottom row of each entry of the correlation matrix shown in Figure 4.7 (i.e. at low area fraction) one can easily observe that the correlation does

qualitatively not change along the x-axis of the entries (i.e. along the elongation of the packed particles), which is consistent with our observations of Figure 4.5. This behaviour can be interpreted as follows: At low area fraction the particles have enough space to almost freely align themselves regardless of the other particles. In such an environment the shape of the particles must not matter, which is indeed what we observe in the correlations. Figure 4.8 shows the actual packings and the corresponding Voronoi tessellations at low volume fraction for the extreme cases of elongation and confirms our explanation.

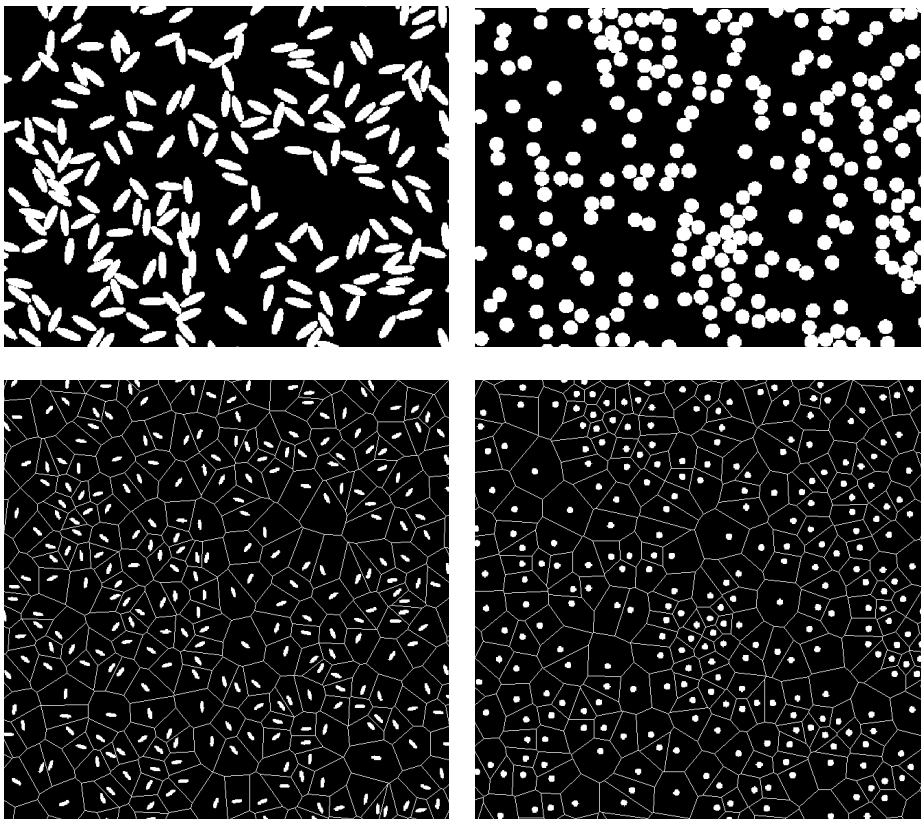

Figure 4.8: Packings at low area fraction. (LEFT) for elongation of particles $e^{-1} = 0.3$; (RIGHT) for elongation of particles $e^{-1} = 1$. The actual packings are shown at the (TOP), the corresponding Voronoi tessellation are shown at the (BOTTOM) where the shape of the particles is reduced and just serves as a reminder of the packed particles, since here the tessellation are constructed solely on the basis of the centre of mass of the particles.

At high area fraction. In the following we want to take a closer look at the correlations for a high area fraction, since in this case the correlations vary along different elongations of the packed particles as one can easily observe in Figure 4.7.

4.3.4 Measures of Elongation

In this section I discuss the perimeter, elongation and standard deviation of the contact lengths. I hereby argue that the rescaled versions of the perimeter and the standard deviation of the contact lengths may be regarded as well as measures of the elongation of the cells.

Cell Perimeter. Let us start with a rather obvious correlation, namely the one between the area and the perimeter of the cells. As one sees in Figure 4.7 there is an overall strong positive correlation between those two measures, i.e. bigger cells tend to have a larger perimeter, so far, so obvious. Nevertheless, the slight decrease of correlation at the upper left corner (i.e. high area fraction, highly elongated packed particles) is interesting and needs to be explained. So let us look at the rescaled version of the perimeter, which is obtained by dividing the original perimeter by the square root of the area in Figure 4.9.

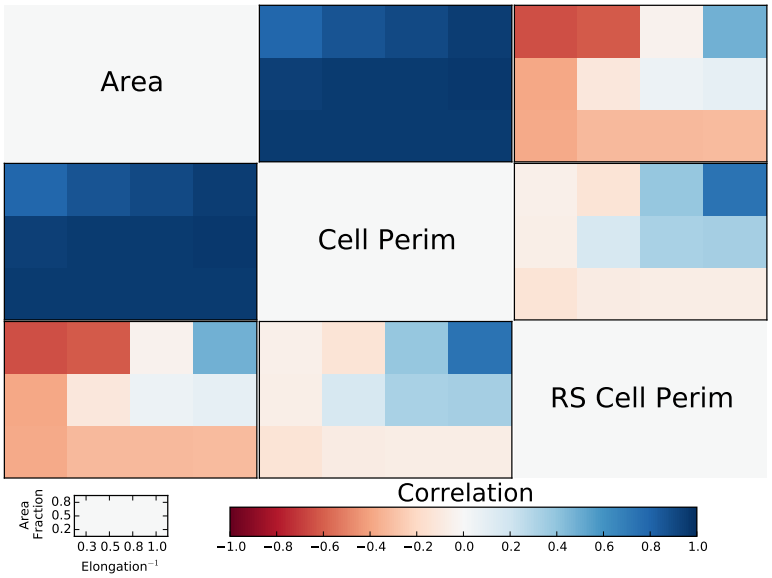

Figure 4.9: Correlation of the area, perimeter and rescaled perimeter of the Voronoi cells. Similar to Figure 4.4 at each *entry* of this correlation matrix, the elongation of the packed particles is put on the x-axis, the area fraction on the y-axis and the value of the correlation is encoded in the color (blue corresponds to positive correlation, red to negative correlation).

Here we observe a qualitative total different behaviour of the correlation between the area and the rescaled cell perimeter than between the area and the non rescaled cell perimeter. Here, for a high area fraction and highly elongated packed particles (i.e. upper left corner), we observe a negative correlation, meaning for bigger cells the rescaled cell perimeter tends to be smaller and thus, the cells more circle-like. In the contrary

for circles as the packed particles (and still at high area fraction) there is a positive correlation, thus, for bigger cells the cells tend to be less circle-like. Since the original cell perimeter carries also the information about the elongation of the cells, that is captured by the rescaled cell perimeter, the slight decrease of correlation at the upper left corner in the correlation between area and actual cell perimeter is explainable as a result of different elongations of the cells.

Let us try to formalise this relationship between the elongation of the cell $e = a/b$ (where a (b) denotes the longer (shorter) semi axis of the fitted ellipse of the Voronoi cell) and the rescaled cell perimeter $\hat{P} = P/\sqrt{A}$ (with P being the non rescaled perimeter and A the cell area). For a circle we would have $a = b$ and thus $e = 1$ and $\hat{P} = 2\pi r/(r\sqrt{\pi}) = 2\sqrt{\pi}$. What changes if the circles transforms to an ellipse? The area of an ellipse reads $A = ab\pi$ and the perimeter is calculated with infinite series which reads up to the seconds order

$$P = 2\pi a \left(1 - \frac{1}{4}\frac{a^2 - b^2}{a^2} + \dots\right). \tag{4.3}$$

Consequently, we can write up to the first order

$$\hat{P} = \frac{p}{\sqrt{A}} = 2\sqrt{\pi}\sqrt{\frac{a}{b}}\left(1 - \frac{1}{4}\frac{a^2 - b^2}{a^2} + \dots\right) \propto \hat{P}|_{\text{circle}} \cdot \sqrt{\frac{a}{b}}. \tag{4.4}$$

The elongation of the ellipses changes of course as $e = e|_{\text{circle}} \cdot a/b$. With the square root being a monotonically increasing function one can now certainly explain the similar behaviour of the rescaled perimeter and the cell elongation regarding the correlation coefficient.

Cell Elongation. Now, let us take a look at the correlation between the area and the elongation of the Voronoi cells. And indeed we observe a qualitative similar correlation between those two measures as between the area and the rescaled cell perimeter, which confirms our calculation above. So at high area fraction, for highly elongated packed particles, bigger cells tend to be more like circles whereas for packed circles, bigger cells tend to be more elongated. Figure 4.10 shows the actual packings for high area fraction and the corresponding Voronoi tessellations for packed circles as well as for highly elongated packed particles. From there, one observers clearly the difference of the packings, whether circles or highly elongated ellipses are being packed. The circles form almost a regular hexagonal grid with equal cell sizes, that is interrupted by a few defections that have a bigger area and a more elongated cell shape than the regular cell. This explains the positive correlation between area and elongation of the cells as well as between area and rescaled perimeter for the upper right corner. In the contrary for highly elongated particles one observes in Figure 4.10 *worm*-like structures in the Voronoi tessellation, where the particles align themselves along their longer semi-axis. These cells tend to be more elongated and smaller in area as the ones next to these *worms*. This explains the negative correlation between area and cell elongation and rescaled cell perimeter, respectively.

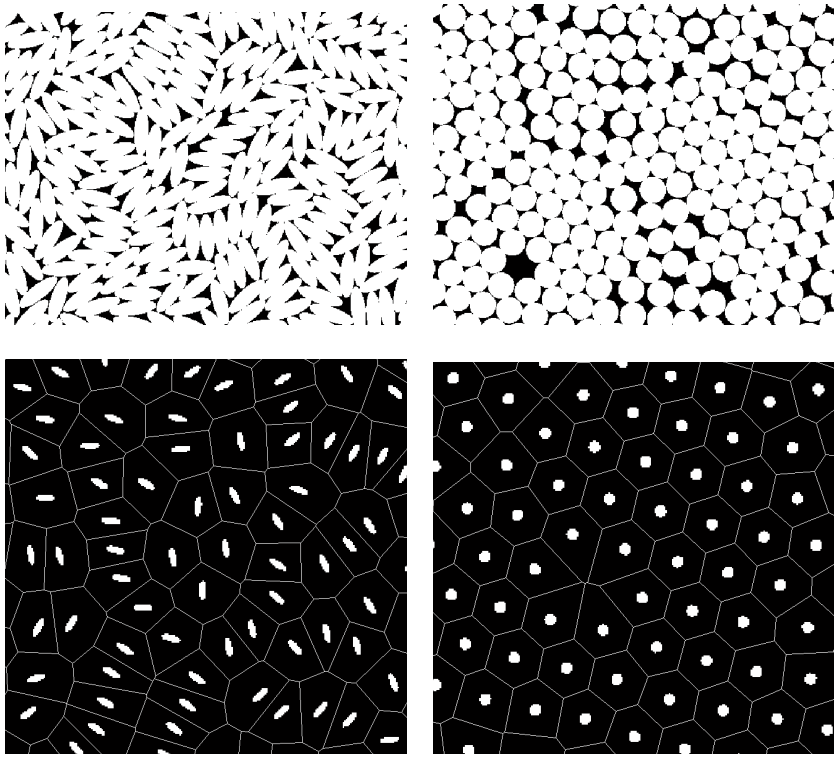

Figure 4.10: Packings at high area fraction. (LEFT) for elongation of particles $e^{-1} = 0.3$; (RIGHT) for elongation of particles $e^{-1} = 1$. The actual packings are shown at the (TOP), the corresponding Voronoi tessellation are shown at the (BOTTOM) where the shape of the particles is reduced and just serves as a reminder of the packed particles, since here the tessellation are constructed solely on the basis of the centre of mass of the particles.

Compare these packing structures also with Figure 4.6 that shows the distributions of the morphological measures at high area fractions. Now it might get intuitively more clear how the more narrow distributions for packed circles emerge from the almost perfect hexagonal grid in comparison to the wider distributions at the *worm*-like structure packing for very elongated packed particles.

Standard Deviation of the Contact Lengths. Let us focus in the following on the standard deviation of the contact lengths of each cell. As one can imagine for highly elongated cells this measures is larger compared to circle like cells. So in that sense we would expect that the standard deviation of the contact lengths servers also as a measure of elongation. And indeed from Figure 4.7 we observe an overall positive correlation between the standard deviation of the contact lengths and the cell elongation. Figure 4.11 shows the discussed measures of elongation, together with the area and the rescaled version of the standard deviation of the contact lengths.

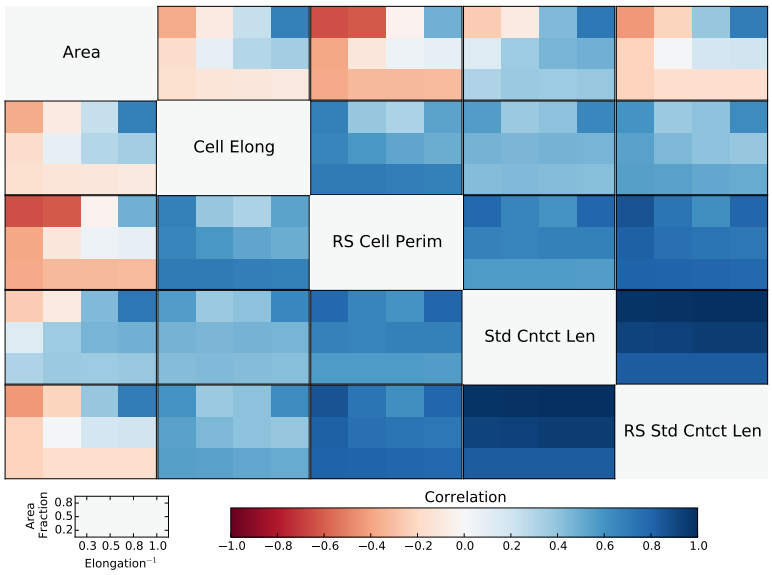

Figure 4.11: Correlation of the area and the measures of elongation: cell elongation, rescaled cell perimeter and the rescaled standard deviation of the contact lengths plus non rescaled standard deviation of the contact length. Similar to Figure 4.4 at each *entry* of this correlation matrix, the elongation of the packed particles is put on the x-axis, the area fraction on the y-axis and the value of the correlation is encoded in the color (blue corresponds to positive correlation, red to negative correlation).

As we can see clearly as an overall trend, all suggested measures of elongation are positively correlated (blue coloured lower-right sub matrix excluding the area). When we compare the correlations of the measures of elongation and the area we observe also a similar behaviour. Note that for high area fraction (upper row in each correlation matrix entry) we observe a clear monotonously increasing gradient from a negative correlation at the upper left corners (i.e. highly elongated packed particles) to a positive correlation at the upper right corners (i.e. packed circles). If we look at the bottom row of these correlations, we see a slight negative correlation except for the correlation between the area and the standard deviation of the contact lengths. Consequently, sufficient criteria from this correlation analyses to determine measures of elongation are

- A monotonically increasing gradient at high area fraction from a negative correlation at elongation of the packed particles of $e^{-1} = 0.3$ to a positive correlation at $e^{-1} = 1.0$, such that $e^{-1} = 0.3$ and 0.5 have negative correlations and $e^{-1} = 0.8$ and 1.0 have positive correlations.

- A slight negative correlation at low area fraction of approximately 0.2.

Accordingly, the non rescaled standard deviation of the contact lengths is not regarded as a measure of elongation from these criteria. But how could we formalize the relationship between the (rescaled) standard deviation of the contact lengths and the elongation of the Voronoi cells? Let us start, similar to the considerations of the relationship between

the rescaled cell perimeter and the cell elongation, with a perfect hexagonal Voronoi cell, such that the fitted ellipse to determine the elongation is a circle. Here, of course, the standard deviation equals zero and the elongation equals one. Let us approximate the transition from a circle to an ellipse, such that the hexagon is stretched at two parallel edges as shown in Figure 4.12. From here we can read that the four short edges of the Voronoi cell have a length of $s = b$ and for the two long edges $l = 2a - b$. It follows for the mean contact length

$$m = \frac{4s + 2l}{6} = \cdots = \frac{2}{3}\Big(a + 2b\Big), \tag{4.5}$$

and for the variance of the contact lengths

$$v = \frac{4(s-m)^2 + 2(l-m)^2}{6} = \cdots = \frac{8}{9}b^2\Big(\frac{a^2}{b^2} - 2\frac{a}{b} + 1\Big). \tag{4.6}$$

The area of the Voronoi cell can be computed as

$$A = \frac{\sqrt{3}}{2}s^2 + \sqrt{3}ls = \cdots = 2\sqrt{3}b^2\Big(\frac{a}{b} - \frac{1}{4}\Big). \tag{4.7}$$

The rescaled standard deviation of the contact length $\hat{\sigma} = \sigma/\sqrt{A} = \sqrt{v/A}$ consequently goes as $\hat{\sigma} \propto \sqrt{a/b}$ in leading order of the elongation a/b. So we have managed to find here as well a formal relationship between the rescaled standard deviation of the contact length and the elongation of the Voronoi cells. The thought might arise to extend this argument as well to the mean m. However, as one can see from Figure 4.7, the mean contact length is highly correlated with the number of neighbours (in contrast to the standard deviation of the contact lengths). Thus, this violates the assumption of the fixed number of six neighbours, we have used during the derivation. Indeed, as we will discuss below, the mean of the contact length is highly connected to the number of neighbours.

4.3.5 Centre of Mass Distance

Let us take a closer look at the correlation between the area and the distance of the centre of mass of the cells and the centre of mass of the packed particles. Obviously one would expect a positive correlation between these two measures, the bigger the cells, the bigger also the centre of mass distance. And indeed the overall impression is a positive correlation, but interestingly at high area fraction it has its maximum for packed circles and almost vanishes for highly elongated packed particles. If we include the rescaled version of the centre of mass distance into the picture as we do with Figure 4.13 we interestingly note that in the contrary to Figure 4.9 there is no qualitative change between the correlation of area and centre of mass distance and rescaled centre of mass distance, respectively. This means that the variation of the centre of mass distance with the area is no direct result of a variation of the area, since in this case there should

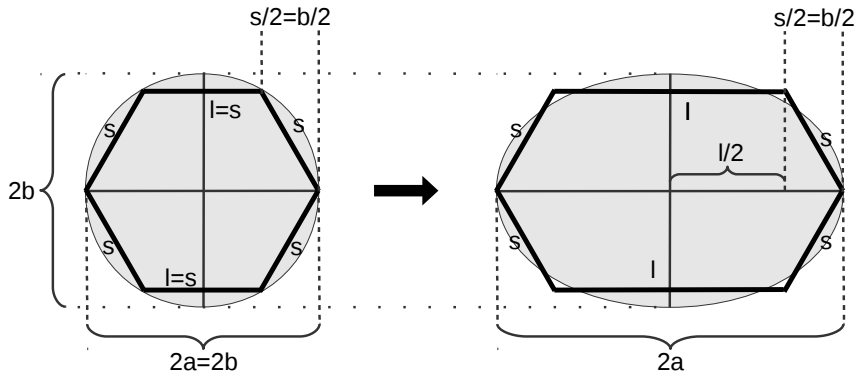

Figure 4.12: Illustration to connect the (rescaled) standard deviation with the elongation of the cell. (LEFT) A perfect hexagonal Voronoi cell with a circle as the fitted ellipse to calculate the elongation. Consequently the semi-axes of the ellipse are equal and equal the shorter and longer edges of the Voronoi tesselation, since they are also equal here: $a = b = s = l$. (RIGHT) The transition to a elongated Voronoi cell. It is assumed the the shorter semi axis b remains constant and the longer semi-axis $a = l/2 + b/2$.

be a difference in the correlation between area and the centre of mass distance and the rescaled centre of mass distance. Instead we conclude that the centre of mass distance increases with bigger cells, as the correlation suggests for a high area fraction and packed circles, is as well a result of the defections of the regular grid (see Figure 4.10) since these defections have greater area and are more elongated. The fact that they have greater area explains the occurrence of the positive correlation here and the fact that they are elongated explains intuitively why they have a bigger centre of mass distance.

This explanation is also supported by the correlation between the elongation of the cells and the centre of mass distance, as one observes in Figure 4.13. Here we have as well a strong positive correlation between the two measures for high area fraction and packed circles. On the other side at high area fraction but for very elongated packed particles the reverse can be observed. Here one finds a negative correlation between the centre of mass distance and the cell elongation, i.e. the more the cells are elongated the shorter their centre of mass distance. If we remember ourselves of the *worm*-like structure at this packing configuration (Figure 4.10) one can conclude that in this sense the typical form of the Voronoi cells in the respective packings occurs only with a low centre of mass distance. Therefore a low centre of mass distance serves to indicate this typical form. For packed circles the typical form of the Voronoi cells are hexagon-like shapes, such that the defections (which are more elongated) have also a higher centre of mass distance. This results in the positive correlation in the upper right corner. For the highly elongated packed particles the typical form are the *worms*, i.e. very elongated cells but with a low centre of mass distance. Consequently the non-*worm*-like cells (which are less elongated) get a higher centre of mass distance, i.e. positive correlation.

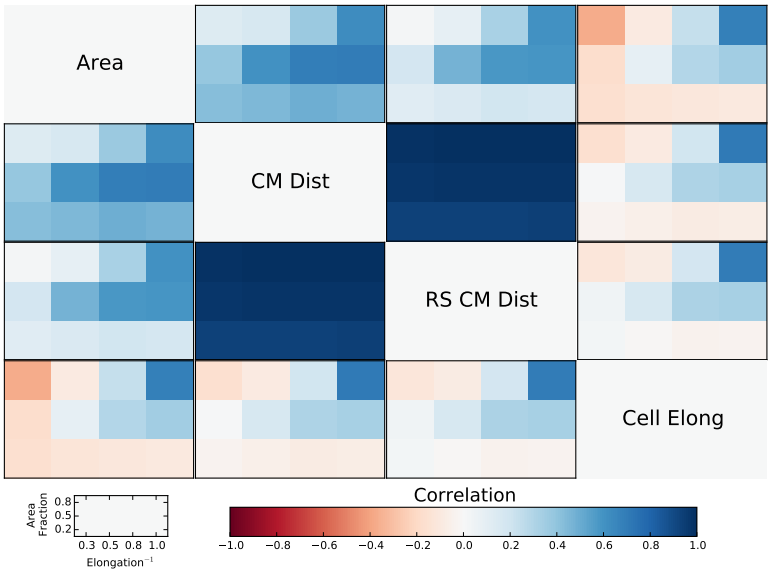

Figure 4.13: Correlations of the area, centre of mass distance, rescaled centre of mass distance and the elongation of the Voronoi cells. Similar to Figure 4.4 at each *entry* of this correlation matrix, the elongation of the packed particles is put on the x-axis, the area fraction on the y-axis and the value of the correlation is encoded in the color (blue corresponds to positive correlation, red to negative correlation).

4.3.6 Lewis' Law

Number of Neighbours. In order to understand the correlations between area and the number of neighbours we first want to briefly cover what is known as Lewis' Law. It describes a linear relationship between the average area of domains in a two-dimensional cellular structure and the number of neighbours of these domains as shown in Figure 4.14. Lewis discovered this behaviour in the 1920s observing cucumber cells. Just recently Kim et al. (2014) provided an explanation for this, arguing that it is a result of the elongations of the cells.

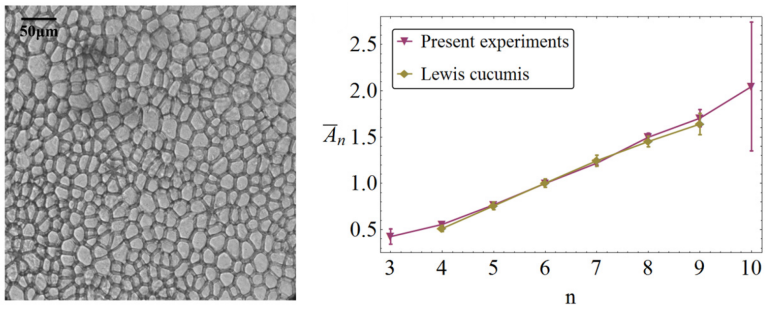

Figure 4.14: Illustration of Lewis' Law. (LEFT) Cucumber cells. (RIGHT) Lewis' Law, i.e. the mean of the area of the cells versus the the number of neighbours. Taken from Kim et al. (2014)

Now observing the correlations between area and the number of neighbours we notice a strong positive correlation at high area fraction for highly elongated particles. This positive correlation almost vanishes while progressing towards more circle-like packed particles which indeed aligns with Kim et al. (2014). We are therefore able to reproduce Lewis' Law as one can see in Figure 4.15.

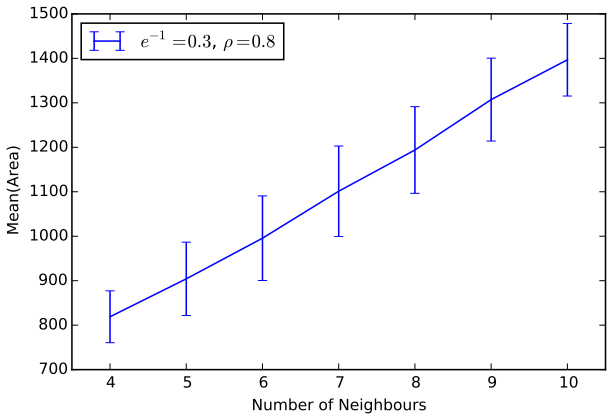

Figure 4.15: Reproduction of Lewis' Law with the highly elongated packed particles ($e^{-1} = 0.3$) at high area fraction ($\rho = 0.8$). On the x-axis the number of neighbours are plotted, the y-axis shows the mean of the area. The error bars correspond to the standard deviation of the area.

Mean of the Contact Lengths. Let us finally focus on the mean of the contact lengths and how it relates to the number of neighbours and therefore also to Lewis' Law. Figure 4.16 shows once more the correlation between the number of neighbours and the mean of the contact lengths that is also plotted in Figure 4.7.

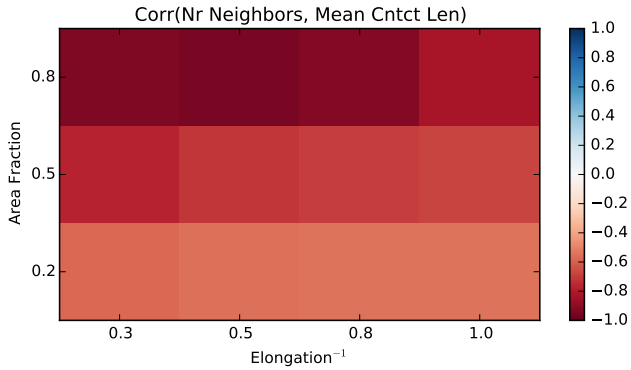

Figure 4.16: Correlation of the number of neighbours and the mean of the contact lengths. The inverse elongation of the packed particles is put on the x-axis, the area fraction on the y-axis and the value of the correlation is encoded in the color (blue corresponds to positive correlation, red to negative correlation).

One clearly observes an overall strong negative correlation, meaning that if the number of neighbours tends to increase than the mean of the contact length tends to decrease. Why is that? And where might the gradient at the high area fraction come from, from a very strong negative correlation at the upper left corner (i.e. highly elongated packed particles) to a slightly less negative correlation at the upper right corner (i.e. packed circles)? Let us first concentrate on the upper left corner. From an observation of the actual packings in Figure 4.10 we see that in this system the particles tend to align next to each other along their greater semi-axis. These alignments form *worm*-like structures which cells, with a comparable higher number of neighbours, are placed next to. This is because the cells in the *worm*-like structures have only a small contact length with these greater cells. On the other hand, the perimeter of these cells next to the *worms* is not increasing with the same factor as they increase their number of neighbours. Hence, the mean of the contact lengths decreases. This explanation is supported by the standard deviations of the these measures divided by their respective means, i.e. we are looking at the relative variations around the mean. Here, this relative standard deviations of the cell perimeter is around only a third of the relative standard deviation of the number of neighbours and the mean contact length, which are approximately equal. This explains the negative correlation which is strongest for the upper left corner (i.e. for highly elongated packed particles at high area fraction) because of the mentioned *worm*-like structures. However, at low area fraction the relative standard deviations of these three measure are about equal. Here we have to take into account the relatively low correlation between the number of neighbours and the perimeter of the cell (see Figure 4.7). Despite the overall trend that a cell with more neighbours has also a greater perimeter there still exit relatively many cells within a given range of perimeter for which their number of neighbours varies enough to explain the negative correlation between the number of neighbours and the mean contact length at low area fraction.

4.3.7 Maximum Correlated Cliques

Let us summarize our results with at least some connection to the network based correlation analysis of Section 3. Figure 4.17 shows for the four extreme corner cases of the phase space the maximal average absolute correlation versus the clique size. A clique, as we know from Section 3.2.1 is a fully connected sub-graph of the whole graph where the vertices are here all non rescaled morphological measures and the edges represent their mutual correlation coefficient. The idea of Network Filtering suggests that a certain maximal amount of a similarity measure (e.g. absolute value of the correlation) is essential to describe the main characteristics of the system (Tumminello et al., 2005). With that in mind we ask of the group of measures which average mutual absolute correlation coefficient is largest.

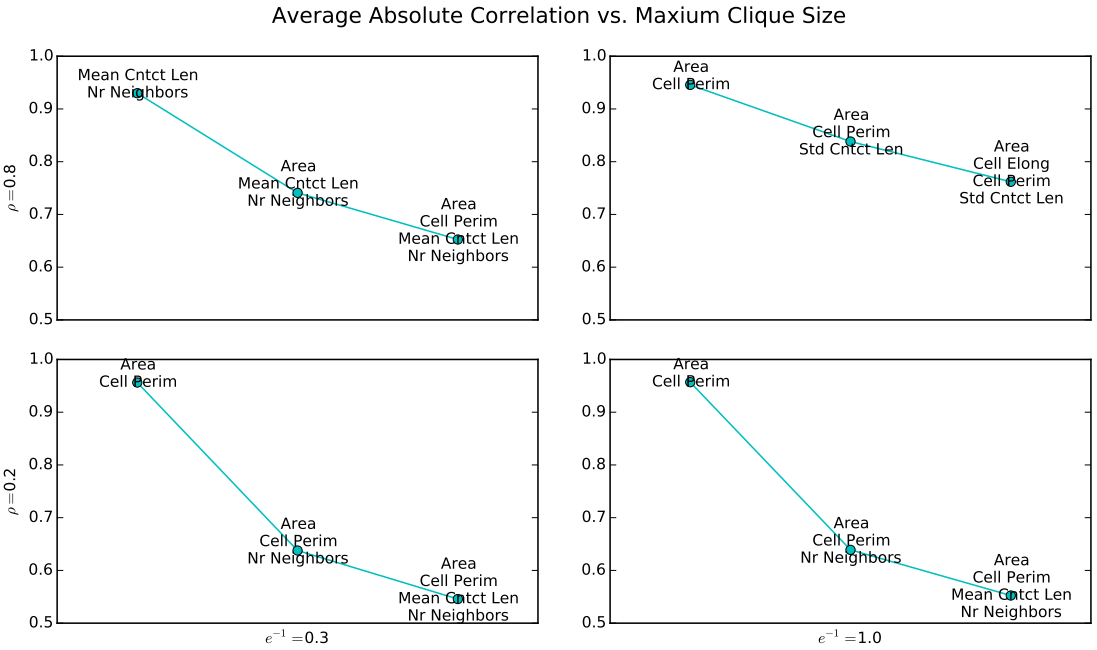

Figure 4.17: Average absolute correlation versus maximum cliques size. A clique is a fully connected sub-graph of the graph made of the morphological measures as vertices and the correlation as weights for their edges. The average absolute correlation is the absolute values of the correlation each edge in the cliques carries to obtain the overall weight of the clique.

From Figure 4.17 we see that for the two plots at low area fraction (bottom) the maximum cliques are identical. Even the actual value of the average maximum correlation is qualitatively equal. This shows us that at the low area fraction the particles are packed randomly without the influence of their shape. This result gets amplified by the observations that the maximum two cliques consist of the obvious relation between the area and the perimeter of the cell which have naturally a high correlation. These get accompanied by the number of neighbours which are followed by the mean of the contact lengths which have also a rather obvious correlation and are influenced by the area as pointed out above.

t the high area fraction (top) we observe an overall different picture. Here, the maximum cliques distinguish themselves clearly depending on the elongation of the particles. So one can state immediately, at high area fraction the elongation of the packed particles matters. Of course, this result is rather obvious as well since only at the high area fraction the particles are so densely packed that they begin to interact with their shapes. But never the less, it is good to have it confirmed by the data. If we continue the analysis with packed circles (top-right) we see the obvious strong correlation between the area and the perimeter of the Voronoi cells as the maximum 2-clique. But they get accompanied further only by measures of elongation like the standard deviation of the contact length and the cell elongation. This indicates how at this configuration the

particles are packaging almost in a regular hexagonal grid but with some defections that make these mutual correlations with the measures of elongation and also the correlation between area and perimeter that strong. Note even further that the overall level of average maximum correlation of these three maximum cliques is higher than at any other configuration. This could point also to the fact of the almost hexagonal grid in which strong interaction and therefore strong correlations are present.

Let us finally observe the configuration of highly elongated packed particles at high area fraction (top-left). Here it strikes that now the maximum 2-clique consists not of the area and the perimeter but instead of the mean of the contact lengths and the number of neighbours. Further these measures are followed by the area which strongly points to the presence of Lewis' Law and the *worm*-like structures. Only in the 4-clique the pair of area and perimeter is included which is also a hint to the different structure of the packed particle system at this configuration.

4.4 Conclusion

To conclude this chapter, we have presented a systematic way how to study the mutual relations between morphological measures of two dimensional random packings along different shapes of the packed particles and area fractions.

The findings of the correlation analysis of the data support the intuitive assumption that for low area fraction the packed systems are independent of the elongation of the packed particles. Further it was possible to relate the correlation data to the actual structure of the packed systems. Moreover, several measures of elongation could be identified: the elongation, the rescaled perimeter and the rescaled standard deviation of the contact lengths of the Voronoi cells. Additionally the centre of mass distance could be interpreted as a measure of the typical form of the Voronoi cells at high area fraction by analysing the correlation between the area, elongation and the centre of mass distance. Next, Lewis' Law was reproduced in our analyses and its connections to the structure of the packings were shown by analysing the correlation of the number of neighbours and the mean of the contact lengths. Finally, I showed how maximum cliques of the averaged absolute correlation, an approach inspired by Network Filtering Techniques (see Chapter 3), covers indeed the important information about the packing system.

5 Conclusion

In this thesis I approached the analysis of general datasets with the focus on the interdependencies of the variables. The guiding question throughout this work asked how one could extract the relevant information of these datasets. In particular two settings were examined at the scale of the ratio of the number of variables and the number of observations.

For high dimensional data (i.e. where the number of variables p is around or smaller than the number of observations T) it is the standard sample covariance matrix itself that poses problems for the proper dealing with these systems. Since at the high dimensional setting it is not invertible or at least error prone a straight forward inversion of this matrix is useless to obtain a reasonable estimate of the parameters of the in general assumed model, the multivariate Gaussian distribution. Here I proposed a novel method, called *LoGo*, that makes use of well known facts from Gaussian Markov Random Fields and Network Filtering Techniques that address this problem by inverting the sample covariance matrix **lo**cally and adding the respective parts to a meaningful **glo**bal estimate of the inverse covariance matrix. Several tests, performed on simulated and real-world data, showed that it is able to outperform the state-of-the-art method to estimate sparse inverse covariance matrices at the high-dimensional data setting. However, still at its infancy, more applications of *LoGo* will necessary to reveal it full potential.

For the low-dimensional data setting (i.e. where the number of observations T is very large compared to the number of variables p) the use of the correlation coefficient, i.e. the covariance for normalized variables, serves usually as a good model of the interdependency relations, so as here. The problem that needed to be addressed in this case was made from the larger phase space which the system lived in. Since we observed the morphological measures (variables) of Voronoi cells (observations) along different elongations of the packed particles and different area fractions we had to deal with more than one correlation matrix to be observed quasi simultaneously (one for each combination of elongation and area fraction) to obtain the whole picture of the systems behaviour. The key here, that addressed this problem properly laid in the visualization of the correlation coefficient. Since it is a bounded measure from -1 to 1 two colors (red and blue) with a prober gradient served to visualize the degree of (anti)correlation up to a sufficient level. Based on this the whole phase space correlation matrix could be plotted such that the whole system (including the transition from low to high area fraction and from highly elongated packed particles to circles) could by analysed. This phase space correlation matrix for example confirmed clearly the intuitive assumption

that at low area fractions the shape of the packed particles does not influence the packing structure.

This work showed, by appearing in both cases, how rich and helpful old and simple concepts like the correlation coefficient still are able to be. One the other hand we can learn explicitly that datasets differ from system to system and from dimension to dimension. There are no tools that fit all possible data matrices out there and the right methods to analyse the data need to be chosen carefully, with a proper understanding of the underlying system and the constraints of the dataset, like its dimensionality.

Bibliography

Albert, R. and A.-L. Barabási (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics 74*(1), 47.

Aste, T. and T. Di Matteo (2008, Feb). Emergence of gamma distributions in granular materials and packing models. *Phys. Rev. E 77*, 021309.

Aste, T., T. Di Matteo, and S. Hyde (2005). Complex networks on hyperbolic surfaces. *Physica A: Statistical Mechanics and its Applications 346*(1), 20–26.

Aste, T., R. Gramatica, and T. Di Matteo (2012). Exploring complex networks via topological embedding on surfaces. *Physical Review E 86*(3), 036109.

Aurell, E. and M. Ekeberg (2012). Inverse ising inference using all the data. *Phys. Rev. Lett. 108*(9), 090201.

Backhaus, J. G. (2012). *Handbook of the History of Economic Thought*, Volume 11 of *The European Heritage in Economics and the Social Sciences*. Springer New York.

Baker, A. (2013). Simplicity. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2013 ed.). The Metaphysics Research Lab, Center for the Study of Language and Information (CSLI), Stanford University.

Banerjee, O., L. El Ghaoui, and A. d'Aspremont (2008, June). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *J. Mach. Learn. Res. 9*, 485–516.

Banerjee, O., L. E. Ghaoui, A. d'Aspremont, and G. Natsoulis (2006). Convex optimization techniques for fitting sparse gaussian graphical models. In *Proceedings of the 23rd international conference on Machine learning*, ICML '06, New York, NY, USA, pp. 89–96. ACM.

Barber, D. (2013). *Bayesian Reasoning and Machine Learning*. Cambridge University Press.

Blair, J. R. and B. Peyton (1992). An introduction to chordal graphs and clique trees. In *Graph theory and sparse matrix computation*, pp. 1–29. Springer.

Boccaletti, S., V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang (2006). Complex networks: Structure and dynamics. *Physics Reports 424*(4), 175–308.

Dempster, A. P. (1972). Covariance selection. *Biometrics 28*(1), 157–175.

Di Matteo, T., T. Aste, S. Hyde, and S. Ramsden (2005). Interest rates hierarchical structure. *Physica A: Statistical Mechanics and its Applications 355*(1), 21–33.

Diestel, R. (2010). *Graph Theory*, Volume 173 of *Graduate Texts in Mathematics*. Heidelberg: Springer.

Donev, A., S. Torquato, and F. H. Stillinger (2005a). Neighbor list collision-driven molecular dynamics simulation for nonspherical hard particles. i. algorithmic details. *Journal of Computational Physics 202*(2), 737–764.

Donev, A., S. Torquato, and F. H. Stillinger (2005b). Neighbor list collision-driven molecular dynamics simulation for nonspherical hard particles.: Ii. applications to ellipses and ellipsoids. *Journal of Computational Physics 202*(2), 765–793.

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

Höffding, W. (1940). Masstabinvariante korrelationstheorie. schriften des math. instituts und des instituts für angewandte mathematik der universitat berlin 5.

Hotelling, H. (1953). New light on the correlation coefficient and its transforms. *Journal of the Royal Statistical Society. Series B (Methodological) 15*(2), 193–232.

Hsieh, C., I. S. Dhillon, P. K. Ravikumar, and M. A. Sustik (2011). Sparse inverse covariance matrix estimation using quadratic approximation. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger (Eds.), *Advances in Neural Information Processing Systems 24*, pp. 2330–2338. Curran Associates, Inc.

Jaynes, E. T. (1957a). Information theory and statistical mechanics. *The Physical Review 106*(4), 620–630.

Jaynes, E. T. (1957b). Information theory and statistical mechanics ii. *The Physical Review 108*(2), 171–190.

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press.

Kim, S., M. Cai, and S. Hilgenfeldt (2014). Lewis' law revisited: the role of anisotropy in size–topology correlations. *New Journal of Physics 16*(1), 015024.

Koller, D. and N. Friedman (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

Kruskal Jr., J. B. (1956, February). On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical Society 7*(1), 48–50.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford:Clarendon.

Lee Rodgers, J. and W. A. Nicewander (1988). Thirteen ways to look at the correlation coefficient. *The American Statistician 42*(1), 59–66.

Lezon, T. R., J. R. Banavar, M. Cieplak, A. Maritan, and N. V. Fedoroff (2006). Using the principle of entropy maximization to infer genetic interaction networks from gene expression patterns. *Proceedings of the National Academy of Sciences 103*(50), 19033–19038.

Mantegna, R. N. (1999). Hierarchical structure in financial markets. *The European Physical Journal B - Condensed Matter and Complex Systems 11*(1), 193–197.

Meinshausen, N. and P. Bühlmann (2006). High dimensional graphs and variable selection with the lasso. *Annals of Statistics 34*(3), 1436–1462.

Murphy, K. P. (2001, May). An introduction to graphical models. Review.

Musmeci, N., T. Aste, and T. D. Matteo (2014, October). Risk diversification: a study of persistence with a filtered correlation-network approach. Papers 1410.5621, arXiv.org.

Oztoprak, F., J. Nocedal, S. Rennie, and P. A. Olsen (2012). Newton-like methods for sparse inverse covariance estimation. In P. Bartlett, F. Pereira, C. Burges, L. Bottou, and K. Weinberger (Eds.), *Advances in Neural Information Processing Systems 25*, pp. 755–763.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 12*, 2825–2830.

Pourahmadi, M. (2013). *High-Dimensional Covariance Estimation: With High-Dimensional Data*. John Wiley & Sons.

Ravikumar, P., M. J. Wainwright, and J. D. Lafferty (2010, 06). High-dimensional ising model selection using ℓ1-regularized logistic regression. *The Annals of Statistics 38*(3), 1287–1319.

Ravikumar, P., M. J. Wainwright, G. Raskutti, and B. Yu (2011). High-dimensional covariance estimation by minimizing ℓ1-penalized log-determinant divergence. *Electronic Journal of Statistics 5*, 935–980.

Rényi, A. (1959). On measures of dependence. *Acta mathematica hungarica 10*(3), 441–451.

Ricci-Tersenghi, F. (2012). The bethe approximation for solving the inverse ising problem: a comparison with other inference methods. *Journal of Statistical Mechanics: Theory and Experiment 2012*(08), P08015.

Roudi, Y., E. Aurell, and J. A. Hertz (2009). Statistical physics of pairwise probability models. *Frontiers in Computational Neuroscience 3*(22), 1–15.

Rue, H. and L. Held (2005). *Gaussian Markov Random Fields: Theory and Applications*, Volume 104 of *Monographs on Statistics and Applied Probability*. London: Chapman & Hall.

Schäfer, J. and K. Strimmer (2005). An empirical bayes approach to inferring large-scale gene association networks. *Bioinformatics 21*(6), 754–764.

Schneidman, E., M. J. Berry, R. Segev, and W. Bialek (2006). Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature 440*(7087), 1007–1012.

Sessak, V. and R. Monasson (2009). Small-correlation expansions for the inverse ising problem. *Journal of Physics A: Mathematical and Theoretical 42*(5), 055001.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B 58*(1), 267–288.

Tumminello, M., T. Aste, T. Di Matteo, and R. N. Mantegna (2005). A tool for filtering information in complex systems. *Proceedings of the National Academy of Sciences of the United States of America 102*(30), 10421–10426.

Tumminello, M., T. Di Matteo, T. Aste, and R. N. Mantegna (2007). Correlation based networks of equity returns sampled at different time horizons. *The European Physical Journal B-Condensed Matter and Complex Systems 55*(2), 209–217.

Wainwright, M. J. and M. I. Jordan (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning 1*(1–2), 1–305.

Weaire, D. and T. Aste (2008). *The pursuit of perfect packing*. CRC Press.

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

Selbstständigkeitserklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Wolfram Barfuß
7. Mai 2015