שערוך מטריצת קוריאנס תחת מודל משולב בעל מבנה גרפי ודרגה לינארית נמוכה

נועם בלום

שערוך מטריצת קוריאנס תחת מודל משולב בעל מבנה גרפי ודרגה לינארית נמוכה

חיבור על מחקר

לשם מילוי חלקי של הדרישות לקבלת תואר מגיסטר למדעים

בהנדסת חשמל

נועם בלום

הוגש לסנט הטכניון – מכון טכנולוגי לישראל

אדר תשע"ח חיפה פברואר 2018

תודות

המחקר נעשה בהנחיית פרופ' רונן טלמון מהפקולטה להנדסת חשמל בטכניון.

תקציר

הבעיה של שערוך מטריצת קווריאנס עוסקת בשערוך שלה מתוך מדגם נתון של נקודות. היא בעלת חשיבות מכרעת בתחומים רבים, בשל האופי הבסיסי של מומנטים שניים בהסתברות וסטטיסטיקה. למרות שהיסודות שלה הונחו עוד בראשית המאה הקודמת, בשנים האחרונות היא זכתה לעניין רב הן מהצד התיאורטי והן מנקודת המבט המעשית. חלק גדול מהעבודות בנושא מדגימות כיצד ניתן לשפר את איכות השערוך ע"י בניית משערך המתאים להנחות מודל מסוימות, וזהו גם הקו המנחה את עבודה זו.

שני מודלים אשר נעשה בהם שימוש נפוץ בעיבוד מידע הם בעלי קשר אינהרנטי למטריצת הקווריאנס ולכן מעניינים במיוחד בהקשר הזה. המודל הראשון הוא של משתנים חבויים לינארים: כל מדידה ממימד גבוה מתקבלת ע"י טרנספורמציה לינארית ממשתנים חבויים בעלי מימד נמוך, בתוספת רעש אופייני למדידה. מודל זה ידוע בשם Factor Analysis, אך למעשה ידוע זה מכבר כי הוא קשור באופן הדוק לשיטה אחרת מוכרת הרבה יותר, PCA (Principal Component Analysis). בהנחה שהשונויות של גורמי הרעש האדיטיביים זהות, משערך ה- ML (נראות מקסימלית, Maximum Likelihood) לפרמטרים של המודל מתקבל בדיוק ע"י ביצוע PCA על המדידות. מטריצת הקווריאנס המתקבלת, בהתאם למודל ההסתברותי של המדידות, היא בעלת מבנה של סכום מטריצה ממימד נמוך ומטריצה אלכסונית.

המודל השני הוא מודל גרפי לא מכוון, הידוע גם בשם שדה מרקובי. מודלים גרפיים הם בעלי שימוש נרחב בכל תחומי הסטטיסטיקה היישומית בשל יכולתם לבטא סוג מסוים של אי-תלויות (אי תלות מותנית) בין המשתנים בצורה קומפקטית. במודל גרפי גאוסי לא מכוון, אי-תלויות אלו מתבטאות בהתאפסות של איברים במטריצה ההופכית לקווריאנס (מטריצת precision), ועל כן גם במקרה זה יש למודל ביטוי ישיר במטריצה. המשערך הידוע בשם Graphical Lasso מתאים למודל זה.

בעבודה זו אנו בוחנים את הבעיה של שערוך מטריצת קווריאנס תחת מודל המורכב משני המודלים הנ"ל בו-זמנית. קיימים מקרים בהם ההנחות של שני המודלים מתקיימות באופן טבעי, למשל, אוסף תמונות פנים: האי-תלויות המותנות נובעות מהמרחק בין הפיקסלים, ובעבר הודגם כי בעזרת PCA ניתן לשמר את רוב המידע המגולם בתמונות גם במימד נמוך.

כפי שהוזכר לעיל, ישנם משערכים המתאימים לכל אחד מן המודלים בנפרד. עם זאת, עד כמה שידוע לנו, לא הוצע משערך המתאים למודל המשותף, וטבעי לשאול האם ניתן

להשתמש בהנחות המודל המשותף כדי לבנות משערך טוב יותר בהשוואה למשערך המשתמש רק בהנחות של אחד המודלים, או ממשערך שאינו משתמש בהן כלל.

כדי לפתח משערך המתאים לשני המודלים גם יחד, יש תחילה לגשר בין הדרכים השונות שבהם מתבטאים מאפייני המודל במטריצת הקווריאנס: בעוד שהמודל הראשון עוסק במטריצה עצמה, השני מתבטא במטריצה ההופכית. לשם כך אנו מפתחים הצגה שקולה למודל הראשון, אשר מתארת אותו במונחים של המטריצה ההופכית. הפירוק לסכום של מטריצה מדרגה נמוכה ומטריצה אלכסונית, שתיהן מוגדרות חיובית, שקול לפירוק אנלוגי של המטריצה ההופכית, להפרש בין שתי מטריצות מוגדרות חיובית שאחת מהן אלכסונית והשניה מדרגה נמוכה.

בעזרת הפירוק הנ"ל, אנו מפתחים משערך מבוסס ML למודל המשותף. הוא מתקבל על ידי הוספת גורם קנס למשערך הגאוסי הרגיל שמטרתו למצוא פתרון בעל דרגה נמוכה, וכן אילוצים הכופים את יתר הנחות המודל. הבעיה המתקבלת היא קמורה ובעלת פונקצית מטרה קמורה חזק, ופתרונה הוא המשערך. לבעיה יש פרמטר יחיד הקבוע את המשקל היחסי של הקנס, וכך את הדרגה של הפתרון המתקבל.

אנחנו מנתחים חלק מהתכונות התיאורטיות שלו כגון קונסיסטנטיות, מדגימים כיצד ניתן לקבוע את ערך פרמטר הקנס המופיע במשערך, ומדגימים את הביצועים המשופרים של המשערך לעומת אלטרנטיבות אחרות בעזרת ניסויים על מידע סינתטי ואמיתי. בנוסף, אנחנו מפתחים אלגוריתם מהיר לפתרון בעיית האופטימיזציה הנדרשת לשערוך, המבוסס על סכמת ADMM (Alternating Directions Method of Multipliers) שזכתה לפופולריות רבה.

Covariance matrix estimation under a combined low-rank and graphical model structure

Noam Bloom

Covariance matrix estimation under a combined low-rank and graphical model structure

Research Thesis

As Partial Fulfillment of the Requirements for
the Degree Master of Science in Electrical Engineering

Noam Bloom

Submitted to the Senate of the Technion—Israel Institute of Technology

Adar 5778 Haifa Febuary 2018

Acknowledgment

This research thesis was done under the supervision of Prof. Ronen Talmon at the Department of Electrical Engineering.

Contents

List of Figures

List of Tables

Abstract

The problem of covariance estimation involves estimating it given a sample of data points. It is of paramount importance in many fields, owing to the fundamental nature of second moments in probability and statistics. Although it has its roots early in the last century, in recent years it has attracted considerable renewed interest from both theoretical and practical standpoints. Two prevalent types of data models are inherently related to the covariance matrix and are therefore of special interest in estimation problems. The first of these is a latent linear factor model, also strongly related to PCA, which is manifested as a low rank component in the covariance matrix. The second is undirected graphical model structure (Markov random field), which is related to presence of zeros in the inverse covariance (precision) matrix. In this work we consider the problem of estimating a covariance matrix subject to both types of models simultaneously. We demonstrate the attractiveness of the combined model and propose a novel estimator to address it. We show that our estimator outperforms other alternatives on both synthetic and real-world data, explore a few of its theoretical properties such as consistency, and develop a fast algorithm for solving the associated optimization problem.

Abbreviations

ADMM Alternating Direction Method of Multipliers

KL Kullback-Leibler

ML Maximum likelihood

MRF Markov Random Field

PCA Principal Component Analysis

PSD Positive Semi-Definite

SDP Semi-Definite Programming

SVD Singular Value Decomposition

Chapter 1

Introduction

1.1 Overview of covariance estimation and literature review

The classic problem of covariance matrix estimation based on samples from the underlying distribution is a of fundamental importance in a wide range of fields. The importance of covariance matrices is understandable given the fundamental nature of second order statistics compared to higher order statistics, especially with regard to the Gaussian distribution, and indeed they play a prominent role in numerous methods in applied statistics, such as PCA, LDA and QDA, regression of multivariate data, analysis of independence and conditional independence relationships and graphical models, and construction of confidence regions, to name a few [Levina et al., 2008]. It is therefore not surprising that the task of estimating them accurately has received considerable attention in both the theoretical and the applied fields.

The standard estimator for the covariance matrix from i.i.d samples is the sample, or empirical, covariance. This estimator can be justified heuristically, and it is also the maximum likelihood estimator in the Gaussian setting (up to a fixed scaling factor, and assuming $p > n$). However, while appealing for its simplicity and natural form, it is known that other estimators often prove superior, either in the MSE sense [James and Stein, 1961, Stein, 1975], or when other loss functions are used to measure estimation error [Haff, 1980].

Recent years have seen an explosive growth in the dimensionality of data across virtually all fields [Donoho et al., 2000]. Some examples include gene expression from microarray data [Schäfer and Strimmer, 2005], financial forecasting [Ledoit and Wolf, 2003], spectroscopic mapping [Lin et al., 2007], fMRI [Derado et al., 2010], portfolio management [Fan et al., 2008], and more.

Classical statistical methods are often ill-suited to handle this rapid growth. In particular, the inadequacy of the sample covariance in the high-dimension / small sample size regime has been known for some time. While the theory of large-sample statistics has traditionally focused more on regimes where either p is fixed or allowed to grow more slowly than n, one classic result [Marchenko and Pastur, 1967] concerns the regime where $\frac{p}{n}$ tends to some constant γ and states that even if the eigenvalues of the population covariance are all equal to 1, the sample covariance eigenvalues may be significantly dispersed, to a degree that depends on γ. Since it is often the eigenvalues of the sample covariance that are of interest, this poses serious questions regarding the applicability of the sample covariance in the high-dimensional setting.

This observation has spurred more rigorous research on both the theoretical and practical issues involving covariance estimation. On the theoretical side, one prominent example is the line of research concerning the so-called spiked model [Johnstone, 2001], which has received attention recently, extending the Marchenko-Pastur law to new types of population models and shedding more light on the behavior of the sample covariance eigenvalues.

One approach for dealing with the eigenvalue dispersion problem involves using either a multiple of the sample covariance, or a linear combination of it with the identity matrix. These ideas sometimes arise naturally from use of various model assumptions and/or optimality criteria, and have been used for some time [Haff, 1980], or [Dey and Srinivasan, 1985] where the inverse is used. Clearly, their potential for improving is the estimation accuracy of the eigenvalues is limited, and the eigenvectors are not changed at all, compared to the plain sample covariance, but nevertheless they are still being investigated. For example, eg in [Ledoit and Wolf, 2004], the authors propose an asymptotically-optimal combination of the sample covariance matrix and the identity matrix, in the sense of minimizing the MSE (expected squared Frobenius norm error). This work has also been

improved upon in several ways in [Chen et al., 2010].

A common theme in recent approaches for covariance matrix estimation, particularly in the small sample size regime, is the use of structural assumptions on the covariance matrix to design regularized estimators that enforce these assumptions. For example, in [Bickel and Levina, 2008], the authors propose banding either the sample covariance, or a factor of the modified cholesky decomposition of the inverse covariance, and demonstrate the consistency of the proposed estimator under certain conditions. Such structure implies that far apart variables are weakly correlated. Various other approaches based on regularizing the cholesky decomposition factors have been proposed, motivated by its interpretation as a regression coefficients and prediction error variances when variables are regressed on their predecessors [Huang et al., 2006]. Other works regularize by employing the lasso [Tibshirani, 1996] and ridge [Hoerl and Kennard, 1970] penalties to the likelihood.

Other works also involve sparsity either in the covariance matrix or its inverse. In [Chaudhuri et al., 2007], the authors consider the problem of estimating a covariance matrix with a prescribed set of zeros at fixed known locations, and propose algorithms for estimating such a matrix. Since zeros in the covariance matrix correspond to (marginal) independence between the variables (or, in the non-Gaussian setting, un-correlation), this approach enforces a certain independence model on the data. In [Bickel et al., 2008], the authors study obtain a sparse covariance estimator by thresholding the sample covariance with a threshold parameter that is selected via cross validation, and discuss convergence properties of their estimator. In contrast to the direct thresholding approach of that work, an l_1-penalty-based thresholding approach is undertaken in [Bien and Tibshirani, 2011], where the authors apply the lasso/l_1-penalty to the Gaussian log-likehood to enforce sparsity of the estimated covariance matrix. Another similar work is [Xue et al., 2012], where the l_1 penalty is again applied, this time to a different objective function, namely, the Frobenius norm distance to the sample covariance, along with a positive-definiteness constraint. The authors present an ADMM algorithm to solve the aforementioned optimization problem, and discuss convergence rates of the estimator under certain conditions.

Another popular line of work involves sparsity of the precision matrix (inverse of the covariance). As will be discussed more in depth in section 3.2, a zero in the precision

matrix corresponds in some models to a conditional independence relation among the variables of the model. Estimating these conditional independence relations from the data is beneficial as they correspond directly to sparsity of the resulting graph and can lead to more interpretable models, and therefore estimating sparse precision matrices is highly appealing in the context of learning graphical models. Thus, this direction has gained considerable interest in recent years. In [Meinshausen and Bühlmann, 2006], the authors use a lasso penalty to estimate the neighbors of each node in the graph, and show the resulting estimator is consistent. [Banerjee et al., 2008] propose an approach based on l_1-penalized maximum likelihood, and present two algorithms for solving the resulting optimization problem. Since then numerous algorithms, analyses and extensions for that problem have surfaced. [Friedman et al., 2008] propose an algorithm, the graphical lasso, exploiting the similarity between the dual formulation of the penalized maximum likelihood problem and the classical lasso regression problem. [Rothman et al., 2008] study a similar penalized problem, and provide convergence rates for the resulting estimator.

We propose a new framework for estimating a covariance matrix, based on a model which draws inspiration from both the sparse precision matrix model, and the probabilistic PCA setting introduced in [Tipping and Bishop, 1999]. At the base of this framework lies a novel (to us) representation of the precision matrix corresponding to a covariance which follows the probabilistic PCA model, that is, a sum of a low rank part and a diagonal part. By utilizing that representation, we are able to incorporate graphical model structure into the model while maintaining a convex formulation for the estimator.

1.2 Thesis structure

This work is organized as follows. In chapter 2, we present scientific background relevant to this work. This includes the sample covariance, the Gaussian likelihood, and Principal Component Analysis (PCA). In chapter 3, we discuss our model and the two models that serve as its basis. In chapter 4 we present an estimator for our model and discuss its properties. In chapter 5 an algorithm for solving the corresponding optimization problem is introduced. Chapter 6 presents experimental results, and Chapter 7 concludes with discussion of future research.

Chapter 2

Scientific Background

2.1 The sample covariance

Let $\{x_i\}_{i=1}^n$ be a given set of i.i.d samples from a Gaussian multivariate distribution on $\mathbb{R}^p$. The sample covariance matrix is given by:

$$S = \frac{1}{n}\sum_i (x_i - \overline{x})(x_i - \overline{x})^T$$

where $\overline{x} = \frac{1}{n}\sum_i x_i$ is the sample mean vector.

$$S_0 = \frac{1}{n-1}\sum_i (x_i - \overline{x})(x_i - \overline{x})^T = \frac{n}{n-1}S$$

This definition has the advantage of satisfying $\mathbb{E}S_0 = \Sigma$ [Bilodeau and Brenner, 2008, chapter 7]. However, we will continue to use the original definition since the log-likelihood function is more easily expressed in terms of it (see section 2.2). When n is large, the difference between S and S_0 is negligible.

Inspecting the definition, it is clear that S is a matrix containing the sample covariances of each pair of variables in the model, since at the k, m'th entry we have

$$S_{\mathrm{km}} = \frac{1}{n}\sum_i \left(x_i^{(k)} - \overline{x}^{(k)}\right)\left(x_i^{(m)} - \overline{x}^{(m)}\right)$$

which is the sample covariance of two scalar variables $X^{(k)}, X^{(m)}$.

It is well-known that in the Guassian setting, $\overline{x}$ and S are independent. $\overline{x}$ is clearly Gaussian, whereas the distribution of nS is known as a Wishart distribution and denoted

$W_p(n-1,\Sigma)$. This distribution is a generalization of the single-variable χ^2 distribution, which is related to the single-variable normal distribution in a similar manner. Another useful property of this distribution is the following [Bilodeau and Brenner, 2008]:

When $n \geq p$, S is invertible with probability 1.

Note:

Throughout this work we will assume the sample mean has already been subtracted and that the underlying random variable satisfies $\mathbb{E}X = 0$. When the underlying distribution is Gaussian, subtracting the mean has the effect of changing the distribution of the sample covariance from $Wp(n)$ to $W_p(n-1)$ [Bilodeau and Brenner, 2008], that is, effectively reducing the number of samples used by 1.

2.2 The Gaussian Likelihood

Recall that the likelihood function corresponding to a parameter θ of interest given i.i.d measurements $\{x_i\}_{i=1}^n$ is $L(\theta;\{x_i\}) = \prod_{i=1}^n p(x_i)$, where $p(x)$ is the density function of the common distribution of $\{x_i\}$. The Maximum likelihood (ML) estimator for θ is obtained by maximizing the likelihood function:

$$\widehat{\theta} = \arg\max L(\theta;\{x_i\})$$

Or, equivalently, the log-likelihood:

$$\widehat{\theta} = \arg\max \log L(\theta;\{x_i\})$$

When the samples are Gaussian i.i.d, as is typically assumed, the log-likelihood can be expressed conveniently as follows.

The (mean-zero) Gaussian density is given by:

$$p(x_i) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}x_i^T\Sigma^{-1}x_i\right)$$

Thus, the log-likelihood function corresponding to n i.i.d measurements is (C is an arbitrary constant):

$$l(\Sigma) = \log \Pi_{i=1}^{n} p(x_i) = C + \sum_{i=1}^{n} \left[-\frac{1}{2} \log |\Sigma| - \frac{1}{2} x_i^T \Sigma^{-1} x_i \right]$$

Since $x_i^T \Sigma^{-1} x_i$ is a scalar quantity, we have $x_i^T \Sigma^{-1} x_i = tr\left(x_i^T \Sigma^{-1} x_i\right)$, and using properties of the trace we have:

$$\begin{aligned} l(\Sigma) &= C - \frac{n}{2} \log |\Sigma| - \frac{1}{2} \sum_{i=1}^{n} \text{trace}\left(x_i^T \Sigma^{-1} x_i\right) = C - \frac{n}{2} \log |\Sigma| - \frac{1}{2} \sum_{i=1}^{n} \text{trace}\left(\Sigma^{-1} x_i x_i^T\right) \\ &= C - \frac{n}{2} \log |\Sigma| - \frac{1}{2} \text{trace}\left(\Sigma^{-1} \sum_{i=1}^{n} x_i x_i^T\right) \end{aligned}$$

Finally, since we are using the likelihood function as an objective for optimization, we may multiply by a constant $\frac{2}{n}$, to obtain:

$$l(\Sigma) = C - \log |\Sigma| - \text{trace}\left(\Sigma^{-1} \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T\right) = C - \log |\Sigma| - \text{trace}\left(\Sigma^{-1} S\right)$$

We will denote the matrix Σ^{-1}, known as the precision matrix, as X. In addition, since we are typically interested in the log-likelihood in the context of optimization, where additive and multiplicative constants are inconsequential, we will write $l_S(X) = -\log |\Sigma| - \text{trace}(XS) = \log |X| - \text{trace}(XS)$. Sometimes, when S is fixed, we will also write $l(X)$ for $l_S(X)$.

2.3 The sample covariance as a maximum likelihood estimator

The log-likelihood function $l(X) = \log |X| - \text{trace}(XS)$ is concave: the log-det function is concave (Boyd & Vandenberge, 3.1.5), and the second term is linear. As such, $\nabla l(X) = 0$ is a necessary and sufficient condition for a point to be a global unconstrained maximum. Using the well-known matrix derivative [Boyd and Vandenberghe, 2004, section A.4.1] $\nabla \log |X| = X^{-1}$, we see that $\nabla l(X) = X^{-1} - S$, and setting the derivative to 0, we have $X^{-1} = S \Leftrightarrow \Sigma = X^{-1} = S$. Thus, the Gaussian maximum likelihood estimator is simply the sample covariance discussed above, which gives another justification for using it.

$$S = \frac{1}{n}\sum_i x_i x_i^T$$

so that S is a sum of n rank-1 matrices. Thus, its maximal rank is n, and when $n < p$ it is not invertible. When $n \geq p$, we have already mentioned that S is invertible w.p. 1.

The sample covariance is a consistent estimator of the population covariance. This easily follows from the law of large numbers. Note that the (m, n)'th entry in S is $S_{\text{mn}} = \frac{1}{n}\sum_i x_i^{(m)} x_i^{(n)}$, where here $x_i^{(n)}$ denotes the $n'th$ component of the $i'th$ sample x_i. Since x_i are i.i.d samples, S_{mn} is obtained as an average over i.i.d samples taken from a random variable which has the same distribution as $x^{(n)}$.

On the other hand, $\Sigma_{\text{mn}} = \left(\mathbb{E}xx^T\right)_{\text{mn}} = \mathbb{E}x_m x_n$, so by the strong law of large numbers, $S_{\text{mn}} \to \Sigma_{\text{mn}}$ w.p. 1 as $i \to \infty$. Since pointwise convergence of the matrix components implies convergence in norm, we have that $\|S - \Sigma\|_2 \to 0$ w.p. 1, ie, the estimator is consistent.

2.4 PCA

PCA is a classic algorithm used extensively in signal processing, machine learning, and virtually all forms of data analysis [Jolliffe, 2002]; its applications are far too many to enumerate. In essence it could be described as a method for linear dimensionality reduction: given a set of high-dimensional data points, it aims to find a lower-dimensional representation obtained by a linear transformation, that in some sense preserves the structure of the original data. The faithfulness criterion can be described either as achieving minimal reconstruction error, or maximum preserved variance, as well shall see below.

More concretely, suppose a set of i.i.d points $\{x_i\}_{i=1}^n \subset \mathbb{R}^p$ sampled from a $N(0, \Sigma)$ distribution on (we assume $\mathbb{E}x = 0$ for simplicity of exposition). In addition, assume first that the covariance matrix Σ is known. We wish to find a linear subspace of $\mathbb{R}^p$ of dimension d, that captures the most variability of the data when projected onto that subspace. In the Gaussian setting, the idea can be visualized graphically as follows: the density of the Gaussian distribution can be visualized by its level sets, which are ellipsoids, and the direction of highest variation corresponds to the longest principal axis

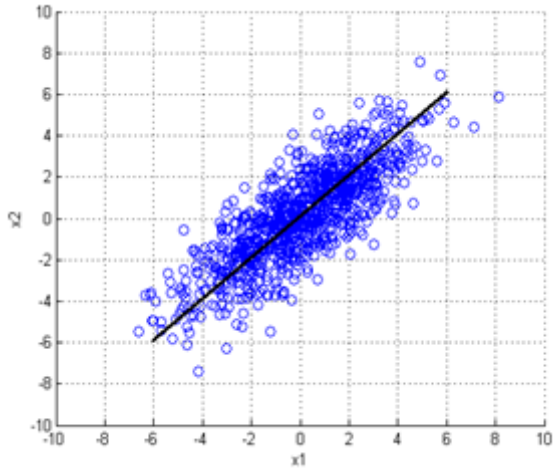

Figure 2.1: 2D Gaussian scatter plot and first principal axis of the ellipsoid

of the ellipsoid. Thus in figure 2.1, the one-dimensional linear subspace that PCA selects is spanned by the plotted black line.

More generally, we wish to find a linear projection $V_r : \mathbb{R}^p \to \mathbb{R}^r$ mapping the original data $\{x_i\}$ to their new representation $\{V_r x_i\}$.

We start with the case $r = 1$. A one-dimensional projection is given by $P = vv^T$ where v is a unit vector. We select v based on the following criterion:

$$\begin{aligned} \min \quad & \mathbb{E}\left\|x - vv^T x\right\|^2 \\ s.t. \quad & \|v\| = 1 \end{aligned}$$

where $x \sim N(0, \Sigma)$. So, the idea is to minimize the error of representing the original vector x by its linear projection $vv^T x$. Simple manipulation yields:

$$\left\|x - vv^T\right\|^2 = \|x\|^2 + \left\|vv^T x\right\|^2 - 2x^T \mathrm{v}\mathrm{v}^T x = \|x\|^2 + \left|v^T x\right|^2 - 2x^T vv^T x = \|x\|^2 - v^T xx^T v$$

We've used the fact that $\|v\| = 1$. Since $\|x\|^2$ doesn't depend on v, we see that:

$$\begin{aligned} \arg\min \quad & \mathbb{E}\left\|x - vv^T x\right\|^2 \\ s.t. \quad & \|v\| = 1 \end{aligned} = \begin{aligned} \arg\max \quad & v^T \underbrace{(\mathbb{E}xx^T)}_{\Sigma} v \\ s.t. \quad & \|v\| = 1 \end{aligned}$$

Note that this expression gives another interpretation for the optimization criterion, namely, maximal projected variance, as $\mathbb{E}\left|v^T x\right|^2$ is the variance of the scalar random

variable $v^T x$.

We claim that the solution to this problem is the largest (normalized) eigenvector of Σ (ie, the eigenvector corresponding to the largest eigenvalue). To see this, let $\Lambda = HDH^T$ be an eigenvalue decomposition so that $HH^T = I$, with $D = diag(\lambda_1, .., \lambda_p)$, $\lambda_1 \geq ... \geq \lambda_p$ (recall that Λ is symmetric so it can be orthogonally diagonalized). The columns of H, h_i, are eigenvectors of Λ.

Now,

$$v^T \Lambda v = v^T HDH^T v^T = \left(H^T v\right)^T D \left(H^T v\right) = \sum_{i=1}^{p} \lambda_i \left(h_i^T v\right)^2 \leq \lambda_1 \sum_{i=1}^{p} \left(h_i^T v\right)^2 = \lambda_1 \|v\| = \lambda_1$$

where we have used the fact that $\sum_{i=1}^{p} \left(h_i^T v\right)^2 = \left\|H^T v\right\|^2 = v^T HH^T v = v^T v = \|v\|^2 = 1$ owing to the orthogonality of H. It follows that the objective function is upper bounded by λ_1. Further, when $v = h_1$, the above inequality holds with equality, because $h_i^T v = 0$ for all $i \neq 1$. We have therefore shown that $\max_{\|v\|=1} v^T \Lambda v = \lambda_1$ and the optimum is attained with $v = h_1$, as claimed.

Thus the best (in terms of the optimization criterion used by PCA) rank-1 projection is given by $x \to vv^T x$, where v is as described above.

$$\begin{aligned} \min \quad & \mathbb{E} \left\|x - (V_k V_k^T x + vv^T x)\right\|^2 \\ s.t. \quad & \|v\| = 1 \\ & v \perp span\{v_1, .., v_k\} \end{aligned}$$

Rewrite the error as:

$$\left\|x - \left(V_k V_k^T x + vv^T x\right)\right\|^2 = \left\|V_k V_k^T x + \left(x - vv^T x\right)\right\|^2 = \left\|V_k V_k^T x\right\|^2 - 2\left(x^T V_k V_k^T x + x^T vv^T V_k V_k^T x\right) + \left\|x -\right.$$

The terms $\left\|V_k V_k^T x\right\|^2$ and $x^T V_k V_k^T x$ do not depend on v, so may be ignored, and the constraint $v^T v_i = 0$ gives $x^T vv^T V_k V_k^T x = 0$. Thus we are left with:

$$\begin{aligned} \min \quad & \mathbb{E} \left\|x - vv^T x\right\|^2 \\ s.t. \quad & \|v\| = 1 \\ & v \perp span\{v_1, .., v_k\} \end{aligned}$$

Which is exactly of the same form as the problem solved earlier, except for the new constraints. The same argument used above shows that the optimum is obtained by is $v = h_{k+1}$.

Thus, for any $1 \leq d \leq p$, the optimal rank-d projector $P_V = VV^T$ (where V is $p \times d$ with orthonormal columns) to a linear subspace, in the sense of minimizing the squared reconstruction error $\mathbb{E}\,\|x - P_V x\|^2$, is obtained when the column space of V is the space spanned by the leading d eigenvectors of Λ. In particular, one can use exactly these eigenvectors as the columns of V.

The matrix V^T serves as a linear dimensionality reduction map, mapping the data x from its original dimension p to a lower dimension d, representing x in the basis of the first d eigenvectors. Selecting d allows one to control the tradeoff between low projected dimension and low reconstruction error.

Note that although we've shown an inductive derivation where each step involves finding a rank-1 projection, it's also possible to arrive at the same result by considering a full rank-d projection at once.

For a more thorough derivation and discussion, see [Jolliffe, 2002].

In the above we've assumed that Σ is known, and in that case PCA is applied by calculating an eigenvalue decomposition $\Sigma = VDV^T$ and using the mapping $x \to V_r^T x$, where V_r is a matrix consisting of the columns of V corresponding to the top-most r eigenvalues. In practice Σ is not known and must be estimated from data. Once $\widehat{\Sigma}$ is estimated, for instance using the sample covariance, the procedure may be applied to $\widehat{\Sigma}$.

Alternatively, one may not calculate $\widehat{\Sigma}$ directly, and instead use a Singular Value Decomposition (SVD) of the data matrix itself. Given centered observations $\{x_i\}_{i=1}^n$ and denoting by X the $n \times p$ matrix which has x_i as its rows, we have:

$$\widehat{\Sigma} = \frac{1}{n} X^T X$$

Write the SVD of X as $X = USV^T$, so that:

$$n\widehat{\Sigma} = VS^T U^T USV^T = VS^T SV^T$$

Since this is an eigenvalue decomposition, the right singular vectors of X are exactly the eigenvectors of $\widehat{\Sigma}$, and the singular values are the squared eigenvalues. Because $\widehat{\Sigma}$ is

Positive Semi-Definite (PSD), the eigenvalues are nonnegative and their order is preserved by the squaring operation, so the order dictated by PCA matches the order of the singular values.

Chapter 3

Our model

3.1 The 'diagonal plus low rank' model and probabilistic PCA

We now discuss in detail the first of two models which will later serve as the basis for our estimator.

The first model is as follows. Let $r < p$ and assume Σ can be expressed the following sum:

$$\Sigma = \underbrace{AA^T}_{C} + \Psi \tag{3.1}$$

where $A \in \mathbb{R}^{p \times r}$ has full rank, and Ψ is a diagonal positive definite matrix.

The setup is depicted pictorially in figure 3.1.

We will further distinguish between the following two cases:

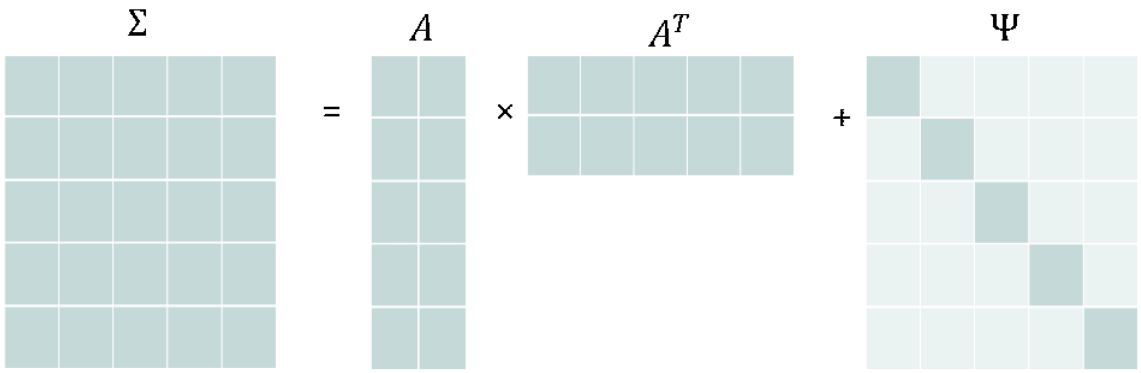

Figure 3.1: Matrices involved in the model

- 'PCA' case, in which $\Psi = \sigma^2 I$ (I is the identity matrix)
- 'Factor analysis' (FA) case, in which Ψ has no further restrictions other than diagonal and positive definite

PCA was discussed in the previous chapter. Factor analysis is a model often associated with PCA, involving linear latent variables which explain the observations [Jolliffe, 2002]. The connection between these two models, and the reasoning behind the names given to the above cases, is best explained by the probabilistic PCA model [Tipping and Bishop, 1999], which casts the PCA algorithm as a maximum likelihood estimation problem.

Consider the following probability model: $x_i \in \mathbb{R}^p$ are observations given by:

$$x_i = At_i + \epsilon_i$$

$t_i \in \mathbb{R}^r, \epsilon_i \in \mathbb{R}^p$ are latent variables, and A is some $p \times r$ matrix. Further, assume $t \sim N(0, I), \epsilon \sim N(0, \Psi)$. For Ψ we consider two possibilities, analogous to the two cases mentioned above:

- PCA model: $\Psi = \sigma^2 I$
- Factor analysis model: $\Psi = diag(\Psi_i)$

Readers familiar with the techniques of factor analysis will immediately recognize that the second case,which we have termed 'factor analysis model', is indeed exactly the model used in the factor analysis setup [Jolliffe, 2002], hence the name. However, at this point it might not be entirely clear why the name 'PCA model' was chosen for the first case. The reason for this is explained by [Tipping and Bishop, 1999], where the authors show that in the PCA model, the maximum likelihood estimator for A, σ^2 is given by:

$$\widehat{\sigma^2} = \frac{1}{p-r} \sum_{j=r+1}^{p} \lambda_j$$

$$\widehat{A} = U_r \left(\Lambda_r - \widehat{\sigma^2} I \right)^{1/2} R,$$

where Λ is a diagonal matrix with entries λ_i, $S = U \Lambda U^T$ is an eigen-decomposition of the sample covariance, $U_r = U(:, 1:r)$, $\Lambda_r = \Lambda(1:r, 1:r)$, and R is an arbitrary orthogonal matrix ($RR^T = I$). The authors then continue to note that estimating A, σ^2 as above

essentially amounts to performing PCA on the data. Thus, we can interpret PCA as a covariance estimation algorithm, subject to a specific model assumption on the covariance structure of the data, which are quite similar to the latent variables assumptions of factor analysis. Under that assumption, PCA is the maximum likelihood estimator.

Finally, the relationship to our model is immediate: subject to the above, the covariance matrix of x is:

$$\Sigma = cov\left(x\right) = Acov\left(t\right)A^T + cov\left(\epsilon\right) = AA^T + \Psi$$

Thus, our 'low rank + diagonal' model, is essentially the same model used in the probabilistic PCA setup. We have ommitted the latent variables because we have no use for them in our context. However, the reader should keep in mind that this model is strongly related to the classic PCA setup and the scenarios in which it is used.

3.2 Graphical model structure

The second model assumption we use to construct our estimator is related to zeros in the precision matrix. To understand why this is important, we first introduce a class of probabilistic models which are directly related to those zeros.

The field of probabilistic graphical models [Koller and Friedman, 2009] revolves around certain types of representations for probability distributions, the properties of which are induced by a corresponding graph. Their strength lies in the observation that in any probabilistic description of a real-world phenomenon, certain redundancies are expected. These make specifying a full unconstrained probability distribution unnecessary, and in fact, often detrimental. Graphical models allow capturing the uncertainty in many real-world phenomena succinctly, and also allow encoding domain-based knowledge or constraints. They have become commonplace tools in many applications of machine learning or applied statistics, such as medical diagnosis, analysis of genomic data, speech recognition and many more [Koller and Friedman, 2009].

Inherent to every graphical model is a factorization of the probability distribution underlying the variables in the model. This factorization is specified indirectly via the graph representing the model, as will be discussed below. The central idea underlying

graphical models is that such a factorization of a joint probability distribution is directly related to a certain type of independence between the variables in the model. While a factorization into terms which involve no common factor, such as $p(x,y,z) = p(x,y)\,p(z)$, encodes full independence between the variables ((x,y) independent of z in this example), a partial factorization of the form $p(x,y,z) = f(x,z)\,g(y,z)\,h(z)$ encodes a weaker type of independence, known as conditional independence. The variables x, y are independent given z, notated $x \perp y \mid z$, if -

$$p(x,y|z) = p(x|z)\,p(y|z)$$

This results in a factorization of $p(x,y,z)$:

$$p(x,y,z) = p(x,y|z)\,p(z) = p(x|z)\,p(y|z)\,p(z)$$

which is of the type mentioned above.

Factorizations like this allow the model to account for redundancies in the representation and make it more succinct, in that fewer parameters are required to represent the distribution, as opposed to a full distribution that doesn't factorize. For instance, in the example above, assuming each one of x, y, z may take N possible values, the probability distribution is specified completely by $N^2 + N^2 + N$ values, instead of the N^3 values that would otherwise be required. Since the factorization is completely determined by the graph corresponding to the model, it alone determines the conditional independence relations present in the model.

We now describe the two types of graphical models which are most commonly used. Let $S = \{X_i\}_{i=1}^n$ be the collection of random variables comprising the model. The first one, known as a Bayesian network, or directed graphical model, represents the factorization of the probability distribution via a directed acyclic graph. Such graph (V, E) induces a factorization of the type $p(x_1, ..x_n) = \prod_{i=1}^n p\left(x_i | x_{\mathrm{par}(i)}\right)$ where $x_{par(i)}$ is the set of parents (direct ancestors) of x_i in the graph.

The other common type of graphical model, and the type we will focus on in the remainder of this section, is called a Markov Random Field (MRF), or an undirected graphical model. Much like its directed counterpart, the factorization is induced by a graph, but as expected, this graph is undirected.

Given an undirected graph representing a MRF, the factorization implied by the graph is as follows: For each maximal clique C_i in the graph, let $S_i \subset S$ be the subset of variables contained in the clique C_i, and let $\phi_i : S_i \to \mathbb{R}^+$ be some positive function (known as a potential function). The probability density of the variables S is then factorized as:

$$p\left(X_1, \ldots, X_n\right) = \frac{1}{Z} \prod_{i=1}^{C} \phi_i(S_i)$$

where $Z = \sum_{X_1} \cdots \sum_{X_n} \prod_{i=1}^{C} \phi_i(S_i)$ is a normalization factor known as the partition function.

Equivalently, we may write factorization in the form of a Gibbs distribution:

$$p\left(X_1, \ldots, X_n\right) = \frac{1}{Z} \exp\left(-\sum_{i=1}^{C} \psi_i\left(S_i\right)\right)$$

where $\psi_i\left(S_i\right) = -\log\left(\phi_i\left(S_i\right)\right)$.

The following basic property of undirected graphical model will be useful to show the connection with the precision matrix.

Theorem [Koller and Friedman, 2009, 4.3.2]:

The variables x_i and x_j are conditionally independent given all other variables if and only if there is no edge connecting x_i and x_j in the graph, ie if x_i and x_j are not neighbors.

As we shall now see, this property is also directly related to the precision matrix of the model, and for that reason we mention it explicitly.

Recall that the (mean-zero) Gaussian density is given by:

$$p\left(x\right) = \frac{1}{\left(2\pi\right)^{\frac{p}{2}} \left|\Sigma\right|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} x^T \Sigma^{-1} x\right)$$

Let us write out the quadratic term explicitly:

$$x^T \Sigma^{-1} x = \sum_{\mathrm{ij}} \Sigma^{-1}_{\mathrm{ij}} x_i x_j = \sum_{i} \Sigma^{-1}_{\mathrm{ii}} x_i^2 + \sum_{i \neq j} \Sigma^{-1}_{\mathrm{ij}} x_i x_j$$

So that:

$$
\begin{aligned}
p(x) &= \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\left[\sum_i \Sigma_{\mathrm{ii}}^{-1} x_i^2 + \sum_{i \neq j} \Sigma_{\mathrm{ij}}^{-1} x_i x_j\right]\right) \\
&= \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\sum_i \Sigma_{\mathrm{ii}}^{-1} x_i^2\right) \exp\left(-\frac{1}{2}\sum_{i \neq j} \Sigma_{\mathrm{ij}}^{-1} x_i x_j\right) \\
&= \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} \prod_i \phi_i(S_i) \prod_{i \neq j} \phi_{\mathrm{ij}}(S_{\mathrm{ij}})
\end{aligned}
$$

Recall that we form the undirected graph corresponding to a given factorization of the density by constructing a corresponding clique for each term in the factorization. In the factorization above, the cliques corresponding to $\phi_i(S_i)$ involve a single node x_i, whereas the cliques corresponding to $\phi_{\mathrm{ij}}(S_{\mathrm{ij}})$ connect two nodes x_i, x_j. It is therefore clear that no edge connects x_i and x_j if and only if $\Sigma_{\mathrm{ij}}^{-1} = 0$, and as discussed above, this is equivalent to $x_i \perp x_j \mid \{x_m\}_{m \neq i,j}$.

Thus we have shown (relying on the theorem above) that edges in the graph, and therefore conditional independence of the type we have considered, corresponds to zeros in the precision matrix. Because the Gaussian distribution is characterized entirely by the mean and the covariance, this argument also shows that in the non-Gaussian, zeros in the precision matrix correspond to conditional uncorrelation.

For the sake of completeness we also give a direct proof of the relationship between the precision matrix and conditional independence in Gaussian undirected graphical models, ie, one which does not rely on the theorem characterizing conditional independence in the graphical model generally.

Assume without loss of generality that $i = 1,\ j = 2$, and further denote $a = x_1, b = x_2, c = \{x_m\}_{m \neq i,j}$ (a vector). The covariance matrix can be partitioned as follows:

$$
\Sigma = \mathbb{E}\begin{pmatrix} a^2 & \mathrm{ab} & ac^T \\ \mathrm{ab} & b^2 & bc^T \\ \mathrm{ac} & \mathrm{bc} & cc^T \end{pmatrix}
$$

Because the model is Gaussian, a and b are conditionally independent given c if and only if their conditional covariance given c, $\rho_{ab|c}$, is 0. Using the well-known expression for the conditional distribution within a multivariate Gaussian distribution [Härdle and Simar, 2007, chapter 5], we have -

$$\rho_{ab|c} = \mathbb{E}ab - \mathbb{E}\left(ac^T\right)\mathbb{E}\left(cc^T\right)^{-1}\mathbb{E}\left(\mathrm{bc}\right)$$

To derive an expression involving the precision matrix, we use the following formula for the inverse of a partitioned matrix: [Horn and Johnson, 2012, 0.7.3]:

For $\Sigma = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$,

$$\Sigma^{-1} = \begin{pmatrix} \left(A_{11} - A_{12}A_{22}^{-1}A_{21}\right)^{-1} & -\left(A_{11} - A_{12}A_{22}^{-1}A_{21}\right)^{-1}A_{12}A_{22}^{-1} \\ -\left(A_{22} - A_{21}A_{11}^{-1}A_{12}\right)^{-1}A_{21}A_{11}^{-1} & \left(A_{22} - A_{21}A_{11}^{-1}A_{12}\right)^{-1} \end{pmatrix}$$

Denote $A_{11} = \mathbb{E}\begin{pmatrix} a^2 & \mathrm{ab} \\ \mathrm{ab} & b^2 \end{pmatrix}$, $A_{22} = \mathbb{E}cc^T$, $A_{21} = \mathbb{E}\begin{pmatrix} \mathrm{ac} & \mathrm{bc} \end{pmatrix}$, $A_{12} = \mathbb{E}\begin{pmatrix} ac^T \\ bc^T \end{pmatrix}$.

Note that the element we are interested in, $\left(\Sigma^{-1}\right)_{\mathrm{ij}} = \left(\Sigma^{-1}\right)_{12}$, is hidden inside the top-left block in the partitioned Σ^{-1} (it is the $(1,2)$ element of that block). Plugging in $A_{11}, A_{12}, A_{22}, A_{21}$ explicitly:

$$\begin{aligned}\left(A_{11} - A_{12}A_{22}^{-1}A_{21}\right)^{-1} &= \left(\mathbb{E}\begin{pmatrix} a^2 & \mathrm{ab} \\ \mathrm{ab} & b^2 \end{pmatrix} - \mathbb{E}\begin{pmatrix} ac^T \\ bc^T \end{pmatrix}\mathbb{E}\left(cc^T\right)^{-1}\mathbb{E}\begin{pmatrix} \mathrm{ac} & \mathrm{bc} \end{pmatrix}\right)^{-1} \\ &= \begin{pmatrix} \mathbb{E}a^2 - \mathbb{E}\left(ac^T\right)\mathbb{E}\left(cc^T\right)^{-1}\mathbb{E}\left(\mathrm{ac}\right) & \mathbb{E}ab - \mathbb{E}\left(ac^T\right)\mathbb{E}\left(cc^T\right)^{-1}\mathbb{E}\left(\mathrm{bc}\right) \\ \mathbb{E}ab - \mathbb{E}\left(bc^T\right)\mathbb{E}\left(cc^T\right)^{-1}\mathbb{E}\left(\mathrm{bc}\right) & \mathbb{E}b^2 - \mathbb{E}\left(bc^T\right)\mathbb{E}\left(cc^T\right)^{-1}\mathbb{E}\left(\mathrm{bc}\right) \end{pmatrix}^{-1}\end{aligned}$$

Denote the elements of this (2×2) matrix by $\begin{pmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{pmatrix}$, so that:

$$\left(\Sigma^{-1}\right)_{12} = \left(\begin{pmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{pmatrix}^{-1}\right)_{12}$$

Using again the formula for the inverse, we have:

$$\left(\Sigma^{-1}\right)_{12} = -\left(v_{11} - v_{12}v_{22}^{-1}v_{21}\right)^{-1}v_{12}v_{22}^{-1}$$

This is a product of 3 terms, the second of which is

$$v_{12} = \mathbb{E}ab - \mathbb{E}\left(ac^T\right)\mathbb{E}\left(cc^T\right)^{-1}\mathbb{E}\left(\mathrm{bc}\right)$$

and this is exactly $\rho_{ab|c}$ as we mentioned above. Since the other two terms are nonzero, $(\Sigma^{-1})_{12} = 0 \Leftrightarrow \rho_{ab|c}$, as claimed.

3.3 The combined model

Now that we have described in some detail the two models that serve as the basic components in our model and estimator, we are ready to address the model itself.

Similar to the viewpoint of [Tipping and Bishop, 1999], we propose to use an ML estimator in the setting of a low rank latent variable model. We have no interest in the latent factors themselves, and only borrow the resulting decomposition of the covariance matrix as in equation 3.1. However, noting that a graphical model structure is stated in terms of the inverse covariance (precision) matrix as discussed earlier, rather than Σ itself, we first bridge the gap by presenting an alternative formulation of the 'low rank + diagonal' model in terms of the precision matrix.

We consider matrices Σ which can be decomposed as $\Sigma = C + \Psi$ where $C \succcurlyeq 0$ has rank $r < n$ and $\Psi \succ 0$ is a diagonal matrix.

Claim:

Σ has the above form if and only if Σ^{-1} has the composition $\Sigma^{-1} = D - L$, where $D = \Psi^{-1}$, $L \succcurlyeq 0$ and L has rank r.

Proof:

We use the following:

(*) If $A \succcurlyeq B \succ 0$ then $A^{-1} \preccurlyeq B^{-1}$ [Horn and Johnson, 2012, 7.7.4]

(**) The well-known Sherman-Morrison formula:

$$(A + UXV)^{-1} = A^{-1} - A^{-1}U\left(X^{-1} + VA^{-1}U\right)^{-1}VA^{-1}$$

(assuming that X, A and $X^{-1} + VA^{-1}U$ are invertible)

($\Rightarrow$) Let $L = \Psi^{-1} - \Sigma^{-1} = \Psi^{-1} - (\Psi + C)^{-1}$.

We know that $\Psi + C \succcurlyeq \Psi$ because $C \succcurlyeq 0$. From (*), it follows that $(\Psi + C)^{-1} \preccurlyeq \Psi^{-1}$, hence $L \succcurlyeq 0$. It remains to show that L has rank r.

Let $C = BB^T$ where B has rank r.

We use (**) with $A = \Psi$, $U = B$, $V = B^T, X = I$. The matrices I, Ψ are clearly invertible. Further, $B^T\Psi^{-1}B \succcurlyeq 0$ because for any vector v,

$$v^T B^T \Psi^{-1} B v = (Bv)^T \Psi^{-1} (Bv) \geq 0$$

owing to $\Psi^{-1} \succcurlyeq 0$. Therefore, $I + B^T\Psi^{-1}B \succcurlyeq I$, so that $I + B^T\Psi^{-1}B$ is invertible. Thus we can apply (**) to obtain:

$$\Sigma^{-1} = (\Psi + C)^1 = \left(\Psi + BB^T\right)^{-1} = \Psi^{-1} - \Psi^{-1}B\left(I + B^T\Psi^{-1}B\right)^{-1} B^T\Psi^{-1}$$

So that $L = \Psi^{-1}B\left(I + B^T\Psi^{-1}B\right)B^T\Psi^{-1}$. Now, because Ψ is invertible we have:

$$\text{rank}\,(L) = \text{rank}\left(B\left(I + B^T\Psi^{-1}B\right)B^T\right) = \text{rank}\left(BB^T + BB^T\Psi^{-1}BB^T\right) = \text{rank}\left(C + C\Psi^{-1}C\right)$$

Letting $T = C + C\Psi^{-1}C = C\,(I + \Psi^{-1}C)$, we clearly have $\text{rank}\,(T) \leq \text{rank}\,(C)$ as $\text{Im}\,(T) \subseteq Im\,(C)$.

We use the following intermediate claim:

Claim:

for $A, B \succcurlyeq 0$,

$$\text{rank}\,(A + B) \geq \text{rank}\,(B)$$

Proof:

It suffices to prove the case where B is of the form $\text{diag}\,(b_1, .., b_r, 0, .., 0)$, because if it isn't, we can let $B = VDV^T$ be an eigen-decomposition with D of the above form, and apply the claim to V^TAV, V^TBV (rank is preserved by matrix conjugation).

So, assume $\text{rank}\,(B) = r$ and B as above, $b_i > 0$. Let $\overline{A}$ be the submatrix obtained by keeping only the first r rows and columns of A. Since A is PSD, so is $\overline{A}$ (since

$x^T\overline{A}x = \overline{x}A\overline{x} \geq 0$ for $\overline{x} = (x_1, .., x_r, 0, .., 0)$). The matrix $\overline{A} + \text{diag}\,(b_1, .., b_r)$ is thus positive definite, so it has rank r. Since this is a sub-matrix of $A + B$, the rank of $A + B$ is at least r, as claimed.

Returning to the original proof, we know that $C \succcurlyeq 0$, and the matrix $C\Psi^{-1}C$ is also PSD, since $v^T C\Psi^{-1}Cv = (\text{Cv})\,\Psi^{-1}\,(\text{Cv}) \geq 0$ owing to $\Psi \succcurlyeq 0$. Therefore, by the result, $\text{rank}\,(T) \geq \text{rank}\,(C)$, and since we've already shown that $\text{rank}\,(T) \leq \text{rank}\,(C)$, in total we have $\text{rank}\,(L) = \text{rank}\,(T) = \text{rank}\,(C) = r$ as claimed.

($\Leftarrow$) Given $D \succ 0$ and $L \succcurlyeq 0$ of rank r, let $C = (D - L)^{-1} - D^{-1}$ and $\Psi = D^{-1}$. Since $L \succcurlyeq 0$, and using (*), we have $D^{-1} \preccurlyeq (D - L)^{-1}$ so that $C \succcurlyeq 0$. It remains to show that C has rank r. Since $C \succcurlyeq 0$, we can now apply the first direction ($\Rightarrow$) to C, Ψ, to deduce that $\Sigma^{-1} = \widehat{D} - \widehat{L}$ with $\widehat{D} = \Psi^{-1} = D$, $\widehat{L} = \Psi^{-1} - (\Psi + C)^{-1}$, and $\text{rank}\left(\widehat{L}\right) = r$. But,

$$\widehat{L} = D - \left(D^{-1} + C\right)^{-1} = D - \left((D - L)^{-1}\right)^{-1} = L$$

so that $\text{rank}\,(L) = r$ as claimed.

Chapter 4

The estimator

4.1 Penalized likelihood estimator

Using the above claim, we now present a novel estimator for the setting of the PCA/factor analysis model, which uses a penalized maximum likelihood formulation. Unlike in [Tipping and Bishop, 1999], we estimate Σ^{-1} directly via an optimization problem, and this fact will allow us later to incorporate the graphical model structure as well.

Recall that our claim from section 3.3 asserts that a PCA-type decomposition (low rank + diagonal) of Σ corresponds to a similar decomposition of Σ^{-1}. Thus we may estimate Σ^{-1} in the form $\Sigma^{-1} = D - L$, where D is a diagonal matrix, $L \succcurlyeq 0$ and L should have low rank. To find such L, we use the well-known trace penalty [Candès and Recht, 2009] to penalize the log-likelihood function. Along with the other constraints, the problem takes the following form:

$$\begin{aligned} \max \quad & l(X) - \rho \cdot \operatorname{trace}(L) \\ & D, L \succcurlyeq 0 \\ \text{s.t.} \quad & D \text{ diagonal} \\ & D - L \succ 0 \\ & X = D - L \end{aligned}$$

or, for the PCA case:

$$\begin{aligned} \max \quad & l(X) - \rho \cdot \operatorname{trace}(L) \\ & D, L \succcurlyeq 0 \\ \text{s.t.} \quad & D = \sigma^2 I \\ & D - L \succ 0 \\ & X = D - L \end{aligned}$$

Where $l(X)$ is the Gaussian likelihood in terms of the precision matrix X:

$$l\left(X\right) = \log|X| - \operatorname{trace}\left(XS\right)$$

and S is the sample covariance matrix.

Note:

The constraint $D - L \succ 0$ ensures that the matrix $X = D - L$, which is our estimator for Σ^{-1}, remains positive definite. It may seem that this constraint is problematic in from an algorithmic viewpoint, as the set $\{D, L | D - L \succ 0\}$ is not closed. However, note that the function $\log|X|$ is a log-barrier function for the semidefinite cone (the determinant approaches 0 as the eigenvalues approach 0), and $l\left(D - L\right) \to -\infty$ as $D - L \to 0$, so this term forces the solution to be obtained in the interior of $\{D, L | D - L \succcurlyeq 0\}$. Thus, the constraint only serves to ensure that X has no negative eigenvalues (since 0 is impossible to have).

Because l is a concave function, as we observed, the above is a convex problem which employs the trace penalty to obtain a solution with a low rank component L. This, in contrast to the ML estimators in [Tipping and Bishop, 1999] which use a fixed rank r for L as part of the problem formulation and algorithm. We discuss later how to choose the penalty parameter ρ.

For the sake of brevity we will not explicitly mention the PCA constraint in what follows, but the reader should keep in mind that the development applies to that case equally well.

4.2 The unconstrained PCA case

When there are no constraints, the above estimator is in fact strongly related to the[Tipping and Bishop, 1999] estimator, at least under the PCA model:

Claim:

Under the PCA model, an optimal solution is obtained when L has the same eigenvectors as the sample covariance matrix S.

Proof:

Recall that the objective function is:

$$f(D,L) = \log\det(D-L) - \text{trace}(S(D-L)) - \rho \cdot \text{trace}(L)$$

It suffice sto show that for for any (D,L) satisfying the constraints, $f(D,L) \leq f(D,L^*)$ where L^* is a matrix with the same eigenvectors as S.

Let $D = \sigma^2 I$ and L be some matrices that satisfy the constraints. We select L^* as follows: Let $S = UEU^T$ be an eigen-decomposition of S where $E = \left[\text{diag}\left(\lambda^{\downarrow}(S)\right)\right]$ (that is, the eigenvectors sorted in descending order). Set $L^* = U[\text{diag}\left(\lambda^{\downarrow}(L)\right)U^T$. Clearly, this matrix has the same eigenvectors as S, and the same eigenvalues of L.

First, we claim that it suffices to show that $\text{trace}(\text{SL}) \leq \text{trace}(SL^*)$. It is known from basic linear algebra that both the determinant and the trace of a matrix depend on its eigenvalues alone, and not on the eigenvectors. In addition, the eigenvalues of $D-L$ are $\sigma^2 - \lambda_i(L)$ (where $\lambda_i(L)$ are the eigenvalues of L), and this also holds irrespective of the eigenvectors of L. As such, the only term in the definition of f which depends on the eigenvectors of L is $-\text{trace}(S(D-L))$. Since L^* and L share the same eigenvalues, we have:

$$f(D,L) - f(D,L^*) = -\text{trace}(S(D-L)) + \text{trace}(S(D-L^*)) = \text{trace}(SL) - \text{trace}(SL^*)$$

Thus, $f(D,L) \leq f(D,L^*)$ is equivalent to $\text{trace}(\text{SL}) \leq \text{trace}(SL^*)$.

Next, we use the following proposition [Bhatia, 2013, III.6.14]:

For any PSD matrices A, B,

$$\operatorname{trace}(AB) \leq \left\langle \lambda^{\downarrow}(A), \lambda^{\downarrow}(B) \right\rangle = \sum_i \lambda_i^{\downarrow}(A)\lambda_i^{\downarrow}(B)$$

The definition of L^* implies that -

$$\operatorname{trace}(SL^*) = \operatorname{trace}\left(UEU^TU[\operatorname{diag}(\lambda(L))]U^T\right) = \operatorname{trace}(E[\operatorname{diag}(\lambda(L))]) = \sum_i E_{\text{ii}}\lambda_i^{\downarrow}(L)$$

where the last equality holds because both E and $[diag(\lambda(L))]$ are diagonal matrices. According to the proposition, we have

$$\operatorname{trace}(SL) \leq \sum_i \lambda_i^{\downarrow}(S)\lambda_i^{\downarrow}(L) = \sum_i E_{\text{ii}}\lambda_i^{\downarrow}(L)$$

$$\operatorname{trace}(SL) \leq \operatorname{trace}(SL^*)$$

as claimed.

Since we now know the optimal solution can be written as $L^* = U\left[\operatorname{diag}\left(\lambda^{\downarrow}(L)\right)\right]U^T$, we can obtain a simpler problem for the eigenvalues alone. Let $\operatorname{diag}\left(\lambda^{\downarrow}(L)\right) = \{x_i\}$. We can rewrite the objective function using the following identities:

$$f(D, L) = \log\det(D - L) - \operatorname{trace}(S(D - L)) - \rho \cdot \operatorname{trace}(L)$$

$$\log\deg(D - L) = \sum_i \log\left(\sigma^2 - x_i\right)$$

$$\operatorname{trace}(S(D - L)) = \sigma^2 \cdot \operatorname{trace}(S) - \sum_i \lambda_i(S)x_i$$

$$\rho \cdot \operatorname{trace}(L) = \rho \cdot \sum_i x_i$$

We therefore obtain the following optimization problem:

$$\begin{aligned} \arg \quad & \max_{x,\sigma^2} \sum_i \log\left(\sigma^2 - x_i\right) - \sigma^2 \cdot \operatorname{trace}(S) + \sum_i \lambda_i(S)\, x_i - \rho \cdot \sum_i x_i \\ & x_i \geq 0, \\ \text{s.t.} \quad & \sigma^2 \geq 0, \\ & \sigma^2 > x_i \end{aligned}$$

With $x = \lambda(L)$. In fact the constraint $x_i \geq 0$ is enough by itself due to the log-term. More compactly,

$$\begin{aligned} \arg \quad & \max_{x,\sigma^2} \sum_i \log\left(\sigma^2 - x_i\right) - \sigma^2 \cdot \operatorname{trace}(S) + \sum_i \left(\lambda_i(S) - \rho\right) x_i \\ \text{s.t.} \quad & x_i \geq 0, \end{aligned}$$

Denote $c_i = \lambda_i(S) - \rho$.

We can use the KKT optimality conditions with Lagrange multipliers $\lambda_i \geq 0$ corresponding to the constraint $x_i \geq 0$:

$$L\left(\sigma^2,\ x,\ \lambda\right) = \sum_i \log\left(\sigma^2 - x_i\right) - \sigma^2 \cdot \operatorname{trace}(S) + \sum_i c_i x_i + \lambda^T x$$

(1) $\frac{\partial L}{\partial \sigma^2} = 0 : \sum_i \frac{1}{\sigma^2 - x_i} - \operatorname{trace}(S) = 0$

(2) $\frac{\partial L}{\partial x_i} = 0 : \frac{-1}{\sigma^2 - x_i} + c_i + \lambda_i = 0$

By Complementary slackness, at the optimum $\lambda_i = 0$ or $x_i = 0$. If $\lambda_i = 0$, from (2):

$$\sigma^2 - x_i = \frac{1}{c_i}$$

Since $\sigma^2 - x_i > 0$ by constraint, it follows that $\lambda_i = 0$ is possible only when $c_i > 0$, that is, $\lambda_i(S) > \rho$. Thus, when $c \leq 0$, then necessarily $x_i = 0$. We conclude that L has a nullity of at least the amount of eigenvalues of S equal to or lower than ρ (but we can't conclude it equal that exactly, since when $c_i > 0$, both $x_i = 0$ and $x_i > 0$ are possible). Thus, when no graphical model constraints are present, the solution to the optimization problem shares many of the characteristics of the standard eigenvalue-decomposition-based estimator: it retains the eigenvectors of the sample covariance, while nulling some of the corresponding eigenvalues. Unlike it, however, the solution doesn't retain the remaining eigenvalues unchanged.

We also briefly derive a simple algorithm for solving the above problem exactly. To that end, let:

$$C_1 = \{i | x_i = 0\}$$

$$C_2 = \{i | \lambda_i = 0\}$$

Then, from (2),

$$\lambda_i = \frac{1}{\sigma^2} - c_i \qquad i \in C_1$$

$$x_i = \sigma^2 - \frac{1}{c_i} \qquad i \in C_2$$

Note that if $c_{i_1} \geq c_{i_2}$ and $\lambda_{i_2} = 0$, then necessarily also $\lambda_{i_1} = 0$ (because $0 \leq \lambda_{i_1} = \frac{1}{\sigma^2} - c_{i_1} \leq \frac{1}{\sigma^2} - c_{i_2} = 0$). We deduce that if we sort $\lambda(S)$ in ascending order (equivalently, $\{c_i\}$), there is an i^* such that $C_2 = \{i^*, i^* + 1, \ \ldots, \ p\}$. Assume henceforth that this was done.

Now, rewrite the equality $\sum_i \frac{1}{\sigma^2 - x_i} - \text{trace}\,(S) = 0$ as follows:

$$\sum_{i \in C_1} \frac{1}{\sigma^2 - x_i} + \sum_{i \in C_2} \frac{1}{\sigma^2 - x_i} - \text{trace}\,(S) = 0$$

And plugging in the above:

$$\sum_{i \in C_1} \frac{1}{\sigma^2} + \sum_{i \in C_2} c_i - \text{trace}\,(S) = 0$$

Letting $m = |C_1|$, we have:

$$\frac{m}{\sigma^2} + \sum_{i \in C_2} c_i - \text{trace}\,(S) = 0$$

$$\sigma^2 = \frac{m}{\text{trace}\,(S) - \sum_{i \in C_2} c_i}$$

We don't know a-priori which indices are in C_2. However, by the remark above, we know that $C_1 = \{1, .., i^* - 1\}$ and $C_2 = \{i^*, i^* + 1, \ \ldots, \ p\}$. Thus, we may, for each

possible $i^* = 1, .., p$, calculate the corresponding quantities m and $\sum_{i \in C_2} c_i$, and hence σ^2 by the above. Then, plugging into $x_i = \sigma^2 - \frac{1}{c_i}$ gives x_i for $i \in C_2$, from which we may calculate the value of the objective function. Maximizing over the possible values of i^* gives the optimal value and the corresponding solution for σ^2 and x.

4.3 Graphical model structure

The main advantage of the above formulation, and indeed the main motivation for using the decomposition in terms of the precision matrix, is the fact that it is allows us to enforce a graphical model structure within the same framework. We assume a desired graphical model is known a-priori as set of constraints $X_{\text{ij}} = 0$ for pairs $(i, j) \in E$, and plug them into the optimization problem:

$$\begin{aligned} \max \quad & l(X) - \rho \cdot \text{trace}(L) \\ & D, L \succcurlyeq 0 \\ & D \text{ diagonal} \\ \text{s.t.} \quad & D - L \succ 0 \\ & X = D - L \\ & X_{ij} = 0, \ (i, j) \in E \end{aligned}$$

or, for the PCA case:

$$\begin{aligned} \max \quad & l(X) - \rho \cdot \text{trace}(L) \\ & D, L \succcurlyeq 0 \\ & D = \sigma^2 I \\ \text{s.t.} \quad & D - L \succ 0 \\ & X = D - L \\ & X_{ij} = 0, \ (i, j) \in E \end{aligned}$$

4.4 Selecting ρ

For arbitrary ρ, denote by $P(\rho)$ the problem with penalty parameter ρ.

Claim:

Let $\rho_0 = \sum_j S_{\text{ij}}$. Then, for every $\rho > \rho_0$, the solution to the optimization problem $P(\rho)$ is $L = 0$ and

- $D = \text{diag}\left(\frac{1}{S_{11}}, \ldots, \frac{1}{S_{\text{pp}}}\right)$ in FA case
- $D = \frac{1}{\alpha} I$ with $\alpha = \frac{1}{p}\sum_{i=1}^{p} S_{\text{ii}}$ in PCA case

Proof:

Denote by $l(X)$ the log-likelihood and by $f(D, L) = l(D - L) - \rho \cdot \text{trace}(L)$ the objective function.

We first show that:

$$\text{trace}(LS) \leq \rho_0 \cdot \text{trace}(L) \tag{4.1}$$

Since $L \succcurlyeq 0$, we have $L_{\text{ij}} \leq \sqrt{L_{\text{ii}} L_{\text{jj}}}$ (this follows e.g. from the C-S inequality for random variables).

We therefore have:

$$\text{trace}(LS) = \sum_{i,j} S_{\text{ij}} L_{\text{ij}} \leq \sum_{i,j} S_{\text{ij}} \sqrt{L_{\text{ii}} L_{\text{jj}}} \leq \frac{1}{2} \sum_{i,j} S_{\text{ij}} (L_{\text{ii}} + L_{\text{jj}})$$

By the definition of $\rho_0 (= \sum_j S_{\text{ij}})$,

$$\sum_{i,j} S_{\text{ij}} L_{\text{ii}} \leq \rho_0 \sum_i L_{\text{ii}} = \rho_0 \cdot \text{trace}(L)$$

So we have 4.1.

Next, we show:

$$l(D - L) \leq l(D) + \rho_0 \cdot \text{trace}(L) \tag{4.2}$$

Writing out,

$$l(D - L) - l(D) = \log|D - L| - \text{trace}((D - L)S) - \log|D| + \text{trace}(DS) = \log\frac{|D - L|}{|D|} + \text{trace}(LS)$$

By a theorem [Horn and Johnson, 2012, 4.3.12], since $L \succcurlyeq 0$ (constraint), $\lambda_i(D-L) \leq \lambda_i(D)$. So:

$$\log \frac{|D-L|}{|D|} = \log \frac{\prod_i \lambda_i(D-L)}{\prod_i \lambda_i(D)} = \log \prod_i \underbrace{\frac{\lambda_i(D-L)}{\lambda_i(D)}}_{\leq 1} \leq 0$$

Using equation 4.1, It follows that:

$$l(D-L) - l(D) \leq \operatorname{trace}(LS) \leq \rho_0 \cdot \operatorname{trace}(L)$$

which is equation 4.2.

Finally, using 4.2,

$$f(D,L) = l(D-L) - \rho \cdot \operatorname{trace}(L) \leq l(D) + \rho_0 \cdot \operatorname{trace}(L) - \rho \cdot \operatorname{trace}(L)$$

$$= l(D) + \underbrace{(\rho_0 - \rho)}_{<0} \cdot \underbrace{(\operatorname{trace}(L))}_{\geq 0} \leq l(D) = f(D,0)$$

The last inequality is strict provided that $\operatorname{trace}(L) > 0$, which is equivalent to $L \neq 0$ since $L \geq 0$. Thus the optimum necessarily satisfies $L = 0$. Expressions for D follow easily.

Claim:

If for some ρ_1 the optimal solution to $P(\rho_1)$ satisfies $L = 0$, then for every $\rho \geq \rho_1$ the optimal solution to $P(\rho)$ also satisfies $L = 0$.

Proof:

Assume $\rho \geq \rho_1$. Let (D, L) be the optimal solution of $P(\rho)$, and $(D_1, L_1) = (D_1, 0)$ be the optimal solution of $P(\rho_1)$. Because $(D_1, 0)$ is feasible for $P(\rho)$, so we have:

$$l(D_1) \leq l(D-L) - \rho \cdot \operatorname{trace}(L)$$

On the other hand, because $\operatorname{trace}(L) \geq 0$ subject to the constraints, we also have:

$$l\left(D-L\right)-\rho\cdot\mathrm{trace}\left(L\right)\leq l\left(D-L\right)-\rho_1\cdot\mathrm{trace}\left(L\right)\leq\max_{D,L}l\left(D-L\right)-\rho_1\cdot\mathrm{trace}\left(L\right)=l(D_1)$$

where we have the omitted the constraints for brevity, and the last equality follows from the definition of $(D_1, 0)$ as the optimum of $P(\rho_1)$.

It follows that the inequalities are equalities throughout and in particular $l\left(D-L\right)-\rho\cdot\mathrm{trace}\left(L\right)=l\left(D-L\right)-\rho_1\cdot\mathrm{trace}\left(L\right)$, so that $\mathrm{trace}\left(L\right)=0$. Because $L\succcurlyeq 0$, it follows that $L=0$ as claimed.

4.4.1 Selecting ρ based on cross-validation

The above results allow us to place an upper bound ρ when selecting its value. However, to select ρ exactly, another method is required.

We propose a method which relies on cross-validation, similar to the way parameters are often selected in the context of machine learning algorithms. To justify that, we first note that maximum likelihood estimation, in general, can be seen as an instance of empirical risk minimization (ERM). Recall that in the context of machine learning, ERM is a meta-algorithm for selecting a hypothesis (eg, classifier) based on the following rule: select the one which achieves the minimal loss, based on some loss function, on the training data. More formally, given a loss function $l\left(x,\ \theta\right)$ and training data $\{x_i\}$, the empirical loss of hypothesis θ is $\sum_{i=1}^{n}l\left(x_i,\theta\right)$, and one selects θ for which this is minimal.

In our setting, given an i.i.d sample $\{x_i\}\sim p\left(x|\theta\right)$ from which a parameter θ is to be estimated, the maximum likelihood estimator is:

$$\widehat{\theta}=\arg\max\log\prod_{i=1}^{n}p\left(x_i|\theta\right)=\arg\max\sum_{i=1}^{n}\log p\left(x_i|\theta\right)$$

Clearly, if we set $l\left(x,\theta\right)=-\log p\left(x_i|\theta\right)$, then ML minimizes this loss function over the data.

This observation motivates us to borrow another idea from machine learning algorithms, namely, using cross-validation (or, more generally, any training/validation data partitioning) for parameter tuning. Thus, we divide the data into a training subset and a validation subset. The estimator is computed using the training data, and then its

performance evaluated on the validation data. This performance measure allows us to compare several competing estimators and select a suitable value for ρ.

Given a candidate value ρ_i, the data is partitioned into $X_{\text{train}} = \{x_i^{\text{trn}}\}$ and $X_{\text{val}} = \left\{x_i^{\text{val}}\right\}$. The training data X_{train} is used to estimate Σ, by forming the sample covariance of X_{train}, and computing an estimator using that sample covariance and the candidate value ρ_i. This gives an estimate $\widehat{\Sigma}_i$. We then evaluate it on the validation set, using the same loss function, by computing the empirical loss: $\text{loss}s_i = \sum_j l\left(x_j^{\text{val}}, \widehat{\Sigma}_i\right) = -\sum_j \log p\left(x_j^{\text{val}}|\widehat{\Sigma}_i\right)$. Up to sign, this is precisely the log-likelihood, and in the Gaussian setting we already know it can be expressed (up to additive and multiplicative constants) as:

$$\text{loss}s_i = \log\left|\widehat{\Sigma}_i\right| + tr\left(\widehat{\Sigma}_i^{-1} S_{\text{val}}\right)$$

Here, S_{val} is the sample covariance of the validation data X_{val}.

In practice, more than a single training/validation subset is used, so the procedure is repeated multiple times (N_{CV}) using different subsets and the losses averaged. We summarize the method below.

1. Select candidate values $\{\rho_i\}$
2. For each $j = 1..N_{\text{CV}}$, sample from $\{x_i\}$ a validation subset X_{val}^j and define the training subset $X_{\text{train}}^j = \{x_i\} - X_{\text{val}}^j$
3. For each i:
    a. For each $j = 1..N_{\text{CV}}$:
        i. Compute an estimator $\widehat{\Sigma}_i$ using ρ_i based on X_{train}^j
        ii. Compute $\text{loss}s_i^j$, the empirical loss of $\widehat{\Sigma}_i$ on X_{val}^j
    b. Compute the average for ρ_i, $\text{loss}s_i = \frac{1}{N_{\text{CV}}} \sum_j \text{loss}s_i^j$
4. Pick $\rho_i = \arg\min \text{loss}s_i$

Note:

The empirical loss can be regarded as a way of approximating the expected loss. Denoting by $p(x)$ the true distribution of the data, we have:

$$\mathbb{E}l\left(x,\theta\right)=\int p\left(x\right)l\left(x,\theta\right)=-\int p\left(x\right)\log p\left(x|\theta\right)=\int p\left(x\right)\log\frac{p\left(x\right)}{p\left(x|\theta\right)}\frac{1}{p\left(x\right)}$$
$$=\int p\left(x\right)\log\frac{p\left(x\right)}{p\left(x|\theta\right)}+\int p\left(x\right)\log\frac{1}{p\left(x\right)}$$

The second term, the entropy of H, doesn't depend on θ. The first one is the Kullback-Leibler (KL)-divergence $D\left(p\left(x\right)||p\left(x|\theta\right)\right)$. Thus, ML essentially aims to minimize the KL-divergence between the true distribution and $p\left(x|\theta\right)$, by selecting θ which minimizes the empirical divergence. Of course, minimizing the expected quantities would require knowledge of the true underlying distribution, so it's impractical.

4.5 Consistency

In this section we show that given sufficiently many i.i.d samples from a fixed covariance which is assumed to the satisfy the model exactly, the algorithm recovers the correct decomposition, at least in the PCA case. We work in the setting of fixed p, and $n\to\infty$.

Denote by $\overline{X},\overline{D},\overline{L}$ the matrices corresponding to the decomposition of the ground truth covariance matrix, ie, for $\Sigma=C+\Psi$,

$$\overline{D}=\Psi^{-1},\ \overline{L}=\Psi^{-1}-\left(\Psi+C\right)^{-1},\ \overline{X}=\Sigma^{-1}$$

We assume throughout this section that Σ,C,Ψ satisfy the model exactly with rank r and constraint set E, and $\Psi=\lambda I$.

Notation:

Consider the problem $P\left(\rho\right)$:

$$\begin{aligned}\max\quad & l(X)-\rho\cdot\operatorname{trace}(L)\\ & D,L\succcurlyeq 0\\ & D=\sigma^2 I\\ \text{s.t.}\quad & D-L\succ 0\\ & X=D-L\\ & X_{ij}=0,\ (i,j)\in E\end{aligned}$$

We denote by $\left(X\left(\rho\right),D\left(\rho\right),L\left(\rho\right)\right)$ the solution to $P\left(\rho\right)$.

When X is a given fixed matrix satisfying $X \succ 0$ and $X_{\text{ij}} = 0, \quad (i,j) \in E$, (*) becomes a problem for D, L alone. We denote this problem by $P(\rho, X)$ and its solution by $D(\rho, X)$, $L(\rho, X)$.

We also consider the following problem for X (*):

$$\begin{array}{ll} \max_X & l_s(X) \\ & X \succ 0 \\ \text{s.t.} & X_{ij} = 0, \ (i,j) \in E \end{array}$$

Denote by $\widehat{X}$ its solution.

Remark:

In fact, a more correct notation would also include the dependence of the solutions on S explicitly. However, most of the claims below do not involve n or S directly, so this is omitted. The reader should keep this dependence in mind during the parts of the discussion that do involve n or S.

The argument of the consistency proof is essentially as follows:

Consider the problem $P(\lambda)$. When $\rho \to 0$, the effect of the trace penalty becomes negligible (assuming, without justification for now, that L is bounded). We know that the solution with $\rho = 0$ must satisfy $D - L = S^{-1}$, and this suggests that as $\rho \to 0$, $D(S,\rho) - L(S,\rho) \to S^{-1}$. Further, when $n \to \infty$, $S \to \Sigma$ (by the law of large numbers, as discuused previously). Thus when $n \to \infty$ and $\rho \to 0$, we expect that $D(S,\rho) - L(S,\rho) \approx \Sigma^{-1}$. Looking back at the optimization problem, we now consider the effect of the trace penalty, and note that under the assumption that $D - L = \Sigma^{-1}$, which we assume satisfies the model assumptions exactly, the trace penalty allows recovery of the correct D, L exactly (this is stated and justified below). If the optimiation problem is reasonably well-behaved, it is not unreasonable to expect that when $D - L \approx \Sigma^{-1}$ (as we argued happens when $n \to \infty$ and $\rho \to 0$), it would recover D and L approximately (and exactly in the limit).

The above argument is essentially correct, but technical difficulties complicate the proof somewhat. In particular, note that when $\rho = 0$ there solution set is not a singleton.

This irregularity makes using the stability results (which allow us to make the tansition from $D - L = \Sigma^{-1}$ to $D - L \approx \Sigma$ in the argument above) more difficult, complicating the proof somewhat.

Claim 1:

The solution to $P(\rho, \Sigma^{-1})$ is $\left(D\left(\rho, \Sigma^{-1}\right), L\left(\rho, \Sigma^{-1}\right)\right) = \left(\overline{D}, \overline{L}\right)$.

Proof:

First, consider the problem $P(\rho, X)$ in general. The constraint $X^{-1} = D - L$ shows that this is in fact a problem for D alone, ie, for σ^2 alone. Also, the constraint $D - L \succ 0$ is satisfied automatically. In addition, since X is fixed, the log-likelihood term in the objective is immaterial, as it depends on X alone. Thus, we are left with an equivalent reduced problem:

$$\begin{array}{cc} \max_{\sigma^2} & -\rho \cdot \operatorname{trace}\left(\sigma^2 I - X\right) \\ \text{s.t.} & \sigma^2 I - X \succcurlyeq 0 \\ & \sigma^2 \geq 0 \end{array}$$

The objective function is maximal when σ^2 is minimal subject to the constraints. The constraint $\sigma^2 I - X \succcurlyeq 0$ is equivalent to $\sigma^2 \geq \lambda_{\max}(X)$, and since $X \succ 0$, the second constraint is satisfied automatically if the first one is. Thus the solution is given by $D = \lambda_{\max}(X) I,\ L = D - X^{-1}$.

Now, set $X = \Sigma^{-1}$. Recall that $\Sigma = C + \Psi = C + \lambda I$, with $\operatorname{rank}(C) = r < p$. Thus:

$$\lambda_{\max}(X) = \lambda_{\max}\left(\Sigma^{-1}\right) = \frac{1}{\lambda_{\min}(\Sigma)} = \frac{1}{\lambda}$$

and the solution is $D = \frac{1}{\lambda} I = \Psi^{-1}$. Together with $L = D - \Sigma^{-1}$ we have exact recovery of $\overline{D}, \overline{L}$ as claimed.

Claim 2:

When $X \to \Sigma^{-1}$, $\left(D\left(\rho, X\right), L\left(\rho, X\right)\right) \to \left(\overline{D}, \overline{L}\right)$.

Proof:

We've seen above that the solution to $P(\rho, X)$ is given by $D = \lambda_{\max}(X) I,\ L = D - X^{-1}$. The function $X \mapsto \lambda_{\max}(X)$ is a continuous function of X Bhatia [2013, VI.1.2 and following discussion], and so is $X \mapsto X^{-1}$ (via the equality $A^{-1} = \frac{1}{\det A} adj A$), so the results follows from claim 1.

Claim 3:

As $\rho \to 0$, $X(\rho) \to \widehat{X}$.

Proof:

The reasoning behind the claim is simple: as $\rho \to 0$, the effect of the term $\rho \cdot \text{trace}(L)$ in the objective of $P(\rho)$ should become negligible, so (regardless of what $D(\rho), L(\rho)$ may be), $X(\rho)$ should approach the solution of (*). However, the term $\rho \cdot \text{trace}(L)$ also depends on L, and if $L(\rho)$ is not bounded uniformly in ρ, the argument breaks down. This slightly complicates the proof: we can't simply ignore this term, but we can approximate it with a term $\rho \cdot \text{trace}\left(\widehat{L}\right)$, where $\widehat{L} = L\left(\rho, \widehat{X}\right)$.

Because $l_s(X)$ is strictly concave on the domain $\{X : X \succ 0\}$ [Borwein and Lewis, 2010, 3.3.3], it is injective, so from continuity of the inverse function there exists δ such that $|l_S(X) - l_S(Y)| < \delta \Rightarrow \|X - Y\| < \epsilon$. Set $\rho_0 = \frac{\delta}{\text{trace}\left(\widehat{L}\right)}$. We show that for every $\rho < \rho_0$, $\left\|X(\rho) - \widehat{X}\right\| < \epsilon$.

First, note that $X(\rho)$ is feasible for (*), for which $\widehat{X}$ is the maximizer, so that:

$$l_S\left(\widehat{X}\right) \geq l_S(X(\rho))$$

As mentioned, set $\widehat{L} = L(\rho, \widehat{X})$ and similarly $\widehat{D} = D\left(\rho, \widehat{X}\right)$. $\left(\widehat{X}, \widehat{D}, \widehat{L}\right)$ is feasible for $P(\rho)$ because $\left(\widehat{D}, \widehat{L}\right)$ is feasible for $P(\rho,\ X)$, and since $(X(\rho), D(\rho), L(\rho)$ is optimal for $P(\rho)$, we have:

$$l_S(X(\rho)) - \rho \cdot \text{trace}(L(\rho)) \geq l_S\left(\widehat{X}\right) - \rho \cdot \text{trace}\left(\widehat{L}\right)$$

Rearranging, we have:

$$l_S\left(\widehat{X}\right) - l_S\left(X\left(\rho\right)\right) \leq -\rho \cdot \text{trace}\left(L\left(\rho\right)\right) + \rho \cdot \text{trace}\left(\widehat{L}\right) \leq \rho \cdot \text{trace}\left(\widehat{L}\right)$$

Together, these inequalities give us:

$$\left|l_s\left(X\left(\rho\right)\right) - l_S(\widehat{X})\right| \leq \rho \cdot \text{trace}\left(\widehat{L}\right) \leq \rho_0 \cdot \text{trace}\left(\widehat{L}\right) = \delta$$

And finally, this implies that

$$\left\|X\left(\rho\right) - \widehat{X}\right\| < \epsilon$$

as claimed.

Claim 4:

As $n \to \infty$, $\widehat{X} \to \overline{X}(= \Sigma^{-1})$ (w.p. 1).

Proof:

Recall $\widehat{X}$ is the solution to:

$$\begin{array}{cc} \max_X & l_s(X) \\ & X \succ 0 \\ \text{s.t.} & X_{ij} = 0,\ (i,j) \in E \end{array}$$

Note that when $S = \Sigma$ we have $\widehat{X} = \overline{X}$: we know that Σ^{-1} is the unconstrained maximum (the standard maximum-likelihood solution), and since we're assuming Σ satisfies the model, this is also the constrained maximum. Therefore, it seems reasonable to think that $S \to \Sigma \Rightarrow \widehat{X} \to \overline{X}$. To justify this, we use a stability theorem from the field of parametric optimization. Before we cite the theorem we start with some preliminary definitions. As will be seen, we require a fairly specialized version of the theorem, so some of the following definitions might seem redundant, but we include them for the purpose of stating the theorem as it appears.

Def: [Bank, 1983, pages 22,25,29]

A function is called quasiconvex if it satisfies

$$f(\lambda x + (1-\lambda)y) \leq \max\{f(x),\ f(y)\}$$

for each x, y in its domain and each $\lambda \in [0,1]$.

Let $(X, d_X), (\Lambda, d_\Lambda)$ be metric spaces.

If $A \subset X$ and $\epsilon > 0$, an ϵ-neighborhood if A is $U_\epsilon A = \left\{ x \in X | \inf_{y \in A} D_X(x,y) < \epsilon \right\}$. An ϵ-neighborhood of $B \subset \Lambda$ will be denoted $V_\epsilon B$.

A set-valued mapping Γ is a function $\Gamma : \Lambda \to 2^X$.

Γ is upper semicontinuous according to Hausdorff (u.s.c-H) at a point λ^0 if for each $\epsilon > 0$, there exists $\delta > 0$ such that

$$\Gamma(\lambda) \subset U_\epsilon \Gamma(\lambda_0) \quad \forall \lambda \in V_\delta\{\lambda^0\}$$

holds.

Γ is lower semicontinuous according to Berge (u.s.c-B) at a point λ^0 if for each open set Ω satisfying $\Omega \cap \Gamma(\lambda^0) \neq \phi$ there exists $\delta > 0$ such that

$$\Gamma(\delta) \cap \Omega \neq \phi \quad \forall \lambda \in V_\delta\{\lambda^0\}$$

holds.

Γ is upper semicontinuous according to Berge (u.s.c-B) at a point λ^0 if for each open set Ω containing $\Gamma(\lambda_0)$ thre exists $\delta > 0$ such that

$$\Gamma(\lambda) \subset \Omega \quad \forall \lambda \in V_\delta\{\lambda^0\}$$

holds.

Γ is closed at a point λ^0 if for each pair of sequences $\{\lambda^t\} \subset \Lambda$ and $\{x^t\} \subset X$, satisfying

$$\lambda^t \to \lambda^0, \quad x^t \to x^0, \quad x^t \in \Gamma(\lambda^t)$$

it follows that $x^0 \in \Gamma(\lambda^0)$.

For a parametric optimization problem (P_λ) of the form

$$\inf \{f(x,\lambda) \mid x \in M(\lambda)\} \qquad \lambda \in \Lambda$$

we define the extreme value function

$$\varphi(\lambda) = \inf \{f(x,\lambda) \mid x \in M(\lambda)\}$$

and the optimal set mapping

$$\psi(\lambda) = \{x \in M(\lambda) \mid f(x,\lambda) = \varphi(\lambda)\}$$

Theorem [Bank, 1983, 4.3.3]

If the following conditions are satisfied:

1. $X = \mathbb{R}^n$
2. $M : \Lambda \to 2^X$ is l.s.c-B at $\lambda^0 \in \Lambda$
3. $\psi(\lambda^0)$ is non-empty and bounded
4. f is lower semicontinuous on $X \times \{\lambda^0\}$ and a point $x^0 \in \psi(\lambda^0)$ exists such that f is upper continuous at (x^0, λ^0)
5. $f(\cdot, \lambda)$ is quasiconvex on X for each fixed $\lambda \in \Lambda$
6. $M(\lambda^0)$ is convex and closed
7. $M(\lambda)$ is convex for each $\lambda \in \Lambda$
8. M is closed at λ^0

Then, φ is continuous at λ^0, and ψ is u.s.c-B at λ^0.

We wish to apply the theorem to our problem, namely:

$$\begin{array}{cc} \max_X & l_s(X) \\ & X \succ 0 \\ \text{s.t.} & \\ & X_{ij} = 0,\ (i,j) \in E \end{array}$$

First note that this is equivalent to -

$$\begin{aligned} \max_X \quad & l_s(X) \\ \text{s.t.} \quad & X \succcurlyeq 0 \\ & X_{ij} = 0, \ (i,j) \in E \end{aligned}$$

because the log-det term in the likelihood function precludes a solution with an eigenvalue of 0.

Now, apply the theorem by setting:

$X = \Lambda = \text{Sym}^{p \times p}$ (the Euclidean space of symmetric matrices),

$x = X$,

$\lambda = S$,

$f(x, \lambda) = l_S(X)$,

$M(\lambda) \equiv M_0 = \{X | X \succcurlyeq 0, \ \ X_{\text{ij}} = 0 \ \ (i,j) \in E\}$, and finally

$\lambda^0 = \Sigma$.

Let us check the conditions one by one to see that they are satisfied. First,

1. This is holds (our problem lies in Euclidean space).

2. In our case, $M(\lambda) = M_0$ is fixed and independent of λ. Thus, the only set containing $\Gamma(\lambda^0)$ is M_0 itself, and the condition for M being closed is satisfied trivially.

3. Because for each fixed S this is a convex optimization problem with a strictly convex objective function l_s, it has a unique minimizer. In other words, $\psi(\lambda^0)$ is a singleton, and in particular it is closed and bounded.

4. l_s is a continuous function, so in particular it is l.s.c and u.s.c at every point in its domain.

5. l_s is convex in X, so in particular it is quasiconvex.

6. M_0 is a closed convex set, being the intersection of the closed convex PSD cone and hyperplanes.

7. See (6) above.

8. Since M_0 is closed, it is easy to see that the condition in the definition of a closed set-valued mapping is satisfied.

Thus we may apply the theorem to conclude that ψ is u.s.c-B at λ^0. However, seeing as $\psi(\lambda)$ is a singleton for each λ as remarked above, the property of being u.s.c-B is equivalent to ψ being a continuous function. We therefore conclude that ψ is continuous. Going back to the notation used before stating the theorem, we have

$$\psi(\Sigma) = \overline{X}$$

$$\psi(S) = \widehat{X}$$

so that $S \to \Sigma \Rightarrow \widehat{X} \to \overline{X}$ follows at once from the continuity of ψ.

Finally, as $n \to \infty$, $S \to \Sigma$ by the strong LLN, from which the result follows.

We can now prove our main result:

Claim 5:

As $n \to \infty$ and $\rho \to 0$, $(X(\rho), D(\rho), L(\rho)) \to \left(\overline{X}, \overline{D}, \overline{L}\right)$.

Proof:

By claim 3, as $\rho \to 0$, $X(\rho) \to \widehat{X}$. By claim 4, as $n \to \infty$, $\widehat{X} \to \overline{X}$, and combining we get $X(\rho) \to \overline{X}$. From claim 2 we now have that $(D(\rho, X(\rho)), L(\rho, X(\rho))) \to \left(\overline{D}, \overline{L}\right)$. However, $D(\rho, X(\rho))$ is simply $D(\rho)$ (this is seen immediately form the definition of $D(\rho)$ and $D(\rho, X)$), and similarly for $L(\rho, X(\rho))$, so that $(D(\rho), L(\rho)) \to \left(\overline{D}, \overline{L}\right)$ as claimed. The convergence of $X(\rho)$ is immediate, as $X(\rho) = D(\rho) - L(\rho)$.

Note:

We've previously described a method to select ρ based on cross-validation. If this method is employed, under the current regime of $n \to \infty$, then the value selected by it will ensure that Σ is estimated correctly in the limit. This is argued as follows. Recall ρ is chosen to minimize

$$\text{loss}_i = \log\left|\widehat{\Sigma}_i\right| + \text{trace}\left(\widehat{\Sigma}_i^{-1} S_{\text{val}}\right)$$

$$D_{\mathrm{KL}}\left(\Sigma_0, \Sigma_1\right) = \frac{1}{2}\left(\operatorname{tr}\left(\Sigma_1^{-1}\Sigma_0\right) - p + \log\frac{|\Sigma_1|}{|\Sigma_0|}\right)$$

It is known that D_{KL} is minimal (and equal to 0) if and if $\Sigma_0 = \Sigma_1$ [Cover and Thomas, 2012, 17.1]. We can rewrite the (limiting) loss in terms of D_{KL} as follows:

$$\mathrm{loss}s_i \to \log\left|\widehat{\Sigma}_i\right| + tr\left(\widehat{\Sigma}_i^{-1}\Sigma\right) = 2D_{\mathrm{KL}}\left(\Sigma, \widehat{\Sigma}_i\right) + p - \log|\Sigma|$$

So, when selecting ρ_i to minimize $\mathrm{loss}s_i$, in the limit, the optimal value is $p - \log|\Sigma|$ and it can be attained when $\widehat{\Sigma}_i = \Sigma$. Since we've already shown that this is achievable by setting $\rho \to 0$, it follows that the cross-validation procedure would select a value of ρ to satisfy $\widehat{\Sigma}_i \to \Sigma$ (either via $\rho \to 0$, or otherwise).

Chapter 5

Algorithm

5.1 Introduction

Consider again the problem:

$$\begin{aligned} \max \quad & l(X) - \rho \cdot \text{trace}(L) \\ & D, L \succcurlyeq 0 \\ & D \text{ diagonal} \\ \text{s.t.} \quad & D - L \succ 0 \\ & X = D - L \\ & X_{ij} = 0, \ (i,j) \in E \end{aligned}$$

This is a convex optimization problem. In fact, it's a Semi-Definite Programming (SDP) problem, and so can be efficiently solved by off-the-shelf solvers when the problem size is small to moderate. However, this approach becomes infeasible when the dimension exceeds a few tens/hundreds, so a different method is desirable. We present here an algorithm based on the Alternating Direction Method of Multipliers (ADMM) framework.

5.2 ADMM overview

The general framework of the ADMM algorithm is as follows [Boyd et al., 2011]:

Consider the general problem:

$$\begin{aligned} \min \quad & f(x) + g(z) \\ \text{s.t.} \quad & Ax + Bz = c \end{aligned}$$

where f, g are convex functions and x, z are vector/matrix variables. Note that this formulation is quite general – for instance, enforcing the constraint $x - z = 0$ and letting $g(z) = \mathbb{I}_C(z)$ be an indicator function of a convex set C (ie 0 for $z \in C$ and ∞ otherwise), the problem becomes a general constrained optimization problem:

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & \text{x}\in\text{C} \end{aligned}$$

The ADMM algorithm proceeds by defining the augmented Lagrangian corresponding to the above problem:

$$L(x, z, y) = f(x) + g(z) + \langle y, Ax + Bz - c \rangle + \frac{\beta}{2} \|Ax + Bz - c\|_2^2$$

Note that this is the 'unaugmented' Lagrangian associated with the problem:

$$\begin{aligned} \min \quad & f(x) + g(z) + \tfrac{\beta}{2} \|Ax + Bz - c\|_2^2 \\ \text{s.t.} \quad & Ax + Bz = c \end{aligned}$$

which is equivalent to the original problem (as the third term in the objective is zero on the feasible set).

Finally, ADMM consists of the following iterations:

$$x^{k+1} = \arg\min_x L(x, z^k, y^k)$$

$$z^{k+1} = \arg\min_z L(x^{k+1}, z, y^k)$$

$$y^{k+1} = y^k + \beta(Ax^{k+1} + Bz^{k+1} - c)$$

The y update can be seen as subgradient ascent applied to the dual problem, where the x, z updates are the corresponding updates of the primal variables. Breaking the x, z update into two separates steps is what allows us in many cases to exploit the structure

of the objective function and achieve simple updates. More details and examples are available at the review paper [Boyd et al., 2011].

Note that wemay enforce a constraint of a general convex set C by adding an indicator function $\mathbb{I}_C(\cdot)$ to either f or g. This effectively moves the constraint to one of the updates but not the other. For instance, if we use $f(x) + \mathbb{I}_C(x)$, the x update becomes:

$$\begin{aligned} x^{k+1} &= \arg\min_x L(x, z^k, y^k) = \arg\min_x f(x) + \mathbb{I}_C(x) + \langle y, Ax + Bz - c\rangle + \frac{\beta}{2}\|Ax + Bz - c\|_2^2 \\ &= \arg\min_{x\in C} f(x) + g(z) + \langle y, Ax + Bz - c\rangle + \frac{\beta}{2}\|Ax + Bz - c\|_2^2 \end{aligned}$$

which is the same as the original update, except for the added constraint $x \in C$. The z update remains unaffected.

Finally, note that when the objective function is separable to more than a pair of functions, eg:

$$\begin{aligned} \min \quad & f(x) + g(z) + h(u) \\ \text{s.t.} \quad & Ax + Bz + Du = c \end{aligned}$$

Then the updates become

$$x^{k+1} = \arg\min_x L(x, z^k, u^k, y^k)$$

$$z^{k+1} = \arg\min_z L\left(x^{k+1}, z, u^k, y^k\right)$$

$$u^{k+1} = \arg\min_u L(x^{k+1}, z^{k+1}, u, y^k)$$

$$y^{k+1} = y^k + \beta(Ax^{k+1} + Bz^{k+1} + Du^{k+1} - c)$$

5.3 Applying ADMM to our problem

We use three variables, D, L, X. The only constraint common to more than a single variable is the linear constraint $X = D - L$, and we tuck the other constraints into the single-variable updates as described above. The augmented Lagrangian is:

$$L\left(X,D,L,Z\right)=\log\det\left(X\right)-\langle S,X\rangle-\rho\cdot\text{trace}\left(L\right)-\langle Z,X-(D-L)\rangle-\frac{\beta}{2}\left\|X-(D-L)\right\|_F^2$$

where we have used the $\langle\rangle$ notation to denote the standard Euclidean inner product in $\mathbb{R}^{m\times n}$, ie, $\langle A,B\rangle=\text{trace}\left(AB\right)=\sum_{\text{ij}}A_{\text{ij}}B_{\text{ij}}$.

5.3.1 D update

We start with the D update step:

$$\begin{aligned}D^{k+1}=\arg\max\log\det\left(X^k\right)-\left\langle S,X^k\right\rangle-\rho\cdot\text{trace}\left(L^k\right)\\-\left\langle Z^k,X^k-\left(D-L^k\right)\right\rangle-\frac{\beta}{2}\left\|X^k-\left(D-L^k\right)\right\|_F^2\end{aligned}$$

We may omit additive terms which do not depend on D to obtain:

$$D^{k+1}=\arg\max\left\langle Z^k,D\right\rangle-\frac{\beta}{2}\left\|X^k-\left(D-L^k\right)\right\|_F^2$$

We use the following completing-the-square identity valid for any three matrices D,A,B:

$$\begin{aligned}\left\|D-A\right\|_F^2+\langle D,B\rangle&=\left\|D\right\|_F^2+\left\|A\right\|_F^2-2\left\langle D,A\right\rangle+\langle D,B\rangle=\left\|D\right\|_F^2+\left\|A\right\|_F^2-2\left\langle D,A-\frac{1}{2}B\right\rangle\\&=\left\|D\right\|_F^2+\left\|A-\frac{1}{2}B\right\|_F^2-\left\|-\frac{1}{2}B\right\|_F^2+2\left\langle A,\frac{1}{2}B\right\rangle-\langle D,A-B\rangle\\&=\left\|D-\left(A-\frac{1}{2}B\right)\right\|_F^2-\left\|\frac{1}{2}B\right\|_F^2+\langle A,B\rangle\end{aligned}$$

So we have:

$$\begin{aligned}
D^{k+1} &= \arg\max \left\langle Z^k, D^k \right\rangle - \frac{\beta}{2} \left\| D - \left(L^k + X^k\right) \right\|_F^2 \\
&= \arg\min \frac{\beta}{2} \left\| D - \left(L^k + X^k\right) \right\|_F^2 - \left\langle Z^k, D \right\rangle \\
&= \arg\min \frac{\beta}{2} \left\| D - \left(L^k + X^k\right) \right\|_F^2 - \frac{\beta}{2} \left\langle \frac{2}{\beta} Z^k, D \right\rangle \\
&= \arg\min \left\| D - \left(L^k + X^k\right) \right\|_F^2 + \left\langle D, -\frac{2}{\beta} Z^k \right\rangle \\
&= \arg\min \left\| D - \left(L^k + X^k + \frac{1}{\beta} Z^k \right) \right\|_F^2 - \left\| \frac{1}{\beta} Z^k \right\|_F^2 + \left\langle L^k + X^k, -\frac{2}{\beta} Z^k \right\rangle \\
&= \arg\min \left\| D - \left(L^k + X^k + \frac{1}{\beta} Z^k \right) \right\|_F^2
\end{aligned}$$

Recalling the constraints and setting $Q_1 = L^k + X^k + \frac{1}{\beta} Z^k$, we arrive at the following sub-problem for D:

$$\begin{aligned}
\min \quad & \|D - Q_1\|_F^2 \\
\text{s.t.} \quad & D \succcurlyeq 0 \\
& D \text{ diagonal, or } D = \sigma^2 I
\end{aligned}$$

where the type of constraint used corresponds again to the PCA or FA case. In both cases, the problem has an immediate closed-form solution, as follows.

Factor analysis case:

The constraint $D \succcurlyeq 0$, for D diagonal, is equivalent to all of the diagonal elements of D being non-negative. Further, owing to D being diagonal, the objective function $\sum_{\text{ij}} \left(d_{\text{ij}} - (Q_1)_{\text{ij}} \right)^2$ depends on D only through $\sum_i \left(d_{\text{ii}} - (Q_1)_{\text{ii}} \right)^2$, so we have the following problem:

$$\begin{aligned}
\min \quad & \textstyle\sum_i \left(d_{\text{ii}} - (Q_1)_{\text{ii}} \right)^2 \\
\text{s.t.} \quad & d_{ii} \geq 0
\end{aligned}$$

Since the objective is separable, each d_{ii} can be optimized separately, and the solution is clearly $d_{\text{ii}} = \max\left((Q_1)_{\text{ii}}, 0 \right)$.

PCA case:

In a similar manner to the above case, we obtain the following problem:

$$\begin{aligned} \min \quad & \textstyle\sum_i \left(\sigma^2 - (Q_1)_{\text{ii}}\right)^2 \\ \text{s.t.} \quad & \sigma^2 \geq 0 \end{aligned}$$

Since the objective function is strictly convex with global minimum at $\frac{1}{p}\sum_i (Q_1)_{\text{ii}}$, the solution is clearly seen to be $\sigma^2 = \max\left(\frac{1}{p}\sum_i (Q_1)_{\text{ii}},\ 0\right)$ and $D = \sigma^2 I$.

5.3.2 L update

The L update, in a similar vein:

$$\begin{aligned} L^{k+1} &= \arg\max \log\det\left(X^k\right) - \left\langle S, X^k\right\rangle \\ &\quad - \rho \cdot \text{trace}\,(L) - \left\langle Z^k, X^k - \left(D^{k+1} - L\right)\right\rangle - \frac{\beta}{2}\left\|X^k - \left(D^{k+1} - L\right)\right\|_F^2 \\ &= \arg\max \left\langle -Z^k, L\right\rangle - \rho \cdot \text{trace}\,(L) - \frac{\beta}{2}\left\|L - \left(D^{k+1} - X^k\right)\right\|_F^2 \\ &= \arg\max \left\langle -Z^k, L\right\rangle - \rho \cdot \langle L, I\rangle - \frac{\beta}{2}\left\|L - \left(D^{k+1} - X^k\right)\right\|_F^2 \\ &= \arg\max \left\langle -Z^k - \rho I, L\right\rangle - \frac{\beta}{2}\left\|L - \left(D^{k+1} - X^k\right)\right\|_F^2 \\ &= \arg\min \frac{\beta}{2}\left[\left\|L - \left(D^{k+1} - X^k\right)\right\|_F^2 + \left\langle \frac{2}{\beta}\left(Z^k + \rho I\right), L\right\rangle\right] \\ &= \arg\min \frac{\beta}{2}\left\|L - \left(D^{k+1} - X^k - \frac{1}{\beta}\left(Z^k + \rho I\right)\right)\right\|_F^2 \\ &= \arg\min \left\|L - \left(D^{k+1} - X^k - \frac{1}{\beta}\left(Z^k + \rho I\right)\right)\right\|_F^2 \end{aligned}$$

Again recalling the constraints and denoting $Q_2 = D^{k+1} - X^k - \frac{1}{\beta}\left(Z^k + \rho I\right)$, we arrive at the sub-problem:

$$\begin{aligned} \min \quad & \|L - Q_2\|_F^2 \\ \text{s.t.} \quad & L \succcurlyeq 0 \end{aligned}$$

This problem is known [Higham, 1988] to have the following closed-form solution:

- Calculate an eigenvalue decomposition $Q_2 = URU$
- Set $L = U\left[R\right]_+ U^T$

Here, $\left[D\right]_+$ is the result of applying the non-negative-part operator elementwise to D, ie, it nulls the negative diagonal elements.

5.3.3 X update

Finally, the X update:

$$\begin{aligned}
X^{k+1} &= \arg\max \log\det(X) - \langle S, X\rangle \\
&\quad - \rho \cdot \operatorname{trace}\left(L^{k+1}\right) - \left\langle Z^k, X - \left(D^{k+1} - L^{k+1}\right)\right\rangle - \frac{\beta}{2}\left\|X - \left(D^{k+1} - L^{k+1}\right)\right\|_F^2 \\
&= \arg\max \log\det(X) - \left\langle S + Z^k, X\right\rangle - \frac{\beta}{2}\left\|X - \left(D^{k+1} - L^{k+1}\right)\right\|_F^2 \\
&= \arg\max \log\det(X) - \frac{\beta}{2}\left[\left\|X - \left(D^{k+1} - L^{k+1}\right)\right\|_F^2 + \left\langle X, \frac{2}{\beta}\left(S + Z^k\right)\right\rangle\right] \\
&= \arg\max \log\det(X) - \frac{\beta}{2}\left[\left\|X - \left(D^{k+1} - L^{k+1} - \frac{1}{\beta}\left(S + Z^k\right)\right)\right\|_F^2\right]
\end{aligned}$$

Denoting $Q_3 = D^{k+1} - L^{k+1} - \frac{1}{\beta}\left(S + Z^k\right)$, we obtain the sub-problem:

$$\begin{aligned}
\max \quad & \log\det(X) - \tfrac{\beta}{2}\|X - Q_3\|_F^2 \\
\text{s.t.} \quad & X \succ 0 \\
& X_{\text{ij}} = 0, \;\; (i,j) \in E
\end{aligned}$$

This problem is considerably harder than the sub-problems for D and L because of the constraints. We address it via a block-coordinate-descent (BCD) scheme wherein each column of X is updated separately while the other columns are considered fixed. To that end, and assuming w.l.g that the last column of X is to be updated, partition both X and Q_3 as follows:

$$X = \begin{pmatrix} Z & x \\ x^T & y \end{pmatrix}$$

$$Q_3 = \begin{pmatrix} V & u \\ u^T & w \end{pmatrix}$$

where $Z, A \in \mathbb{R}^{p\times p}, x, b \in \mathbb{R}^p, y, c \in \mathbb{R}$.

We use the following formula for the determinant of a partitioned matrix [Boyd and Vandenberghe, 2004, A.5.5]:

$$\det X = (\det Z)\left(y - x^T Z^{-1} x\right)$$

So that:

$$\log\det X = \log\det Z + \log\left(y - x^T Z^{-1} x\right)$$

Similarly, $\|X - Q_3\|_F^2$ is decomposed corresponding to the partition:

$$\|X - Q_3\|_F^2 = \|Z - V\|_F^2 + (y - w)^2 + 2\|x - u\|^2$$

Further, the constraint $X \succ 0$ is equivalent to $Z \succ 0$ and $y - x^T Z^{-1} x > 0$ [Horn and Johnson, 2012, 7.7.6], and the constraint $X_{\text{ij}} = 0 \;\; (i,j) \in E$ is easily expressed using the parititioned variables Z, x (we will omit the exact statement for the sake of brevity and simply write $(Z,x) \in E$ with a slight abuse of notation. Note that y cannot be zero due to the PD constraint.)

Combining the above, we may rewrite the optimization problem in partitioned form:

$$\begin{aligned} \max \quad & \log\det Z + \log\left(y - x^T Z^{-1} x\right) - \tfrac{\beta}{2}\left[\|Z - V\|_F^2 + (y - w)^2 + 2\|x - u\|^2\right] \\ & Z \succ 0 \\ \text{s.t.} \quad & y - x^T Z^{-1} x > 0 \\ & (Z,x) \in E \end{aligned}$$

Now, as mentioned above, within the BCD scheme we hold Z fixed while updating x, y. Also note that the constraint $y - x^T Z^{-1} x > 0$ is effectively enforced by optimizing the objective function since it contains a corresponding log-barrier term. We therefore have the following problem for a single column update:

$$\begin{aligned} \max \quad & \log\left(y - x^T Z^{-1} x\right) - \tfrac{\beta}{2}\left[(y - w)^2 + 2\|x - u\|^2\right] \\ \text{s.t.} \quad & x \in E \end{aligned}$$

This problem is simple enough and has low enough dimension (p) to solve with an off-the-shelf solver.

Finally, if the constraint $x \in E$ forces some elements of x to be zero, denote by $\widetilde{x}$ the part of x corresponding to the remaining elements, and similarly for $\widetilde{u}$, and also denote the sub-matrix of Z^{-1} corresponding to the nonzero elements of x by $\widetilde{Z^{-1}}$. It then holds that:

$$x^T Z^{-1} x = \widetilde{x}^T \widetilde{Z^{-1}} \widetilde{x}$$

$$\|x - u\|^2 = \|\widetilde{x} - \widetilde{u}\|^2$$

So that we may we may transform the problem to an equivalent unconstrained problem:

$$\max_{x,y} \log\left(y - x^T Z^{-1} x\right) - \frac{\beta}{2}\left[(y - w)^2 + 2\|x - u\|^2\right]$$

in which we have omitted the tildes for brevity.

5.3.4 Z update

The Z update is the simplest, since it doesn't require solving an optimization sub-problem:

$$Z^{k+1} = Z^k + \left(X^{k+1} - \left(D^{k+1} - L^{k+1}\right)\right)$$

5.3.5 Putting everything together

We initialize the algorithm with $D = \left(\frac{1}{p}\sum_i S_{\text{ii}}\right)^{-1} I$ in the PCA case or $D = \sum_i \text{diag}\left(\frac{1}{S_{\text{ii}}}\right)$ in the FA case, and $L = 0, X = D - L$.

Our termination criterion is based on change in X, D, L compared to the previous iteration. That is, we calculate:

$$\delta = \frac{\left\|\left[L^{k+1}\ D^{k+1}\ X^{k+1}\right] - \left[L^k\ D^k\ X^k\right]\right\|_F}{\left\|\left[L^k\ D^k\ X^k\right]\right\|_F}$$

and terminate if δ is lower than some threshold, or if a maximal number of iteraitons has passed.

To recap, the algorithm consists of the following steps:

- Input: β, ρ
- Init:
  - PCA case: $D = \left(\frac{1}{p}\sum_i S_{\text{ii}}\right)^{-1} I, L = 0, X = D - L$
  - FA case: $D = \sum_i \text{diag}\left(\frac{1}{S_{\text{ii}}}\right), L = 0, X = D - L$

- For $k = 1, 2..$
  - D update:
    - PCA case: $D^{k+1} = \sigma^2 I$ with $\sigma^2 = \max\left(\frac{1}{p}\sum_i (Q_1)_{\text{ii}},\ 0\right)$
    - FA case: $D^{k+1} = \sum_i \text{diag}\,(d_{ii})$ with $d_{\text{ii}} = \max\left((Q_1)_{\text{ii}}, 0\right)$
  - L update:
    - Calculate an eigenvalue decomposition $Q_2 = URU^T$
    - $L^{k+1} = U\left[R\right]_+ U^T$
  - X update:
    - Set $X^{k+1} = X^k$
    - For $m = 1, .., p$:
      - Set w, u, Z corresponding to columm m of X^{k+1}
      - Solve:

$$\max_{x,y} \log\left(y - x^T Z^{-1} x\right) - \frac{\beta}{2}\left[(y-w)^2 + 2\left\|x-u\right\|^2\right]$$

      - Set column m of X^{k+1} to $[x; y]$
  - Z update:
    - $Z^{k+1} = Z^k + \beta(X^{k+1} - \left(D^{k+1} - L\,(k+1)\right)$
  - Check termination criteria

Chapter 6

Experimental results

6.1 Synthetic results

We perform synthetic experiments where samples are drawn from a $N(0, \Sigma)$ distribution and input to the estimator. The process is described in greater detail below.

The following parameters come into play when running an experiment:

p – problem dimension

r – rank of L

s – target sparsity of Σ^{-1}

n – number of samples used to estimate the covariance

N – number of times to run the estimator

N_{CV} – number of cross validation subsets used

Each experiment is comprised of the following steps:

1. Select the ground truth covariance matrix
2. Perform parameter tuning via cross-validation to obtain ρ (or other parameters for other algorithms, see below)
3. For $i = 1, .., N$
    a. Draw n samples
    b. Run the estimation algorithm to obtain $error_i$
4. Calculate the mean and standard deviation of $\{error_i\}_{i=1}^{N}$

Generating a ground truth matrix that satisfies the model assumptions exactly is not entirely straightforward. We employ an iterative scheme as follows. We start with a random $p \times r$ matrix A. This matrix will serve to define L via $L = AA^T$. Iteratively, we pick a location (i, j) in A to zero, that satisfies $A_{\text{ij}} \neq 0$, and such that setting $A_{\text{ij}} = 0$ would not introduce a row or column of zeros in A. We proceed until the target sparsity of Σ^{-1} (equivalently, of $L = AA^T$) is reached, or until a maximal number of iterations has passed.

Once the target sparsity is reached, we test whether L is block-diagonal by traversing the graph defined by L and partitioning it into connected components. Having block-diagonal L means the problem is decomposable into sub-problems of lower rank than our intended rank r. Therefore, if more than a single component is found, meaning that L is block-diagonal, we start over by drawing A again and repeating the process.

Once A is set, we select D to satisfy the constraints $D \succcurlyeq L$ and $D \succcurlyeq 0$. In the FA model, we set D to be the solution to the following optimization problem:

$$\begin{aligned} \min \quad & \text{trace}(D) \\ \text{s.t.} \quad & D \succcurlyeq L \end{aligned}$$

In the PCA model (that we use for the experiments presented here) D is also a multiple of the identity, so the above reduces to $D = \lambda_{max}(L)$.

Finally we make $X = D - L$ more stable by increasing all of the eigenvalues of D by $\lambda_{\min}(D)$. We then set $\Sigma = (D - L)^{-1}$, $\Psi = D^{-1}$, and $C = \Sigma - \Psi$.

Once the ground truth matrices are available, we perform parameter tuning by cross-validation. This is done by performing the steps outlined in chapter 4, which we repeat here briefly: First, a set of values to test for the parameter of interest is selected. We then draw a set of $K = n \times \frac{N_{\text{CV}}}{N_{\text{CV}}-1}$ samples, and run the estimator (using a candidate parameter value) N_{CV} times using random subset of size n from the drawn samples. Each time, the remaining $K - n = \frac{n}{N_{\text{CV}}-1}$ samples are used to test the estimator. This process is repeated for each candidate parameter value, and the averaged values of the N_{CV} errors are used to select which of the candidate parameters is selected. We typically use $N_{\text{CV}} = 3$.

After the algorithm's parameters have been chosen using the validation set, we draw n samples and apply the algorithm, using the selected parameter value. To calculate error

statistics, this is repeated N times, to obtain N estimation errors. Typically, a value between $N = 30$ and $N = 50$ is used.

The above is repeated for each estimator we test, using the same samples for both the cross validation and the error calculation.

When applying our algorithm, we select the constraint set E used by first running the graphical lasso on the same data used for cross-validation, and use its output to determine which elements of Σ^{-1} are constrained to be zero in our problem. This set is then fixed when the test errors are calculated.

The performance of our algorithm is compared to three other algorithms. Two of them are algorithms suited for each of the underlying models individually – the graphical lasso and the probabilistic PCA estimator, and the third is a naïve algorithm taking both models into account.

The first of these is the graphical lasso, which is defined as the solution to the following sparsity-penalized ML problem:

$$\widehat{\Sigma^{-1}} = \arg\max l_S(X) - \rho \|X\|_1$$

The parameter ρ is selected by cross-validation.

The second is the probabilistic PCA algorithm [Tipping and Bishop, 1999] already mentioned. It estimates each component of the covariance separately, and given by the following:

$$\widehat{\sigma^2} = \frac{1}{p-r} \sum_{j=r+1}^{p} \lambda_j$$

$$\widehat{A} = U_r \left(\Lambda_r - \widehat{\sigma^2} I \right)^{1/2},$$

where Λ is a diagonal matrix with entries λ_i, $S = U \Lambda U^T$ is an eigen-decomposition of the sample covariance and $U_r = U(:, 1:r)$, $\Lambda_r = \Lambda(1:r, 1:r)$. The rank r to use is selected by cross-validation.

The third method is a naïve combination of both of the above methods. It applies first the graphical lasso to obtain an estimate $\widehat{\Sigma^{-1}}$. It then applies the probabilistic PCA

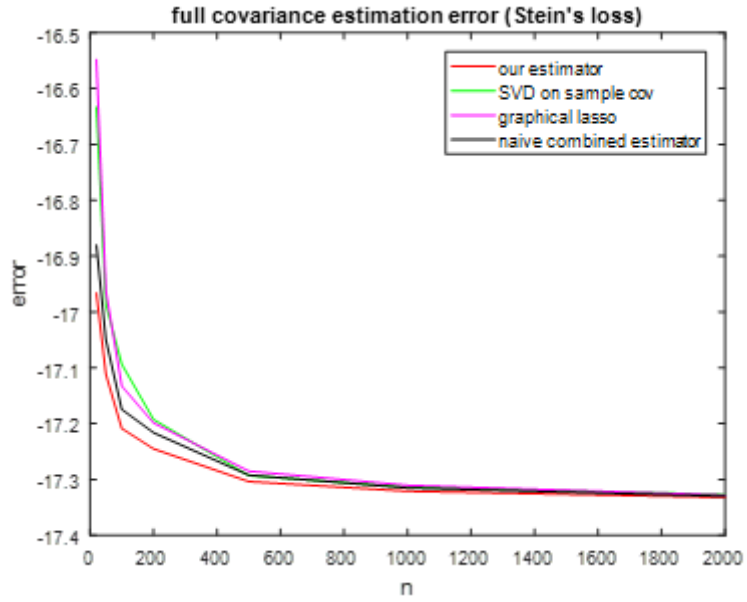

Figure 6.1: Synthetic experiment results

decomposition to $\left(\widehat{\Sigma^{-1}}\right)^{-1}$, rather than to the sample covariance as usual. Both the rank and ρ are selected by cross-validation.

We use the KL divergence to compare the algorithms (also known as Stein's loss in this context). Figure 6.1 shows a plot of the error collected over 7 experiments, using different values of n as shown on the x axis. In these experiments the value of $p = 10$, $r = 3$ and $s = 0.5$ are used.

6.2 ADMM convergence

The second result demonstrates the convergence of the ADMM algorithm to the true solution. We perform a synthetic experiment (as detailed in section 6.1), first using cvx [Grant et al., 2008] to solve the optimization problem, and then using our algorithm. We plot the (normalized) distances between the outputs of cvx, which we consider as the true solution to the optimization problem, and the outputs of our algorithm in each iteration. The normalized distances are defined by:

$$d_L = \frac{\left\|\widehat{L}_{\text{ADMM}} - \widehat{L}_{\text{cvx}}\right\|_F}{\left\|\widehat{L}_{\text{cvx}}\right\|_F}$$

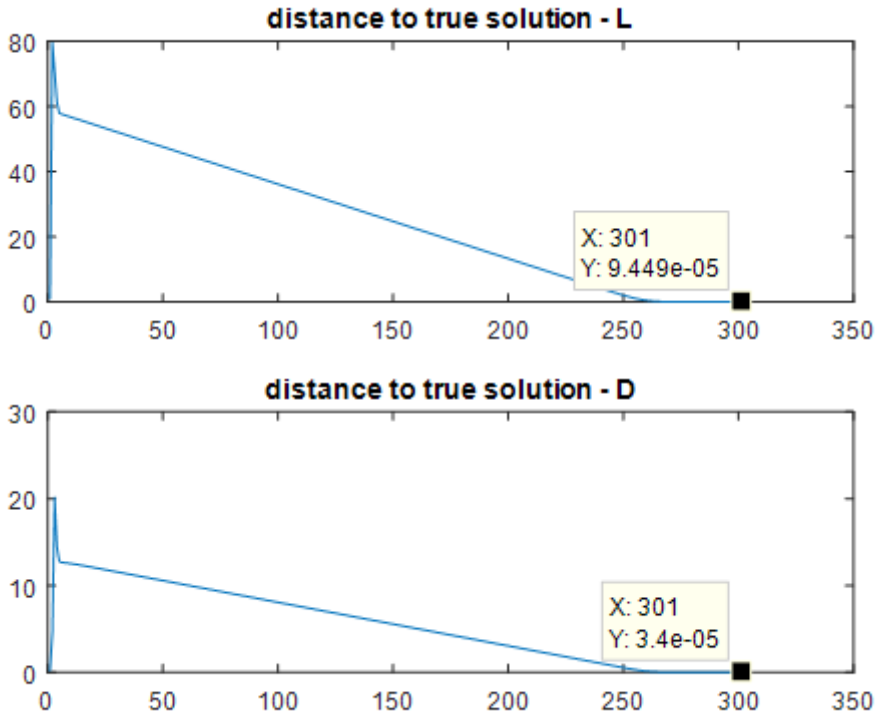

Figure 6.2: ADMM convergence

$$d_D = \frac{\left\| \widehat{D}_{\text{ADMM}} - \widehat{D}_{\text{cvx}} \right\|_F}{\left\| \widehat{D}_{\text{cvx}} \right\|_F}$$

The parameters used in the experiment are $p = 10$, $n = 100$, $r = 3$, $s = 0.3$. We use $\beta = 0.5$ as the step size for the algorithm.

Figure 6.2 shows the errors in each iteration. The runtime measured in the test was $T_{\text{cvx}} = 10.7[sec]$ for the cvx solver, and $T_{\text{ADMM}} = 0.4[sec]$ for our algorithm (for 300 iterations).

6.3 Image dataset

As a practical-inspired example, we run the algorithm on a dataset of face images. Such images are known [Turk and Pentland, 1991] to have a low rank structure amenable to analysis using PCA. In addition, as we now demonstrate, such data is expected to have a graphical model structure, exhibiting conditional independence of far-away pixels given all other pixels in the image. The reasoning behind this is simple – When pixels (x, y) are far

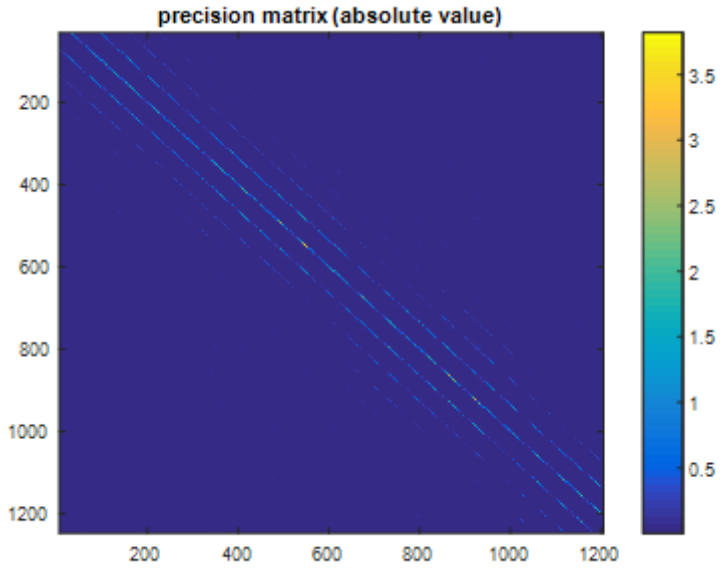

Figure 6.3: Precision matrix of face image dataset

apart, the information contained in the surrounding and intermediate area between them (Z) is sufficient to cover all probability-related knowledge on the target pixels themselves. That is, y adds no new information related to x, so that $p(x|Z, y)$ and $p(x|Z)$ coincide. When they are adjacent, however, there is still useful information in the value of y that is not contained in Z. Clearly, the farther away the pixels, the more likely it is that they are independent.

The validity of the above argument is demonstrated in figure 6.3, which shows a part of the precision matrix of the empirical data covariance. The images are each represented in column-stack notation for the purpose of vectorizing them, so that the immediate four neighbors of a pixel at location i in column-stack representation (and therefore in the plotted precision matrix) are present at locations $i-1, i+1, i-ImSize, i+ImSize$ (where ImSize is the dimension of the image). Slightly farther neighbors are similarly located at $i-2, i+2, i-imSize+1, i-imSize-1, i-2\times imSize$, etc. In the figure, it is exactly those neighbors that have a high value in the precision matrix, whereas the other values are close to 0.

To test the performance of our algorithm, we apply it in the following scenario: The covariance of the dataset is calculated once using the complete data. This serves as the ground truth which the estimated covariance is compared against. For estimation, we

Algorithm	**Error (Stein loss)**
Probabilistic PCA	-42.5 ± 0.7
Constreained graphical Lasso	-52.1 ± 0.6
Naïve combined algorithm	-50.3 ± 0.8
Our estimator	**-53.5 ± 0.4**

Table 6.1: Image dataset results

sample a subset of the images and compute the covariance from that. Unlike in section 6.1, the input constraint set E used in the algorithm is supplied in the advance and extracted from the full covariance. Thus, this method is impractical for real-world applications involving such data, but it demonstrates that knowledge of the locations of expected zeros can improve estimation accuracy. To make the comparison fair, we compare to a variant of the graphical lasso algorithm, instead of the standard version, that also has the input E supplied to it in advance, and uses it as constraints in place of the sparsity penalty. This version is used both in the standalone graphical lasso and when used in conjunction with the probabilistic PCA decomposition in the naïve combination. The fact our algorithm outperforms these algorithms shows that it is not just the inclusion of E that allows our algorithm to achieve superior performance, but rather, the fact that it exploits the low-rank structure as well, and does so in a better way than the naïve combined estimator. The results are shown in table 6.1.

Chapter 7

Future work

We list several possible directions for future research.

- First, while we proved consistency of the estimator for the PCA case, the FA case remains a challenge, and it would be interesting to pursue a consistency result for that case as well. If we were to pursue the same strategy employed in the consistency proof that we presented for the PCA case, then the only point where the proof would diverge is in proving claim 1. The reduced problem in this case is:

$$\begin{aligned} \max_D \quad & -\rho \cdot \operatorname{trace}(D - X) \\ \text{s.t.} \quad & D - X \succcurlyeq 0 \\ & D \succcurlyeq 0 \end{aligned}$$

  And one must show that this problem recovers the correct D.

- We have proven that when ρ is larger than some computable threshold value, the rank of L in the obtained solution is 0. This fits well with the intended purpose of the trace penalty, which is to encourage the solution to have low rank, and the paramter ρ controls how severe this penalty is. It would be desirable to quantify how ρ affects the rank of the obtained solution more thoroughly than the criterion that we gave. That is, obtain additional thresholds for ρ that would allow one to guarantee the rank of L is at most i, for any desired rank i.

- The behavior of the ADMM-based algorithm we developed depends on the step size parameter β. While ADMM in general is not extremely sensitive to this value [Boyd et al., 2011], it seems that our implementation, possibly owing to the fact that the

intermediate update of X is not full minimization, is somewhat more sensitive. It would therefore be good to have a criterion for selecting β.

- Our synthetic experiments demonstrate the superior performance of the algorithm. However, we have no theoretical results to justify this behavior. The preferred type of error bound would take into account some measure of the size of E. When no constraints are given, the performance is not expected to be significantly better than other estimators, but when extra information is available in the shape of a large constraint set, the bound should reflect that.

- Finally, it would be nice to find more data sets which show the applicability of the algorithm in real-world scenarios.

Bibliography

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

Bernd Bank. Non-linear parametric optimization. 1983.

Rajendra Bhatia. *Matrix analysis*, volume 169. Springer Science & Business Media, 2013.

Peter J Bickel and Elizaveta Levina. Regularized estimation of large covariance matrices. *The Annals of Statistics*, pages 199–227, 2008.

Peter J Bickel, Elizaveta Levina, et al. Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604, 2008.

Jacob Bien and Robert J Tibshirani. Sparse estimation of a covariance matrix. *Biometrika*, 98(4):807–820, 2011.

Martin Bilodeau and David Brenner. *Theory of multivariate statistics*. Springer Science & Business Media, 2008.

Jonathan Borwein and Adrian S Lewis. *Convex analysis and nonlinear optimization: theory and examples*. Springer Science & Business Media, 2010.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717, 2009.

Sanjay Chaudhuri, Mathias Drton, and Thomas S Richardson. Estimation of a covariance matrix with zeros. *Biometrika*, 94(1):199–216, 2007.

Yilun Chen, Ami Wiesel, Yonina C Eldar, and Alfred O Hero. Shrinkage algorithms for mmse covariance estimation. *IEEE Transactions on Signal Processing*, 58(10):5016–5029, 2010.

Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

Gordana Derado, F DuBois Bowman, and Clinton D Kilts. Modeling the spatial and temporal dependence in fmri data. *Biometrics*, 66(3):949–957, 2010.

Dipak K Dey and C Srinivasan. Estimation of a covariance matrix under stein's loss. *The Annals of Statistics*, pages 1581–1591, 1985.

David L Donoho et al. High-dimensional data analysis: The curses and blessings of dimensionality. *AMS Math Challenges Lecture*, 1:32, 2000.

Jianqing Fan, Yingying Fan, and Jinchi Lv. High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147(1):186–197, 2008.

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

Michael Grant, Stephen Boyd, and Yinyu Ye. Cvx: Matlab software for disciplined convex programming, 2008.

LR Haff. Empirical bayes estimation of the multivariate normal covariance matrix. *The Annals of Statistics*, pages 586–597, 1980.

Wolfgang Härdle and Léopold Simar. *Applied multivariate statistical analysis*, volume 22007. Springer, 2007.

Nicholas J Higham. Computing a nearest symmetric positive semidefinite matrix. *Linear algebra and its applications*, 103:103–118, 1988.

Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2nd edition, 2012.

Jianhua Z Huang, Naiping Liu, Mohsen Pourahmadi, and Linxu Liu. Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, 93(1):85–98, 2006.

William James and Charles Stein. Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 361–379, 1961.

Iain M Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Annals of statistics*, pages 295–327, 2001.

Ian T Jolliffe. Principal component analysis. In *Principal component analysis*, pages 115–128. Springer, 2nd edition, 2002.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Olivier Ledoit and Michael Wolf. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of empirical finance*, 10(5): 603–621, 2003.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411, 2004.

Elizaveta Levina, Adam Rothman, and Ji Zhu. Sparse estimation of large covariance matrices via a nested lasso penalty. *The Annals of Applied Statistics*, pages 245–263, 2008.

Fa-Hsuan Lin, Shang-Yueh Tsai, Ricardo Otazo, Arvind Caprihan, Lawrence L Wald, John W Belliveau, and Stefan Posse. Sensitivity-encoded (sense) proton echo-planar spectroscopic imaging (pepsi) in the human brain. *Magnetic resonance in medicine*, 57 (2):249–257, 2007.

Vladimir Alexandrovich Marchenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik*, 114(4):507–536, 1967.

Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, pages 1436–1462, 2006.

Adam J Rothman, Peter J Bickel, Elizaveta Levina, Ji Zhu, et al. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

Juliane Schäfer and Korbinian Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, 4(1), 2005.

Charles Stein. Estimation of a covariance matrix, rietz lecture. In *39th Annual Meeting IMS, Atlanta, GA, 1975*, 1975.

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

Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

Matthew A Turk and Alex P Pentland. Face recognition using eigenfaces. In *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR'91., IEEE Computer Society Conference on*, pages 586–591. IEEE, 1991.

Lingzhou Xue, Shiqian Ma, and Hui Zou. Positive-definite l1-penalized estimation of large covariance matrices. *Journal of the American Statistical Association*, 107(500): 1480–1491, 2012.