# Covariance estimation : The GLM and regularization perspectives

Leo Davy, Martin Gjorgjevski

December 16, 2021

**Abstract**

The goal of this report is to introduce the topic "Covariance estimation : the GLM and regularization perspectives"surveyed in [Pou11] by Mohsen Pourahmadi. We discuss the difficulties in covariance estimation in the high dimensional setting, where the number of samples is comparable or much smaller than the number of covariates. The sample covariance matrix is a poor estimator in this case. By introducing the Generalized Linear Model (GLM) framework, we see that it is possible to remove the notorious constraint of positive definiteness i.e. to have models for estimation of covariance matrices with unconstrained parameters. Several matrix decompositions are discussed, each with varying degree of statistical interpretability and parameter constraints. We then discuss regularization approaches (e.g. shrinkage, thresholding, penalization). We conclude with Bayesian approaches and its relationship with regularization techniques.

## Contents

# 1 Introduction, covariance estimation in a simple setting

## 1.1 Interest of covariance estimation

In Statistics, probably the most commonly investigated topic is *mean estimation*, or *regression analysis*. However, in many situations, knowing the mean is simply not enough and one is interested in a more precise description of the data, and in some situations the covariance structure is the most important[Car03]. Hence, we consider here the topic of covariance estimation.

Estimating accurately and efficiently a covariance matrix is an important topic in itself since, for instance, it is in the covariance matrix that one can naturally find informations such as dependence between variables. Also, in certain problems the interest of knowing the mean is very limited and one is interested in knowing *how* the data will change (e.g. time-series analysis[Box+15], climate forecasting[CBK21], drug testing[DCS88], ...). For instance, during a medical experiment on a new drug, or on a vaccine, a statistician is given a lot of measurements for a small number of people and has to be able to compute and compare the effect of one treatment, compared to another one or depending on the characteristics of the patient receiving the treatment.

More recently covariance estimation has gained a lot of interest with the development of areas such as *Machine Learning* or the *Big-Data* paradigm, where one has potentially a huge dataset, where each data sample is of very high-dimension[Pou13]. In such situations, covariance estimations turns out to be a real challenge for practitioners since having theoretical methods is not sufficient, they must also run as fast as possible. Another challenge with such datasets with a large size and varying content, is that it is unlikely that the whole dataset can be described using a simple (linear) distribution with just a few parameters. Hence, methods which need hypothesis on the distribution of the dataset could be very risky and lead to erroneous estimations. This brings the need to estimate the covariance matrix with as few hypothesis as possible and the smallest risk in prediction. Last but not least, another difficulty in such datasets is that some part of the data can be missing (e.g. for measures on a population through time as in medical surveys on a large number of people, some individuals or some of the data might not have been measured). Also, it is common for such datasets that some of the recorded data, instead of not having been measured, might have been wrongly measured. Hence, there is a strong need for these methods to be robust with respect to missing data and outliers.

## 1.2 Standard methods for covariance estimation

Let $Y = (Y_1, \cdots, Y_n)$ a sample of size $n$ from a mean zero normal population where each $Y_i \in \mathbb{R}^p$ with covariance matrix $\Sigma$. Then, it can be shown([And03] Theorem 3.2.1 or the first part of the course with Clément Marteau) that the maximum likelihood estimator of $\Sigma$ is given by

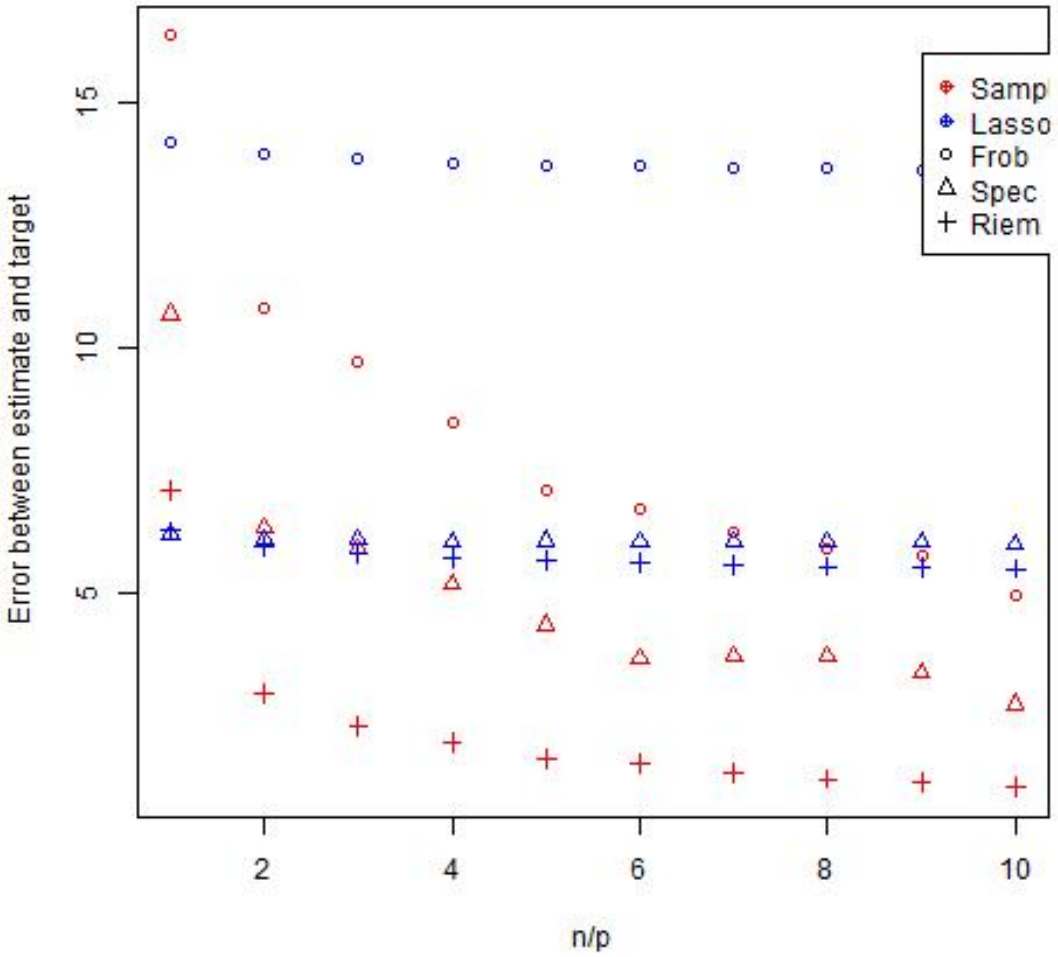
$$S_{ML} = \frac{1}{n}\sum_{i=1}^{n} Y_i Y_i^* = \frac{1}{n} Y Y^*$$

and is a positive-definite when there exists a set of $p$ affinely independent observations. This happens with probability 1 when $n \geq p$. $S_{ML}$ is also an unbiased estimator of $\Sigma$. In what follows we refer to $S = S_{ML}$ as a sample covariance matrix. This estimator is (strongly) consistent by the strong law of large numbers and performs well in the case when the number of features is low and we have a lot of data.

## 1.3 Numerical example

In this section we will numerically study the performance of covariance estimation using the sample covariance and the Lasso obtained covariance. In order to do so we will for several choices of $n$ and $p$ generate a fixed covariance matrix, then compute the estimators $S$ and $\hat{\Sigma}_{\text{Lasso}}$, which is introduced later. With these estimators we compute their average error using three different metrics on a number $N_{\text{sim}}$ of simulations.
The metrics used are the one given by the Frobenius norm (i.e. $||S-\Sigma||_F$ and $||\hat{\Sigma}_{\text{Lasso}}-\Sigma||_F$), the spectral norm (replace $\cdot_F$ by $\cdot_2$) then the distance on the Riemannian metric of covariance matrices. The reason for investigating the error using those three metrics is that they do not give us the same information (they induce some geometry on the space under investigation). The first one is the most uniform,

it is blind to the location of the coefficients in the matrix and since it squares them it will tend to penalize overestimates. The spectral norm is the standard operator norm. We also display results in a more intrinsic way specific to symmetric positive definite matrices[Smi05] using the *distcov* function imple-




mented in R[sha].

From this graph we can draw several conclusions.

- In the very low dimensional regime ($\frac{n}{p} >> 1$) the sample covariance converges to the estimate for any metric used (it is consistent)
- In the intermediate regime ($\frac{n}{p} \sim 1$) then the Lasso estimate outperforms the sample covariance estimate for any of the metric used.
- The Lasso estimate doesn't gain much from a large sample of data
- In low dimensions for a simple linear model the sample covariance is a good approximation of the covariance matrix

Three questions that we have not yet explored and are the main subject of discussion in the article are the following :

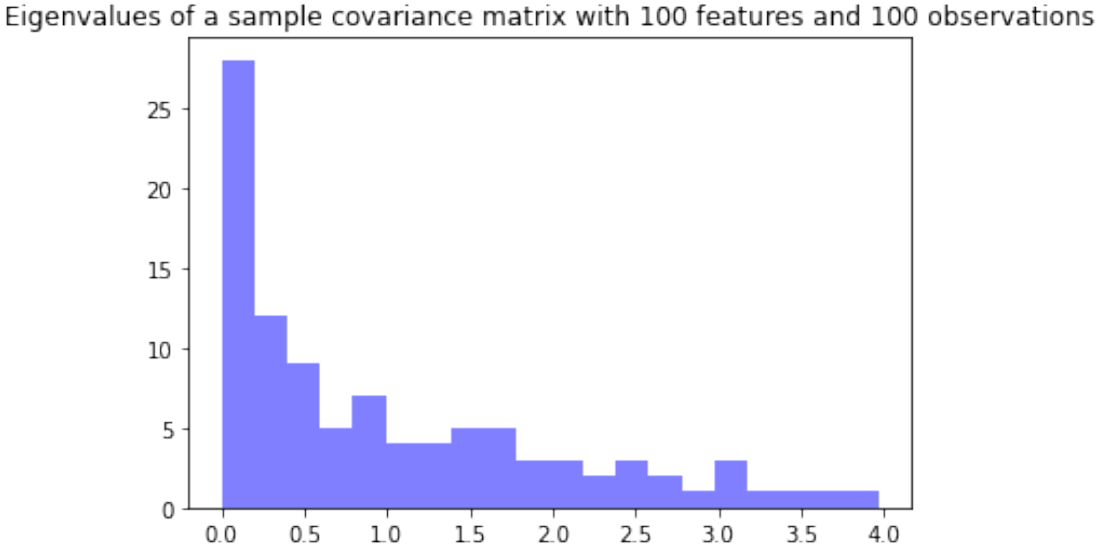
- Is it possible to perform covariance estimation in more general models (i.e. going further than the $Y = X\beta$ model and the Gaussian distribution assumption) ?
- Is it possible to perform covariance estimation in a data-based way (i.e. without assumption on the distribution) ?

- What happens in the high dimensional ($p > n$) regime ?

# 2 Problems in high dimensional settings

## 2.1 Marchenko Pastur law

Suppose $X_1, ..., X_n$ is sampled from a gaussian $N(0, \Sigma)$ law for $\Sigma$ a positive definite p by p matrix. The scatter matrix is a p by p positive definite matrix given by $S = \sum_{i=1}^{n} X_i X_i^*$. We say that $S$ follows a Wishart distribution denoted $W_p(\Sigma, n)$ where $p$ denotes the number of covariates and $n$ is called the degree of freedom. The Wishart distribuiton $W_p$ is a probability distribution over the set of positive semidefinite $p$ by $p$ matrices. We sample a matrix from $W_{100}(I, 100)$. A histogram of the eigenvalues of $S/n$ is shown on the figure below, as well as the extreme eigenvalues of $S/n$.




The largest and smallest eigenvalues in this simulation were $\lambda_{max} = 3.96969$ and $\lambda_{min} = 0.00019$. This spectra is far from the spectra of the identity matrix, i.e. we see that the largest eigenvalue is biased upwards and the smallest is biased downwards. Questions about the distribution and limiting behavior of large symmetric matrices are the starting point of the field of random matrix theory. The first result is the so called Wigner semicircle law, which states that the limiting distribution of the eigenvalues of Wigner matrices is the Wigner semicircular distribution. An analog for covariance matrices is a result due to Marchenko and Pastur[MP67], which states that if $p/n \to \gamma$ then the limiting distribution $G$ of the eigenvalues of $S/n$ has density

$$g(t) = \frac{\gamma}{2t\pi}\sqrt{(b-t)(t-a)}1_{\{t\in(a,b)\}}$$

with $a = (1-\gamma^{-1/2})^2$, $b = (1+\gamma^{-1/2})^2$ (and a point mass at 0 $G(0) = (1-\frac{1}{\gamma})_+$). In the case when $n = p$, this gives that we expect to see that the largest eigenvalue around 4 and the smallest eigenvalue around 0, which explains our simulation results. The technique of correcting for this bias is known as shrinkage and we will explore it in the following chapters.

## 2.2 Precision matrices and Gaussian Graphical models

Often we are interested in the inverse of covariance matrix, which is called precision matrix. One example comes from undirected graphical models that are used to model conditional independence. In such models, we associate a graph to a joint law, with vertices being variables and edges are omitted if two variables are independent conditionally on all the other variables. In the Gaussian case, for example, the precision matrix appears in the density and it is easy to see that two variables $X_i$ and $X_j$ will be conditionally independent (given the other variables) if and only if the precision matrix $\Theta = \Sigma^{-1}$ has 0 in the $(i, j)$-th entry. Important problem in graphical models is neighbourhood selection, i.e.

edge estimation. An easy computation, shows that if $X$ is Gaussian vector with covariance matrix $\Sigma$ then the least square error coefficients are given by the the precision matrix $\Theta$, that is

$$X_k = -\sum_{j \neq k} \frac{\theta_{kj}}{\theta jj} X_j + \epsilon_k$$

with $\epsilon_k$ Gaussian and independent of $\{X_i : i \neq k\}$. Here, we are more interested in whether the entry $\theta_{ij} = 0$ or not rather than the actual value of the precision matrix. Meinhausen and Buhlman [MB06] proposed the lasso estimator of the form

$$\hat{\theta}_a = argmin_{\{\theta:\theta_a=0\}}(||X_a - X\theta||_2^2 + \lambda||\theta||_1)$$

in order to predict the neighbourhood of $a$, where $\lambda$ is a tuning parameter which determines the sparsity of the obtained estimator.
While in theory we can recover $\Sigma^{-1}$ from $\Sigma$ in $O(p^3)$ operations, this is computationally challenging for large values of p. Some problems arise when we try to estimate the precision matrix by taking an inverse of the estimated covariance matrix, because the latter is not invertible in the high dimensional regime. Indeed, writing $X$ for the $p$ by $n$ matrix which has the sample $X_i$ in the $i$-th column, the sample covariance matrix can be written as $\frac{1}{n}XX^*$ which is not invertible, because it has rank at most $n < p$. Therefore there is a need for estimators that are guaranteed to be positive definite (in particular invertible). An example of such an estimator is the penalized likelihood estimator [MB06],[FHT07]:

$$\hat{\Theta} = argmax_{\{\Theta \succ 0\}} \log\det\Theta - Tr(S\Theta) - \lambda||\Theta||_1.$$

as shown by [BGd07] such an estimate is always invertible when $\lambda > 0$. We denote this covariance estimate with $\hat{\Sigma}_{lasso} = (\hat{\Theta})^{-1}$.

### 2.3 Empirical Bayes estimators

Another sense in which the sample covariance matrix is not optimal comes from decision theory. The Empirical Bayes Estimator, which is of the form

$$\hat{\Sigma} = \frac{n}{p}(S + ut(nu)C)$$

$u = \frac{1}{tr(S^{-1}C)}$, $0 \leq t$ is a decreasing, absolutely continuous function and $C$ is a known positive definite matrix. It can be shown that such estimators outperform the sample covariance in the sense that they have smaller risk $R_1$ associated to the loss function

$$L_1(\hat{\Sigma}, \Sigma) = tr(\hat{\Sigma}\Sigma^{-1}) - \log\det(\hat{\Sigma}\Sigma^{-1})$$

(also known as Kullback Liebler loss). This is true for other loss functions as well, with different parameters [Haf80]

## 3 Generalized Linear Models

### 3.1 GLM for mean estimation

The goal of this section is to introduce Generalized Linear Models (GLM). Let us recall the Linear model.

$$Y = X^*\beta + \varepsilon$$

where $Y$ is the measured data, $X$ is a (design) matrix of predictors (also known as features, covariates or explanatory variables), $\beta$ are (unknown) parameters and $\epsilon$ (the response error) is a centered Gaussian variable with covariance matrix $\sigma I$, independent of the data. The conditional expectation of Y given X is

$$\mathbb{E}(Y|X) = X^*\beta = \mu.$$

Generalized Linear Models are a way to extend this model by introducing a link function $g$ such that we have:

$$g(\mathbb{E}(Y|X)) = X^*\beta = g(\mu).$$

Having this link function allows to consider models that wouldn't even make sense in linear regression like in the case where the responses have a finite or discrete support. A very well known method for prediction of 0-1 valued variables is the logistic regression model. In it, we assume

$$E(Y|X) = \frac{e^{X^*\beta}}{1 + e^{X^*\beta}}$$

The link function g in this case is the logit function $g(p) = \log(\frac{p}{1-p})$ for $0 < p < 1$. The main idea is that with the transformation $g^{-1}$ the output $g^{-1}(X^*\beta)$ is guaranteed to be in the interval $(0, 1)$, which is a requirement for a variable taking values in $\{0, 1\}$.

**Definition 1.** *A family of distributions $\{P_\theta : \theta \in \Theta\}$, $\Theta \subset \mathbb{R}^k$ is said to be an exponential family if there exists real valued functions :*

- *$\eta$ and $B$ depending on $\theta$*
- *$T$ and $h$ depending on $x$*

*such that the pdf of $P_\theta$ is*

$$p_\theta(x) = \exp(\eta(\theta)T(x) - B(\theta))h(x).$$

Examples of distributions belonging to an exponential family are Bernoulli, Binomial (with fixed $n$), Poisson, Gaussian, Gamma, $\chi^2$, inverse Wishart ...

An exponential family can also be written in canonical form as

$$f_\theta(x) = \exp(\frac{x^*\theta - b(\theta)}{\phi} + c(x, \phi))$$

for some known functions $b(\cdot)$ and $c(\cdot)$, and some parameter $\phi \in \mathbb{R}$ called dispersion parameter. From this definition one can derive simple formulas for the mean and variance :

$$\mathbb{E}(Y) = \mu = b'(\theta), \qquad \mathbb{V}(Y) = b''(\theta)\phi.$$

**Definition 2.** *A GLM $M(\Theta, g)$ assumes that the response $Y$ is distributed according to an exponential family indexed by $\Theta$, and that there is a link function $g$ such that*

$$\mathbb{E}(Y|X) := \mu = g^{-1}(X^*\beta).$$

*where $X$ is a vector of covariates.*

In this framework estimating the mean is equivalent to estimating the distribution of $Y$, since the distribution is essentially parameterized by the mean. An advantage of exponential family distributions in GLM is that the log-likelihood $l(\theta)$ is relatively easy to compute (under suitable assumptions) and is strictly concave whenever $\phi > 0$[1]. Maximum likelihood estimation is a well-posed problem and optimization techniques such as Newton-Raphson, Fischer-Scoring and Iteratively reweighted least squares can be introduced in order to recover the parameter $\beta$.Hence, the problem of mean estimation for GLM seems to be solved and the article of Mohsen Pourahmadi surveys the situation of (co)variance estimation in GLM.

## 3.2 Linear and Log-Linear Covariance models

There are various ways of modeling covariance matrices. We discuss different parameterizations of the space of positive definite matrices with varying statistical interpretability and constraints.

[1] This can be seen from the identity $\mathbb{V}(Y) = b''(\theta)\phi$, since $\mathbb{V} > 0$ so is $b''$ and then the second derivative of $\log \circ f_\theta$ is strictly negative.

**Linear Covariance Model** To explain the appeal of unconstrained parameterization, we first consider the Linear covariance model. Let $U_1, ..., U_q$ be positive definite p by p matrices. The Linear covariance model is parameterized by

$$\Theta = \{(u_1, ..., u_q) : U_0 + u_1 U_1 + ...u_q U_q \text{ is positive definite } \}$$

This parameter space is highly constrained. Usually one assumes that the matrices are linearly independent, and that $U_0$ is orthogonal to the other matrices with respect to the Hilbert-Schmidt product. The assumption is that we are given a sample vectors with covariance matrix $\Sigma$ and that $\Sigma$ belongs in the linear span of the matrices $U_i$, $1 \leq i \leq q$. The goal is to estimate the coefficients in the expansion of $\Sigma$ along $U_i$. The model is quite general, for example if we let $E_{i,j}$ be the matrix whose $(i,j)$-th entry is 1 and all others are zero then $U_0 = I_p$, $U_{i,j} = E_{i,j} + E_{j,i}$ models all covariance matrices. If we make the assumption that the sample vectors come from a Gaussian distribution then the corresponding model is called a linear Gaussian covariance model. We denote by $\Theta(U_0, ..U_q)$ the set of matrices corresponding to the parameters in $\Theta$. It was shown that the MLE estimator for the parameters can be computed explicitly in terms of the entries of the sample covariance matrix if and only if $\Sigma \in \Theta(U_0, ..., U_q)$ implies $\Sigma^2 \in \Theta(U_0, ..., U_q)$ [Sza80]. It is also known that the likelihood is convex function with high probability. More precisely, it was shown by Zwiernik, Uhler and Richards [ZUR16] that the random convex set

$$\Delta_{2S_n} = \{v \in \Theta | 0 \prec \Sigma_v \prec 2S_n\}$$

contains the global MLE, the least squares estimator, and the true data generating parameter with high probability (here $A \prec B$ means that $B - A$ is positive definite)

**Log-Linear Covariance Models** The idea here is to use link function on the set of covariance matrices, instead of the mean as in the classical GLM approach. This model is intimately related with the spectral decomposition. Since the space of matrices is a Banach space, it is easy to define the exponential of a matrix via its Taylor series. We note that if P is orthogonal, K arbitrary, then $(P^*KP)^m = P^*K^mP$ and so we get the identity $exp(P^*KP) = P^*exp(K)P$. Using the spectral decomposition, $\Sigma = P^*KP$ with $P$ orthogonal and $K$ diagonal, we see that we can define the logarithm of $\Sigma$ as $\log(\Sigma) = P^* \log(K)P$, where $\log(K)$ is the diagonal matrix with $(\log(K))_{i,i} = \log((K)_{i,i})$.This argument essentially shows that the exponential map maps the symmetric matrices into positive definite matrices. The benefit of doing this is that we have removed the constraint on the parameters on our model, i.e. now we can assume that we model $A = \log(\Sigma)$ instead of $\Sigma$, with

$$\log(\Sigma) = U_0 + v_1 U_1 + .... + v_r U_r$$

with unconstrained parameters $v_1, ...v_r$.

## 3.3 Matrix decompositions

Three different matrix decompositions will be investigated in this section, namely the variance-correlation, spectral and Cholesky decompositions.

### 3.3.1 Variance-correlation

The *variance-correlation* decomposition

$$\Sigma = DRD$$

where $D$ is the diagonal matrix of standard deviations and $R$ is the correlation matrix. This decomposition has a strong appeal since the coefficients are readily statistically interpretable. Indeed in order to know the self variation of each coefficient of the response it is enough to look into $D$ and for the interactions between the variables it is enough to look at the coefficients in $R$.
However it is not a satisfactory one for the GLM since its entries are constrained. More precisely, although the logarithm of the matrix of standard deviation is unconstrained (the coefficients can vary freely in $\mathbb{R}$); the matrix $R$ still has to be positive definite (and the diagonal entries equal to 1).
Although it is not an appropriate decomposition for the GLM it can sometimes be a convenient decomposition. Such a case is given by the Gaussian graphical models where one is interested in finding

conditional dependencies of the variables. Those can be found in the off-diagonal coefficients of the precision matrix $\Sigma^{-1}$. Applying the variance-correlation decomposition to the precision matrix results in

$$\Sigma^{-1} = \tilde{D}\tilde{R}\tilde{D}.$$

It is shown in the paper that $(\tilde{R}, \tilde{D})$ also have direct statistical interpretations where $\tilde{R}$ contains partial correlations (for pairs of variables after removing the effects of every other variable) and $\tilde{D}$ contains the variance of predicting a variable given the rest.

### 3.3.2 Spectral

The *spectral decomposition*

$$\Sigma = P\Lambda P^* = \sum_{i=1}^{p} \lambda_i e_i e_i^*$$

is standard matrix decompositions and is naturally directly statistically interpretable in terms of principal components. The entries of $\Lambda$ corresponds to the variance of the principal components whereas the column vectors of $P$ give the principal components themselves.
However this matrix decomposition is not satisfactory for our purpose in the GLM since the matrix $P$ has a strong constraint (orthogonality). Despite this constraint the spectral decomposition can still be of use by applying the matrix logarithm to the spectral decomposition

$$\log \Sigma = P \log \Lambda P^*.$$

Under this decomposition, $\log \Sigma$ is a symmetric matrix with unconstrained entries. The issue that arises comes from the fact that the logarithm is a very non-linear transformation on the entries of $\Sigma$ and this makes the statistical interpretation very difficult as we have seen from the discussion of the log-linear covariance model.

### 3.3.3 Cholesky, Regression based proof

We give a very instructive proof of *Cholesky's decomposition* theorem based on the idea of regression. We begin with a useful lemma.

**lemma** If $A$ is a p by p lower triangular matrix with 1's on the main diagonal, then so is $A^{-1}$.

*Proof.* Write $A = I - L$, with $L$ lower triangular and 0's on the main diagonal. Clearly $L$ is nilpotent, that is $L^p = 0$. One easily checks that $(I - L)(I + L + ... + L^{p-1}) = I - L^p = I$. Each of the matrices $L^k$ for $k = 1, ..., p-1$ is lower triangular with 0's on the main diagonal. It follows that $A^{-1} = I + L + ... + L^{p-1}$ has the desired form. □

**Theorem (Variant of Cholesky decomposition)** Let $\Sigma$ be positive definite p by p matrix. Then there exist unique matricies $L$ and $D$ such that $L$ is lower triangular with 1's on the main diagonal, and $D$ is diagonal with positive entries such that $\Sigma = LD^2L^*$.

*Proof.* Consider a p-dimensional Gaussian vector $X$ with mean 0 and covariance matrix $\Sigma$. We want coefficients $c_{t,j}$, $j = 1, 2, .., t-1$ such that

$$E((X_t - \sum_{j=1}^{t-1} c_{t,j} X_j) X_i) = 0$$

for $i = 1, 2, ..., t-1$.Note that such scalars exist and are unique, since essentially this is equivalent to finding an orthogonal projection of $X_t$ onto the space spanned by $X_1, ..., X_{t-1}$. Equivalently, we want

$$\sum_{j=1}^{t-1} c_{t,j} E X_i X_j = E X_t X_i$$

This system of equations can be written in matrix form as $Ac = b$ with $A = (EX_iX_j)_{i,j=1,..t-1}$, $c = (c_{t,1}, ..., c_{t,t-1})^*$, $b = (EX_tX_1, ..., EX_tX_{t-1})^*$. Also, we can easily compute the variance of $\epsilon_t$:

$$E\epsilon_t^2 = E(\epsilon_t X_t) = EX_t^2 - \sum_{j=1}^{t-1} c_{t,j} EX_tX_j$$

By construction, $\epsilon_t$ is independent of $X_1, ...X_{t_1}$ and therefore of $\epsilon_{t-1}$, so that the vector $\epsilon = (\epsilon_1, ...\epsilon_p)^*$ has diagonal covariance matrix. On the other hand, we have the system of equations

$$X_t - \sum_{j=1}^{t-1} c_{t,j} X_j = \epsilon_t$$

which can be written in matrix form as

$$LX = \epsilon$$

with $L$ a lower triangular matrix with 1's on the main diagonal. Taking covariances gives:

$$L\Sigma L^* = Cov(LX) = Cov(\epsilon) = D^2$$

Finally, using the lemma we conclude. Uniqueness easily follows from our construction. □

This proof is interesting in its own right, but also shows how we can model covariance matrices with unconstrained and statistically interpretable parameters. The lower triangular matrix $L$ besides the diagonal constraint is completely unconstrained i.e. the subdiagonal entries can be chosen freely. For the diagonal matrix $D$, the entries can be thought of as the least square error of regressing $X_t$ on its predccessors (in fact it can easily be shown that our construction gives $\epsilon_t$ = least square error of approximating $X_t$ by its predecessors). The parameters in $L$ are called autoregressive parameters, while the parameters in $D$ are called inovation variances. The idea is that these parameters can be estimated from data, and the estimated matrices $\hat{L}$ and $\hat{D}$ will be such that $\hat{L}(\hat{D})^2\hat{L}^*$ will be positive definite[CT10b]; [Pou99].
Having obtained this decomposition from a regression perspective it is natural to employ the classical regression machinery to perform this covariance estimation.
However a drawback of this decomposition is that we have to decide of an order on the variables $Y_t$ in order to make this analysis, so, even though it is adapted for many situations (e.g. temporal/longitudinal data) it might sometimes be inappropriate (e.g. categorical data).

# 4 Regularization

## 4.1 Difficulties of the high-dimensional setting

We've seen that in the high dimensional setting $p > n$, the sample covariance matrix is a poor estimate for the covariance matrix because it fails to be invertible (i.e. it's only positive semidefinite), it has bias on the extreme eigenvalues (due to Marchenko-Pastur's law). In model selection for the linear model, we have seen that adding a penalty reduces overfitting (i.e. helps us distinguish between relevant and irrelevant features). Adding the penalty to the loss function forces us to take model complexity into account. There are various choices of penalties which have their benefits and drawbacks. The similar idea of regularization has been studied for the covariance estimation problem. The estimators we consider here are of the form:

$$argmin_{\hat{\Sigma}} DF(\hat{\Sigma}, \Sigma) + pen(\hat{\Sigma})$$

where $DF$ is a data-fidelity (or loss function) term (Kullback Liebler distance $L_1(\hat{\Sigma}, \Sigma) = tr(\hat{\Sigma}\Sigma^{-1}) - \log\det(\hat{\Sigma}\Sigma^{-1})$ or Frobenius norm $L_2(\hat{\Sigma}, \Sigma) = tr(\hat{\Sigma}\Sigma^{-1} - I)^2$ ) and $pen$ is a penalty function (e.g. $\ell^1$ penalty on a subset of coefficients).

We discuss shrinkage estimators which are meant to reduce the bias of the eigenvalues of the sample covariance matrix.

In the same spirit as in signal processing or compressed sensing à la Donoho or Candès-Tao a natural constraint to consider sparsity. However, since a sparsity constraint is not convex[2] in order to compute estimates it will be turned in a $\ell^1$ penalty resulting in LASSO type estimators. It is known that the $\ell^1$ norm promotes sparsity. The approach in which we impose penalties on the coefficients of the (maximum likelihood) estimators is called the *penalized likelihood approach*.

Another way to add constraints is by having strong priors on the shape of the covariance matrix. For instance if we have ordered samples which are known not to correlate if they are far apart, this results in a covariance matrix with zero coefficients away from the diagonal. A natural way for an estimator to apply this constraint is *banding*. Similar techniques that will be discussed are *tapering* and *thresholding*.

## 4.2 Shrinkage

The first improvements for covariance matrix estimation were proposed by Stein (1975) from the (natural but not desired) property that the largest (resp. smallest) eigenvalue of the sample covariance matrix is larger (resp. smaller) than their true value[3]. In order to counter this phenomenon he considered estimators that left the eigenvectors of $S$ unchanged acting only on the eigenvalues. Hence estimators of the form:

$$\hat{\Sigma}(S) = P\Phi(\lambda)P^*$$

where $P$ is the orthogonal matrix of eigenvectors of $S$, $\lambda = (\lambda_1, \cdots, \lambda_n)$ the ordered eigenvalues of $S$ and $\Phi(\lambda) = diag(\varphi_1(\lambda), \cdots, \varphi_p(\lambda))$.

For instance under the loss function

$$L_1(\hat{\Sigma}, \Sigma) = Tr(\hat{\Sigma}\Sigma^{-1}) - \log|\hat{\Sigma}\Sigma^{-1}| - p$$

the corresponding choices of $\varphi_i$ are

$$\varphi_i = \frac{n\lambda_i}{\alpha_i}$$

where

$$\alpha_i(\lambda) = n - p + 1 + 2\lambda_i \sum_{i \neq j} \frac{1}{\lambda_i - \lambda_j}$$

as proved in Lin Perlman (1984). This estimator is called Stein-Haff estimator. We have

$$\frac{\lambda_j}{\varphi_j} = 1 - \frac{p}{n} + \frac{1}{n} + 2\sum_{i \neq j} \frac{\lambda_j}{\lambda_j - \lambda_i}$$

From here we can see that if $p/n$ is large or if there are eigenvalues that are near to $\lambda_j$, $\varphi_j$ and $\lambda_j$ will be quite different. It is also possible to get negative values for $\varphi_i$ so that the estimated matrix is not positive definite.

Another direction of improvement has been proposed by Ledoit and Wolf [LW04] in order to obtain an estimator that is simultaneously well conditioned and more accurate than the sample covariance matrix. In this article they investigate estimators of the form

$$\rho_1 I + \rho_2 S$$

and they minimize quadratic risk by an appropriate choice of coefficients $\rho_1$ and $\rho_2$. They prove that there exists an optimal choice of those coefficients depending only on four parameters that depend on the true covariance matrix. This strong reduction on the number of parameters matters in practice since it reduces the number of assumptions on the data distribution. They also give consistent estimators of those parameters without dependence on the covariance matrix.

The estimator they obtain in this way is consistent, more accurate and has smaller condition number[4] than the sample covariance matrix. They also establish superiority of their estimator compared to Stein-Haff, Empirical Bayesian and minimax estimators.

[2] In the sense that a $||\cdot||_0$ penalty is not convex.
[3] Detailed in Ledoit and Wolf [LW04]
[4] ratio of largest to smallest eigenvalue

## 4.3 Penalization

A more general approach is that of *penalized likelihood* where the loss function is $-l$ the opposite of the log-likelihood and various choices can be done for the penalty function. For instance it has been shown by Warton [War08] that the approach of Ledoit and Wolf [LW04] amounts to a penalized likelihood with a penalty on the diagonal entries of the precision matrix $\hat{\Sigma}^{-1}$.

Another approach pursued in [Hua+06] is to penalize the coefficients of the Cholesky factors (which contains LASSO as a special case). An advantage of these methods is that under suitable choices they may allow to preserve (or induce) sparsity in the resulting matrix[5] by applying an $\ell^1$ penalty (as in [BGd07]; [YL07]; [FHT07]).

Several algorithms have been developed to compute those estimators, the fastest being the graphical lasso (glasso) of Friedman, Hastie and Tibshirani [FHT07]. An important feature of this algorithm, as shown by Banerjee, El Ghaoui and d'Aspremont[BGd07], is that the estimate is guaranteed to be positive definite.

Other algorithms discussed are the Sparse Pseudo-Likelihood Inverse Covariance Estimation (SPLICE) algorithm of Rocha, Zhao and Yu [RZY08] and the SPACE (Sparse PArtial Correlation Estimation) algorithm of Peng et al. [Pen+09]. Those algorithms directly impose sparsity constraints on the precision matrix. In order to do that they use a reparameterization of the correlation matrix (as in the previous section) which they solve using a least squares estimation. It is also noted that this differs from the approach of Meinshausen and Buhlmann [MB06] where $p$ linear regression problems are solved, here they are all solved simultaneously.

An important notion for these algorithms is the *sparsistency*. Indeed, instead of trying to estimate accurately the whole precision matrix, one is more interested in identifying which coefficients are of importance. Hence, sparsistency refers to the property that all zero entries are identified as zero asymptotically.

## 4.4 Elementwise shrinkage

The last estimators we will discuss in these sections are of a different nature than the previous ones as they will act elementwise on the sample covariance matrix. A clear advantage of those is that the number of computations is minimal, however there are two major drawbacks. The first one is that in order to produce the right estimate they need to be properly tuned. Fine tuning is performed through cross validation which is computationally expensive. Another drawback, which can create serious difficulties (e.g. for inversion) is that since positive-definiteness is a global property of the matrix, acting elementwise on the sample covariance matrix may result in a non-positive-definite matrix.

The first of those techniques we consider is banding where the estimator is the following:

$$B_k(S) = (S_{i,j}\mathbb{1}_{\{|i-j|<k\}})_{i,j}$$

which keeps only the coefficients less than $k$ away from the diagonal. This estimator is clearly ideal when $B_k(\Sigma) = \Sigma$ which can be realistic in many scenarios. An instance of such a case could be that of moving process where the innovations depends on the past for times at most $k$. More formally, if $y_1, \cdots, y_n$ is an inhomogeneous moving average process:

$$y_t = \sum_{j=t-k-1}^{t-1} \theta_{t,j}\varepsilon_j$$

where the $\varepsilon_j$ are i.i.d. with mean zero and finite variances.

Another assumption we might have on the data could that the covariance matrix has some shape $R$ (e.g. with 0-1 entries) which is not necessarily concentrated around the diagonal. This leads to the idea of tapering the sample covariance matrix with $R$, i.e. replacing $S$ the sample covariance matrix by:

$$S \times R = (S_{i,j}R_{i,j})_{i,j}$$

where $\times$ denotes the Schur (or elementwise) product of matrices. An important feature of the Schur product is that the Schur product of two positive definite matrices is positive definite.

[5] Usually the precision matrix as in (sparse) Gaussian Graphical models

We can then note that banding corresponds to tapering by the matrix $R_k = (\mathbb{1}\{|i-j| < k\})_{i,j}$ and also that the previously mentioned property of Schur products doesn't apply to banding since $R_k$ is not positive definite whenever $k \geq 2$.

*Thresholding* is the last elementwise operator discussed. In practice it can be often observed (e.g. in the sample covariance matrix of a sparse matrix) that although most coefficients are not zero, many of the coefficients that are wrongly estimated as non-zero are still small with respect to estimated coefficients of non zero true coefficients. Hence a simple operation that can be applied consists of thresholding by setting coefficients less than a specified coefficient to zero and keeping the others untouched, i.e.:

$$T_\alpha(S) = (S_{i,j}\mathbb{1}_{|S_{i,j}|\geq\alpha})_{i,j}.$$

As in banding, a clear advantage is the simplicity, both in implementation and computationally, but drawbacks are in keeping positive-definiteness and estimating the appropriate coefficient $\alpha$.

However, although difficulties aforementioned appear when using banding, tampering and thresholding, a good thing is that it has been shown by Bickel and Levina [BL08b]; [BL08a] for banding and thresholding that they produce consistent estimators of the covariance matrix for certain classes of matrices[6] provided $\log p/n \to 0$. For

# 5 A bayesian approach

## 5.1 Priors on the spectral decomposition

We have seen that the spectral decomposition allows for unconstrained parametrization. Recall that if $\Sigma$ is a covariance matrix, with spectral decomposition $PDP^*$ then $A = \log\Sigma = P\log DP^*$ is a symmetric matrix such that $e^A = \Sigma$. Recall also that exp takes the symmetric matrices into positive definitive matrices in a one-to-one and onto manner. The idea is that the entries of the matrix $A$ on and above the main diagonal are completely unconstrained, and we can stack them into a vector of length $q = p(p+1)/2$, which we denote with $vec(A)$. Now we can consider various distributions on $\mathbb{R}^q$ which will induce a distribution on the space of symmetric p by p matrices. This is the prior distribution in the Bayesian framework. One possibility for the prior is to assume that the stacked vector has normal distribution with specified mean and covariance matrix. An interesting property of this choice of prior is that it is rotationally invariant. Indeed, if $X_1, ...X_n$ are sampled from a distribution with covariance matrix $\Sigma$ and $G$ is orthogonal then $GX_1, ...GX_n$ have covariance matrix $G\Sigma G^*$, and as discussed in the section spectral decomposition, this will give that the matricial logarithm of the covariance matrix of the rotated variables is $GAG^*$. Now this is linear function of $vec(A)$. Hence assuming $vec(A)$ is Gaussian gives that also the vector $vec(GAG^*)$ is Gaussian, being a linear function of $vec(A)$. The log-likelihood in this case takes the form (up to a constant)

$$-\frac{1}{2}(tr(A) + tr(S\exp(-A)))$$

This is easy to derive from the likelihood of a multivariate normal distribution using properties of the exponential map such as $\log\det\Sigma = trA$ and $\exp(-A) = (\exp A)^{-1}$. However this is not so easy to maximize, because the exponential map is quite complicated to understand. For example, in general it is not true that $e^Ae^B = e^{A+B}$. We know that this likelihood is uniquely maximized for $A = \log S$ from the MLE for multivariate normal and the fact that we've essentially constructed a link function between unconstrained and constrained parameters. The difficulty lies in recovering $A$ from $S$, due to the complicated nature of the exponential map. A second order approximation is proposed in [LH92]. This model lacks statistical interpretability and is rarely used in practice.

## 5.2 Priors on Correlation matrices

Recall that the variance-correlation decomposition of a positive definite matrix $\Sigma$ is given by $(S, R)$ where $\Sigma = SRS$ with $S$ diagonal matrix with positive entries and $R$ is a correlation matrix. Since $S$ is diagonal, it is natural to consider a Gaussian prior $p(S)$ on the entries of $\log S$. This prior is of the form $N(\eta, \Lambda)$ where $\Lambda$ is diagonal matrix. The prior on the pair $(S, R)$ is then given in the form of

[6] Respectively, the matrices have to be "bandable" and satisfy an appropriate notion of sparsity (see [BL08b]; [BL08a])

$p(S,R) = p(S)p(R|S)$. Modeling the conditional prior $p(R|S)$ is more challenging. An approach taken in [BMM00] is to assume that that $S$ and $R$ are independent, which amounts to modeling separately the priors of $S$ and $R$. The entries of the correlation matrix $R$ are supported on $[-1,1]$ and therefore one possibility is to assume uniform prior, i.e. $p(R) \propto 1$. Such a prior however, does not result in a uniform prior on the marginals $r_{ij}$, due to the shape of the space of correlation matrices. An alternative is to put a prior such that the marginals $r_{ij}$ are uniformly distributed on $[-1,1]$. This is accomplished by assuming that $\Sigma$ has standard Inverse Wishart (IW) distribution $W_p^{-1}(I,v)$ with $v \geq p$, which has density

$$f_p(\Sigma|v) \propto det(\Sigma)^{-\frac{1}{2}(v+p+1)} \exp(-\frac{1}{2} tr(\Sigma^{-1})$$

It can be shown that in that case $R$ has distribution

$$f_p(R|v) \propto (detR)^{\frac{1}{2}(v-1)(p-1)-1} \prod_{i=1}^{p} (detR_{ii})^{-v/2}$$

where $R_{ii}$ is a principal submatrix of $R$ obtained by deleting the $i$-th row and the $i$-th column. Moreover, the marginals $r_{ij}$ have the following distribution

$$f_p(r_{ij}|v) = (1-\rho_{ij}^2)^{(v-p-1)/2}$$

for $|\rho_{ij}| \leq 1$. This is the $Beta((v-p+1)/2, (v-p+1)/2)$ distribution which is uniform when $v = p+1$.

# 6 Conclusion

The author of the survey "Covariance estimation: The GLM and regularization perspectives", Mohsen Pourahmadi has summarized the progress and development in covariance estimation in the last several decades. The intent of that survey is to point out the literature of different approaches on the problem for new and aspiring researchers. It has been a challenging task for us to balance between selecting results from the massive amount of work done on this subject, being technically precise and presenting the main ideas clearly. We have seen that in the high dimensional regime the natural idea of estimating the covariance matrix by the sample covariance matrix does not really work and should be avoided. We've also seen that by using standard matrix decompositions, one can model covariance matrices with unconstrained parameters, which allow for Bayesian point of view by putting priors on those parameters. We have also seen that regularization techniques such as shrinkage, likelihood penalization and even elementwise operations such as banding, tapering and thresholding can be used for estimation in the case when the ratio data to features $(n/p)$ is moderate or large.

# References


[And03] T.W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley Series in Probability and Statistics. Wiley, 2003. ISBN: 9780471360919.

[BGd07] Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. *Model Selection Through Sparse Maximum Likelihood Estimation*. 2007. arXiv: 0707.0704 [cs.AI].

[BL08a] Peter J. Bickel and Elizaveta Levina. "Covariance regularization by thresholding". In: *The Annals of Statistics* 36.6 (2008), pp. 2577–2604. DOI: 10.1214/08-AOS600.

[BL08b] Peter J. Bickel and Elizaveta Levina. "Regularized estimation of large covariance matrices". In: *The Annals of Statistics* 36.1 (2008), pp. 199–227. URL: https://doi.org/10.1214/009053607000000758.

[BMM00] John Barnard, Robert McCulloch, and Xiao-Li Meng. "Modeling covariance in terms of standard deviations and correlation, with application to shrinkage". In: *Statistica Sinica* 10.4 (2000), pp. 1281–1311. ISSN: 10170405, 19968507. URL: http://www.jstor.org/stable/24306780.

[Box+15] G.E.P. Box et al. *Time Series Analysis: Forecasting and Control*. Wiley Series in Probability and Statistics. Wiley, 2015. ISBN: 9781118674925. URL: https://books.google.fr/books?id=rNt5CgAAQBAJ.

[Car03] Raymond J. Carroll. "Variances Are Not Always Nuisance Parameters". In: *Biometrics* 59.2 (2003), pp. 211–220. ISSN: 0006341X, 15410420. URL: http://www.jstor.org/stable/3695498.

[CBK21] Diego S. Carrió, Craig H. Bishop, and Shunji Kotsuki. "Empirical determination of the covariance of forecast errors: An empirical justification and reformulation of hybrid covariance models". In: *Quarterly Journal of the Royal Meteorological Society* 147.736 (2021), pp. 2033–2052. DOI: https://doi.org/10.1002/qj.4008.

[CT10a] Changgee Chang and Ruey Tsay. "Estimation of covariance matrix via the sparse Cholesky factor with lasso". In: *Journal of Statistical Planning and Inference* 140 (Dec. 2010), pp. 3858–3873. DOI: 10.1016/j.jspi.2010.04.048.

[CT10b] Changgee Chang and Ruey S. Tsay. "Estimation of covariance matrix via the sparse Cholesky factor with lasso". In: *Journal of Statistical Planning and Inference* 140.12 (2010). Special Issue in Honor of Emanuel Parzen on the Occasion of his 80th Birthday and Retirement from the Department of Statistics, Texas AM University, pp. 3858–3873. ISSN: 0378-3758. DOI: https://doi.org/10.1016/j.jspi.2010.04.048.

[DCS88] M. Davidian, R. J. Carroll, and W. Smith. "Variance Functions and the Minimum Detectable Concentration in Assays". In: *Biometrika* 75.3 (1988), pp. 549–556. ISSN: 00063444. URL: http://www.jstor.org/stable/2336606.

[FHT07] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *Sparse inverse covariance estimation with the lasso*. 2007. arXiv: 0708.3517 [stat.ME].

[Haf80] L. Haff. "Empirical Bayes Estimation of the Multivariate Normal Covariance Matrix". In: *The Annals of Statistics* 8 (May 1980). DOI: 10.1214/aos/1176345010.

[Hua+06] Jianhua Z. Huang et al. "Covariance Matrix Selection and Estimation via Penalised Normal Likelihood". In: *Biometrika* 93.1 (2006), pp. 85–98. ISSN: 00063444. URL: http://www.jstor.org/stable/20441262.

[LH92] Tom Leonard and John S. J. Hsu. "Bayesian Inference for a Covariance Matrix". In: *The Annals of Statistics* 20.4 (1992), pp. 1669–1696. DOI: 10.1214/aos/1176348885.

[LW04] Olivier Ledoit and Michael Wolf. "A well-conditioned estimator for large-dimensional covariance matrices". In: *Journal of Multivariate Analysis* 88.2 (2004), pp. 365–411. ISSN: 0047-259X. DOI: https://doi.org/10.1016/S0047-259X(03)00096-4.

[MB06] Nicolai Meinshausen and Peter Bühlmann. "High-dimensional graphs and variable selection with the Lasso". In: *The Annals of Statistics* 34.3 (2006), pp. 1436–1462. URL: https://doi.org/10.1214/009053606000000281.

[MP67] V A Marčenko and L A Pastur. "Distribution of eigenvalues for some sets of random matrices". In: *Mathematics of the USSR-Sbornik* 1.4 (Apr. 1967), pp. 457–483. DOI: 10.1070/sm1967v001n04abeh001994.

[Pen+09] Jie Peng et al. "Partial Correlation Estimation by Joint Sparse Regression Models". In: *Journal of the American Statistical Association* 104.486 (2009), pp. 735–746. DOI: 10.1198/jasa.2009.0126.

[Pou11] Mohsen Pourahmadi. "Covariance Estimation: The GLM and Regularization Perspectives". In: *Statistical Science* 26.3 (2011), pp. 369–387. ISSN: 08834237. URL: http://www.jstor.org/stable/23059137.

[Pou13] M. Pourahmadi. *High-Dimensional Covariance Estimation*. John Wiley Sons, Ltd, 2013. ISBN: 9781118573617. DOI: https://doi.org/10.1002/9781118573617.fmatter.

[Pou99] Mohsen Pourahmadi. "Joint Mean-Covariance Models with Applications to Longitudinal Data: Unconstrained Parameterisation". In: *Biometrika* 86.3 (1999), pp. 677–690. ISSN: 00063444. URL: http://www.jstor.org/stable/2673662.

[RMC08] Bala Rajaratnam, Hélène Massam, and Carlos M. Carvalho. "Flexible covariance estimation in graphical Gaussian models". In: *The Annals of Statistics* 36.6 (2008), pp. 2818–2849. DOI: 10.1214/08-AOS619. URL: https://doi.org/10.1214/08-AOS619.

[RZY08] Guilherme V. Rocha, Peng Zhao, and Bin Yu. "A path following algorithm for Sparse Pseudo-Likelihood Inverse Covariance Estimation (SPLICE)". In: *arXiv: Methodology* (2008).

[sha] shapes. *distcov: Compute a distance between two covariance matrices*. URL: https://www.rdocumentation.org/packages/shapes/versions/1.1-13/topics/distcov.

[Smi05] Steven Thomas Smith. "Covariance, Subspace, and Intrinsic Cramér-Rao Bounds". In: (June 2005). DOI: 10.1109/tsp.2005.845428. URL: https://doi.org/10.1109/tsp.2005.845428.

[Sza80] Ted H. Szatrowski. "Necessary and Sufficient Conditions for Explicit Solutions in the Multivariate Normal Estimation Problem for Patterned Means and Covariances". In: *Annals of Statistics* 8 (1980), pp. 802–810.

[War08] David I Warton. "Penalized Normal Likelihood and Ridge Regularization of Correlation and Covariance Matrices". In: *Journal of the American Statistical Association* 103.481 (2008), pp. 340–349. DOI: 10.1198/016214508000000021.

[YL07] Ming Yuan and Yi Lin. "Model selection and estimation in the Gaussian graphical model". In: *Biometrika* 94.1 (Mar. 2007), pp. 19–35. ISSN: 0006-3444. DOI: 10.1093/biomet/asm018.

[ZUR16] Piotr Zwiernik, Caroline Uhler, and Donald Richards. *Maximum Likelihood Estimation for Linear Gaussian Covariance Models*. 2016. arXiv: 1408.5604 [math.ST].