ESTIMATION OF JOINTLY CONSTRAINED MEAN-COVARIANCE OF MULTIVARIATE NORMAL DISTRIBUTION

A Dissertation

by

ANUPAM KUNDU

Submitted to the Graduate and Professional School of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Mohsen Pourahmadi
Committee Members,	Anirban Bhattacharya
	Xianyang Zhang
	Grigoris Paouris
Head of Department,	Brani Vidakovic

August 2022

Major Subject: Statistics

ABSTRACT

Estimation of the mean vector and covariance matrix is of central importance in the analysis of multivariate data. In the framework of generalized linear models, usually the variances are certain functions of the means with the normal distribution being an exception. We study some implications of functional relationships between covariance and the mean by focusing on the maximum likelihood and Bayesian estimation of the mean-covariance under the joint constraint $\boldsymbol{\Sigma\mu} = \boldsymbol{\mu}$ for a multivariate normal distribution. It can be viewed as a multivariate counterpart of the classical estimation problem in the $N(\theta, \theta^2)$ distribution. In addition to the usual inference challenges under such non-linear constraints among the parameters (curved exponential family), one has to deal with the basic requirements of symmetry and positive definiteness when estimating a covariance matrix. I have tackled these issues in two ways and verified the solutions using extensive simulation studies.

In the first case, we derive the non-linear likelihood equations for the constrained maximum likelihood estimator of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and solve them using iterative methods. Generally, the MLE of covariance matrices computed using iterative methods do not satisfy the constraints. We propose a novel algorithm to modify such (infeasible) estimators or any other (reasonable) estimator. The key step is to re-align the mean vector along the eigenvectors of the covariance matrix using the idea of regression. In using the Lagrangian function for constrained MLE (Aitchison and Silvey, 1958), the Lagrange multiplier entangles with the parameters of interest and presents another computational challenge. We handle this by either iterative or explicit calculation of the Lagrange multiplier. The existence and nature of location of the constrained MLE are explored within a data-dependent convex set using recent results from random matrix theory.

In the second case, a novel structured covariance is proposed through reparameterization of the spectral decomposition of $\boldsymbol{\Sigma}$ involving its eigenvalues and $\boldsymbol{\mu}$ which lets us study some implications of the functional relationships between covariance and the mean by focusing on the maximum likelihood and Bayesian estimation of the mean-covariance under the joint constraint $\boldsymbol{\Sigma\mu} = \boldsymbol{\mu}$ for

a multivariate normal distribution. This is designed to address the challenging issue of positive-definiteness and to reduce the number of covariance parameters from quadratic to linear function of the dimension. We propose a fast (noniterative) method for approximating the maximum likelihood estimator by maximizing a lower bound for the profile likelihood function, which is concave. We use normal and inverse gamma priors on the mean and eigenvalues, and approximate the maximum aposteriori estimators by both MH within Gibbs sampling and a faster iterative method.

DEDICATION

To my mother, my father, grandparents, relatives and some close friends

ACKNOWLEDGMENTS

First and foremost, I want to acknowledge my parents for the struggle they had to go through in supporting my education. Their constant support, interest in academia, essential life lessons, and advice motivated me throughout my education. I am grateful to my thesis advisor Prof. Mohsen Pourahmadi for his valuable guidance and cooperation in various aspects of my research and for being supportive during the last five years. His unique style of mentoring helped me a lot during this journey. I want to take this opportunity to thank all my teachers from primary and high school, my professors from bachelor and masters programs at the Indian Statistical Institute, Kolkata (especially Prof. Subir Kumar Bhandari for his valuable guidance), and my professors at the Texas A&M University. I also want to thank my senior, Dr. Riddhi Pratim Ghosh, for his persistent encouragement during this journey.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a dissertation committee consisting of Professors Mohsen Pourahmadi, Xianyang Zhang and Anirban Bhattacharya of the Department of Statistics and Professor Grigoris Paouris of the Department of Mathematics.

All work conducted for the thesis dissertation was completed by the student independently.

Funding Sources

Graduate study was supported by a fellowship from Texas A&M University.

NOMENCLATURE

AR(1)	Autoregressive Model of Order 1
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
CASE	Covariate Assisted Screening Estimates
ESAG	Elliptically Symmetric Angular Gaussian
Gram-Schmidt	An Algorithm for Orthogonalization
HQIC	Information Criterion by Hannan and Quinn (1979)
KL	Kullback-Leibler Divergence
K-Means	An Algorithm for Clustering
L_2	Euclidean Norm
LASSO	Least Absolute Shrinkage and Selection Operator
MAP	Maximum Aposteriori
MCMC	Markov Chain Monte Carlo
MF	Distance between Symmetric Positive Definite Matrices by Förstner and Moonen (2003)
MH	Metropolis-Hastings Algorithm
MLE	Maximum Likelihood Estimator
NIW	Normal-Inverse Wishart
NR	Newton- Raphson Algorithm
PCA	Principal Component Analysis
p.d.,pd	Positive Definite Matrix
R^2	Coefficient of Determination

S&C	Method by Strydom and Crowther (2012)
A&S	Method by Aitchison and Silvey (1958)
SIW	Shrinkage-Inverse Wishart
SMLE	Standard MLE
UMVU	Uniformly Minimum-Variance Unbiased

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. LITERATURE REVIEW

In this day and age of information boom, multivariate data is collected from everywhere. More recently, the proliferation of the use of buzzwords like "Big Data", "Data Science", "Machine Learning" in almost every area of business shows us the importance of these concepts in popular scientific discourse. Big Data is a blanket term for very large and complex multivariate data sets where application of traditional data management techniques become obsolete. Data Science is the scientific pursuit of extracting knowledge by using various methods to analyze the data sets, big or small. Multivariate data analysis helps us to get a deeper understanding of the data through a whole range of statistical techniques. For example, in pharmaceutical industry we look at the progression of disease by studying various clinical endpoints through biological measurements of human body, in food industry we study various market response parameters using product quality and marketing parameters etc. Since these data sets almost always have more than two variables in them, the study of multivariate analysis becomes very important. These concepts have wide applications in various aspects of scientific research e.g. climate science (Lakshmanan et al., 2015), geology, financial analysis (Engle et al., 2019), economics (Athey and Imbens, 2019), drug development (Gaurav et al., 2021) etc.

1.1 Classical Statistical Literature:

Over the last hundred years, the research for the analysis of data sets is growing rapidly. Although the main focus is on continuous data, a significant portion of analysis is done for discrete data sets also. Bishop et al. (2007) has shown a great collection of methods related to this field in an organized way. There are various topics such as contingency tables, discrete distributions (e.g.multinomial, negative-multinomial, Dirichlet-multinomial, Dirichlet-negative multinomial etc.) studied extensively in this area. Categorical data sets are also studied using various techniques like logistic regression, ordered logistic regression, bounded integer models in clinical trial, transportation data sets etc. On the other hand, continuous data is abundant in real life and

has been studied using various statistical techniques (e.g. Regression analysis see, Rao (1973), Bibby et al. (1979) Montgomery et al. (2021) etc.). Although various continuous distributions are studied for vectors (e.g. multivariate normal, multivariate t distribution, multivariate pareto distribution etc.), there are other matrix distributions (e.g. Wishart, inverse-Wishart, shrinkage-inverse-Wishart, inverse-matrix gamma distribution etc.) which are used for analysis and estimation in random matrix theory especially covariance estimation. In this thesis our problem mainly revolves around the study of multivariate normal distributions and the estimation of its parameters.

Modeling multivariate data using various discrete or continuous distributions, involves estimation of unknown parameters, which is an essential part for drawing statistical inference (Bibby et al., 1979, chapter 4). For example, if we model the data set using a normal distribution, the estimation of the parameters of the distribution i.e. $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ becomes a necessity. There are various types of estimation performed in statistical literature and maximum likelihood estimation (MLE), a point estimate for the unknown parameters, is the most common of them all. However in the Bayesian set up the posterior distributions of the parameters help us draw inference on them with maximum aposteriori (MAP) being an example of a Bayesian point estimator.

"Covariance matrix is, arguably, the second most important object in all of statistics" (Ledoit and Wolf, 2020). In this chapter we present a thorough literature review of the estimation of the covariance matrix. Mean and covariance estimation of multivariate normal distribution are essential in almost every area of classical multivariate statistics (Bibby et al., 1979). The range of modern applications include economics (Ledoit and Wolf, 2004a), genetics (Schäfer and Strimmer, 2005), astrophysics (Hamimeche and Lewis, 2009), environmental sciences (Eguchi et al., 2010) and climatology (Guillot et al., 2015).

The literature for estimation of covariance matrix is vast and has seen an exponential growth over the years. Let us assume that $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ is a sample of size n from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution. We know that the unconstrained maximum likelihood estimator (MLE) of the parameters for this

set up is $(\widehat{\mu}, \widehat{\Sigma}) = (\bar{x}, S)$ where

$$S = \frac{1}{n}\sum (x_i - \bar{x})(x_i - \bar{x})^\top = \frac{1}{n}A(\bar{x})$$

see (Bibby et al., 1979, §4.2.2). The unconstrained MLE for normal distribution has been studied extensively and we know the distributions of the MLE are:

$$\bar{x} \sim N_p\left(\mu, \frac{1}{n}\Sigma\right), \text{ and } nS \sim W_p(\Sigma, n-1)$$

where W denotes the Wishart distribution and $\bar{x}$ and S are independent. When p is fixed and small and n grows large, S becomes a reliabale and undisputed estimator (Anderson, 2003). But estimation of covariance matrix using sample covariance matrix suffers from several problems, see Jobson and Korkie (1980). Stein (1956) showed that when p and p/n is large S performs poorly and proposed an estimator based on eigenvalue shrinkage. Another large class of estimators is proposed by Ledoit and Wolf (2004b).

It is well-known that estimation of a covariance matrix alone is a daunting task because of some standard constraints e.g. 1. the number of unknown covariance parameters growing quadratically with the dimension and 2. positive definiteness. Dempster (1972) stated that, "The computational ease with which this abundance of parameters can be estimated should not be allowed to obscure the probable unwisdom of such estimation from limited data." In order to alleviate the problems, various structure of covariance matrices have been studied in the literature, for example: compound symmetry, Huynh-Feldt Structure, AR(1), One-Dependent Covariance Structure etc. which reduces the number of unknown parameters in the model, see (Pourahmadi, 2013, chapter 3.3) for a detailed discussion. Principal component analysis (PCA) had been developed for reducing the dimension of the data although it suffers from the problem of interpretability. Many strategies are developed to bypass the notorious positive definiteness e.g. spectral, Cholesky decomposition and factor models see Chiu et al. (1996); Pourahmadi (1999); Fan et al. (2008). Some excellent unconstrained parametrisations of the covariance matrix are proposed by Anderson (1973), Pinheiro

and Bates (1996), Pourahmadi (1999), and many more. For example, modeling the unconstrained parameters of the precision matrix (inverse of covariance matrix) goes along with the approach of generalized linear models (GLM) and includes ideas like covariance selection using linear covariance models (LCM) of Dempster (1972) and graphical models in Cox and Wermuth (2014) where certain entries of the precision matrix are set to zero, see Dempster (1972) and Pourahmadi (2013). Yuan and Lin (2007), Banerjee et al. (2008) and Friedman et al. (2008) frame this as sparse estimation problem described by Bien and Tibshirani (2011).

The idea of selecting a sparse covariance matrix by identifying zeroes in its inverse (Dempster, 1972), or setting a specific structure (e.g. compound symmetry, AR(1) etc.) of the covariance matrix can be viewed as a constraint on the covariance matrix while proposing estimators. Exposition of approaches to compute MLE of multivariate normal distribution under various constraints on the parameters has a long history in the statistical literature. For example Bibby et al. (1979, §4.2.2.2 & 4.2.2.3) shows constraints like $\boldsymbol{\mu} = c_0\boldsymbol{\mu}_0$, $\boldsymbol{R\mu} = \boldsymbol{r}$, $\boldsymbol{\Sigma} = c_0\boldsymbol{\Sigma}_0$ etc. which are straight forward to solve. Anderson and Olkin (1985) considers cases of estimation of means for rank constraint and linear constraint, estimation of variance for fixed correlation (Styan, 1973). A big chunk of work has been produced for covariances when certain entries are zero with recent examples being Butte et al. (2000), Grzebyk et al. (2004), Bien and Tibshirani (2011), Mao et al. (2012). Chaudhuri et al. (2007) proposes an algorithm for covariance graph model having nice convergence properties with similar studies described in Edwards (2012, §7.4), Wermuth et al. (2006) and many more. Estimation of covariance matrix with a constraint on the condition number are studied in Won and Kim (2006), Aubry et al. (2012). More generalized constraints are studied in Aitchison and Silvey (1958), Jamshidian and Bentler (1993), Luo et al. (2016) and many others.

1.2 Bayesian Statistical Literature:

In statistics there are two paradigms in solving the fundamental questions: frequentist and Bayesian. Upto this point the discussion of literature mainly falls under the frequentist paradigm. Gelman and Shalizi (2013) has described the difference in the following words: “Schools of statistical inference are sometimes linked to approaches to the philosophy of science. Classical statistics

- as exemplified by Fishers p-values, Neyman-Pearson hypothesis tests, and Neymans confidence intervals - is associated with the hypothetico-deductive and falsificationist view of science". Hypothesis devised by scientists can be tested and rejected but never really established in the same way. In Bayesian statistics we start with a prior distribution, observe data and formulate inference based on posterior distribution following an inductive pattern of learning rather than employment of tests and attempted falsification. The main characteristic of Bayesian statistics is their explicit use of probability for uncertainty quantification (Gelman et al., 2013).

Bayesian statistical conclusions about an unknown parameter θ are made in terms of probability statements. These statements are conditional on the observed data written as $p(\theta \mid y)$. It is on the fundamental level of conditioning where Bayesian inference departs from the statistical inference based on retrospective evaluation of the procedure to estimate the unknown parameter θ over the distribution of data conditional on the true unknown θ. There is a vast literature that deals with the fundamentals of this philosophy of statistics, see Gelman et al. (2013) for further details. The Bayesian paradigm can be used to deal with the same problems as the frequentist paradigm - once such example is estimation. In our case, the joint estimation of mean and covariance matrix of a normal distribution under a special type of constraint is the problem of interest. Naturally, we will discuss the literature for covariance estimation in this regard.

In the Bayesian context, the literature for covariance estimation is vast and growing. In the early days of Bayesian statistics, Sir Harold Jeffery proposed a non informative prior distribution on the parameter space by proposing a density function proportional to the square root of the determinant of the Fisher information matrix in 1946. Another traditional prior that has been used widely is the inverse-Wishart prior, see (Gelman et al., 2013, Chapter 3.6). This is described in section 2.6.2. The main motivation for using inverse Wishart prior is the conjugacy property which makes the calculation of the posterior distribution very easy. With the improvement in Bayesian computation, various non conjugate priors are proposed, see Yang and Berger (1994), Daniels and Kass (2001), Wong et al. (2003),Hoff (2009b). However, the priors used on the covariance matrix is broadly proposed in three different directions - 1) priors on spectral decomposition 2) generalized inverse

Wishart prior and regression-based approach and 3) priors on correlation matrix. We are going to describe each of these classes in the next few paragraphs and also describe some very recent works.

1.2.1 Prior on Spectral Decomposition:

After the seminal work of Stein (1956), a lot of work has been dedicated in shrinking the eigenvalues of the sample covariance matrix in order to achieve a lower risk (Johnstone and Lu, 2009). Examples of these type of work include Dey et al. (1985), Lin and Perlman (1985), Haff et al. (1991), Yang and Berger (1994), Daniels and Kass (1999), Hoff (2009a) etc.

The prior on unconstrained parametrization of the covariance matrix, which uses spectral decomposition, can be divided intro three broad classes with the goal of shrinking either some functions of the off diagonal entries or the correlation matrix towards a a common value. The three classes are: log matrix prior, reference prior (non-informative) and prior on eigenvectors.

1.2.1.1 Log Matrix Prior:

The log matrix prior by Leonard and Hsu (1992) places multivariate normal prior on $log(\mathbf{\Sigma})$.
Pro: It is flexible, yield more general empirical and hierarchical Bayes smoothing and inference.
Con: It introduces a large number of hyperparameters, is non conjugate and lacks interpretability (Brown et al., 1994).

1.2.1.2 Reference Prior:

The common noninformative prior is Jefferey's prior i.e. $p(\mathbf{\Sigma}) = \mid \mathbf{\Sigma} \mid^{-(p+1)/2}$ which fails to shrink the eigenvalues properly (Jeffreys, 1998). The reference prior of Yang and Berger (1994) has the form

$$p(\mathbf{\Sigma}) = c\left[\mid \mathbf{\Sigma} \mid [\lambda_i - \lambda_j]^{-1}\right]$$

Pro: It provides better shrinkage compared to Jeffery's prior because it puts more mass near the region of equality of the eigenvalues. Daniels and Kass (2001) has shown that it performs well as compared to Haff et al. (1991) in terms of risk and also with Daniels and Kass (1999) for non-diagonal, ill-conditioned matrix.

Con: It requires computation of high dimensional posterior expectations. It provides a uniform prior on the orthogonal $\boldsymbol{P}$ matrix. It suffers if the true matrix is diagonal.

An alternative non informative prior with closed form posterior is proposed by Rajaratnam et al. (2008).

1.2.1.3 Prior on Eigenvectors:

Daniels and Kass (1999) prior is designed to shrink the eigenvectors by reparametrizing the orthogonal matrix in terms of $p(p-1)/2$ Givens angles Van Loan and Golub (1996). Hoff (2009a) introduced matrix Bingham distributions as priors on the orthogonal matrices.

Pro: It reduces the small sample risk effectively. Performs really well if the true matrix is diagonal.

Con: It lacks interpretability of the Givens angles.

1.2.2 The Generalized Inverse Wishart Prior and Regreesion-based Approach:

The earliest use of Cholesky decomposition or regression dissection for covariance matrix can be traced back to Bartlett (1934) and Liu (1993). Although the parameters are statistically interpretable, they are not permutation invariant. Brown et al. (1994) has shown that a regression dissection of the inverse Wishart (IW) distribution makes it possible to define a generalized inverted Wishart(GIW) prior for general covariance matrices which offers considerably more flexibility compared to inverse Wishart (Daniels and Pourahmadi (2002), Rajaratnam et al. (2008)) due to the presence of many parameters to control the variability. Daniels and Pourahmadi (2002) refined the GIW prior in such a way that the restrictions on hyperparameters are removed from normal and inverse Wishart distributions. The details of the advantages of this prior in the analysis of real longitudinal data can be found in Daniels (2005).

The inverse Wishart prior has a tendency to increase the variability of the eigenvalues of the covariance matrix. Berger et al. (2020) has proposed a new class of priors called the shrinkage inverse Wishart prior (SIW) on the covariance matrix using the ideas of Yang and Berger (1994) which solves this problem. This is described in section 2.6.3 and motivated us in developing a prior for our constrained covariance estimation problem.

Estimation of covariance matrix using regression-based approach is performed by focusing on priors on the regression parameters, for example see, Daniels and Pourahmadi (2002), Smith and Kohn (2002), Fox and Dunson (2011) and many more. Tibshirani (1996) gives us the idea of construction of a bona fide prior by exponentiation of the penalty function.

1.2.3 Priors on Correlation Matrix:

Barnard et al. (2000) has used the variance-correlation factorization to provide a lognormal prior on variances and uniform priors on off diagonal entries of correlation matrix independently. Another interesting example where the variance correlation decomposition is used along with the Cholesky decomposition of the correlation matrix is Lan et al. (2017). Here the correlations are expressed as a product of vectors on unit sphere and spherical distributions are used to provide a flexible prior on the correlation matrix.

The decomposition of the covariance matrix in variance correlation matrices has a wide application in network analysis especially in Gaussian Graphical models where the sparse precision matrix is modeled, for example see Baladandayuthapani et al. (2014).

Although we have described the three broad classes of priors in this discussion, constrained estimation of covariance matrix is not very well studied in the literature. We will use the idea of putting prior on spectral decomposition in our case because it helps us model the covariance matrix in a special structure and provide a Bayesian framework for constrained estimation in our case.

2. INTRODUCTION

In the literature review, we have described various constraints for the normal distribution that are already studied in the literature. Some study is done on generalized constraints for example see Aitchison and Silvey (1958), Strydom and Crowther (2012).

2.1 Motivation and Goal:

In this thesis, our goal is to study and resolve new challenges which appear when one attempts to jointly estimate the mean vector and the covariance matrix of a multivariate normal distribution (Paine et al., 2018) under a new type of constraints. The idea of looking into a constraint like (2.2.1) come from the discussion of elliptically symmetric angular Gaussian distribution studied by Paine et al. (2018). This is a distribution suitable for spherical data which arise in many scientific disciplines like shape analysis, geology, meteorology (e.g. Mardia and Jupp (2000)), text analysis (e.g. Hamsici and Martinez (2007)), etc. Our idea was to look at the possiblility of using such a constraint without the spherical nature and provide a generalization to it which can be used in a much larger context. Since our constraint is similar without the spherical nature of the distribution, we can always project them in the corresponding $\mathbb{R}^p$ (Mardia and Jupp, 2000, §3.5.6) and fit a normal model with the constraint (2.2.1). This is helpful in developing multivariate techniques like discriminant analysis, clustering, etc. for spherical data. We have applied it on earth's historic magnetic pole data by Schmidt (1976). This type of a constraint is not studied in the literature to the best of our knowledge.

2.2 Description of the constraint:

The constraint is described below.

$$\boldsymbol{\Sigma}\boldsymbol{\mu} = \boldsymbol{\mu}, \qquad |\boldsymbol{\Sigma}| = 1, \tag{2.2.1}$$

It is interesting to note that the first constraint forces the mean vector $\boldsymbol{\mu}$ to be an eigenvector of $\boldsymbol{\Sigma}$ corresponding to the eigenvalue one, and the second constrains the product of the remaining eigenvalues. The first constraint turns out to be more consequential for statistical inference due to the entanglement (nonlinearity) of the mean-covariance parameters and that $\boldsymbol{\mu}$ as an eigenvector is identifiable up to a constant. Nevertheless, the two together will definitely impact the estimators and the shape of the contour plots of a multivariate normal density function as gleaned from the spectral decomposition of the covariance matrix

$$\boldsymbol{\Sigma} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^\top = \sum_{i=1}^{p} \lambda_i \boldsymbol{P}_i \boldsymbol{P}_i^\top = \sum_{i=1}^{p-1} \lambda_i \boldsymbol{P}_i \boldsymbol{P}_i^\top + \boldsymbol{\mu}\boldsymbol{\mu}^\top, \tag{2.2.2}$$

where $\boldsymbol{D} = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_{p-1}, 1)$ is the diagonal matrix of ordered eigenvalues other than 1 and $\boldsymbol{P} = [\boldsymbol{P}_1, \boldsymbol{P}_2, \ldots, \boldsymbol{P}_p]$ is the corresponding orthogonal matrix of eigenvectors. The second constraint is less stringent and can be achieved by a rescaling. Though these constraints arise in the context of directional data analysis (Paine et al., 2018), they seem to resonate with some of the deeper issues in the classical statistical estimation theory.

2.3 Challenges

It is well-known that constraint or functional relationship among the parameters of a distribution can be the source of computational and inferential challenges. Interestingly, presence of the "quadratic" term $\boldsymbol{\mu}\boldsymbol{\mu}^\top$ in (2.2.2) suggests similarity with the classical inference problems for the $N(\theta, \theta^2)$ distribution where it is known that the minimal sufficient statistic $T(X) = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ is not complete and the UMVU estimators may not exist, see Keener (2011), Chapter 5, for other interesting examples. More generally, the setup is within the multivariate curved exponential family (Efron et al., 1975) where the parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ satisfy the constraints in (2.2.1).

2.4 Statistical Interpretation and Prevalence of the Constraint

In this section we interpret the constraints in the context of factor and error-in-variable models, and then point out that the mean-covariance of the multinomial distributions do not satisfy the constraints.

Let us consider a factor model with a single factor of the form (Rao, 1973, §8f.4)

$$\boldsymbol{X}_i = \boldsymbol{\mu} + \boldsymbol{\mu} w_i + \boldsymbol{\epsilon}_i \tag{2.4.1}$$

where $w_i \overset{iid}{\sim} N(0,1)$ and $\boldsymbol{\epsilon}_i \overset{iid}{\sim} N_p(0, \boldsymbol{\Sigma}_\epsilon)$ are uncorrelated. Note that the mean vector $\boldsymbol{\mu}$ appears as the loading matrix and w_i is the common factor. The covariance matrix of $\boldsymbol{X}_i$ is as in (2.2.2):

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_\epsilon + \boldsymbol{\mu}\boldsymbol{\mu}^\top.$$

This factor model interpretation can also be expanded and viewed as the error-in-variable model in the context of multivariate regression (Molstad et al., 2020) where the error covariance matrix and the regression coefficient matrix are parametrically connected. Our model can be viewed as a special case of the envelop models in Cook and Zhang (2015). It also encourages modeling the covariance matrix as a parsimonious quadratic function of the mean vector similar to Hoff and Niu (2012).

To get a feel for the prevalence of the first constraint involving both the mean vector and the covariance matrix we show that multinomial distributions do not satisfy the constraints as a non-example. In addition, we explore the role of an "intermediate", seemingly less stringent, constraint of the form $\boldsymbol{\Sigma b} = \boldsymbol{\mu}$ where $\boldsymbol{b}$ is ideally a vector independent of the parameters. However, such a $\boldsymbol{b}$ may not always exist as shown in the following example.

Suppose $\boldsymbol{Y} \sim$ multinomial$(n, q_1, q_2, \ldots, q_p)$ where $\sum_{i=1}^{p} q_i = 1$. Then, $Var(Y_i) = nq_i(1-q_i)$ and $Cov(Y_i, Y_j) = -nq_iq_j$ for $i \neq j$, and the mean and covariance matrix have the form

$$\boldsymbol{\mu} = n\boldsymbol{q}, \qquad \boldsymbol{\Sigma}(\boldsymbol{Y}) = \text{diag}(\boldsymbol{q}) - \boldsymbol{q}\boldsymbol{q}^\top$$

where $\boldsymbol{q} = (q_1, q_2, \ldots, q_p)^\top$. The covariance matrix is positive semi-definite with one eigenvalue 0 corresponding to the eigenvector $\mathbb{1} = (1, \ldots, 1)^\top$. We note that $\boldsymbol{\Sigma\mu} \neq \boldsymbol{\mu}$, and there does not exist a vector $\boldsymbol{b}$ such that $\boldsymbol{\Sigma b} = \boldsymbol{\mu}$. For example, in the one– dimensional case $b = 1 - q_1$ depends on

the parameter. More generally, the class of Dirichlet distributions is another example of this kind which do not satisfy the constraints.

A similar type of constraint arises when one deals with discrete Markov chain in statistical literature. In that situation the transition matrix $\boldsymbol{P}$ (Hoel et al., 1986) has the property $\boldsymbol{P}\mathbb{1} = \mathbb{1}$ which is similar to our constraint. But here the transition matrix may not be symmetric. Also, stationary distribution π of the chain satisfies $\pi\boldsymbol{P} = \pi$ which has an application in the PageRank algorithm of Google.

2.5 Frequentist Approach for Joint Estimation and the Idea of "Post-Processing"

Though an explicit formula for the MLE of θ in $N(\theta, \theta^2)$ is given in Khan (1968), finding explicit formula for the MLE in our setup seems to be out of reach. A Lagrange multiplier method for computing constrained MLE and its asymptotic distribution for general distributions satisfying certain regularity conditions is given in Aitchison and Silvey (1958). Unfortunately, computing the Lagrange multiplier in our setup is not straightforward, perhaps due to implicit nonlinearity in the first constraint, and requires special attention. Existing methods such as that of Strydom and Crowther (2012) and Aitchison and Silvey (1958) are described in 3.1.2 and 3.1.3 respectively.

2.5.1 Post Processing

We know that in statistical literature, parameter spaces are sometimes defined by constraints on the usual parameter spaces, which in turn helps us specify priors in Bayesian context Lee et al. (2020). When the dimension of support is lower than the original parameter space, the constrained prior distribution can be a solution to many problems especially like ours. Lee et al. (2020) has proposed a two step process in this context: first, computation of a posterior distribution ignoring the constraint and second, project the posterior sample to the desired parameter space. This might seem to be unrelated to our situation since our approach is frequentist in this section, but upon closer look our idea is very similar. We are too first computing the estimator (without fully ignoring the constraint unlike (Lee et al., 2020)) and later use the algorithm to "post process" the estimator to project it in our desired parameter space. Therefore we will use the term modified estimator and

post-processed estimator interchangeably as they refer to the same idea. This idea has been used in the literature. For example, ordered finite-dimensional parameter space (Dunson and Neelon, 2003), the continuous monotone function space (Lin and Dunson, 2014), measurable monotone function space (Chakraborty and Ghosal, 2021) etc. They showed that the convergence rate of the projected posterior distribution is at least that of the original. Bashir et al. (2019) used post processing to capture the sparsity structure of the precision matrix and Bondell and Reich (2012) conducted variable selection in high dimensional regression model.

2.5.2 Frequentist Post Processing

It turns out that the presumed MLEs obtained from the iterative methods invariably do not satisfy the constraint in (2.2.1), and in some cases the covariance estimator is neither symmetric nor positive-definite. It is a genuine challenge to have the MLE of the covariance matrix to satisfy (2.2.1), in addition to being symmetric and positive definite. So it makes sense to use the idea of post processing of the estimator.

A novel algorithm (see section 3.3) is developed where starting with any pair of mean-covariance estimators, they are modified so as to satisfy the conditions in (2.2.1). The key conceptual idea is to re-align the given mean vector to be in the space spanned by the orthogonal eigenvectors of the given covariance matrix estimator. We re-interpret this as a regression problem with the given mean as the response vector and the eigenvectors as predictors with the associated variable selection step. The modified eigenspace is formed using the Gram-Schmidt orthogonalization process starting with the given estimate of mean to ensure that the estimate is an eigenvector of the estimated covariance matrix.

2.6 Bayesian Approach

The Bayesian paradigm needs prior proposition and posterior calculation for the joint estimation of the unknown parameters in the model which is our goal. It is difficult to propose a prior jointly on mean-covariance that satisfy the constraint under consideration. So we separate the parameters through a reparametrization, that makes it possible to propose priors easily. In order to

do so, we propose a simple and novel structured covariance model based on the spectral decomposition (2.6.1) which addresses the core challenges in covariance estimation under the specific constraint in equation (2.2.1).

2.6.1 Structured Covariance Model:

Let $\boldsymbol{u} = \frac{\mu}{\|\mu\|}$ be the direction of the mean vector so that $\boldsymbol{\mu} = c_0\boldsymbol{u}$ where $c_0 \in \mathbb{R}^+$ and $\boldsymbol{u}$ lies on a unit sphere; c_0 can be interpreted as the radius of the sphere on which the mean vector lie. Our structured covariance model is

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{u}, \lambda_1, \ldots, \lambda_{p-1}) = \boldsymbol{P}(\boldsymbol{u})\boldsymbol{D}\boldsymbol{P}^\top(\boldsymbol{u}) \tag{2.6.1}$$

where $\boldsymbol{D} = \text{diag}\,(1, \lambda_1, \ldots, \lambda_{p-1}) = \text{diag}\,(1, \boldsymbol{\lambda})$ is the matrix of eigenvalues, and for a given value of the mean direction $\boldsymbol{u}$, the orthogonal matrix of eigenvectors is $\boldsymbol{P}(\boldsymbol{u}) = [\boldsymbol{u}, \boldsymbol{V}(\boldsymbol{u})]$ where $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{u}) \in R^{p\times(p-1)}$ is a known matrix function so that orthogonality of $\boldsymbol{P}$ is ensured. A simple and prominent example of such a $\boldsymbol{V}$ is obtained by an application of Gram-Schmidt procedure (Trefethen and Bau III, 1997) to the set $\{\boldsymbol{u}, \boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_{p-1}\}$ where $\boldsymbol{e}_i, i = 1, 2, \ldots, p-1$ denote the canonical basis of $\mathscr{R}^p$ and $u_p \neq 0$. With $\boldsymbol{V}$ assumed known in (2.6.1), the unknown parameters then are $(\boldsymbol{\mu}, \boldsymbol{\lambda})$, the vectors of mean and the $(p-1)$ eigenvalues, so that the number of parameters drastically reduces from quadratic to linear i.e. $(2p-1)$ in the dimension. Moreover, the first constraint is automatically satisfied, since using $\boldsymbol{\mu} \perp \boldsymbol{V}_i(\boldsymbol{u}) = \boldsymbol{V}_i$ (columns of the $\boldsymbol{V}$ matrix) for $i = 1, 2, \ldots, (p-1)$, it follows that

$$\boldsymbol{\Sigma}\boldsymbol{\mu} = \boldsymbol{P}(\boldsymbol{u})\boldsymbol{\Lambda}\boldsymbol{P}^\top(\boldsymbol{u})\boldsymbol{\mu} = \sum_{i=1}^{p-1} \lambda_i \boldsymbol{V}_i\boldsymbol{V}_i^\top\boldsymbol{\mu} + \boldsymbol{u}\boldsymbol{u}^\top\boldsymbol{\mu} = c_0\boldsymbol{u}\boldsymbol{u}^\top\boldsymbol{u} = \boldsymbol{\mu}. \tag{2.6.2}$$

Next, we will provide two concrete examples to elucidate the idea behind the model and its components, particularly the eigenvectors as functions of the mean vector.

Example 1: Consider the case of equal means $\boldsymbol{\mu} = c\mathbb{1}$, then using Gram-Schmidt procedure (Tre-

fethen and Bau III, 1997), it follows that $\boldsymbol{P}(\boldsymbol{u})$ is of the form:

$$\boldsymbol{P}(\boldsymbol{u}) = \begin{bmatrix} \mathbb{1}_p & \boldsymbol{z}_p^{(p)} & \boldsymbol{z}_{p-1}^{(p)} & \boldsymbol{z}_{p-2}^{(p)} & \dots & \boldsymbol{z}_2^{(p)} \end{bmatrix} \tag{2.6.3}$$

where

$$\begin{aligned} \boldsymbol{z}_s^{(p)} &= \left(0, \dots, 0, \boldsymbol{z}_s^{(s)}\right)^\top \qquad \forall s \in \{2, 3 \dots, p\} \\ \boldsymbol{z}_s^{(s)} &= \left(\frac{s-1}{\sqrt{s(s-1)}}, \frac{-1}{\sqrt{s(s-1)}}, \dots, \frac{-1}{\sqrt{s(s-1)}}\right)^\top \end{aligned} \tag{2.6.4}$$

and $\mathbb{1}_p = \left(\frac{1}{\sqrt{p}}, \frac{1}{\sqrt{p}}, \dots, \frac{1}{\sqrt{p}}\right)$. The c drops due to normalization. Choosing the eigenvalues to be $\boldsymbol{D} = \text{diag}\left(1, \frac{1-\rho}{1+(p-1)\rho}, \dots, \frac{1-\rho}{1+(p-1)\rho}\right)$, the covariance matrix Σ is compound symmetry i.e.:

$$\boldsymbol{\Sigma} = \frac{1}{1+(p-1)\rho} \begin{bmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & \vdots & \dots & \vdots \\ \rho & \rho & \dots & 1 \end{bmatrix}$$

We note that for $p = 2$, the other eigenvalue of $\boldsymbol{\Sigma}$ is $\frac{1-\rho}{1+\rho}$.

Example 2: Let us take

$$\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_2)$$

where only the first two entries are different from each other. Then using the construction method described in the paragraph following (2.6.1), the $\boldsymbol{P}$ matrix takes the form:

$$\boldsymbol{P}(\boldsymbol{u}) = \begin{bmatrix} \boldsymbol{u} & \boldsymbol{w}_1 & z_{p-1}^{(p)} & z_{p-2}^{(p)} & \dots z_2^{(p)} \end{bmatrix} \tag{2.6.5}$$

where $z_s^{(p)}$'s are same as in Example 1 and

$$\boldsymbol{w}_1 = ((p-1)\mu_2, -\mu_1, \ldots, -\mu_1)\,/(\|\boldsymbol{\mu}\|).$$

If we choose the diagonal entries to be $(1, \lambda_1, \lambda_2, \ldots, \lambda_2)$, then the covariance matrix $\boldsymbol{\Sigma}$ will have the following format.

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{12} & \ldots & \sigma_{12} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} & \ldots & \sigma_{23} \\ \sigma_{12} & \sigma_{23} & \sigma_{22} & \ldots & \sigma_{23} \\ \ldots & \ldots & & & \ldots \\ \sigma_{12} & \sigma_{23} & \sigma_{23} & \ldots & \sigma_{22} \end{bmatrix}$$

where $\sigma_{11} = \dfrac{\mu_1^2 + \lambda_1(p-1)\mu_2^2}{\mu_0^2}, \quad \sigma_{12} = \dfrac{\mu_1\mu_2(1-\lambda_1)}{\mu_0^2}$

$$\sigma_{22} = \frac{(1+\lambda_2(p-2))\mu_0^2 - \mu_1^2(1-\lambda_1)}{\mu_0^2(p-1)}, \quad \sigma_{23} = \frac{(1-\lambda_2)\mu_0^2 - (1-\lambda_1)\mu_1^2}{\mu_0^2(p-1)}$$

and $\mu_0^2 = \|\mu\|^2$.

Example 3: As a generalization of Example 1 and 2, take

$$\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3, \ldots, \mu_3)^\top$$

where the first three entries of the mean vector are different and rest of them are the same as the third entry of the vector. Then, using the construction method described in the paragraph following (2.6.1), the $\boldsymbol{P}$ matrix takes the form:

$$\boldsymbol{P}(\boldsymbol{u}) = \begin{bmatrix} \boldsymbol{u} & \boldsymbol{w}_1 & \boldsymbol{w}_2 & z_{p-2}^{(p)} & \ldots z_2^{(p)} \end{bmatrix} \tag{2.6.6}$$

where $z_s^{(p)}$'s are same as in Example 1, and

$$\begin{aligned}\boldsymbol{w}_1 &= \left(\mu_0^2, -\mu_1\mu_2, -\mu_1\mu_3, \ldots, -\mu_1\mu_3\right)^\top /(\mu_0\|\mu\|)\\ \boldsymbol{w}_2 &= (0, (p-2)\mu_3, -\mu_2, \ldots, -\mu_2)^\top /((p-2)\mu_0),\end{aligned} \tag{2.6.7}$$

where $\mu_0^2 = \mu_2^2 + (p-2)\mu_3^2$. When $p = 3$, $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ reduce to the $\tilde{\xi}_1$ and $\tilde{\xi}_2$ in Paine et al. (2018, §2.3), so that this example is a generalization of their construction to the case of $p > 3$ or more general spherical data.

Our goal is to compute the maximum likelihood estimator and provide a Bayesian framework for estimation of the parameters in (2.6.1). In the Bayesian context, the maximum aposteriori (MAP) does not have a closed form, even though the posterior distribution obtained by using a Gaussian prior for the mean vector and inverse gamma on the eigenvalues does. We have suggested this prior by extending the shrinkage inverse Wishart prior of Berger et al. (2020) to the case of nonzero means. Here we are going to describe the standard normal-inverse Wishart prior and shrinkage inverse Wishart priors to illustrate the reason for the proposition of such a prior.

2.6.2 Normal-Inverse Wishart Priors:

For the parameters of a multivariate normal distribution the most popular prior is normal-inverse Wishart described in Gelman et al. (2013, Section 3.6) and Barnard et al. (2000) with the hyperparameters $(\boldsymbol{\mu}_0, \kappa_0; \nu_0, \boldsymbol{\Lambda}_0)$:

$$\boldsymbol{\mu} \,|\, \boldsymbol{\Sigma} \sim N_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}/\kappa_0), \qquad \boldsymbol{\Sigma} \sim \pi_{IW}(\nu_0, \boldsymbol{\Lambda}_0^{-1}),$$

where ν_0 and $\boldsymbol{\Lambda}_0$ are the degrees of freedom and the scale matrix. The suggested values for the hyperparameters in Gelman and Hill (2006) are $\boldsymbol{\Lambda}_0 = \boldsymbol{I}_p$ and $\nu_0 = p + 1$. The remaining hyperparameters are the prior mean, $\boldsymbol{\mu}_0$ and κ_0 on the $\boldsymbol{\Sigma}$ scale. Due to conjugacy, the posterior density has the following parameters:

$$\boldsymbol{\mu}_n = \frac{\kappa_0}{\kappa_0 + n}\boldsymbol{\mu}_0 + \frac{n}{\kappa_0 + n}\bar{\boldsymbol{x}},$$
$$\boldsymbol{\Lambda}_n = \boldsymbol{\Lambda}_0 + \boldsymbol{A} + \frac{n\kappa_0}{\kappa_0 + n}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)^\top. \tag{2.6.8}$$

As a point estimator posterior mode or the maximum aposteriori probability (MAP) estimator for $\boldsymbol{\Sigma}$ is generally the most popular choice (Murphy, 2012, §5.2.1). The posterior marginal distribution of the mean parameter $\boldsymbol{\mu}$ is multivariate $t_{\nu_n - p+1}\left(\boldsymbol{\mu}_n, \frac{\boldsymbol{\Lambda}_n}{\kappa_n(\nu_n - p+1)}\right)$ with the MAP estimator $\boldsymbol{\mu}_n$, and the MAP estimator for $\boldsymbol{\Sigma}$ is given by

$$\widehat{\boldsymbol{\Sigma}}_{map} = \frac{\boldsymbol{\Lambda}_n}{\nu_n + p + 2},$$

where $\nu_n = \nu_0 + n$ (proof is shown in appendix 1).

2.6.3 Shrinkage Inverse Wishart Priors

The most natural prior for mean and covariance i.e. normal-inverse Wishart reviewed above is known to overdisperse the eigenvalues of the posterior covariance estimator. In fact, Yang and Berger (1994) describes that the normal-inverse Wishart prior has the term $\prod_{i<j}(\lambda_i - \lambda_j)$ in the density where λ_i's are the ordered eigenvalues of $\boldsymbol{\Sigma}$, thus forcing the eigenvalues apart and increasing the variability (Berger et al., 2020). This is one major motivation for using the normal-shrinkage inverse Wishart prior.

The shrinkage-inverse Wishart (SIW) prior has the same density as inverse Wishart prior save the extra term $\prod_{i<j}(\lambda_i - \lambda_j)$ in the denominator:

$$\pi_{SIW}(\boldsymbol{\Sigma} \mid \nu_0, b, \boldsymbol{\Lambda}_0^{-1}) \propto |\boldsymbol{\Sigma}|^{-\frac{\nu_0 + p + 1}{2}} \frac{\exp\left[-\frac{1}{2}Tr(\boldsymbol{\Lambda}_0\boldsymbol{\Sigma}^{-1})\right]}{\prod_{i<j}(\lambda_i - \lambda_j)^b}, \tag{2.6.9}$$

where ν_0 is a real constant, $b \in [0, 1]$ and $\boldsymbol{\Lambda}_0$ is a positive semi-definite matrix. It is interesting to notice that $b = 0$ corresponds to some common priors like inverse Wishart, reference, Jefferey's

etc. which also contain the term $\prod_{i<j}(\lambda_i - \lambda_j)$, see Berger et al. (2020).

1. Constant Prior, $\pi_C(\boldsymbol{\Sigma}) = 1$ corresponding to $\nu_0 = -(p+1)$, $b = 0$ and $\boldsymbol{\Lambda}_0 = \mathbf{0}$
2. Jeffrey's Prior, $\pi_J(\boldsymbol{\Sigma}) = \mid \boldsymbol{\Sigma} \mid^{-\frac{p+1}{2}}$ corresponding to $\nu_0 = 0$, $b = 0$ and $\boldsymbol{\Lambda}_0 = \mathbf{0}$.
3. Reference Prior, $\pi_R(\boldsymbol{\Sigma}) = \mid \boldsymbol{\Sigma} \mid^{-1} \left[\prod_{i<j}(\lambda_i - \lambda_j)\right]^{-1}$ corresponding to $\nu_0 = -(p-1)$, $b = 1$ and $\boldsymbol{\Lambda}_0 = \mathbf{0}$.
4. Modified Reference Prior, $\pi_{MR}(\boldsymbol{\Sigma}) = \mid \boldsymbol{\Sigma} \mid^{-(1-1/2p)}$ corresponding to $\nu_0 = 1 - (p + 1/p)$, $b = 1$ and $\boldsymbol{\Lambda}_0 = \mathbf{0}$.
5. Inverse Wishart Prior, $\pi_{IW}(\boldsymbol{\Sigma} \mid \nu_0, \boldsymbol{\Lambda}_0^{-1})$ corresponding to $b = 0$,

The posterior density for $(\boldsymbol{D}, \boldsymbol{P})$, using the one-to-one transformation from $\boldsymbol{\Sigma}$ to $\boldsymbol{D} = \text{diag}(\lambda_1, \ldots, \lambda_p)$ and the orthogonal eigenvector matrix $\boldsymbol{P}$, turns out to be

$$\pi_{SIW}(\boldsymbol{D}, \boldsymbol{P} \mid \nu_0, b, \boldsymbol{\Lambda}_0^{-1}) \propto \mid \boldsymbol{\Sigma} \mid^{-\frac{\nu_0+p+1}{2}} \frac{\exp\left[-\frac{1}{2}Tr(\boldsymbol{\Lambda}_0 \boldsymbol{P}\boldsymbol{D}^{-1}\boldsymbol{P}^\top)\right]}{\prod_{i<j}(\lambda_i - \lambda_j)^{(b-1)}} \mathbb{1}_{\lambda_1 \geq \lambda_2 \geq \ldots \lambda_p}. \tag{2.6.10}$$

with the Jacobian being,

$$\left| \frac{\partial \boldsymbol{\Sigma}}{\partial(\boldsymbol{D}, \boldsymbol{P})} \right| = \prod_{i<j}(\lambda_i - \lambda_j)$$

Note that the posterior is zero whenever the eigenvalues are close together so that effectively it forces the eigenvalues apart. Moreover, when $b = 1$, the term in question goes away retaining the conjugacy property, even with the additional normal prior on the nonzero mean (Proof is shown in the Appendix 2) and the same posterior parameters as before.

It is interesting to note that a joint prior on $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ can be constructed and the normal-shrinkage inverse Wishart can be parametrized using the five hyperparameters $(\boldsymbol{\mu}_0, \kappa_0; \nu_0, b, \boldsymbol{\Lambda}_0)$:

$$\boldsymbol{\mu} \mid \boldsymbol{\Sigma} \sim N_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}/\kappa_0), \qquad \boldsymbol{\Sigma} \sim \pi_{SIW}(\nu_0, b, \boldsymbol{\Lambda}_0^{-1}) \tag{2.6.11}$$

The sampling scheme of $\boldsymbol{\Sigma}$ from shrinkage-inverse Wishart using the 'new method' is described in appendix B.2. It helps us generate a pair $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ jointly from the posterior i.e.

$$\pi_{NSIW}(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \boldsymbol{\mu}_n, \kappa_n, \nu_n, \boldsymbol{\Lambda}_n) \propto \mid \boldsymbol{\Sigma} \mid^{-\frac{\nu_n+p}{2}-1} \frac{\exp\left[-\frac{1}{2}\left\{(\boldsymbol{\mu}-\boldsymbol{\mu}_n)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_n) + \mathrm{Tr}\left(\boldsymbol{\Lambda}_0\boldsymbol{\Sigma}^{-1}\right)\right\}\right]}{\prod_{i<j}\left(\lambda_i - \lambda_j\right)}.$$

The next important consequence of choosing $b = 1$, from (2.6.10), is that the conditional distribution of the eigenvalues given the orthogonal matrix $\boldsymbol{P}$ is an ordered inverse gamma distribution (Berger et al., 2020, §3.2). This plays a central role in developing a computational scheme for computing the MAP.

Recall that the structured covariance model in (2.6.1) has its orthogonal matrix $\boldsymbol{P}$ determined by a non-zero mean vector. This requires working with nonzero mean vectors in contrast to the mean zero framework in Berger et al. (2020), while maintaining the conditional distribution of the eigenvalues given the mean vector as an inverse gamma distribution, see also Berger et al. (2020). This motivates us to select an inverse gamma prior for the eigenvalues as in (4.3.1) which, fortunately, in turn implies that the conditional distribution of eigenvalues is the inverse gamma (see section 4.3.2).

3. FREQUENTIST SOLUTION USING LAGRANGE MULTIPLIER

3.1 Constrained Maximum Likelihood Estimation

The Lagrange multiplier method (Aitchison and Silvey, 1958) as described in the previous section is used to incorporate the constraints for finding the MLE of the parameters of a multivariate distribution. We derive the likelihood equations, present three iterative methods and study some of their computational and statistical properties. The first method is described in this section and the other two methods i.e. Aitchison and Silvey (1958) and explicit calculation of Lagrange multiplier is described the previous section on literature review. Curiously, the MLEs first appear to be explicit and have closed-forms, but on closer inspection they actually depend on the random Lagrange multipliers and hence disqualified as bona fide statistical estimators. This realization calls attention to estimating the Lagrange multiplier using iterative methods in conjunction with the MLE. Such coupling of estimation of the main and the nuisance parameters makes the task of computing the constrained MLE and study of their convergence much more challenging as shown in this section.

Let $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ be a sample of size n from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is a positive-definite matrix. If $\boldsymbol{X}$ is the $n \times p$ data matrix, then the log-likelihood of the multivariate normal distribution is proportional to

$$l(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\boldsymbol{X}) \propto -\frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}). \tag{3.1.1}$$

The MLE of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ ignoring the constraints is $(\bar{\boldsymbol{x}}, \boldsymbol{S})$, the familiar sample mean and sample covariance matrix, which evidently do not satisfy the conditions in (2.2.1). However, the log-likelihood function generally is not concave under constraints on the covariance matrix and may have multiple local maxima. For $n > p$ we set $\boldsymbol{A}(\boldsymbol{\mu}) = \sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\top$ and note that $\boldsymbol{A}(\bar{\boldsymbol{x}}) = n\boldsymbol{S}$.

Theorem 1. *The Lagrangian function for MLE under the intermediate constraint* $\boldsymbol{\Sigma b} = \boldsymbol{\mu}$ *(ex-*

pressed in terms of the inverse covariance matrix) is:

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \boldsymbol{X}) = l(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \boldsymbol{X}) + \alpha_1 \left(\mid \boldsymbol{\Sigma}^{-1} \mid -1 \right) - \boldsymbol{\alpha}_2^{\top} \left(\boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} - \boldsymbol{b} \right) \tag{3.1.2}$$

where α_1 *and* $\boldsymbol{\alpha}_2$ *are the Lagrange multipliers.*

(a) Under the solo constraint $\mid \boldsymbol{\Sigma} \mid = 1$ *(*$\boldsymbol{\alpha}_2 = \mathbf{0}$*), the MLE* $\widehat{\boldsymbol{\mu}}_{mle} = \bar{\boldsymbol{x}}$ *is the sample mean and* $\widehat{\boldsymbol{\Sigma}}_{mle} = \frac{A(\bar{x})}{|A(\bar{x})|^{1/p}}$ *is a shape matrix.*

(b) If $\mid \boldsymbol{\Sigma} \mid = 1$ *and* $\boldsymbol{\Sigma b} = \boldsymbol{\mu}$ *as in (3.1.2), then the constrained MLE satisfies*

$$\boldsymbol{\mu} = \left(\bar{\boldsymbol{x}} - \frac{1}{n} \boldsymbol{\alpha}_2 \right), \qquad \boldsymbol{\Sigma} = \frac{\left[\boldsymbol{A}(\boldsymbol{\mu}) + 2\boldsymbol{\alpha}_2 \boldsymbol{\mu}^{\top} \right]}{n + 2\alpha_1}, \tag{3.1.3}$$

$$\boldsymbol{\Sigma b} = \boldsymbol{\mu}, \qquad \mid \boldsymbol{\Sigma} \mid = 1. \tag{3.1.4}$$

(c) Under both constraints in (2.2.1), the MLE satisfies

$$\boldsymbol{\mu} = \left(\bar{\boldsymbol{x}} - \frac{1}{n} (\boldsymbol{I} - \boldsymbol{\Sigma}) \boldsymbol{\alpha}_2 \right), \qquad \boldsymbol{\Sigma} = \frac{\left[\boldsymbol{A}(\boldsymbol{\mu}) + 2\boldsymbol{\alpha}_2 \boldsymbol{\mu}^{\top} \right]}{n + 2\alpha_1} \tag{3.1.5}$$

$$\boldsymbol{\Sigma \mu} = \boldsymbol{\mu}, \qquad \mid \boldsymbol{\Sigma} \mid = 1$$

The proof is provided in the Appendix 1. Unlike the closed-form solution in (a), computing the MLE in (b) and (c) is more challenging and involves both α_1 and $\boldsymbol{\alpha}_2$. Thus, one may resort to iterative methods for solving for the (random) Lagrange multipliers, which must go through all the four steps (equalities) to complete one iteration. To highlight the role of the intermediate constraint, we note that in Theorem 1.(b), every parameter can be expressed in terms of $\boldsymbol{\alpha}_2$ due to the intermediate constraint $\boldsymbol{\Sigma b} = \boldsymbol{\mu}$, so that the iterations will be over $\boldsymbol{\alpha}_2$ only, see Section 3.1.1.1 for details. By contrast, the case in Theorem 1.(c) under $\boldsymbol{\Sigma \mu} = \boldsymbol{\mu}$ is much more challenging, at least, due to the presence of $\boldsymbol{\Sigma}$ in $\boldsymbol{\mu}$. These observations serve as strong motivations for considering the alternative method of explicit calculation of the Lagrange multiplier $\boldsymbol{\alpha}_2$ in Section 3.1.2. In view

of Theorem 1 (a), from here on we focus mostly on the first constraint and deemphasize the second constraint $|\,\boldsymbol{\Sigma}\,| = 1$ which is achievable through a scale change.

3.1.1 Algorithms for Computing Constrained MLE:

In spite of the apparent closed forms in (3.1.3) and (3.1.5), these can not be implemented or viewed as bona fide estimators because of their dependence on the Lagrange multipliers α_1 and $\boldsymbol{\alpha}_2$. Here, first we propose a natural iterative method for computing the Lagrange multipliers leading to statistically viable estimators of the mean and the covariance matrix. Then, explicit calculation of the Lagrange multipliers as in Aitchison and Silvey (1958) and Strydom and Crowther (2012) is pursued and its role on the convergence of the iterative methods is studied.

3.1.1.1 Solving (3.1.3) for α_2

Knowing $\boldsymbol{\alpha}_2$ in (3.1.3), determines all the other unknown quantities. To emphasize dependence on $\boldsymbol{\alpha}_2$, we set $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\alpha}_2)$ and denote the numerator of $\boldsymbol{\Sigma}$ by

$$\boldsymbol{U}(\boldsymbol{\alpha}_2) = \boldsymbol{A}\,[\boldsymbol{\mu}(\boldsymbol{\alpha}_2)] + 2\boldsymbol{\alpha}_2\boldsymbol{\mu}^{\top}(\boldsymbol{\alpha}_2).$$

From the second constraint in (3.1.4) it follows that $|\,\boldsymbol{U}(\boldsymbol{\alpha}_2)\,|^{1/p} = n + 2\alpha_1$. Replacing the numerator by $\boldsymbol{U}(\boldsymbol{\alpha}_2)$ and the denominator by $|\,\boldsymbol{U}(\boldsymbol{\alpha}_2)\,|^{1/p}$ in the right hand side of the second identity of (3.1.3) leads to

$$\boldsymbol{\Sigma}(\boldsymbol{\alpha}_2) = \frac{\boldsymbol{U}(\boldsymbol{\alpha}_2)}{|\,\boldsymbol{U}(\boldsymbol{\alpha}_2)\,|^{1/p}},$$

which is a function of $\boldsymbol{\alpha}_2$. Substituting $\boldsymbol{\Sigma}(\boldsymbol{\alpha}_2)$ in the first expression of (3.1.4), we obtain

$$\boldsymbol{\mu}(\boldsymbol{\alpha}_2) = \frac{\boldsymbol{\Sigma}(\boldsymbol{\alpha}_2)b}{|\,\boldsymbol{\Sigma}(\boldsymbol{\alpha}_2)\,|^{1/p}}, \tag{3.1.6}$$

where further replacing $\boldsymbol{\mu}(\boldsymbol{\alpha}_2), \boldsymbol{\Sigma}(\boldsymbol{\alpha}_2), \boldsymbol{U}(\boldsymbol{\alpha}_2)$ and $n + 2\alpha_1$ in terms of $\boldsymbol{\alpha}_2$ one obtains the following after some algebraic manipulation:

$$\boldsymbol{\alpha}_2 = \frac{|\,\boldsymbol{\Sigma}(\boldsymbol{\alpha}_2)\,|^{1/p}\,\bar{\boldsymbol{x}} - (n-1)\boldsymbol{S}b - (1/n^2)\boldsymbol{\alpha}_2\boldsymbol{\alpha}_2^\top b}{2\left(\bar{\boldsymbol{x}}^\top b - \frac{\boldsymbol{\alpha}_2^\top b}{n}\right) - \frac{|\boldsymbol{\Sigma}(\boldsymbol{\alpha}_2)|^{1/p}}{n}} = f(\boldsymbol{\alpha}_2). \tag{3.1.7}$$

This being nonlinear in $\boldsymbol{\alpha}_2$ suggests using the iterations:

$$\boldsymbol{\alpha}_2^{(k+1)} = f(\boldsymbol{\alpha}^{(k)}), k = 0, 1, 2\ldots, \text{ with } \boldsymbol{\alpha}_2^{(0)} = \bar{\boldsymbol{x}},$$

for solving it.

Although the intermediate constraint seems similar to (2.2.1), in the next section it is demonstrated that the latter is much harder to work with in that one needs to iterate over the α_1 as well.

3.1.1.2 Solving (3.1.5) for α_1nd $\boldsymbol{\alpha}_2$

After replacing $\boldsymbol{\mu}$ from the first identity, which involves $\boldsymbol{\Sigma}$, the second equation in (3.1.5) reveals that $\boldsymbol{\Sigma}$ is a nonlinear function of $\boldsymbol{\alpha}_2$. This is different from Theorem 1.(b) in that not all parameters can be expressed in terms of a single parameter (like $\boldsymbol{\alpha}_2$). Thus, one may resort to iterative methods involving the four parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha_1, \boldsymbol{\alpha}_2)$ where the updates for the (k+1)-th iteration is done in the following order :

$$\begin{aligned}
\alpha_1^{(k+1)} &= \frac{1}{2}\left(|\boldsymbol{A}\left(\boldsymbol{\mu}^{(k)}\right) + 2\boldsymbol{\alpha}_2^{(k)}\boldsymbol{\mu}^{(k)\top}|^{1/p} - n\right), \qquad \boldsymbol{\Sigma}^{(k+1)} = \frac{\left[\boldsymbol{A}(\boldsymbol{\mu}^{(k)}) + 2\boldsymbol{\alpha}_2^{(k)}\boldsymbol{\mu}^{(k)\top}\right]}{n + 2\alpha_1^{(k)}} \\
\boldsymbol{\alpha}_2^{(k+1)} &= \frac{1}{2}\left[\left(n + 2\alpha_1^{(k)}\right)\boldsymbol{\Sigma}^{(k)} - \boldsymbol{A}\left(\boldsymbol{\mu}^{(k)}\right)\right]\boldsymbol{\mu}^{(k)}, \qquad \boldsymbol{\mu}^{(k+1)} = \boldsymbol{\Sigma}^{(k)}\left(\bar{\boldsymbol{x}} - \frac{1}{n}\left(\boldsymbol{I} - \boldsymbol{\Sigma}^{(k)}\right)\boldsymbol{\alpha}_2^{(k)}\right).
\end{aligned} \tag{3.1.8}$$

Our suggested initial values are $(\boldsymbol{\Sigma}^{(0)}, \boldsymbol{\alpha}_2^{(0)}, \boldsymbol{\mu}^{(0)}) = (\boldsymbol{S}, \bar{\boldsymbol{x}}, \bar{\boldsymbol{x}})$, and for $k = 0$ we compute $\alpha_1^{(1)}$ using $(\boldsymbol{\mu}^{(0)}, \boldsymbol{\alpha}_2^{(0)})$ from the first equation above. But for the updates $\boldsymbol{\Sigma}^{(1)}$ and $\boldsymbol{\alpha}_2^{(1)}$, we need the value of $\alpha_1^{(0)}$. In order to avoid the confusion, we simply choose $\alpha_1^{(0)} = \alpha_1^{(1)}$ for the first iteration, then use $(\alpha_1^{(1)}, \boldsymbol{\alpha}_2^{(1)}, \boldsymbol{\Sigma}^{(1)}, \boldsymbol{\mu}^{(1)})$ and repeat the process.

3.1.1.3 Common Challenges with Iterative Methods for Computing MLE of $\boldsymbol{\Sigma}$

An estimate of a covariance matrix from iterative methods is usually asymmetric and not necessarily positive definite. The first issue is addressed by replacing the estimator with $\frac{1}{2}\left(\boldsymbol{\Sigma}+\boldsymbol{\Sigma}^{\top}\right)$, producing an estimator of the form $\boldsymbol{A}+\boldsymbol{a}\boldsymbol{b}^{\top}+\boldsymbol{b}\boldsymbol{a}^{\top}$ where $\boldsymbol{a},\boldsymbol{b}\in\mathbb{R}^{p}$ and a positive definite matrix $\boldsymbol{A}$. Ensuring positive definiteness of a matrix of this form is difficult and discussed in the following lemma, its is presented in the Appendix.

Lemma 1. *Let* $\boldsymbol{a},\boldsymbol{b}\in\mathbb{R}^{p}$ *and* $\boldsymbol{A}$ *be a positive definite matrix. Then,*

(a) The non-zero eigenvalues of $(\boldsymbol{a}\boldsymbol{b}^{\top}+\boldsymbol{b}\boldsymbol{a}^{\top})$ *are* $\boldsymbol{a}^{\top}\boldsymbol{b}\pm\|\boldsymbol{a}\|\|\boldsymbol{b}\|$

(b) The matrix $\boldsymbol{M}=\boldsymbol{A}+(\boldsymbol{a}\boldsymbol{b}^{\top}+\boldsymbol{b}\boldsymbol{a}^{\top})$ *has at most one negative eigenvalue.*

To ensure positive-definiteness, Lemma 1.(b) suggests replacing the smallest eigenvalue of $\boldsymbol{M}$ by $\left(\prod_{j=1}^{p-1}\lambda_{j}\right)^{-1}$ where λ_{j}'s are the ordered eigenvalues of $\boldsymbol{M}$. This is justified by noting that according to Weyl's inequality (Bhatia, 2007)

$$\lambda_{p-1}(\boldsymbol{M})\geq\lambda_{p}(\boldsymbol{A})>0.$$

In addition, there are a number of existence and convergence problems related to Theorem 1.(c). These are dealt with partially by relying on more explicit calculations of the Lagrange multipliers under the first constraint only using the methods of Aitchison and Silvey (1958) and Strydom and Crowther (2012).

3.1.2 Explicit Calculation of the Lagrange Multiplier:

Iterative computation of the Lagrange multipliers along with the parameters of interest as above can be the source of several convergence problems. We present a method from Strydom and Crowther (2012) which computes the Lagrange multiplier through a Taylor series expansion of the constraint function.

Note that our mean-covariance constraint can be written either as a scalar function or vector function of the parameters. We start with expressing the constraint $\boldsymbol{\Sigma}\boldsymbol{\mu}=\boldsymbol{\mu}$ as the scalar function

$h : \mathbb{R}^{p^2+p} \to \mathbb{R}$ of the natural parameter vector $\boldsymbol{m}$ of a multivariate normal distribution:

$$h(\boldsymbol{m}) = \left[\boldsymbol{m}_2 - \boldsymbol{m}_1 \otimes \boldsymbol{m}_1 - \text{vec}(\boldsymbol{I}_p)\right]^\top (\mathbb{1} \otimes \boldsymbol{m}_1) = 0, \tag{3.1.9}$$

where $\boldsymbol{m}^\top = \left[\boldsymbol{\mu}^\top, \text{vec}(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top)^\top\right]^\top = \left[\boldsymbol{m}_1^\top, \boldsymbol{m}_2^\top\right]^\top$.

Using the Taylor's expansion of $h(\boldsymbol{m})$ around T, the sufficient statistics of the exponential family (the normal distribution in our case) leads to the following explicit formula for the Lagrange multiplier:

$$\boldsymbol{\alpha}_2 = -\left[\nabla h(\boldsymbol{m})^\top \nabla \boldsymbol{m}(\boldsymbol{\theta}) \nabla h(T)\right]^{-1} h(T), \tag{3.1.10}$$

where $\boldsymbol{\theta}$ is the canonical parameter for the multivariate normal distribution, see Appendix B. Substituting this in (A.2.1) leads to the identity

$$\boldsymbol{m} = T(\boldsymbol{X}) - \boldsymbol{V}\nabla h(\boldsymbol{m}) \frac{h(T)}{[\nabla h(\boldsymbol{m})^\top \boldsymbol{V} \nabla h(T)]} \quad \text{with } \nabla \boldsymbol{m}(\boldsymbol{\theta}) = \boldsymbol{V}. \tag{3.1.11}$$

It can be used iteratively via a "double iteration" over T and $\boldsymbol{m}$, see Strydom and Crowther (2012, §2), with the initial values chosen as the observed canonical statistics for both T and $\boldsymbol{m}$, see Algorithm 2 in Appendix B.

As usual positive-definiteness and symmetry of the covariance estimate are not guaranteed. Nevertheless, its performance in terms of the Frobenius risk in the simulation studies is better than the standard MLE procedure of Section 3.1.1. This can be attributed to the explicit calculation of Lagrange multiplier.

3.1.3 The Aitchison and Silvey (1958) Method

For investigating the asymptotic distribution of the MLE and its iterative computation (Aitchison and Silvey, 1958), it is common to confine attention to a ball or neighbourhood of the true parameter value. More concretely, we consider the set $U_\epsilon = \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| < \epsilon\}$ where $\boldsymbol{\theta}_0$ is

the true parameter value for the parameter vector $\boldsymbol{\theta} = \left(\boldsymbol{\mu}^\top, \text{vec}(\boldsymbol{\Sigma})^\top\right)^\top$ of a multivariate normal distribution.

For the vector-valued constraint function

$$h(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \boldsymbol{\Sigma}\boldsymbol{\mu} - \boldsymbol{\mu},$$

its first derivative denoted by $\boldsymbol{H}_{\boldsymbol{\theta}}^1$ is the $(p + p^2) \times p$ full-rank matrix:

$$\boldsymbol{H}_{\boldsymbol{\theta}}^1 = \begin{bmatrix} \frac{\partial h}{\partial \boldsymbol{\mu}} \\ \frac{\partial h}{\partial \boldsymbol{\Sigma}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma} - \boldsymbol{I} \\ \boldsymbol{\mu} \otimes \boldsymbol{I} \end{bmatrix}.$$

The notations $\boldsymbol{H}_{\hat{\boldsymbol{\theta}}}^1$ and $\boldsymbol{H}_{\boldsymbol{\theta}_0}^1$, with obvious interpretation, are used as needed next. The partitioned matrix $\boldsymbol{E} = \begin{bmatrix} \boldsymbol{B}_{\boldsymbol{\theta}_0} & -\boldsymbol{H}_{\boldsymbol{\theta}_0}^1 \\ -\boldsymbol{H}_{\boldsymbol{\theta}_0}^{1\top} & \boldsymbol{0} \end{bmatrix}$ is non singular (Aitchison and Silvey, 1958, Lemma 3) where $\boldsymbol{B}_{\boldsymbol{\theta}_0} = \begin{pmatrix} \boldsymbol{\Sigma}^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1} \end{pmatrix}$, and its inverse is given by

$$\boldsymbol{E}^{-1} = \begin{bmatrix} \boldsymbol{P}_{\boldsymbol{\theta}} & \boldsymbol{Q}_{\boldsymbol{\theta}} \\ \boldsymbol{Q}_{\boldsymbol{\theta}}^\top & \boldsymbol{R}_{\boldsymbol{\theta}} \end{bmatrix}$$

where

$$\boldsymbol{R}_{\boldsymbol{\theta}} = -\left(\boldsymbol{H}_{\boldsymbol{\theta}}^{1\top} \boldsymbol{B}_{\boldsymbol{\theta}}^{-1} \boldsymbol{H}_{\boldsymbol{\theta}}^1\right)^{-1} = -\left[(\boldsymbol{\Sigma} - \boldsymbol{I})\boldsymbol{\Sigma}(\boldsymbol{\Sigma} - \boldsymbol{I}) + (\boldsymbol{\mu}^\top \boldsymbol{\Sigma}\boldsymbol{\mu})\boldsymbol{\Sigma}\right]^{-1}$$
$$\boldsymbol{Q}_{\boldsymbol{\theta}} = -\boldsymbol{B}_{\boldsymbol{\theta}} \boldsymbol{H}_{\boldsymbol{\theta}}^1 \boldsymbol{R}, \quad \boldsymbol{P}_{\boldsymbol{\theta}} = \boldsymbol{B}_{\boldsymbol{\theta}}^{-1}\left[\boldsymbol{I} - \boldsymbol{H}_{\boldsymbol{\theta}}^1 \boldsymbol{Q}^\top\right].$$

It follows from Lemmas 1 and 2 of Aitchison and Silvey (1958) that, under some regularity conditions on the density and the constraint function, the solution to the equation $\frac{\partial L}{\partial \boldsymbol{\theta}} = 0$ (first derivative of the Lagrangian function) exists within the set U_ϵ almost surely and it maximizes the likelihood function subject to the constraint $h(\boldsymbol{\theta}) = 0$. We denote the constrained maximum

likelihood estimator by $\hat{\boldsymbol{\theta}}_n(x)$ and $\hat{\boldsymbol{\alpha}}_{2n}(x)$ for the parameters and Lagrange multiplier, respectively. Then, the following joint asymptotic normality of the estimators of the parameter vector and the Lagrange multiplier (Aitchison and Silvey, 1958)is useful for developing test statistics for testing various constraints:

$$\begin{bmatrix} \sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) \\ \frac{1}{\sqrt{n}}\hat{\alpha}_{2n} \end{bmatrix} \rightarrow N\left(\mathbf{0}, \begin{bmatrix} \boldsymbol{P}_\theta & \mathbf{0} \\ \mathbf{0} & -\boldsymbol{R}_\theta \end{bmatrix}\right). \tag{3.1.12}$$

Some of the requisite regularity conditions for the above results are verified in the Appendix B.1 using the fact that for multivariate normal distribution all the moments exist (Chacón and Duong, 2015). The rest is verified in Luo et al. (2016) for sufficiently large n.

Next, expressing the Taylor series expansion of the first derivative of the Lagrangian function in matrix form, one arrives at the following iterative method (Aitchison and Silvey, 1958), abbreviated as the A&S method, for computing the MLE:

$$\begin{bmatrix} \widehat{\boldsymbol{\theta}}^{(j+1)} \\ \frac{1}{n}\widehat{\boldsymbol{\alpha}}_2^{(j+1)} \end{bmatrix} = \begin{bmatrix} \widehat{\boldsymbol{\theta}}^{(j)} \\ \frac{1}{n}\widehat{\boldsymbol{\alpha}}_2^{(j)} \end{bmatrix} + \begin{bmatrix} \boldsymbol{P}_{1\boldsymbol{\theta}} & \boldsymbol{Q}_{1\boldsymbol{\theta}} \\ \boldsymbol{Q}_{1\boldsymbol{\theta}}^\top & \boldsymbol{R}_{1\boldsymbol{\theta}} \end{bmatrix} \begin{bmatrix} \frac{1}{n}\frac{\partial l(\boldsymbol{\theta}|\boldsymbol{X})}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^{(j)}} + \boldsymbol{H}_{\widehat{\boldsymbol{\theta}}^{(j)}}\frac{1}{n}\widehat{\boldsymbol{\alpha}}_2^{(j)} \\ h(\widehat{\boldsymbol{\theta}}^{(j)}) \end{bmatrix} \tag{3.1.13}$$

where $\begin{bmatrix} \boldsymbol{P}_{1\boldsymbol{\theta}} & \boldsymbol{Q}_{1\boldsymbol{\theta}} \\ \boldsymbol{Q}_{1\boldsymbol{\theta}}^\top & \boldsymbol{R}_{1\boldsymbol{\theta}} \end{bmatrix}$ is the inverse of $\begin{bmatrix} \boldsymbol{B}_{\widehat{\boldsymbol{\theta}}^{(j)}} & -\boldsymbol{H}_{\widehat{\boldsymbol{\theta}}^{(j)}}^{1} \\ -\boldsymbol{H}_{\widehat{\boldsymbol{\theta}}^{(j)}}^{1\top} & \mathbf{0} \end{bmatrix}$ for $j = 0$. An important point to note here is that in the A&S method, the coefficient matrix in the right-hand-side stays the same through the iterations and has to invert a matrix only once.

3.2 Existence and Uniqueness of the Constrained MLE

In this section we study existence and uniqueness of the constrained MLE when the search is limited to convex subsets of the parameter space. It is based on the intuition that if the true parameter belongs to a predetermined random set with high probability (Zwiernik et al., 2017), then an iterations restricted to this set will move closer to the true parameter.

Recall that with the constraint $\boldsymbol{\Sigma\mu} = \boldsymbol{\mu}$, the Lagrangian function is

$$\begin{aligned} L(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \boldsymbol{X}) &= l(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \boldsymbol{X}) + \boldsymbol{\alpha}_2^\top (\boldsymbol{\Sigma\mu} - \boldsymbol{\mu}) \\ &= -\frac{n}{2} \log \mid \boldsymbol{\Sigma} \mid -\frac{1}{2} \mathrm{Tr}\left[\boldsymbol{A}(\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}\right] + \alpha_2^\top (\boldsymbol{\Sigma\mu} - \boldsymbol{\mu}) \\ &= -\frac{n}{2} \log \mid \boldsymbol{\Sigma} \mid -\frac{n}{2} Tr\left[\boldsymbol{S}\boldsymbol{\Sigma}^{-1}\right] - \frac{n}{2} Tr\left[(\bar{\boldsymbol{x}} - \boldsymbol{\mu})(\bar{\boldsymbol{x}} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}\right] + \alpha_2^\top (\boldsymbol{\Sigma\mu} - \boldsymbol{\mu}). \quad (3.2.1) \end{aligned}$$

It is not concave under the constraint on the covariance matrix and may have multiple local maxima. However, we show that the Lagrangian function is concave in any direction in a predefined set of the form $\Delta_{\boldsymbol{A}} = \{\boldsymbol{\Sigma} : \boldsymbol{0} \prec \boldsymbol{\Sigma} \prec \boldsymbol{A}\}$, see Zwiernik et al. (2017). Let $\mathbb{S}^p$ denotes the set of all $p \times p$ real symmetric matrices as a subset of $\mathbb{R}^{\frac{p(p+1)}{2}}$ and $\mathbb{S}^p_{\succ 0}$ denotes the open convex cone in $\mathbb{S}^p$ of positive definite matrices. The following lemma establishes concavity of the profiled Lagrangian function where the mean parameter is estimated by the sample mean for a fixed value of $\boldsymbol{\alpha}_2$.

Lemma 2. *For a given value of $\boldsymbol{\alpha}_2$ and the mean vector $\boldsymbol{\mu}$ estimated by $\bar{\boldsymbol{x}}$, the Lagrangian function $L : \mathbb{S}^p \to \mathbb{R}$ in (3.2.1) is strictly concave in $\boldsymbol{\Sigma}$ in the region $\Delta_{2\boldsymbol{S}}$.*

The proof is given in the Appendix (B.1.4). The strict concavity of the Lagrangian function also guarantees that the covariance matrix where the Lagrangian attains its maximum is unique.

Lemma 3. *If $\boldsymbol{\Sigma}_{max} = arg\max_{\boldsymbol{\Sigma} \in \Delta_{2\boldsymbol{S}}} L(\bar{\boldsymbol{x}}, \boldsymbol{\Sigma})$, then $\boldsymbol{\Sigma}_{max}$ is unique in $\Delta_{2\boldsymbol{S}}$.*

Proof. *Suppose there are two matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ in $\Delta_{2\boldsymbol{S}}$, which maximize the Lagrangian function for a given $\boldsymbol{\alpha}_2$. Then, for the matrix $\boldsymbol{\Sigma}(t) = (1-t)\boldsymbol{\Sigma}_1 + t\boldsymbol{\Sigma}_2$, $t \in [0, 1]$, we have*

$$L(\bar{\boldsymbol{x}}, \boldsymbol{\Sigma}(t)) \geq (1-t)L(\bar{\boldsymbol{x}}, \boldsymbol{\Sigma}_1) + tL(\bar{\boldsymbol{x}}, \boldsymbol{\Sigma}_2) = L(\bar{\boldsymbol{x}}, \boldsymbol{\Sigma}_1)$$

so that $\boldsymbol{\Sigma}(t)$'s also maximizes the Lagrangian function. Therefore there is a direction in which the Lagrangian is not strictly concave contradicting lemma 2. So if the maximizer exists within $\Delta_{2\boldsymbol{S}}$, it is unique.

To analyze the probability that Δ_{2S} contains the true covariance matrix, we rely on the known fact that (Bibby et al., 1979, Theorem 3.4.1) a sample covariance matrix $\boldsymbol{S}$ based on a random sample of $n \geq p$ observations from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, follows a Wishart distribution i.e. $n\boldsymbol{S} \sim \pi_W(n-1, \Sigma)$ and also $\boldsymbol{W}_{n-1} = n\boldsymbol{\Sigma}^{-1/2}\boldsymbol{S}\boldsymbol{\Sigma}^{-1/2} \sim \pi_W(n-1, \boldsymbol{I}_p)$. Then, the probability that $\boldsymbol{\Sigma} \in \Delta_{2S}$ is expressed as follows:

$$\begin{aligned} P\left[\boldsymbol{\Sigma} \in \Delta_{2S}\right] &= P\left[2\boldsymbol{S} - \boldsymbol{\Sigma} \succ \boldsymbol{0}\right] = P\left[2\boldsymbol{\Sigma}^{-1/2}\boldsymbol{S}\boldsymbol{\Sigma}^{-1/2} - \boldsymbol{I}_p \succ \boldsymbol{0}\right] \\ &= P\left[\frac{2}{n}W_{n-1} \succ \boldsymbol{I}_p\right] = P\left[\boldsymbol{W}_{n-1} \succ \frac{n}{2}\boldsymbol{I}_p\right] \\ &= P\left[\lambda_p(\boldsymbol{W}_{n-1}) > \frac{n}{2}\right]. \end{aligned}$$

Interestingly, the probability that the true parameter $\boldsymbol{\Sigma}$ lies within the set Δ_{2S} is independent of $\boldsymbol{\Sigma}$ and is equal to the probability that $\lambda_p(\boldsymbol{W}_{n-1}) > \frac{n}{2}$ where $\boldsymbol{W}_{n-1} \sim \pi_W(n-1, \boldsymbol{I}_p)$. It is known that (Zwiernik et al., 2017) this probability gets closer to 1 as $n, p \to \infty$,$n/p \to \gamma^* < 6 + 4\sqrt{2}$. Thus, for big enough dataset we expect an iterative algorithm, when restricted to this random set, will eventually converge to the constrained MLE.

3.3 An Algorithm for Enforcing the Constraints

Most estimators presented so far do not necessarily satisfy the constraints. In order to obtain an estimator in the desired constrained parameter space (where the constraint is satisfied), we need to post process or modify the estimator even though we did not ignore the constraints in the first part of the computation process described in the previous section. In this section, starting with any reasonable estimators for $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (like those in Sections 3.1), we present an algorithm for modifying them so as to satisfy both constraints in (2.2.1). The notation $M_0 = (\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$ is used from here on to denote any such pre-estimate of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $M_i, i = 1, 2, 3$ for its gradual modifications.

3.3.1 Scale Modifications of the Mean and Covariance Matrix

The modification process starts by the task of modifying the given covariance matrix estimator to accommodate the mean vector estimate. For $p = 3$, a slightly different reparameterization of the covariance matrix is developed in Paine et al. (2018).

Lemma 4. *Given $\tilde{\boldsymbol{\mu}} \in \mathbb{R}^p$ and $\tilde{\boldsymbol{\Sigma}}$ any $p \times p$ positive-definite covariance matrix with the spectral decomposition $\boldsymbol{PDP}^\top$ as in (2.2.2). Set $\boldsymbol{P}_p^* = \frac{\tilde{\boldsymbol{\mu}}}{||\tilde{\boldsymbol{\mu}}||}$ and apply the Gram - Schmidt orthonormalization process to the set of vectors $\{\boldsymbol{P}_p^*, \boldsymbol{P}_{p-1}, \ldots, \boldsymbol{P}_2, \boldsymbol{P}_1\}$ to obtain $\{\boldsymbol{P}_p^*, \ldots, \boldsymbol{P}_2^*, \boldsymbol{P}_1^*\}$. Then, the modified covariance matrix*

$$\widehat{\boldsymbol{\Sigma}}^* = \sum_{j=1}^{p-1} \frac{\lambda_j}{\lambda_{pr}} \boldsymbol{P}_j^* \boldsymbol{P}_j^{*\top} + \boldsymbol{P}_p^* \boldsymbol{P}_p^{*\top} \qquad \textit{where } \lambda_{pr} = \left(\prod_{k=1}^{p-1} \lambda_k \right)^{\frac{1}{p-1}} \tag{3.3.1}$$

satisfies the conditions in (2.2.1).

We denote this estimator by M_1. In Lemma 4, $\tilde{\boldsymbol{\mu}}$ is effectively forced to become an eigenvector corresponding to the eigenvalue 1 of a modified covariance matrix estimator, i.e. $\widehat{\boldsymbol{\Sigma}}^* \tilde{\boldsymbol{\mu}} = \tilde{\boldsymbol{\mu}}$. It turns out that estimators obtained by this simple-minded modification, and inspired by basic linear algebra do not perform well. This is somewhat expected as only the covariance estimator is modified and the mean vector is left intact.

In view of the simultaneous constrains on the mean vector and the covariance matrix, their joint modification seems a natural idea to consider. Next, the mean vector is forced in the direction (span) of the eigenvectors of the covariance estimator. This is implemented by entertaining regression-like models for the given mean vector with the eigenvectors serving as covariates. First, we consider simple linear regressions by choosing a single eigenvector and estimating the corresponding regression coefficient c i.e. $\tilde{\boldsymbol{\mu}} = c\boldsymbol{P}_i$ for some eigenvector P_i.

Lemma 5. *Given $\tilde{\boldsymbol{\mu}} \in \mathbb{R}^p$ and $\tilde{\boldsymbol{\Sigma}}$ a $p \times p$ positive-definite covariance matrix with spectral decomposition $\boldsymbol{PDP}^\top$. Define*

$$c_{0i} = \arg\min_{c\in\mathbb{R}} \|\tilde{\boldsymbol{\mu}} - c\boldsymbol{P}_i\|^2 = \frac{\langle \boldsymbol{P}_i, \tilde{\boldsymbol{\mu}}\rangle}{||\boldsymbol{P}_i||^2} \qquad \textit{and} \qquad i_0 = \arg\min_i \left(1 - \frac{\lambda_i}{c_{0i}^2}\right)^2. \tag{3.3.2}$$

Then, the modified mean-covariance estimators

$$\widehat{\boldsymbol{\mu}}^* = c_{0i_0}\boldsymbol{P}_{i_0}, \quad \widehat{\boldsymbol{\Sigma}}^* = \sum_{\substack{j\neq i_0\\ j=1}}^{p} \frac{\lambda_j}{\lambda_{pr}} \boldsymbol{P}_i\boldsymbol{P}_i^\top + \widehat{\boldsymbol{\mu}}^*\widehat{\boldsymbol{\mu}}^{*\top} \textit{ where } \lambda_{pr} = \left(\prod_{\substack{j\neq i_0\\ j=1}} \lambda_j\right)^{\frac{1}{p-1}} \tag{3.3.3}$$

satisfy (2.2.1).

We refer to the estimator $(\widehat{\boldsymbol{\mu}}^*, \widehat{\boldsymbol{\Sigma}}^*)$ as M_2 in the sequel. The intuition behind the method for selecting i_0 is that from $\widehat{\boldsymbol{\mu}}^*\widehat{\boldsymbol{\mu}}^* = c_{0i}^2\boldsymbol{P}_i\boldsymbol{P}_i^\top$ it is desirable to have the eigenvalue corresponding to $\widehat{\boldsymbol{\mu}}^*$ to be as close as possible to one of the λ_i's. Thus, it is reasonable that λ_i/c_{0i}^2 should be as close to 1 as possible. More details about such selection can be found in Appendix (5).

Modifying the initial estimator jointly using (3.3.3) we obtain $(\widehat{\boldsymbol{\mu}}^*, \widehat{\boldsymbol{\Sigma}}^*)$. Since the covariance matrix is not modified too much it is likely that the mean will suffer too much while the estimator of the covariance will not. In light of this intuition we need to have a balance for joint estimation while satisfying the constraint.

3.3.2 The Modification Algorithm: Multiple Regression

In this section we consider a full-fledged multiple linear modeling of $\tilde{\boldsymbol{\mu}}$ on $\boldsymbol{P}_i$'s. It amounts to a generalization of Lemma 5 and involves variable selection in the context of multiple regression. The details are organized in the following Algorithm 1, where the task is to divide the eigenvectors (regressors) into two groups. We rely on the maximum distance between the consecutive terms of ordered absolute values of the regression coefficients in the saturated model. A viable alternative for this is the 2-means clustering algorithm applied to absolute values of the entries of the vector c of regression coefficients. The estimator from this algorithm is denoted by M_3.

Algorithm 1 Modifying an Estimator to Satisfy (2.2.1)

1: Start with a given $(\tilde{\mu}, \tilde{\Sigma})$ and its spectral decomposition as in (2.2.2)

2: **Variable (Basis) Selection:** Write $\tilde{\mu} = \sum_{j=1}^{p} c_i \boldsymbol{P}_i = \boldsymbol{P}\boldsymbol{c}$ where $c = (c_1, c_2, \ldots, c_p)$.

- Simple Clustering : Viewing the c_i's as weights, select those P_i's which has largest absolute weight by ordering absolute values of c_i's and find out the biggest gap. Let the index set of the group with higher absolute value of c_i be $\mathscr{S} = \{i_1, i_2, \ldots, i_{j_0}\}$ where j_0 is its cardinality.

  OR

- Cluster c_i's by applying K-means clustering with $K = 2$ (Hartigan and Wong, 1979) on absolute values of c_i's.

3: **Regress $\tilde{\mu}$ on the span of columns of $\boldsymbol{P}_{j_0} = [\boldsymbol{P}_{i_1}, \boldsymbol{P}_{i_2}, \ldots, \boldsymbol{P}_{i_{j_0}}]$:**

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\beta} \left\| \tilde{\mu} - \boldsymbol{P}_{j_0}\boldsymbol{\beta} \right\|^2, \quad \widehat{\boldsymbol{\mu}}^* = \boldsymbol{P}_{j_0}\widehat{\boldsymbol{\beta}} \tag{3.3.4}$$

4: **Orthogonalization to accommodate $\widehat{\mu}^*$:** Let $g = \arg\max\{|\ \widehat{\beta}_{k^*}\ | : k^* = i_1, i_2, \ldots, i_{j_0}\}$. Apply the Gram-Schmidt process on $\left\{\boldsymbol{P}_{i_1}, \ldots, \boldsymbol{P}_{i_{g-1}}, \widehat{\boldsymbol{\mu}}^*, \boldsymbol{P}_{i_{g+1}}, \ldots, \boldsymbol{P}_{i_{j_0}}\right\}$ to obtain $\left\{\widehat{\boldsymbol{\mu}}^*, \boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_{j_0-1}\right\}$ with $\widehat{\boldsymbol{\mu}}^*$ as the starting vector.

5: Set,

$$\widehat{\boldsymbol{\Sigma}}^* = \widehat{\boldsymbol{\mu}}^*\widehat{\boldsymbol{\mu}}^{*\top} + \sum_{\substack{j \neq \{i_1, \ldots, i_{j_0}\} \\ j=1}}^{p} \lambda_j \boldsymbol{P}_j \boldsymbol{P}_j^{\top} + \sum_{k=1}^{j_0-1} \lambda_{i_k}^{'} \boldsymbol{b}_k \boldsymbol{b}_k^{\top}$$

estimate $\lambda_{i_k}^{'}$ by

$$\widehat{\lambda}_{i_k} = \boldsymbol{b}_k^{\top} \boldsymbol{\Sigma} \boldsymbol{b}_k, \qquad k = 1, 2, \ldots, j_0 - 1. \tag{3.3.5}$$

(The proof of this step is presented in Appendix 7).

6: Let $\lambda_{pr} = \left(\prod_{\substack{j \notin \mathscr{S} \\ j=1}}^{p} \lambda_j . \prod_{k=1}^{j_0} \widehat{\lambda}_{i_k} \right)^{\frac{1}{p-1}}$. The modified estimator $(\widehat{\boldsymbol{\mu}}^*, \widehat{\boldsymbol{\Sigma}}^*)$ is given by

$$\widehat{\boldsymbol{\mu}}^* = \boldsymbol{P}_{j_0}\widehat{\boldsymbol{\beta}}$$

$$\widehat{\boldsymbol{\Sigma}}^* = \sum_{\substack{j\neq\{i_1,\ldots,i_{j_0+1}\}\\ j=1}}^{p} \frac{\lambda_j}{\lambda_{pr}}\boldsymbol{P}_j\boldsymbol{P}_j^\top + \sum_{k=1}^{j_0}\frac{\widehat{\lambda}_{i_k}}{\lambda_{pr}} b_k b_k^\top + \widehat{\boldsymbol{\mu}}^*\widehat{\boldsymbol{\mu}}^{*\top}. \tag{3.3.6}$$

Our goal is to compute the maximum likelihood estimator and provide a Bayesian framework for estimation of the parameters in (2.6.1). It turns out that computing the maximum likelihood estimator is challenging due to the nonlinearity and intractability of $\boldsymbol{P}(\boldsymbol{u})$. However, the MLE of the eigenvalues and c_0 have closed forms as a function of the mean direction, thus making it possible to compute its profile likelihood. At this stage, we approximate the MLE of the mean direction by first finding a lower bound for the concave profile log-likelihood function of the mean direction (for given $\boldsymbol{\lambda}$) and then maximizing it. In the Bayesian context, the maximum aposteriori (MAP) does not have a closed form, even though the posterior distribution obtained by using a Gaussian prior for the mean vector and inverse gamma on the eigenvalues does. These priors are suggested by extending the shrinkage inverse Wishart prior of Berger et al. (2020) to the case of nonzero means. We show in Section 4.3.2 that one can generate from the posterior distribution quite easily using Metropolis-Hastings within Gibbs sampling. But if we are only interested in point estimation (e.g. MAP), then we propose a simpler way of computing it. The idea is to follow the calculation of the approximate MLE by providing a lower bound for the posterior and maximize it using a modified version of Newton's method.

4.1 MLE of μ and λ in Model (2.6.1)

Let $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ be a sample of size n from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is parameterized as in (2.6.1) with $\boldsymbol{\mu} = c_0\boldsymbol{u}$ and $\boldsymbol{X}$ is the $n \times p$ data matrix. Then, the log-likelihood can be written as

$$l(\boldsymbol{u}, c_0, \boldsymbol{\lambda} \mid \boldsymbol{X}) \propto -\frac{n}{2}\sum_{i=1}^{p-1}\log(\lambda_i) - \frac{1}{2}\left[\sum_{j=1}^{n}(c_0\boldsymbol{u} - \boldsymbol{x}_i)^\top \boldsymbol{\Sigma}^{-1}(c_0\boldsymbol{u} - \boldsymbol{x}_i)\right] \tag{4.1.1}$$

$$\propto -\frac{n}{2}\sum_{i=1}^{p-1}\log(\lambda_i) - \frac{1}{2}\mathrm{Tr}\left[\boldsymbol{D}^{-1}\boldsymbol{B}(\boldsymbol{X}, c_0, \boldsymbol{u})\right]$$

$$\propto -\frac{n}{2}\sum_{i=1}^{p-1}\log(\lambda_i) - \frac{1}{2}\left[\boldsymbol{B}(\boldsymbol{X}, c_0, \boldsymbol{u})_{11} + \sum_{i=1}^{p-1}\frac{\boldsymbol{B}(\boldsymbol{X}, c_0, \boldsymbol{u})_{(i+1)(i+1)}}{\lambda_i}\right], \tag{4.1.2}$$

where $\boldsymbol{B}(\boldsymbol{X}, c_0, \boldsymbol{u}) = \boldsymbol{P}^\top(\boldsymbol{u})\boldsymbol{A}(c_0\boldsymbol{u})\boldsymbol{P}(\boldsymbol{u})$ and $\boldsymbol{A}(\boldsymbol{\mu}) = \sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\top$.

Evidently, maximization of the log-likelihood function with respect to the mean direction is more challenging than those with respect to c_0 and $\boldsymbol{\lambda}$. This suggests and we propose approximating the MLE of the parameters using the following three (non-iterative) steps:

1. Differentiate the log-likelihood function to obtain the MLE of c_0 and λ_i's as a function of the mean direction vector $\boldsymbol{u}$.

2. Compute the profile log-likelihood.

3. Find a lower bound for the profile log-likelihood and maximize it to obtain the approximate MLE of the mean direction $\boldsymbol{u}$. Use the latter to compute the approximate MLE of c_0 and the eigenvalues.

Next, we provide further details about implementing these three steps.

Step 1: For a given $\boldsymbol{u}$, differentiating the log-likelihood in (4.1.1) with respect to c_0 (Bibby et al., 1979, §4.2.9) and in (4.1.2) with respect to λ_i, we obtain

$$\widehat{c}_0 = \frac{\boldsymbol{u}^\top \boldsymbol{P}(\boldsymbol{u})\boldsymbol{D}^{-1}\boldsymbol{P}^\top(\boldsymbol{u})\bar{\boldsymbol{x}}}{\boldsymbol{u}^\top \boldsymbol{P}(\boldsymbol{u})\boldsymbol{D}^{-1}\boldsymbol{P}^\top(\boldsymbol{u})\boldsymbol{u}}, \quad \text{and} \quad \widehat{\lambda}_i = \frac{\boldsymbol{B}(\boldsymbol{X}, c_0, \boldsymbol{u})_{(i+1)(i+1)}}{n}. \tag{4.1.3}$$

The expression for the $\widehat{c}_0$ in (4.1.3) reduces to

$$\widehat{c}_0 = \frac{\boldsymbol{e}_1^\top \boldsymbol{D}^{-1}\boldsymbol{P}^\top(\boldsymbol{u})\bar{\boldsymbol{x}}}{\boldsymbol{e}_1^\top \boldsymbol{D}^{-1}\boldsymbol{e}_1} = \frac{\frac{1}{\lambda_1}\boldsymbol{e}_1^\top \boldsymbol{P}^\top(\boldsymbol{u})\bar{\boldsymbol{x}}}{\frac{1}{\lambda_1}} = \boldsymbol{u}^\top \bar{\boldsymbol{x}}, \tag{4.1.4}$$

and note that the estimates of the eigenvalues are always positive.

Step 2: Substituting for $\widehat{c}_0$ and $\widehat{\lambda}_i$ in the log-likelihood (4.1.2), the profile likelihood of the mean direction $\boldsymbol{u}$ turns out to be

$$l(\boldsymbol{u}, \widehat{c}_0, \widehat{\boldsymbol{\lambda}} \mid \boldsymbol{X}) \propto -\frac{n}{2}\sum_{i=1}^{p-1}\log\left(\frac{\boldsymbol{B}(\boldsymbol{X}, \widehat{c}_0, \boldsymbol{u})_{(i+1)(i+1)}}{n}\right) - \frac{1}{2}\left[\boldsymbol{B}(\boldsymbol{X}, \widehat{c}_0, \boldsymbol{u})_{11} + n(p-1)\right] \tag{4.1.5}$$

Denoting the columns of $\boldsymbol{V}$ by $\boldsymbol{V}_i$ for $i = 1, 2, \ldots, (p-1)$, then the diagonal entries of the $\boldsymbol{B}(\boldsymbol{X}, \widehat{c}_0, \boldsymbol{u})$ are

$$\left\{\boldsymbol{B}(\boldsymbol{X}, \widehat{c}_0, \boldsymbol{u})_{11}, \boldsymbol{B}(\boldsymbol{X}, \widehat{c}_0, \boldsymbol{u})_{22}, \ldots, \boldsymbol{B}(\boldsymbol{X}, \widehat{c}_0, \boldsymbol{u})_{p,p}\right\} \tag{4.1.6}$$

$$= \left\{\boldsymbol{u}^\top \boldsymbol{A}(\widehat{c}_0 \boldsymbol{u})\boldsymbol{u}, \boldsymbol{V}_1^\top \boldsymbol{A}(\widehat{c}_0 \boldsymbol{u})\boldsymbol{V}_1, \ldots, \boldsymbol{V}_{p-1}^\top \boldsymbol{A}(\widehat{c}_0 \boldsymbol{u})\boldsymbol{V}_{p-1}\right\}. \tag{4.1.7}$$

Since $\boldsymbol{u} \perp \boldsymbol{V}_i$ for $i = 1, 2, \ldots, p-1$, one may further simplify the diagonal entries of $\boldsymbol{B}(\boldsymbol{X}, \widehat{c}_0, \boldsymbol{u})$ to the following:

$$\begin{aligned} \boldsymbol{V}_i^\top \boldsymbol{A}(\widehat{c}_0 \boldsymbol{u})\boldsymbol{V}_i &= \boldsymbol{V}_i^\top \left[\boldsymbol{A}(\boldsymbol{0}) - n\widehat{c}_0 \bar{\boldsymbol{x}} \boldsymbol{u}^\top - n\widehat{c}_0 \boldsymbol{u} \bar{\boldsymbol{x}}^\top + n\widehat{c}_0^2 \boldsymbol{u}\boldsymbol{u}^\top\right] \boldsymbol{V}_i = \boldsymbol{V}_i^\top \boldsymbol{A}(\boldsymbol{0})\boldsymbol{V}_i \\ \boldsymbol{u}^\top \boldsymbol{A}(\widehat{c}_0 \boldsymbol{u})\boldsymbol{u} &= \boldsymbol{u}^\top \boldsymbol{A}(\bar{\boldsymbol{x}})\boldsymbol{u}. \end{aligned} \tag{4.1.8}$$

In spite of this simplification, the profile log-likelihood is hard to differentiate as a function of $\boldsymbol{u}$, in general, when we are not assuming any specific form for the matrix $\boldsymbol{V}$.

Step 3: We find a workable lower bound for (4.1.5) and maximize it with respect to the mean direction to obtain an approximate MLE for $\boldsymbol{u}$. Alternatively, this amounts to assuming that the first sum in the profile likelihood (2.7) is constant. The expression in equation (4.1.8) is a quadratic form of the matrix $\boldsymbol{A}(\boldsymbol{0}) = \sum_{j=1}^{n} \boldsymbol{x}_j \boldsymbol{x}_j^\top$ and can be bounded by its largest eigenvalue (Rao, 1973, §1f.2.1). Consequently, using (4.1.8) and that columns of $\boldsymbol{V}$ are orthonormal simplify the profile log-likelihood (4.1.5) and we arrive at the following lower bound:

$$\begin{aligned} l(\boldsymbol{u}, \widehat{c}_0, \boldsymbol{\lambda} \mid \boldsymbol{X}) &\propto -\frac{n}{2} \sum_{i=1}^{p-1} \log\left[\frac{\boldsymbol{V}_i^\top \boldsymbol{A}(\boldsymbol{0})\boldsymbol{V}_i}{n}\right] - \frac{1}{2}\left[\boldsymbol{u}^\top \boldsymbol{A}(\bar{\boldsymbol{x}})\boldsymbol{u} + n(p-1)\right] \\ &\geq -\frac{n(p-1)}{2} \log\left[\frac{\lambda_1\{\boldsymbol{A}(\boldsymbol{0})\}}{n}\right] - \frac{1}{2}\left[\boldsymbol{u}^\top \boldsymbol{A}(\bar{\boldsymbol{x}})\boldsymbol{u} + n(p-1)\right] = h(\boldsymbol{u}) \end{aligned} \tag{4.1.9}$$

Evidently, the lower bound denoted by $h(\boldsymbol{u})$ is maximized by minimizing the quadratic expression inside the second bracket. Since $\boldsymbol{u} \neq 0$ and h is a quadratic function of $\boldsymbol{u}$, the maximum

occurs when $\boldsymbol{u}$ is the eigenvector corresponding to the smallest eigenvalue of the matrix $\boldsymbol{A}(\bar{\boldsymbol{x}})$. As such the approximate MLE of $\boldsymbol{u}$ is unique only up to a sign.

Even though the computed mean direction $\widehat{\boldsymbol{u}}$ is not the exact MLE, it is still a good approximation as confirmed by the simulation results in Table 5.4 of Section 5.2. The approximate maximum likelihood estimate of the eigenvalues and the constant c_0 are obtained by plugging in the approximate MLE of $\boldsymbol{u}$ in (4.1.3) and (4.1.4). Fortunately, this idea of approximatiing MLE can also be replicated in our posterior MAP estimator approximation developed in the next section, but the maximization of the lower bound is not as straightforward and requires employing a version of Newton's iterative method.

We can find out the distribution of the estimates. We know that (Bibby et al., 1979, §3.4.2)

$$\boldsymbol{A}(c_0\boldsymbol{u}) \mid c_0, \boldsymbol{u} \sim \text{Wishart}_p\left(\boldsymbol{\Sigma}, n\right) \tag{4.1.10}$$

From equation 4.1.6 and Bibby et al. (1979, §Theorem 3.4.2), we can also show that

$$\begin{gathered} n\widehat{\lambda}_i = \boldsymbol{P}_i(\boldsymbol{u})^\top \boldsymbol{A}(c_0\boldsymbol{u})\boldsymbol{P}_i(\boldsymbol{u}) \mid c_0, \boldsymbol{u} \sim \text{Wishart}_1\left(\boldsymbol{P}_i(\boldsymbol{u})^\top \boldsymbol{\Sigma}\boldsymbol{P}_i(\boldsymbol{u}), n\right) \\ \left.\frac{n\widehat{\lambda}_i}{\boldsymbol{P}_i(\boldsymbol{u})^\top \boldsymbol{\Sigma}\boldsymbol{P}_i(\boldsymbol{u})} \right| c_0, \boldsymbol{u} \sim \chi^2_n \end{gathered} \tag{4.1.11}$$

The estimate of $\boldsymbol{u}$ is the eigenvector corresponding to smallest eigenvalue of the matrix $\boldsymbol{A}(\bar{\boldsymbol{x}}) = n\boldsymbol{S}$ where $\boldsymbol{S}$ is the MLE of the covariance matrix for data coming from a normal distribution. Finding the exact distribution of $\boldsymbol{u}$ is challenging. But Anderson (2003, §Theorem 13.3.3) has a nice result, which is useful in this regard. The statement is the following:

Theorem 2. *Suppose* $\boldsymbol{A} \sim Wishart_p(\boldsymbol{I}, n)$ *and* $\boldsymbol{P} = \left[\boldsymbol{P}_1, \boldsymbol{P}_2, \ldots, \boldsymbol{P}_p\right]$ *is the matrix of normalized eigenvectors of* $\boldsymbol{A}$ *with* $P_{1i} \geq 0$. *Then* $\boldsymbol{C} = \boldsymbol{P}^\top$ *has the conditional Haar invariant distribution and* $\boldsymbol{C}$ *is distributed independently of the eigenvalues.*

However, the asymptotic distribution of the MLE approximation $\widehat{\boldsymbol{u}}$ which is the eigenvector

corresponding to the smallest eigenvalue of the matrix $\boldsymbol{A}(\bar{\boldsymbol{x}}) = n\boldsymbol{S}$ is known, see (Anderson, 2003, §Theorem 13.5.1). From this theorem we can say that

$$\sqrt{n}\left(\widehat{\boldsymbol{u}} - \boldsymbol{P}_p^*\right) \to N_p\left(\mathbf{0}, \sum_{k=1,k\neq p}^{p} \frac{\lambda_p \lambda_k}{\left(\lambda_p - \lambda_k\right)^2} \boldsymbol{P}_k^* \boldsymbol{P}_k^{*\top}\right)$$

when the data generating covariance matrix $\boldsymbol{\Sigma}$ has $\boldsymbol{P}_1^*, \boldsymbol{P}_2^*, \ldots, \boldsymbol{P}_p^*$ as the eigenvectors corresponding to the ordered eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p$.

4.2 A Structured Covariance Model Without Joint Constraint

In this section, we modify our problem to better understand the structured covariance model by moving away from the joint constraint and model the covariance matrix directly under a similar constraint. Here the constraint is on the covariance matrix only i.e.

$$\boldsymbol{\Sigma}\boldsymbol{u} = \boldsymbol{u} \tag{4.2.1}$$

where $\boldsymbol{u}$ is some unknown vector and not the mean unlike our original setup. For the sake of simplicity we assume that the mean vector $\boldsymbol{\mu} = 0$ and $\boldsymbol{u}$ is NOT related to $\boldsymbol{\mu}$ anymore. We are going to show that under this set up we can still use the structured covariance modeling with the covariance matrix $\boldsymbol{\Sigma}$ being a function of $\boldsymbol{u}$ and compute the MLE with almost identical computation. In that sense, this can be seen as a corollary of the computation of section 4.1.

To emphasize the difference, the model in (2.6.1) still holds in this setup under the constraint (4.2.1). The orthogonal matrix $\boldsymbol{P}$ calculated in the same way as before as a function of $\boldsymbol{u}$ because the structured covariance model treats $\boldsymbol{u}$ as a vector without using any property of $\boldsymbol{\mu}$. In this sense the constraint is no longer a joint constraint and we are left with a functional covariance model.

4.2.1 MLE of u and λ in Model (4.2.1)

Let $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ be a sample of size n from $N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is parameterized as in (2.6.1) with $\boldsymbol{X}$ being the $n \times p$ data matrix. Unfortunately the task of maximization of the likelihood function is difficult since it involves computation of the derivative (a nonstandard quadratic form due to the dependence of $\boldsymbol{\Sigma}$ on $\boldsymbol{u}$) of $\boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{u})\boldsymbol{x}$ with respect to $\boldsymbol{u}$. Then, the log-likelihood can be written as

$$l(\boldsymbol{u}, \boldsymbol{\lambda} \mid \boldsymbol{X}) \propto -\frac{n}{2}\sum_{i=1}^{p-1}\log(\lambda_i) - \frac{1}{2}\left[\sum_{j=1}^{n} \boldsymbol{x}_i^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{x}_i\right] \tag{4.2.2}$$

$$\propto -\frac{n}{2}\sum_{i=1}^{p-1}\log(\lambda_i) - \frac{1}{2}\mathrm{Tr}\left[\boldsymbol{D}^{-1}\boldsymbol{B}(\boldsymbol{X}, \boldsymbol{u})\right]$$

$$\propto -\frac{n}{2}\sum_{i=1}^{p-1}\log(\lambda_i) - \frac{1}{2}\left[\boldsymbol{B}(\boldsymbol{X}, \boldsymbol{u})_{11} + \sum_{i=1}^{p-1}\frac{\boldsymbol{B}(\boldsymbol{X}, \boldsymbol{u})_{(i+1)(i+1)}}{\lambda_i}\right], \tag{4.2.3}$$

where $\boldsymbol{B}(\boldsymbol{X}, \boldsymbol{u}) = \boldsymbol{P}^\top(\boldsymbol{u})\boldsymbol{A}(\boldsymbol{0})\boldsymbol{P}(\boldsymbol{u})$ and $\boldsymbol{A}(\boldsymbol{0}) = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top = n\boldsymbol{S}$ as before, where $\boldsymbol{S}$ is the unconstrained MLE of the covariance matrix of normal distribution.

Evidently, maximization of the log-likelihood function with respect to $\boldsymbol{u}$ is more challenging than that with respect to $\boldsymbol{\lambda}$. This suggests and we propose approximating the MLE of the parameters using the following three (non-iterative) steps:

1. Differentiate the log-likelihood function to obtain the MLE of λ_i's as a function of $\boldsymbol{u}$.

2. Compute the profile log-likelihood.

3. Find a lower bound for the profile log-likelihood and maximize it to obtain the approximate MLE of $\boldsymbol{u}$. Use the latter to compute the approximate MLE of the eigenvalues.

Next, we provide further details about implementing these three steps.

Step 1: For a given $\boldsymbol{u}$, differentiating the log-likelihood in (4.2.3) with respect to λ_i, we obtain

$$\widehat{\lambda}_i = \frac{\boldsymbol{B}(\boldsymbol{X}, \boldsymbol{u})_{(i+1)(i+1)}}{n}. \tag{4.2.4}$$

Note that the estimates of the eigenvalues are always non-negative.

Step 2: Substituting for $\widehat{\lambda}_i$ in the log-likelihood (4.2.3), the profile likelihood of $\boldsymbol{u}$ turns out to be

$$l(\boldsymbol{u}, \widehat{\boldsymbol{\lambda}} \mid \boldsymbol{X}) \propto -\frac{n}{2} \sum_{i=1}^{p-1} \log\left(\frac{\boldsymbol{B}(\boldsymbol{X}, \boldsymbol{u})_{(i+1)(i+1)}}{n}\right) - \frac{1}{2}\left[\boldsymbol{B}(\boldsymbol{X}, \boldsymbol{u})_{11} + n(p-1)\right] \tag{4.2.5}$$

Denoting the columns of $\boldsymbol{V}$ by $\boldsymbol{V}_i$ for $i = 1, 2, \ldots, (p-1)$, then the diagonal entries of the $\boldsymbol{B}(\boldsymbol{X}, \boldsymbol{u})$ are

$$\left\{\boldsymbol{B}(\boldsymbol{X}, \boldsymbol{u})_{11}, \boldsymbol{B}(\boldsymbol{X}, \boldsymbol{u})_{22}, \ldots, \boldsymbol{B}(\boldsymbol{X}, \boldsymbol{u})_{p,p}\right\} = \left\{\boldsymbol{u}^\top \boldsymbol{A}(\boldsymbol{0})\boldsymbol{u}, \boldsymbol{V}_1^\top \boldsymbol{A}(\boldsymbol{0})\boldsymbol{V}_1, \ldots, \boldsymbol{V}_{p-1}^\top \boldsymbol{A}(\boldsymbol{0})\boldsymbol{V}_{p-1}\right\}.$$

In spite of this, the profile log-likelihood is hard to differentiate as a function of $\boldsymbol{u}$, in general, when we are not assuming any specific form for the matrix $\boldsymbol{V}$.

Step 3: We find a workable lower bound for (4.2.5) and maximize it with respect to $\boldsymbol{u}$ to obtain an approximate MLE for $\boldsymbol{u}$. Alternatively, this amounts to assuming that the first sum in the profile likelihood (4.2.6) is constant. The expression $\boldsymbol{V}_i^\top \boldsymbol{A}(\boldsymbol{0})\boldsymbol{V}_i$ is a quadratic form of the matrix $\boldsymbol{A}(\boldsymbol{0}) = \sum_{j=1}^{n} \boldsymbol{x}_j \boldsymbol{x}_j^\top$ and can be bounded by its largest eigenvalue (Rao, 1973, §1f.2.1). Consequently, we arrive at the lower bound of the profile log-likelihood in (4.2.5) as follows:

$$\begin{aligned} l(\boldsymbol{u}, \widehat{c}_0, \boldsymbol{\lambda} \mid \boldsymbol{X}) &\propto -\frac{n}{2} \sum_{i=1}^{p-1} \log\left[\frac{\boldsymbol{V}_i^\top \boldsymbol{A}(\boldsymbol{0})\boldsymbol{V}_i}{n}\right] - \frac{1}{2}\left[\boldsymbol{u}^\top \boldsymbol{A}(\boldsymbol{0})\boldsymbol{u} + n(p-1)\right] \\ &\geq -\frac{n(p-1)}{2} \log\left[\frac{\lambda_1\{\boldsymbol{A}(\boldsymbol{0})\}}{n}\right] - \frac{1}{2}\left[\boldsymbol{u}^\top \boldsymbol{A}(\boldsymbol{0})\boldsymbol{u} + n(p-1)\right] = h(\boldsymbol{u}) \end{aligned} \tag{4.2.6}$$

Evidently, the lower bound denoted by $h(\boldsymbol{u})$ is maximized by minimizing the quadratic expression inside the second bracket. Since $\boldsymbol{u} \neq 0$ and h is a quadratic function of $\boldsymbol{u}$, the maximum occurs when $\boldsymbol{u}$ is the eigenvector corresponding to the smallest eigenvalue of the matrix $\boldsymbol{A}(\boldsymbol{0})$. As such the approximate MLE of $\boldsymbol{u}$ is unique only up to a sign.

Even though the computed value of $\widehat{\boldsymbol{u}}$ is not the exact MLE, it is still a good approximation. The approximate maximum likelihood estimate of the eigenvalues are obtained by plugging in the approximate MLE of $\boldsymbol{u}$ in (4.2.4).

In this context, we can find out the distribution of MLE approximation of the unknown parameters in the model. We know that $\boldsymbol{x}_i \sim N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$ which tells us that

$$\begin{aligned} \boldsymbol{A}(\boldsymbol{0}) &= \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top \sim \text{Wishart}_p(\boldsymbol{\Sigma}, n) \\ n\widehat{\lambda}_i = \boldsymbol{P}_i^\top(\boldsymbol{u})\boldsymbol{A}(\boldsymbol{0})\boldsymbol{P}_i(\boldsymbol{u}) \mid \boldsymbol{u} &\sim \text{Wishart}_1\left(\boldsymbol{P}_i^\top(\boldsymbol{u})\boldsymbol{\Sigma}\boldsymbol{P}_i(\boldsymbol{u}), n\right) \\ \frac{n\widehat{\lambda}_i}{\boldsymbol{P}_i(\boldsymbol{u})^\top \boldsymbol{\Sigma} \boldsymbol{P}_i(\boldsymbol{u})} \,\Bigg|\, \boldsymbol{u} &\sim \chi^2_n \end{aligned} \tag{4.2.7}$$

The main difference between this computation and the one in Section 4.1 are the following:

1. In Section 4.1 $\boldsymbol{u}$ is a function of the mean direction $\boldsymbol{\mu}$. Here $\boldsymbol{\mu} = 0$ and $\boldsymbol{u}$ is an unknown quantity not related to $\boldsymbol{\mu}$.

2. The expression of MLE of λ_i: Previously it was a function of the diagonal entries of matrix $\boldsymbol{B}(\boldsymbol{X}, c_0, \boldsymbol{u})$. Here the corresponding matrix is $\boldsymbol{B}(\boldsymbol{X}, \boldsymbol{u})$. Since $\boldsymbol{\mu} = 0$, there is no c_0 here and the $\boldsymbol{B}$ matrix is a function of $\boldsymbol{A}(\boldsymbol{0})$ instead of $\boldsymbol{A}(c_0\boldsymbol{u})$.

3. The $\boldsymbol{P}$ matrix remains the same as before.

4. The simplification in equation (4.1.8) is no longer required and the first diagonal entry of the $\boldsymbol{B}$ matrix involves $\boldsymbol{A}(\boldsymbol{0})$ instead of $\boldsymbol{A}(\bar{\boldsymbol{x}})$.

5. Finally the estimate of $\boldsymbol{u}$ is the eigenvector corresponding to the lowest eigenvalue of the matrix $\boldsymbol{A}(\boldsymbol{0})$ instead of $\boldsymbol{A}(\bar{\boldsymbol{x}})$.

4.3 Bayesian Estimation of (μ, λ) in 2.6.1

Here we propose a new prior on $(\boldsymbol{\mu}, \boldsymbol{\lambda})$ in 2.6.1 and provide the details of calculation of MAP estimators in two ways.

4.3.1 The Mean-Eigenvalue Priors

A convenient prior distribution on the mean vector is the multivariate normal and the inverse gamma on the ordered eigenvalues. The latter prior comes naturally from the shrinkage inverse Wishart prior of (Berger et al., 2020, §3.2) as the conditional distribution of eigenvalues given the matrix of eigenvectors (see section 2.6.3 for detailed discussion), see also Hoff (2009a) (section 3.3). More specifically, our proposed prior is:

$$\boldsymbol{\mu} \mid \boldsymbol{D} \sim N_p(\boldsymbol{\mu}_0, \boldsymbol{D}/\kappa_0) \text{ and } \boldsymbol{\lambda}_i \sim \text{Inverse-gamma}(a-1, c_i/2), \tag{4.3.1}$$

where a, c_i's are the hyperparameters. Setting $\boldsymbol{H}_0 = \text{diag}(1, c_1, c_2, \ldots, c_{p-1})$, then the posterior distribution has the form

$$p(\boldsymbol{\mu}, \boldsymbol{\lambda} \mid \boldsymbol{X}) \propto \left(\prod_{i=1}^{p-1} \lambda_i \right)^{-\frac{n+1+2a}{2}} \exp\left[-\frac{1}{2} Tr\left\{ \boldsymbol{D}^{-1} \boldsymbol{H}_N \right\} \right] \tag{4.3.2}$$

with $\boldsymbol{H}_N = \boldsymbol{B}(\boldsymbol{X}, \boldsymbol{\mu}) + \kappa_0(\boldsymbol{\mu} - \boldsymbol{\mu}_0)(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^\top + \boldsymbol{H}_0$. Here $\boldsymbol{B}(\boldsymbol{X}, \boldsymbol{\mu})$ has the same expression as $\boldsymbol{B}(\boldsymbol{X}, c_0, \boldsymbol{u})$ which appeared in Section 2.

One needs to generate from this posterior distribution when computing the Bayesian point estimates, credible intervals and various other quantities. We show that Metropolis-Hastings within Gibbs sampling is possible here and suitable for our goals (Brooks et al., 2011, §1.12.10).

4.3.2 Gibbs Sampling

The Gibbs sampler (Gelfand and Smith, 1990) method is a numerical technique for sampling from the joint posterior distribution. Given an initial vector, the Gibbs sampling proceeds by sampling from the conditional posterior distribution. More generally, if the full conditional posterior distribution in any Gibbs step is of a non-standard form, using MH (Metropolis-Hastings) step is convenient (Brooks et al., 2011). The technique is known as Metropolis-Hastings within Gibbs Sampling or alternatively as single-component Metropolis-Hastings sampling as follows:

1. Since $p(\boldsymbol{D} \mid \boldsymbol{\mu}, \boldsymbol{X}) \propto p(\boldsymbol{\mu}, \boldsymbol{\lambda} \mid \boldsymbol{X})$ has an Inverse-Gamma distribution, the orderd eigenvalues

are generated from independent Inverse Gamma $\left(\frac{n+2a-1}{2}, c_i^*/2\right)$ where c_i^* is the i-th diagonal entry of $\boldsymbol{H}_N$ matrix.

2. Generate from $p(\boldsymbol{\mu} \mid \boldsymbol{D}, \boldsymbol{X})$ using the MH algorithm with a proposal distribution q selected as

$$q\left(. \mid \boldsymbol{\mu}^{(i-1)}\right) = N_p\left(\boldsymbol{\mu}^{(i-1)}, \frac{1}{n}\boldsymbol{P}\left(\boldsymbol{\mu}^{(i-1)}\right)\boldsymbol{D}\boldsymbol{P}^{\top}\left(\boldsymbol{\mu}^{(i-1)}\right)\right).$$

More concretely, the steps of MH within Gibbs algorithm in our context are as follows:

Algorithm 2 MH within Gibbs for generating samples from the posterior distribution

1: Start with $\boldsymbol{\mu}^{(0)} = \bar{\boldsymbol{x}}$.
2: **Repeat** s **times:** $j-th$ step
3: Generate $\boldsymbol{\lambda}_i^{(j)} \sim IG\left(\frac{n+2a-1}{2}, c_i^*/2\right)$ to form $\boldsymbol{D}^{(j)}$.
4: Start with $\boldsymbol{\mu}^{(j-1)}$, **Repeat MH Step** l **times:** $k-th$ step
5: Generate $\boldsymbol{\mu}^* \sim q\left(. \mid \boldsymbol{\mu}^{(k-1)}\right)$
6: Calculate

$$r\left(\boldsymbol{\mu}^*, \boldsymbol{\mu}^{(k-1)}\right) = \min\left\{1, \frac{p\left(\boldsymbol{\mu}^*, \boldsymbol{\lambda}^{(j)} \mid \boldsymbol{X}\right) q\left(\boldsymbol{\mu}^{(k-1)} \mid \boldsymbol{\mu}^*\right)}{q\left(\boldsymbol{\mu}^* \mid \boldsymbol{\mu}^{(k-1)}\right) p\left(\boldsymbol{\mu}^{(k-1)}, \boldsymbol{\lambda}^{(j)} \mid \boldsymbol{X}\right)}\right\}$$

7: Generate $u \sim U(0,1)$
8: Set $\boldsymbol{\mu}^{(k)} = \boldsymbol{\mu}^*$ if $u < r\left(\boldsymbol{\mu}^*, \boldsymbol{\mu}^{(k-1)}\right)$ else $\boldsymbol{\mu}^{(k)} = \boldsymbol{\mu}^{(k-1)}$
9: Set $\boldsymbol{\mu}^{(j)} = \boldsymbol{\mu}^{(l)}$
10: Collect $(\boldsymbol{\mu}^{(j)}, \boldsymbol{\lambda}^{(j)})$ for $j = 1, 2, \ldots, s$
11: **End**

4.3.3 Approximation of MAP through a Lower Bound

The Gibbs sampling is computationally challenging and the time complexity increases exponentially with the dimensions. We follow the steps Section 2 to approximate the MAP estimator

using a lower bound for the log posterior density. The main difference is in the maximization step of the lower bound as a function of $\boldsymbol{u}$ where it is not as straightforward as the MLE case and does not have a closed-form. So, we resort to an iterative modified Newton-Raphson algorithm. Using $\boldsymbol{\mu} = c_0\boldsymbol{u}$ where $c_0 \in \mathbb{R}$ as before we arrive at

$$\log p(\boldsymbol{u}, c_0, \boldsymbol{\lambda} \mid \boldsymbol{X}) \propto -t\sum_{i=1}^{p-1}\log\lambda_i - \frac{1}{2}\left[f(c_0) + \operatorname{Tr}(\boldsymbol{D}^{-1}\boldsymbol{H}_0)\right] \tag{4.3.3}$$

$$\propto -t\sum_{i=1}^{p-1}\log\lambda_i - \frac{1}{2}\left[(\boldsymbol{H}_N)_{11} + \sum_{i=1}^{p-1}\frac{(\boldsymbol{H}_N)_{i+1,i+1}}{\lambda_i}\right] \tag{4.3.4}$$

where $t = \frac{n+1+2a}{2}$ and $f(c_0) = \sum_{i=1}^{n}(\boldsymbol{x}_i - c_0\boldsymbol{u})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - c_0\boldsymbol{u}) + \kappa_0(c_0\boldsymbol{u} - \boldsymbol{\mu}_0)^\top\boldsymbol{D}^{-1}(c_0\boldsymbol{u} - \boldsymbol{\mu}_0)$. Following similar calculations as in (4.1.3), the estimates of c_0 and the eigenvalues in terms of the mean direction are as follows:

$$\widehat{c_0} = \frac{n\boldsymbol{u}^\top\bar{\boldsymbol{x}} + \kappa_0\boldsymbol{u}^\top\boldsymbol{D}^{-1}\boldsymbol{\mu}_0}{n + \kappa_0\boldsymbol{u}^\top\boldsymbol{D}^{-1}\boldsymbol{u}}, \qquad \widehat{\lambda}_i = \frac{(\boldsymbol{H}_N)_{i+1,i+1}}{2t} \tag{4.3.5}$$

where $\boldsymbol{H}_N = \boldsymbol{B}(\boldsymbol{X}, c_0, \boldsymbol{u}) + \kappa_0(c_0\boldsymbol{u} - \boldsymbol{\mu}_0)^\top\boldsymbol{D}^{-1}(c_0\boldsymbol{u} - \boldsymbol{\mu}_0) + \boldsymbol{H}_0$. Here the expression of $\widehat{c_0}$ and $\widehat{\lambda}_i$ are intertwined. So we are doing the following calculation with the aim to propose an iterative algorithm for the MAP estimators. We will use equation (4.3.5) and (4.3.12) as our updating equation.

Using the expression of $\widehat{\lambda}_i$ from equation (4.3.5) in the log posterior likelihood function we obtain the following analogue of the profile likelihood in (4.1.5):

$$\log p(\boldsymbol{u}, c_0, \widehat{\boldsymbol{\lambda}} \mid \boldsymbol{X}) \propto -t\sum_{i=1}^{p-1}\log\left(\frac{(\boldsymbol{H}_N)_{i+1,i+1}}{2t}\right) - \frac{1}{2}\left[(\boldsymbol{H}_N)_{11} + (p-1)t\right] \tag{4.3.6}$$

where the diagonal entries of $\boldsymbol{H}_N$ are

$$(\boldsymbol{H}_N)_{i+1,i+1} = \begin{cases} \boldsymbol{u}^\top\boldsymbol{A}(c_0\boldsymbol{u})\boldsymbol{u} + \kappa_0(c_0\boldsymbol{u} - \boldsymbol{\mu}_0)_1^2 + (\boldsymbol{H}_0)_{1,1} & \text{for } i = 0 \\ \boldsymbol{V}_i^\top\boldsymbol{A}(0)\boldsymbol{V}_i + \kappa_0(c_0\boldsymbol{u} - \boldsymbol{\mu}_0)_{i+1}^2 + (\boldsymbol{H}_0)_{i+1,i+1} & \text{for } i \neq 0 \text{ and } \forall c_0 \end{cases}. \tag{4.3.7}$$

This follows from the expression of the matrix $\boldsymbol{B}(\boldsymbol{X}, c_0, \boldsymbol{u})$ in equation in (4.1.8) and the definition of $\boldsymbol{H}_N$. We observe the following

$$\begin{aligned}(\boldsymbol{H}_N)_{i+1,i+1} &\leq \lambda_1\{\boldsymbol{A}(\boldsymbol{0})\} + \kappa_0\|c_0\boldsymbol{u} - \boldsymbol{\mu}_0\|^2 + (\boldsymbol{H}_0)_{i+1,i+1} \\ &= m_i + \kappa_0\|c_0\boldsymbol{u} - \boldsymbol{\mu}_0\|^2 \quad \text{for } i \neq 0 \text{ and } \forall c_0 \in \mathbb{R}. \end{aligned} \tag{4.3.8}$$

For a given value of c_0, equation (4.3.6), provides a lower bound $h(\boldsymbol{u})$, which we will maximize to approximate $\boldsymbol{u}$. This is equivalent to minimizing $-h(\boldsymbol{u})$. We apply Newton - Raphson algorithm (Lange, 2013) to obtain an update for $\boldsymbol{u}$. The calculation of derivatives are shown below.

$$\log p(\boldsymbol{u}, c_0, \widehat{\boldsymbol{\lambda}} \mid \boldsymbol{X}) \geq h(\boldsymbol{u}),$$

$$-h(\boldsymbol{u}) = t\sum_{i=1}^{p-1}\log\left[\frac{m_i + \kappa_0\|c_0\boldsymbol{u} - \boldsymbol{\mu}_0\|^2}{2t}\right] + \frac{1}{2}\left[\boldsymbol{u}^\top\boldsymbol{A}(c_0\boldsymbol{u})\boldsymbol{u} + \kappa_0\|c_0\boldsymbol{u} - \boldsymbol{\mu}_0\|_1^2 + m_1\right], \tag{4.3.9}$$

$$-\nabla h(\boldsymbol{u}) = t\sum_{i=1}^{p-1}\frac{4t\kappa_0 c_0(c_0\boldsymbol{u} - \boldsymbol{\mu}_0)}{m_i + \kappa_0\|c_0\boldsymbol{u} - \boldsymbol{\mu}_0\|^2} + [\boldsymbol{A}(0)\boldsymbol{u} - c_0 n\bar{\boldsymbol{x}} + \kappa_0 c_0(c_0\boldsymbol{u} - \boldsymbol{\mu}_0)], \tag{4.3.10}$$

$$\begin{aligned}-\nabla^2 h(\boldsymbol{u}) = t\sum_{i=1}^{p-1}&\frac{4t\kappa_0 c_0\left[(m_i + \kappa_0\|c_0\boldsymbol{u} - \boldsymbol{\mu}_0\|^2)\boldsymbol{I} - 2(c_0\boldsymbol{u} - \boldsymbol{m}\boldsymbol{u}_0)(c_0\boldsymbol{u} - \boldsymbol{\mu}_0)^\top\right]}{\{m_i + \kappa_0\|c_0\boldsymbol{u} - \boldsymbol{\mu}_0\|^2\}^2}, \\ &+\left[\boldsymbol{A}(0) + \kappa_0 c_0^2(c_0\boldsymbol{u} - \boldsymbol{\mu}_0)(c_0\boldsymbol{u} - \boldsymbol{\mu}_0)^\top\right] \end{aligned} \tag{4.3.11}$$

Thus, the updating equation for the mean direction is

$$\boldsymbol{u}^{(k+1)} = \boldsymbol{u}^{(k)} - \left[\nabla^2 h(\boldsymbol{u})\right]^{-1}\nabla h(\boldsymbol{u}) \tag{4.3.12}$$

Due to the concavity of the lower bound, Newton's method is an appealing choice in our case. However there are two potential problems with Newton's method (Lange, 2013, §10.3). First, it may be computationally expensive to invert the second derivative matrix in each step. Second, the Newton's method is not really a ascent algorithm in the sense that $h\left(\boldsymbol{u}^{(k+1)}\right) > h\left(\boldsymbol{u}^{(k)}\right)$ for a concave function. The second problem can be remedied by modifying the increment such that it is

a partial step in the ascent direction. Let us denote:

$$\boldsymbol{v} = \left[\nabla^2 h(\boldsymbol{u})\right]^{-1} \nabla h(\boldsymbol{u}) \tag{4.3.13}$$

The idea is to take a sufficiently short increment in the direction of $\boldsymbol{v}$. If $\left[\boldsymbol{u}^{(k)} - \alpha\boldsymbol{v}\right]$ shows increment in $h(\boldsymbol{u})$ value, we update our mean vector in the iteration, otherwise we look at $\left[\boldsymbol{u}^{(k)} - \alpha^j\boldsymbol{v}\right]$ for $j = 1, 2, \ldots$ until we observe an increment. Due to the good performance of the MLE, we use the MLE of c_0 and $\boldsymbol{u}$ as our initial values which make the convergence fast. Based on (4.3.5) and (4.3.12), an iterative algorithm for computing the approximate MAP is summarized in the following algorithm:

Algorithm 3 MAP Approximation

Initialize: Start with $c_0^{(0)} =$ MLE of c_0, $\boldsymbol{u}^{(0)} =$ MLE of $\boldsymbol{u}$ and $\lambda_i^{(0)} = \frac{H_N\left(c^{(0)}, \boldsymbol{u}^{(0)}\right)_{i+1,i+1}}{2t}$
For $k \to (k+1)$
    Update $c^{(k+1)}$ and $\lambda_i^{(k+1)}$ from equation (4.3.5)
    For $j \to (j+1)$
        $\boldsymbol{u}^{(j)} = \boldsymbol{u}^{(k)}$ and $\boldsymbol{v}^{(j)} = \left[\nabla^2 h(\boldsymbol{u}^{(j)})\right]^{-1} \nabla h(\boldsymbol{u}^{(j)})$
        For $l \to (l+1)$
            $\boldsymbol{u}^{(j+1)} = \boldsymbol{u}^{(j)} - \alpha^l \boldsymbol{v}^{(j)}$
            If $h(\boldsymbol{u}^{(j+1)}) > h(\boldsymbol{u}^{(j)})$
                Accept the update of $\boldsymbol{u}$
                Break
            Else Change $\alpha^l \to \alpha^{l+1}$
        If $\|\boldsymbol{u}^{(j+1)} - \boldsymbol{u}^{(j)}\| < \epsilon$
            $\boldsymbol{u}^{(k+1)} = \boldsymbol{u}^{(j+1)}$
            Break
    If $h(\boldsymbol{u}^{k+1}) < h(\boldsymbol{u}^{(k)})$
    Break
Get the final value of c_0, $\boldsymbol{\lambda}$ and $\boldsymbol{u}$.

18: **End**

This algorithm produces the MAP approximation of the eigenvalues and the mean vector. The continuity of the estimate of the covariance matrix, shown in Lemma 6, ensures that the covariance estimate will converge with the convergence of the mean direction vector $\boldsymbol{u}$.

Lemma 6. *If $\boldsymbol{V}(\boldsymbol{u})$, described after (2.6.1) is continuous, then the approximate MAP of the covariance matrix $\boldsymbol{\Sigma}$ is also a continuous function of $\boldsymbol{u}$.*

Proof. *If $\boldsymbol{V}(\boldsymbol{u})$ is a continuous function of $\boldsymbol{u}$, then each column of $\boldsymbol{V}$, $\boldsymbol{V}_j$ is also continuous. This implies diagonal entries of $\boldsymbol{H}_N$ and in turn the MAP estimate of the eigenvalues are also a continuous function of $\boldsymbol{u}$. The claim follows since the MAP of $\boldsymbol{\Sigma}$ given by*

$$\widehat{\boldsymbol{\Sigma}} = \boldsymbol{u}\boldsymbol{u}^\top + \sum_{i=1}^{p-1} \widehat{\lambda}_i \boldsymbol{V}_i \boldsymbol{V}_i^\top = \boldsymbol{u}\boldsymbol{u}^\top + \sum_{i=1}^{p-1} \frac{(\boldsymbol{H}_N)_{i+1,i+1}}{n} \boldsymbol{V}_i \boldsymbol{V}_i^\top \tag{4.3.14}$$

is sum of continuous functions of $\boldsymbol{u}$.

5. SIMULATIONS

In this section, we are going to describe the simluation studies for the methods described in section 3 and section 4 both in two subsections.

5.1 Simulation Experiments for Iterative Methods with Lagrange Multiplier in Section 3

Through several simulation experiments, we assess the performance of the following three iterative methods and our modified estimators: 1. Standard MLE, 2. Standard MLE with explicit calculation of Lagrange multiplier (denoted by S&C), 3. The Aitchison and Silvey (1958) iterative method (denoted by A&S) as described in Section 3.

5.1.1 The Simulation Set up:

We have taken sample size and dimension to be $(n, p) = (50, 5), (50, 25), (100, 10), (300, 30)$. Risks are approximated by averaging the losses for 100 independent replications in each of the four combinations of (n, p). In all cases the data generation mechanism and the risk function are kept the same, we have used Frobenius loss as our default loss function and calculated Stein's loss in some specific cases.

For the parameters of the Gaussian distributions used for data generation we take the entries of the mean vector $\boldsymbol{\mu}$ to be values of independent standard Gaussian variables. For the covariance matrix, we start with $\boldsymbol{\Psi} = \boldsymbol{L}\boldsymbol{L}^\top$ where $\boldsymbol{L}$ is a lower triangular matrix with the diagonal entries generated from $N(5, 1)$ and standard normal for the off-diagonal entries. The larger diagonal entries of $\boldsymbol{L}$ ensure positive-definiteness of $\boldsymbol{\Sigma}$. Since such $(\boldsymbol{\mu}, \boldsymbol{\Psi})$ do not necessarily satisfy conditions (2.2.1), the covariance matrix is modified first by applying (3.3.1) to $(\boldsymbol{\mu}, \boldsymbol{\Psi})$.

The performance of the estimators is assessed using the scaled L_2 risk (Ledoit and Wolf, 2004b, §3.1):

$$R(\boldsymbol{\mu}, \widehat{\boldsymbol{\mu}}^*) = \mathbb{E}\left[\frac{1}{p}\left\|\widehat{\boldsymbol{\mu}}^* - \boldsymbol{\mu}\right\|_{\mathscr{F}}^2\right] \qquad , \qquad R(\boldsymbol{\Sigma}, \widehat{\boldsymbol{\Sigma}}^*) = \mathbb{E}\left[\frac{1}{p}\left\|\widehat{\boldsymbol{\Sigma}}^* - \boldsymbol{\Sigma}\right\|_{\mathscr{F}}^2\right],$$

where $\widehat{\boldsymbol{\mu}}^*$ and $\widehat{\boldsymbol{\Sigma}}^*$ are the final modified estimators described in Section 3.3.

5.1.2 Simulation Results from the Three Iterative Methods:

5.1.2.1 The Standard MLE:

The iterative method for computing the maximum likelihood estimator described in Section 3.1.1 does not always converge. Since convergence of the four sets of parameters simultaneously is unlikely, the convergence criterion used here is to stop iterations if at least two of the parameters converge. In most cases the iterations for $\boldsymbol{\mu}$ and α_1 converge, but the rate of decrease of Frobenius risk for estimating $\boldsymbol{\Sigma}$ is slow in successive iteration. In the simulations we have taken the maximum number of iterations to be 1000. When the convergence does not happen after 1000 iterations, we take the output at the 1000-th iteration as the estimator and pass it through the Algorithm 1 for M_3 to arrive at the final estimator. This method referred to as the standard MLE (SMLE), involves iterative updating of the Lagrange multipliers. In contrast, the next two iterative methods involve exact calculation of the Lagrange multiplier.

5.1.2.2 The S&C Method:

The S&C method is described in Section 3.1.2. It does not guarantee the positive definiteness of the estimate of the covariance matrix. Thus, we only take the cases where the estimate is positive definite for the risk calculation, otherwise the corresponding simulation run is ignored (see Table 5.1)

Table 5.1: No of Times the Estimate is Positive Definite

n	**p**	**No of cases with positive definite covariance estimate**
50	5	90
50	25	100
100	10	99
300	30	100

Moreover, the method does not guarantee exact satisfaction of the constraints, so we apply the Algorithm 1 to the estimates using M_3 with two types of clustering, they produce similar results with K-Means clustering performing slightly better. Since convergence is a recurring issue, we have taken the maximum number of iterations in both the loops of the "double iteration" to be 100, and the value of ϵ to be 0.1. From the Table 5.2 we can see that the S&C method is losing very little while achieving the satisfaction of the joint constraint (2.2.1).

5.1.2.3 The A&S Method:

The iterative method for calculation of the constrained MLE described in Section 3.2 operates inside a closed ball of radius $\delta = \|\boldsymbol{\theta}^{(0)}\|$ around the true parameter. Hence choosing a good initial value for the iterations to run is essential and here we chose $\boldsymbol{\theta}^{(0)} = \left(\bar{x}^\top, \text{vec}(S)^\top\right)^\top$.

Suppose in the i-th stage we have the value of the parameter vector to be $\boldsymbol{\theta}_0^{(i)} = \boldsymbol{a}$ and in the $(i+1)$-th step it moves to a point $\boldsymbol{\theta}_0^{(i+1)} = \boldsymbol{b}$ outside the ball. Then, we find the point $\boldsymbol{c} = (1-t)\boldsymbol{\theta}^{(0)} + t\boldsymbol{b}$ with $t = \frac{\delta}{\|\boldsymbol{\theta}^{(0)} - \boldsymbol{b}\|}$ resides on the ball, and continue the iteration with the new point $\boldsymbol{c}$ instead of $\boldsymbol{b}$. This can be seen from the Figure 5.1.

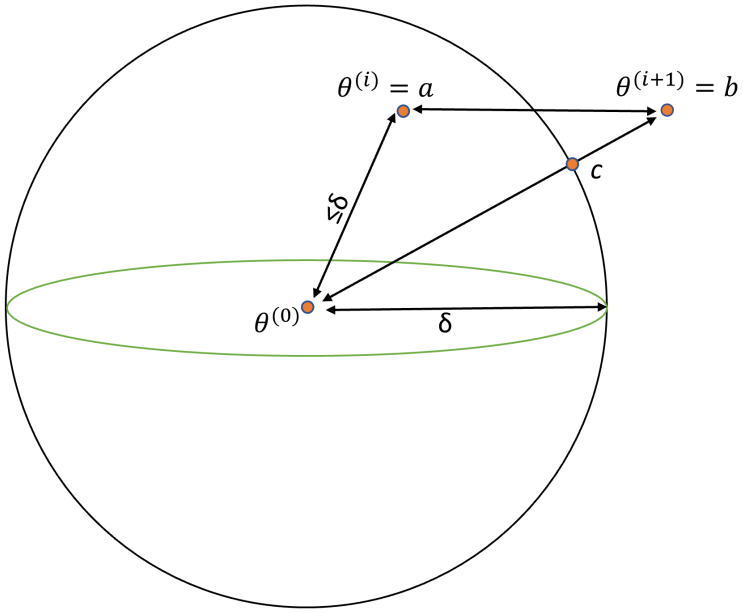

Figure 5.1: This pictorial representation shows how we update when the iteration goes outside the ball in Aitchison and Silvey (1958) method

In each iteration we symmetrize the update for the covariance matrix. We take only those simulation runs where iterations produce a positive definite output. The interesting part is that the positive definiteness is achieved after symmetrization in almost all. The performance is close to the S&C method as can be seen from Table 5.2.

Table 5.2: Risks for the three iterative methods of finding constrained MLE, modified by Algorithm 1 (M3) with K-Means. SMLE: standard MLE; S&C is the method of Strydom and Crowther (2012), and A&S denotes the method of Aitchison and Silvey (1958).

			Mean		**Sigma - Frobenius**	
Method	n	p	MLE	*M3*	MLE	*M3*
SMLE	50	5	0.5313	0.5463	1.2632	0.4212
S&C			0.4817	0.5206	0.3553	0.3057
A&S			0.4389	0.6296	1.2246	0.6097
SMLE	50	25	0.7998	0.8	6.1632	1.5567
S&C			0.7913	0.8158	1.073	1.5678
A&S			0.1691	0.8643	2.4889	2.3021
SMLE	100	10	0.6709	0.673	2.3789	0.4617
S&C			0.6468	0.6849	0.3797	0.3963
A&S			0.2814	0.6797	1.5451	0.8660
SMLE	300	30	0.8071	0.8072	4.9507	0.5507
S&C			0.8014	0.8238	0.5117	0.5477
A&S			0.1575	0.9639	2.1536	0.8122

5.1.3 An Example: Estimates of the Historic Position of Earths Magnetic Pole

The dataset collected by Schmidt (1976) contains the site mean direction estimates of the Earth's historic magnetic pole from 33 different sites in Tasmania. The longitude and latitudes from the data set is transformed to $X_1, X_2, \ldots, X_{33}$ on a three dimensional unit sphere (Preston and Paine, 2017). The angualr gaussian distribution family is the marginal directional component of a multivariate normal distribution with ESAG distribution as a subfamily. Paine et al. (2018) provided strong evidence in favor of ESAG distribution which satisfy the constraint over isotropic angular gaussian distribution while analysing this dataset. This inspire us to make normality as-

sumption under the constraint similar to ESAG distribution disregaring the spherical nature of the tansformed dataset. The constrained maximum likelihood estimate calculated using the numerical method by Aitchison and Silvey (1958) with 1000 iterations is:

$$\mu = \begin{bmatrix} -0.593 & 0.167 & 0.787 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 0.670 & 0.235 & -0.299 \\ 0.235 & 2.333 & -0.106 \\ -0.299 & -0.106 & 0.797 \end{bmatrix}$$

which is comparable to the maximum likelihood estimate calculated using eeliptically symmetric angular Gaussian distribution with a specific parametrization in three dimension (Paine et al., 2018). One main advantage is that our calculation is not restricted to three dimension.

5.2 Simulation Experiments for the Estimation of the Structured Covariance Model in Section 4

In this section, we perform several simulations to assess the performance of the structured covariance model described in Section 4. We focus on mainly on the following three cases:

1. MLE approximation through lower bound maximization (see section 4.1) of the profile likelihood function,

2. MAP estimator with normal-inverse gamma prior (see section 4.3.1) approximated through Gibbs sampling in section 4.3.2, and

3. MAP approximation (see section 4.3.3) through posterior lower bound

by computing their risks. Then the three estimators are compared with the MAP estimator with normal-inverse Wishart prior (see section 2.6.2). We have used RStudio 1.3.1093 and R 4.0.3 on a 64 bit 4 Core Windows 10 laptop for all our simulations.

5.2.1 The Simulation Set up:

We have kept the data generation mechanism and the risk function same as in Section 5.1 with the sample sizes and dimensions to be $n = 50, 100, 300$ and $p = 3, 5, 10$, respectively. The risks

are approximated by averaging the losses for 100 independent replications in each of the nine combinations of (n, p). The parameters of the Gaussian distributions used for data generation are selected in the following way:

- the entries of the mean vector $\boldsymbol{\mu}$ are taken to be independent standard Gaussian variables,
- the covariance matrix is generated from $\boldsymbol{\Psi} = \boldsymbol{L}\boldsymbol{L}^\top$ where

$$L = \begin{cases} L_{ij} \sim N(0,1) & \text{for } i \neq j, \\ L_{ij} \sim N(5,1) & \text{for } i = j. \end{cases}$$

  The larger diagonal entries of $\boldsymbol{L}$ ensure positive-definiteness of the modified covariance matrix $\boldsymbol{\Sigma}$.

- Since such $(\boldsymbol{\mu}, \boldsymbol{\Psi})$ does not necessarily satisfy conditions (2.2.1), the covariance matrix $\boldsymbol{\Sigma}$ is constructed from applying Lemma 4 of Kundu and Pourahmadi (2020) on $(\boldsymbol{\mu}, \boldsymbol{\Psi})$.

The performance of the estimators is assessed using the Frobenius (scaled L_2) risk as in Ledoit and Wolf (2004b, §3.1):

$$R(\boldsymbol{\mu}, \widehat{\boldsymbol{\mu}}^*) = \mathbb{E}\left[\frac{1}{p}\left\|\widehat{\boldsymbol{\mu}}^* - \boldsymbol{\mu}\right\|_{\mathscr{F}}^2\right] \qquad , \qquad R(\boldsymbol{\Sigma}, \widehat{\boldsymbol{\Sigma}}^*) = \mathbb{E}\left[\frac{1}{p}\left\|\widehat{\boldsymbol{\Sigma}}^* - \boldsymbol{\Sigma}\right\|_{\mathscr{F}}^2\right],$$

where $\widehat{\boldsymbol{\mu}}^*$ and $\widehat{\boldsymbol{\Sigma}}^*$ are the final estimators in each of the cases (see Tables C.1 and C.2). However, we have also used Förstner and Moonen (2003) distance metric to assess the performance of the estimate of the covariance matrix. The distance between two symmetric and positive definite matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ is defined by (denoted by "MF" in the following tables):

$$d(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2) = \sqrt{\sum_{i=1}^{p} ln^2 \lambda_i(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2)}$$

where $\lambda_i(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2)$ are the eigenvalues of the matrix $\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1}$.

In this thesis we are dealing with Gaussian distribution. Hence for assessing the performance of the estimator we have also calculated the Kullback-Leibler divergence between the original distribution and the estimated one. The formula for KL divergence between $P_1 = N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $P_2 = N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ is as follows Duchi (2007):

$$D(P_1 \,||\, P_2) = \frac{1}{2}\left[\log\left(\frac{|\,\boldsymbol{\Sigma}_2\,|}{|\,\boldsymbol{\Sigma}_1\,|}\right) - n + \mathrm{Tr}\left(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\right) + (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)\right]$$

5.2.2 Simulation Results:

Here we provide the details of the simulation studies for the MLE approximation, MAP estimator computed through Gibbs sampling and using a lower bound of the posterior density. We have used normal-inverse Wishart as a yardstick to compare the risk of our estimators (MLE approximation and the two MAP's).

5.2.2.1 Comparison between constrained and unconstrained MLE

The normal-inverse Wishart prior is very common in statistical literature, it is still in the Bayesian paradigm. From a frequentist perspective, the unconstrained MLE of Gaussian distribution i.e. $(\bar{\boldsymbol{x}}, \boldsymbol{S})$ might be another standard for comparing our estimator. Therefore, we have first computed the ratio of risks (L_2 loss for mean and Förstner and Moonen (2003) loss for $\boldsymbol{\Sigma}$) between the unconstrained MLE and constrained approximate MLE.

From this simulation (see Table 5.3), we can see that the constrained estimation of mean vector performs really well and the constrained estimator of the covariance matrix performs equivalent for smaller values of n. However, when n is very large, the sample covariance is generally a good estimator and the constrained estimator for covariance suffers.

Table 5.3: Risk ratio of MLE approximation (through a lower bound) relative to unconstrained MLE i.e. $(\bar{\boldsymbol{x}}, \boldsymbol{S})$

n	**p**	$\boldsymbol{\mu}$ L_2	$\boldsymbol{\Sigma}$ **(MF)**	**Time(sec)**
50	3	0.1578	0.9238	7.12
50	5	0.118	1.0297	7.46
50	10	0.1571	1.1425	7
100	3	0.1242	1.0483	6.79
100	5	0.1183	1.3464	7
100	10	0.1389	1.5521	7.41
300	3	0.1425	1.3915	7.56
300	5	0.1206	2.1531	7.28
300	10	0.1335	2.6936	7.54

5.2.2.2 MLE and MAP Approximation using a Lower Bound

The structured covariance matrix defined in equation (2.6.1) does not allow us to calculate the maximum likelihood estimate and the MAP estimator directly due to the intractability of the likelihood function. Hence, we approximate them through a lower bound described in Sections 4.1 and 4.3.3, respectively. The approximation of MLE performs well.

A modified version of the Newton-Raphson method (described in the paragraph following equation (4.3.12)) converges relatively fast i.e. two to five iterations for MAP when initialized with the approximation of MLE of c_0 and $\boldsymbol{u}$.

Table 5.4: Risk ratio of MLE and MAP approximation (through a lower bound) relative to the MAP of normal-inverse Wishart (using L_2 norm for mean, MF norm for covariance and KL Divergence)

		MLE Approx. Risk Ratio		**MAP Approx. Risk Ratio**		**Kullback - Leibler Divergence**		**Time (Sec.)**
n	**p**	μ	Σ **(MF)**	μ	Σ **(MF)**	**MLE**	**MAP**	
50	3	0.233	0.7734	0.2796	0.8658	0.4631	0.6956	6.53
50	5	0.2169	0.8137	0.2169	0.9489	0.4531	0.7469	6.35
50	10	0.3615	0.8177	0.3615	0.9447	0.3684	0.7221	6.5
100	3	0.2019	0.9955	0.2143	1.0466	0.7066	0.846	6.33
100	5	0.2057	1.0999	0.2166	1.1676	0.8775	1.1166	6.49
100	10	0.3396	1.2577	0.3396	1.2849	1.0195	1.3628	6.62
300	3	0.2005	1.2115	0.2005	1.2248	1.1582	1.2089	7.1
300	5	0.1942	1.9514	0.1942	1.9401	2.8434	2.9394	6.66
300	10	0.3434	2.4899	0.3434	2.4244	4.3709	4.547	6.91

The performance of these approximations (both MLE and MAP), assessed by the risk, are better than the Gibbs sampling technique discussed in the next section with the added benefit of being significantly faster (see the last column of Table 5.4). The usual observations like the decline of risk as n increases and p decreases are generally true. One interesting observation is that the performance of MLE approximation is not significantly different from the MAP approximation.

5.2.2.3 Gibbs Sampling

The risk of the MAP estimator is computed through Gibbs sampling as described in Section 4.3.2. Let us denote the initial values to be $\left(\boldsymbol{\mu}^{(0)}, \boldsymbol{\lambda}^{(0)}\right)$. We have chosen $\boldsymbol{\mu}^{(0)} = \boldsymbol{\mu}_0 = \bar{\boldsymbol{x}}, \boldsymbol{H}_0 = \boldsymbol{I}, \kappa_0 = 1.5$ and $a = (p+1)$. The first iteration takes in $\boldsymbol{\mu}^{(0)}$ and calculates the value of $\boldsymbol{H}_N$ as described after equation (4.3.2). Using this, we can generate the value of $\boldsymbol{\lambda}^{(1)}$ from the distribution $p(\boldsymbol{D} \mid \boldsymbol{\mu}, \boldsymbol{X})$ (see step 1 of Gibbs sampling in section 4.3.2) which by construction makes sure that

each eigenvalue is nonnegative. We can easily see that the initial value of $\boldsymbol{\lambda}$ i.e. $\boldsymbol{\lambda}^{(0)}$ does not affect the Gibbs sampling.

In the MCMC, the chain for the eigenvalues depend on the previous iteration only through the generation of mean vector. This is happening because the parameters for the conditional distribution of the eigenvalues are $(n+2a-1)/2$ and $c_i^*/2$ where c_i^*'s are the diagonal entries of the $\boldsymbol{H}_N$ matrix and hence a function of the mean vector.

To illustrate the algorithm, we have generated pairs of $(\boldsymbol{\mu}, \boldsymbol{\lambda})$ from the posterior distribution by MH within Gibbs sampling. We find out the posterior maximizing $(\boldsymbol{\mu}, \boldsymbol{\lambda})$ pair by exhaustive computation of posterior density at each pair. This is possible because the posterior density has a closed form barring the normalizing constants. We take the value of the mean vector from the posterior maximizing pair, set it as our MAP estimate for the mean vector discarding the corresponding $\boldsymbol{\lambda}$ and make a final update on $\boldsymbol{\lambda}$. Given the estimate of the mean vector, the posterior conditional distribution of $\boldsymbol{\lambda}$ is inverse gamma and has a known mode. Using this, we can calculate the best eigenvalue vector $\widehat{\boldsymbol{\lambda}}$ to be $\widehat{\lambda}_i = \frac{c_i^*}{n+1+2a}$ for $i = 1, 2, \ldots, (p-1)$. The newly calculated value $\widehat{\lambda}$, accompanied by the MAP estimate of $\boldsymbol{\mu}$ constitute our MAP estimator $\left(\widehat{\mu}, \widehat{\lambda}\right)$. The performance of such an estimator is assessed through a simulation study (see Table 5.5).

We have generated 100 samples from the posterior distribution using the MH within Gibbs algorithm described in section 4.3.2. We repeat this experiment 100 times and approximate the risk. The performance of the Gibbs sampling is displayed in the following Table 5.5. We can observe that the performance is improving as the number of observations (n) increase. For low dimensional cases the risk of the MAP estimator of the covariance matrix calculated from Gibbs sampling is equivalent and better as compared to its normal-inverse Wishart counterpart.

The time complexity of the MAP approximation increases with the increase in dimension demonstrated by Figure 5.2. The acceptance rate of the MH step within Gibbs sampling is reasonable and decreases with in increment of dimension.

Table 5.5: The risk ratio of MAP approximation (from MH within Gibbs sampling) relative to the MAP of normal-inverse Wishart (using L_2 norm for mean and MF norm for covariance)

		Normal - Inverse Gamma Risk Ratio			**AccProb**	**Time (Sec.)**
n	**p**	μ (L_2)	Σ (**MF**)	**KL Div**		
50	3	1.1173	0.7851	0.7	0.4135	980.52
50	5	1.0824	0.981	1.0976	0.2822	2029.15
50	10	1.2423	0.8042	0.774	0.1218	13574.59
100	3	1.0542	0.9613	0.9677	0.4352	910.92
100	5	1.1677	0.9824	0.9179	0.2908	1901.41
100	10	1.1963	1.2124	1.6632	0.1245	12183.56
300	3	1.1661	0.8392	1.0116	0.4287	1256.81
300	5	1.5729	1.4729	2.2182	0.2895	3215.97
300	10	1.5533	3.0386	10.9107	0.1294	9385

Uncertainty Quantification: In Gibbs sampling we did a small experiment to better understand the uncertainty in the model using the credible interval in a two dimensional model. We select $n = 100$, and $p = 2$. We know the structure of the covariance matrix for $p = 2$ from Appendix C.2 when we select the mean vector to be $\boldsymbol{\mu} = (1, 1)^\top$ and $\rho = 0.2$. We generate 1000 samples from the posterior distribution to find out the following credible interval in Table 5.6:

Table 5.6: Credible Interval for $p = 2$

	μ_1	μ_2	λ
Model Values	1	1	0.6667
Lower Bound	0.7840	0.7668	0.5402
Upper Bound	1.150	1.1245	0.9274

Time for Gibbs Sampling In order to understand how the time to run the Gibbs sampler related to the dimension of the data, we have run a small experiment. In this experiment we have generated 100 data points of various dimensions and looked at the time taken to find out the MAP estimator. We have generated 10 samples from the posterior distribution and repeated the experiment three times. Under this set up the time taken to run the Gibbs sampler is plotted below in Figure 5.2.

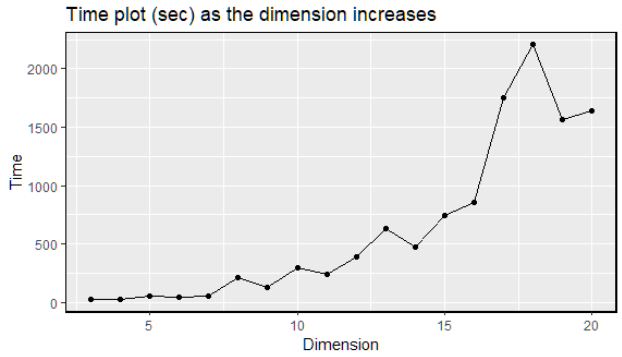

Figure 5.2: Time taken to run the Gibbs sampler and calculate the estimate

5.2.2.4 An Example: Estimates of the Historic Position of Earths Magnetic Pole

The dataset collected by Schmidt (1976) contains the site mean direction estimates of the Earth's historic magnetic pole collected from 33 different sites in Tasmania. The longitude and latitudes from the data set is transformed to $X_1, X_2, \ldots, X_{33}$ on a three dimensional unit sphere (Preston and Paine, 2017). The angualr gaussian distribution family is the marginal directional

component of a multivariate normal distribution with ESAG distribution as a subfamily. Paine et al. (2018) provided strong evidence in favor of ESAG distribution which satisfy the constraint over isotropic angular Gaussian distribution while analyzing this dataset. This inspire us to make normality assumption under the constraint similar to ESAG distribution disregarding the spherical nature of the transformed dataset. The constrained maximum likelihood estimate under the structured covariance model is:

$$\widehat{\boldsymbol{u}}^{\top} = \begin{bmatrix} -0.44 & 0.32 & 0.75 \end{bmatrix}, \quad \widehat{\boldsymbol{\Sigma}} = \begin{bmatrix} 0.95 & -0.01 & -0.03 \\ -0.01 & 1.06 & -0.03 \\ -0.03 & -0.03 & 1.00 \end{bmatrix}$$

which are comparable to the maximum likelihood estimate calculated using elliptically symmetric angular Gaussian distribution with the parametrization in three dimension (Paine et al., 2018) i.e. $\widehat{\boldsymbol{u}}^{\top} = (-0.56, 0.24, 0.79)$ and $\widehat{\Sigma} = \begin{bmatrix} 1.31 & -0.55 & 0.40 \\ -0.55 & 0.85 & -0.34 \\ 0.40 & -0.34 & 1.39 \end{bmatrix}$

REFERENCES

Aitchison, J. and Silvey, S. (1958). Maximum-likelihood estimation of parameters subject to restraints. *The annals of mathematical Statistics*, pages 813–828.

Anderson, T. (2003). An introduction to multivariate statistical analysis (wiley series in probability and statistics). *July 11*.

Anderson, T. W. (1973). Asymptotically efficient estimation of covariance matrices with linear structure. *The Annals of Statistics*, pages 135–141.

Anderson, T. W. and Olkin, I. (1985). Maximum-likelihood estimation of the parameters of a multivariate normal distribution. *Linear algebra and its applications*, 70:147–171.

Athey, S. and Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, 11:685–725.

Aubry, A., De Maio, A., Pallotta, L., and Farina, A. (2012). Maximum likelihood estimation of a structured covariance matrix with a condition number constraint. *IEEE Transactions on Signal Processing*, 60(6):3004–3021.

Baladandayuthapani, V., Talluri, R., Ji, Y., Coombes, K. R., Lu, Y., Hennessy, B. T., Davies, M. A., and Mallick, B. K. (2014). Bayesian sparse graphical models for classification with application to protein expression data. *The annals of applied statistics*, 8(3):1443.

Banerjee, O., El Ghaoui, L., and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *The Journal of Machine Learning Research*, 9:485–516.

Barnard, J., McCulloch, R., and Meng, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, pages 1281–1311.

Bartlett, M. S. (1934). On the theory of statistical regression. *Proceedings of the Royal Society of Edinburgh*, 53:260–283.

Bashir, A., Carvalho, C. M., Hahn, P. R., and Jones, M. B. (2019). Post-processing posteriors over

precision matrices to produce sparse graph estimates. *Bayesian Analysis*, 14(4):1075–1090.

Berger, J. O., Pericchi, L. R., Ghosh, J., Samanta, T., De Santis, F., Berger, J., and Pericchi, L. (2001). Objective bayesian methods for model selection: introduction and comparison. *Lecture Notes-Monograph Series*, pages 135–207.

Berger, J. O., Sun, D., Song, C., et al. (2020). Bayesian analysis of the covariance matrix of a multivariate normal distribution with a new class of priors. *Annals of Statistics*, 48(4):2381–2403.

Bhatia, R. (2007). *Perturbation bounds for matrix eigenvalues*. SIAM.

Bibby, J., Kent, J., and Mardia, K. (1979). *Multivariate analysis*. Academic Press, London.

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

Bishop, Y. M., Fienberg, S. E., and Holland, P. W. (2007). *Discrete multivariate analysis: theory and practice*. Springer Science & Business Media.

Bondell, H. D. and Reich, B. J. (2012). Consistent high-dimensional bayesian variable selection via penalized credible regions. *Journal of the American Statistical Association*, 107(500):1610–1624.

Breaux, H. J. (1967). On stepwise multiple linear regression. Technical report, ARMY BALLISTIC RESEARCH LAB ABERDEEN PROVING GROUND MD.

Brooks, S., Gelman, A., Jones, G., and Meng, X.-L. (2011). *Handbook of markov chain monte carlo*. CRC press.

Brown, P. J., Le, N. D., and Zidek, J. V. (1994). Inference for a covariance matrix. *Aspects of Uncertainty: A Tribute to DV Lindley*, pages 77–92.

Butte, A. J., Tamayo, P., Slonim, D., Golub, T. R., and Kohane, I. S. (2000). Discovering functional relationships between rna expression and chemotherapeutic susceptibility using relevance networks. *Proceedings of the National Academy of Sciences*, 97(22):12182–12186.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480.

Chacón, J. E. and Duong, T. (2015). Efficient recursive algorithms for functionals based on higher order derivatives of the multivariate gaussian density. *Statistics and Computing*, 25(5):959–974.

Chakraborty, M. and Ghosal, S. (2021). Convergence rates for bayesian estimation and testing in monotone regression. *Electronic Journal of Statistics*, 15(1):3478–3503.

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

Chiu, T. Y., Leonard, T., and Tsui, K.-W. (1996). The matrix-logarithmic covariance model. *Journal of the American Statistical Association*, 91(433):198–210.

Cook, R. D. and Zhang, X. (2015). Foundations for envelope models and methods. *Journal of the American Statistical Association*, 110(510):599–611.

Cox, D. R. and Wermuth, N. (2014). *Multivariate dependencies: Models, analysis and interpretation*. Chapman and Hall/CRC.

Daniels, M. J. (2005). A class of shrinkage priors for the dependence structure in longitudinal data. *Journal of statistical planning and inference*, 127(1-2):119–130.

Daniels, M. J. and Kass, R. E. (1999). Nonconjugate bayesian estimation of covariance matrices and its use in hierarchical models. *Journal of the American Statistical Association*, 94(448):1254–1263.

Daniels, M. J. and Kass, R. E. (2001). Shrinkage estimators for covariance matrices. *Biometrics*, 57(4):1173–1184.

Daniels, M. J. and Pourahmadi, M. (2002). Bayesian analysis of covariance matrices and dynamic models for longitudinal data. *Biometrika*, 89(3):553–566.

Dellaportas, P., Forster, J. J., and Ntzoufras, I. (2002). On bayesian model and variable selection using mcmc. *Statistics and Computing*, 12(1):27–36.

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

Dey, D. K., Srinivasan, C., et al. (1985). Estimation of a covariance matrix under stein's loss. *The Annals of Statistics*, 13(4):1581–1591.

Doornik, J. A. (2009). Econometric model selection with more variables than observations. Tech-

nical report, Working paper, Economics Department, University of Oxford.

Duchi, J. (2007). Derivations for linear algebra and optimization. *Berkeley, California*, 3(1):2325–5870.

Dunson, D. B. and Neelon, B. (2003). Bayesian inference on order-constrained parameters in generalized linear models. *Biometrics*, 59(2):286–295.

Edwards, D. (2012). *Introduction to graphical modelling*. Springer Science & Business Media.

Efron, B. et al. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency). *The Annals of Statistics*, 3(6):1189–1242.

Eguchi, N., Saito, R., Saeki, T., Nakatsuka, Y., Belikov, D., and Maksyutov, S. (2010). A priori covariance estimation for co2 and ch4 retrievals. *Journal of Geophysical Research: Atmospheres*, 115(D10).

Engle, R. F., Ledoit, O., and Wolf, M. (2019). Large dynamic covariance matrices. *Journal of Business & Economic Statistics*, 37(2):363–375.

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

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911.

Fernandez, C., Ley, E., and Steel, M. F. (2001). Benchmark priors for bayesian model averaging. *Journal of Econometrics*, 100(2):381–427.

Flom, P. L. and Cassell, D. L. (2007). Stopping stepwise: Why stepwise and similar selection methods are bad, and what you should use. *NorthEast SAS Users Group (NESUG): Statistics and Data Analysis*.

Förstner, W. and Moonen, B. (2003). A metric for covariance matrices. In *Geodesy-the Challenge of the 3rd Millennium*, pages 299–309. Springer.

Fox, E. and Dunson, D. (2011). Bayesian nonparametric covariance regression. *arXiv preprint arXiv:1101.2017*.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the

graphical lasso. *Biostatistics*, 9(3):432–441.

Gaurav, D., Rodriguez, F. O., Tiwari, S., and Jabbar, M. (2021). Review of machine learning approach for drug development process. In *Deep Learning in Biomedical and Health Informatics*, pages 53–77. CRC Press.

Gelfand, A. E. and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409.

Gelman, A. and Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

Gelman, A. and Shalizi, C. R. (2013). Philosophy and the practice of bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 66(1):8–38.

Gelman, A., Stern, H. S., Carlin, J. B., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian data analysis*. Chapman and Hall/CRC.

George, E. I. and McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

Grzebyk, M., Wild, P., and Chouanière, D. (2004). On identification of multi-factor models with correlated residuals. *Biometrika*, 91(1):141–151.

Guillot, D., Rajaratnam, B., Emile-Geay, J., et al. (2015). Statistical paleoclimate reconstructions via markov random fields. *The Annals of Applied Statistics*, 9(1):324–352.

Haff, L. et al. (1991). The variational form of certain bayes estimators. *The Annals of Statistics*, 19(3):1163–1190.

Hamimeche, S. and Lewis, A. (2009). Properties and use of cmb power spectrum likelihoods. *Physical Review D*, 79(8):083012.

Hamsici, O. C. and Martinez, A. M. (2007). Spherical-homoscedastic distributions: The equivalency of spherical and normal distributions in classification. *Journal of Machine Learning Research*, 8(7).

Hannan, E. J. and Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2):190–195.

Hartigan, J. A. and Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108.

Hoel, P. G., Port, S. C., and Stone, C. J. (1986). *Introduction to stochastic processes*. Waveland Press.

Hoff, P. D. (2009a). A hierarchical eigenmodel for pooled covariance estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):971–992.

Hoff, P. D. (2009b). Simulation of the matrix bingham–von mises–fisher distribution, with applications to multivariate and relational data. *Journal of Computational and Graphical Statistics*, 18(2):438–456.

Hoff, P. D. and Niu, X. (2012). A covariance regression model. *Statistica Sinica*, pages 729–753.

Hurvich, C. M. and Tsai, C. (1990). The impact of model selection on inference in linear regression. *The American Statistician*, 44(3):214–217.

Jamshidian, M. and Bentler, P. M. (1993). A modified newton method for constrained estimation in covariance structure analysis. *Computational Statistics & Data Analysis*, 15(2):133–146.

Jeffreys, H. (1998). *The theory of probability*. OUP Oxford.

Jobson, J. D. and Korkie, B. (1980). Estimation for markowitz efficient portfolios. *Journal of the American Statistical Association*, 75(371):544–554.

Johnstone, I. M. and Lu, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–693.

Ke, T., Jin, J., and Fan, J. (2014). Covariance assisted screening and estimation. *Annals of statistics*, 42(6):2202.

Keener, R. W. (2011). *Theoretical statistics: Topics for a core course*. Springer.

Kundu, A. and Pourahmadi, M. (2020). Mle of jointly constrained mean-covariance of multivariate normal distributions. *arXiv preprint arXiv:2012.11826*.

Kuo, L. and Mallick, B. (1998). Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 65–81.

Lakshmanan, V., Gilleland, E., McGovern, A., and Tingley, M. (2015). Machine learning and data mining approaches to climate science. In *Proceedings of the 4th International Workshop on Climate Informatics*. Springer.

Lan, S., Holbrook, A., Fortin, N. J., Ombao, H., and Shahbaba, B. (2017). Flexible bayesian dynamic modeling of covariance and correlation matrices.

Lange, K. (2013). Elementary optimization. In *Optimization*, pages 1–17. Springer.

Ledoit, O. and Wolf, M. (2004a). Honey, i shrunk the sample covariance matrix. *The Journal of Portfolio Management*, 30(4):110–119.

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

Ledoit, O. and Wolf, M. (2020). The power of (non-) linear shrinking: A review and guide to covariance matrix estimation. *Journal of Financial Econometrics*.

Lee, K., Lee, K., and Lee, J. (2020). Post-processed posteriors for banded covariances. *arXiv preprint arXiv:2011.12627*.

Lehmann, E. L. and Casella, G. (2006). *Theory of point estimation*. Springer Science & Business Media.

Leonard, T. and Hsu, J. S. (1992). Bayesian inference for a covariance matrix. *The Annals of Statistics*, 20(4):1669–1696.

Lin, L. and Dunson, D. B. (2014). Bayesian monotone regression using gaussian process projection. *Biometrika*, 101(2):303–317.

Lin, S. P. and Perlman, M. (1985). A monte carlo comparison of four estimators of a covariance matrix. *Multivariate Analysis*, pages 411–429.

Liu, C. (1993). Bartlett s decomposition of the posterior distribution of the covariance for normal monotone ignorable missing data. *Journal of Multivariate Analysis*, 46(2):198–206.

Luo, H., Bouchard-Côté, A., Freue, G. C., and Gustafson, P. (2016). The constrained maximum likelihood estimation for parameters arising from partially identified models. *arXiv preprint arXiv:1607.08826*.

Mao, Y., Kschischang, F., and Frey, B. J. (2012). Convolutional factor graphs as probabilistic models. *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, Eds. M. Chickering & J. Halpern*, pages 374–381.

Mardia, K. V. and Jupp, P. E. (2000). *Directional statistics*, volume 494. John Wiley & Sons.

Matthews, G. and Crowther, N. (1995). A maximum likelihood estimation procedure when modelling in terms of constraints. *South African Statistical Journal*, 29(1):29–50.

Molstad, A. J., Weng, G., Doss, C. R., and Rothman, A. J. (2020). An explicit mean-covariance parameterization for multivariate response linear regression. *Journal of Computational and Graphical Statistics*, pages 1–24.

Montgomery, D. C., Peck, E. A., and Vining, G. G. (2021). *Introduction to linear regression analysis*. John Wiley & Sons.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

O'Hara, R. B., Sillanpää, M. J., et al. (2009). A review of bayesian variable selection methods: what, how and which. *Bayesian analysis*, 4(1):85–117.

Paine, P., Preston, S. P., Tsagris, M., and Wood, A. T. (2018). An elliptically symmetric angular gaussian distribution. *Statistics and Computing*, 28(3):689–697.

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

Pinheiro, J. C. and Bates, D. M. (1996). Unconstrained parametrizations for variance-covariance matrices. *Statistics and computing*, 6(3):289–296.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika*, 86(3):677–690.

Pourahmadi, M. (2013). *High-dimensional covariance estimation: with high-dimensional data*, volume 882. John Wiley & Sons.

Preston, S. and Paine, P. (2017). *Analysis of spherical data with ESAG*.

Rajaratnam, B., Massam, H., and Carvalho, C. M. (2008). Flexible covariance estimation in graphical gaussian models. *The Annals of Statistics*, 36(6):2818–2849.

Rao, C. R. (1973). *Linear statistical inference and its applications*. Wiley New York, second edition.

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

Schmidt, P. (1976). The non-uniqueness of the australian mesozoic palaeomagnetic pole position. *Geophysical Journal International*, 47(2):285–300.

Smith, M. and Kohn, R. (2002). Parsimonious covariance matrix estimation for longitudinal data. *Journal of the American Statistical Association*, 97(460):1141–1153.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the third Berkeley symposium on mathematical statistics and probability*, pages 197–206.

Strydom, H. F. and Crowther, N. A. S. (2012). Maximum likelihood estimation for multivariate normal samples: theory and methods. *South African Statistical Journal*, 46(1):115–153.

Styan, G. P. (1973). Hadamard products and multivariate statistical analysis. *Linear algebra and its applications*, 6:217–240.

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

Trefethen, L. N. and Bau III, D. (1997). *Numerical linear algebra*, volume 50. Siam.

Van Loan, C. F. and Golub, G. (1996). *Matrix computations (Johns Hopkins studies in mathematical sciences)*. The Johns Hopkins University Press.

Wermuth, N., Cox, D. R., and Marchetti, G. M. (2006). Covariance chains. *Bernoulli*, 12(5):841–862.

Won, J. H. and Kim, S.-J. (2006). Maximum likelihood covariance estimation with a condition number constraint. In *2006 Fortieth Asilomar Conference on Signals, Systems and Computers*, pages 1445–1449. IEEE.

Wong, F., Carter, C. K., and Kohn, R. (2003). Efficient estimation of covariance selection models.

Biometrika, 90(4):809–830.

Yang, R. and Berger, J. O. (1994). Estimation of a covariance matrix using the reference prior. *The Annals of Statistics*, pages 1195–1211.

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

Zwiernik, P., Uhler, C., and Richards, D. (2017). Maximum likelihood estimation for linear gaussian covariance models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1269–1292.

APPENDIX A

FIRST APPENDIX

A.1 Proofs of Results:

1. **Proof of Theorem 1:**

(a) By differentiating the Lagrangian in 1.(a)

$$\begin{aligned} L(\boldsymbol{X};\boldsymbol{\mu},\boldsymbol{\Sigma}) &= l(\boldsymbol{X};\boldsymbol{\mu},\boldsymbol{\Sigma}) + \alpha_1(\mid \boldsymbol{\Sigma}^{-1} \mid -1) \\ &= c + \frac{n}{2}\log \mid \boldsymbol{\Sigma}^{-1} \mid -\frac{1}{2} Tr\left[\boldsymbol{A}(\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}\right] + \alpha_1\left(\mid \boldsymbol{\Sigma}^{-1} \mid -1\right) \end{aligned} \tag{A.1.1}$$

with respect to $\boldsymbol{\mu}$ and setting to zero leads to $\widehat{\boldsymbol{\mu}}_{mle} = \bar{\boldsymbol{x}}$. Differentiation with respect to α_1 gives us the condition $\mid \boldsymbol{\Sigma}^{-1} \mid = 1$. Using this and setting derivative with respect to $\boldsymbol{\Sigma}^{-1}$ to 0, leads to

$$\begin{aligned} \frac{\partial L(\boldsymbol{X};\boldsymbol{\mu},\boldsymbol{\Sigma})}{\partial \boldsymbol{\Sigma}^{-1}} &= \frac{1}{2}\left[n\boldsymbol{\Sigma} - \boldsymbol{A}(\boldsymbol{\mu}) + 2\alpha_1 \mid \boldsymbol{\Sigma}^{-1} \mid \boldsymbol{\Sigma}\right] \\ &= \frac{1}{2}\left[n\boldsymbol{\Sigma} - \boldsymbol{A}(\boldsymbol{\mu}) + 2\alpha_1\boldsymbol{\Sigma}\right] \\ \widehat{\boldsymbol{\Sigma}} &= \frac{\boldsymbol{A}(\boldsymbol{\mu})}{\mid \boldsymbol{A}(\boldsymbol{\mu}) \mid^{1/p}} \end{aligned} \tag{A.1.2}$$

Now substituting for the MLE of $\boldsymbol{\mu}$ we obtain $\widehat{\boldsymbol{\Sigma}}_{mle} = \frac{A(\bar{x})}{|A(\bar{x})|^{1/p}}$.

(b) The Lagrangian in 1.(b) is:

$$L(\boldsymbol{X};\boldsymbol{\mu},\boldsymbol{\Sigma}) = c + \frac{n}{2}\log \mid \boldsymbol{\Sigma}^{-1} \mid -\frac{1}{2} Tr\left[\boldsymbol{A}(\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}\right] + \alpha_1(\mid \boldsymbol{\Sigma}^{-1} \mid -1) - \boldsymbol{\alpha}_2^{\top}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - b\right) \tag{A.1.3}$$

Rewriting the condition as $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} = \boldsymbol{b}$ and $|\boldsymbol{\Sigma}^{-1}| = 1$ in the Lagrangian is necessary for

taking the derivative with respect to $\boldsymbol{\Sigma}^{-1}$ in accordance with the standard practice in the unconstrained case (Bibby et al., 1979, §4.2.2). Differentiating with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}^{-1}$ we have:

$$\begin{aligned}\frac{\partial L(\boldsymbol{X};\boldsymbol{\mu},\boldsymbol{\Sigma})}{\partial \boldsymbol{\mu}} &= n\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{\mu})-\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}_2 \\ \frac{\partial L(\boldsymbol{X};\boldsymbol{\mu},\boldsymbol{\Sigma})}{\partial \boldsymbol{\Sigma}^{-1}} &= \frac{1}{2}\left[(n+2\alpha_1)\boldsymbol{\Sigma}-\boldsymbol{A}(\boldsymbol{\mu})-2\boldsymbol{\alpha}_2\boldsymbol{\mu}^{\top}\right]\end{aligned} \tag{A.1.4}$$

and obtain the MLEs as the solution of the following equations:

$$\begin{aligned}\boldsymbol{\mu} &= \bar{\boldsymbol{x}}-\frac{1}{n}\boldsymbol{\alpha}_2, \qquad & \boldsymbol{\Sigma} &= \frac{\boldsymbol{A}(\boldsymbol{\mu})+2\boldsymbol{\alpha}_2\boldsymbol{\mu}^{\top}}{n+2\alpha_1} \\ |\,\boldsymbol{\Sigma}^{-1}\,| &= 1, \qquad & \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} &= \boldsymbol{b}\end{aligned}$$

(c) The Lagrangian

$$L(\boldsymbol{X};\boldsymbol{\mu},\boldsymbol{\Sigma}) = c+\frac{n}{2}\log|\,\boldsymbol{\Sigma}^{-1}\,|-\frac{1}{2}Tr\left[\boldsymbol{A}(\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}\right]+\alpha_1(|\,\boldsymbol{\Sigma}^{-1}\,|-1)-\boldsymbol{\alpha}_2^{\top}\left(\boldsymbol{I}_p-\boldsymbol{\Sigma}^{-1}\right)\boldsymbol{\mu} \tag{A.1.5}$$

and its derivatives with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}^{-1}$ are:

$$\begin{aligned}\frac{\partial L(\boldsymbol{X};\boldsymbol{\mu},\boldsymbol{\Sigma})}{\partial \boldsymbol{\mu}} &= n\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{\mu})-(\boldsymbol{I}_p-\boldsymbol{\Sigma}^{-1})\boldsymbol{\alpha}_2 \\ \frac{\partial L(\boldsymbol{X};\boldsymbol{\mu},\boldsymbol{\Sigma})}{\partial \boldsymbol{\Sigma}^{-1}} &= \frac{1}{2}\left[(n+2\alpha_1)\boldsymbol{\Sigma}-\boldsymbol{A}(\boldsymbol{\mu})-2\boldsymbol{\alpha}_2\boldsymbol{\mu}^{\top}\right]\end{aligned} \tag{A.1.6}$$

and obtain the MLE as the solution of the following equations:

$$\boldsymbol{\mu} = \bar{\boldsymbol{x}} - \frac{1}{n}(\boldsymbol{I}_p - \boldsymbol{\Sigma})\boldsymbol{\alpha}_2, \qquad \boldsymbol{\Sigma} = \frac{\boldsymbol{A}(\boldsymbol{\mu}) + 2\boldsymbol{\alpha}_2\boldsymbol{\mu}^\top}{n + 2\alpha_1}$$
$$| \boldsymbol{\Sigma}^{-1} | = 1, \qquad \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} = \boldsymbol{\mu}$$

2. **Proof of Lemma 1**

(a) Follows from simple algebra and the definition of eigenvalue:

$$(\boldsymbol{a}\boldsymbol{b}^\top + \boldsymbol{b}\boldsymbol{a}^\top)\left(\frac{\boldsymbol{a}}{\|\boldsymbol{a}\|} + \frac{\boldsymbol{b}}{\|\boldsymbol{b}\|}\right) = \left(\boldsymbol{a}^\top\boldsymbol{b} + \|\boldsymbol{a}\|\|\boldsymbol{b}\|\right)\left(\frac{\boldsymbol{a}}{\|\boldsymbol{a}\|} + \frac{\boldsymbol{b}}{\|\boldsymbol{b}\|}\right)$$
$$(\boldsymbol{a}\boldsymbol{b}^\top + \boldsymbol{b}\boldsymbol{a}^\top)\left(\frac{\boldsymbol{a}}{\|\boldsymbol{a}\|} - \frac{\boldsymbol{b}}{\|\boldsymbol{b}\|}\right) = \left(\boldsymbol{a}^\top\boldsymbol{b} - \|\boldsymbol{a}\|\|\boldsymbol{b}\|\right)\left(\frac{\boldsymbol{a}}{\|\boldsymbol{a}\|} - \frac{b}{\|\boldsymbol{b}\|}\right)$$

(b) Let $\boldsymbol{B} = \boldsymbol{B}(\boldsymbol{a}, \boldsymbol{b}) = \boldsymbol{a}\boldsymbol{b}^\top + \boldsymbol{b}\boldsymbol{a}^\top$, and $\lambda_j(\boldsymbol{M})$ be the j-th largest eigenvalue of $\boldsymbol{M}$ with $\boldsymbol{M} = \boldsymbol{A} + \boldsymbol{B}$. We apply Weyl's Inequality (Bhatia, 2007, Theorem 8.2) to obtain

$$\begin{aligned}
\lambda_j(\boldsymbol{M}) = \lambda_j(\boldsymbol{B} + \boldsymbol{A}) &\geq \lambda_j(\boldsymbol{B}) + \lambda_p(\boldsymbol{A}) \\
&\geq \lambda_p(\boldsymbol{B}) + \lambda_p(\boldsymbol{A}) \\
&= (\boldsymbol{a}^\top\boldsymbol{b} - \|\boldsymbol{a}\|\|\boldsymbol{b}\|) + \lambda_p(\boldsymbol{A}), \qquad \text{(By applying the first part.)}
\end{aligned}$$

where $\lambda_p(\boldsymbol{A}) > 0$. Since $\boldsymbol{B}$ is a rank two matrix its at most two non-zero eigenvalues are $(\boldsymbol{a}^\top\boldsymbol{b} \pm \|\boldsymbol{a}\|\|\boldsymbol{b}\|)$. By applying Cauchy-Schwartz inequality it follows that these non-zero eigenvalues belong to the range $\left(\boldsymbol{a}^\top\boldsymbol{b} + \|\boldsymbol{a}\|\|\boldsymbol{b}\|\right) \in [0, 2\|\boldsymbol{a}\|\|\boldsymbol{b}\|]$ and $\left(\boldsymbol{a}^\top\boldsymbol{b} - \|\boldsymbol{a}\|\|\boldsymbol{b}\|\right) \in [-2\|\boldsymbol{a}\|\|\boldsymbol{b}\|, 0]$. This tells us that $\lambda_1(\boldsymbol{B}) \geq 0$ and $\lambda_j(\boldsymbol{B}) = 0$ for $j = 2, 3, \ldots, p-1$. Weyl's inequality (Bhatia, 2007, Theorem 8.2) for $j = 2, 3, \ldots, p-1$, gives us

$$\lambda_j(\boldsymbol{M}) \geq \lambda_j(\boldsymbol{B}) + \lambda_p(\boldsymbol{A}) = \lambda_p(\boldsymbol{A}) > 0$$

and $\lambda_1(\boldsymbol{M}) > 0$ trivially.

This cannot be said for the lowest eigenvalue of $\boldsymbol{M}$ i.e. for $j = p$, we cannot say whether $(\boldsymbol{a}^\top \boldsymbol{b} - \|\boldsymbol{a}\|\|\boldsymbol{b}\|) + \lambda_p(\boldsymbol{A})$ is positive or not. It depends on $\lambda_p(\boldsymbol{A})$. Therefore except for the smallest eigenvalue all other eigenvalues of $\boldsymbol{M}$ are positive completing the proof.

3. **Verification of Conditions** $\mathscr{F}1 - \mathscr{F}4$ and $\mathscr{H}1 - \mathscr{H}3$ (Aitchison and Silvey, 1958)

Checking the conditions amounts to calculation of second derivative matrix of the likelihood function, which in turn verifies the existence of the third derivative as one of the conditions. Here we present the details of these calculations.

First Derivative:

$$\frac{\partial l}{\partial \boldsymbol{\mu}} = n\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}), \qquad \frac{\partial l}{\partial \boldsymbol{\Sigma}} = -\frac{1}{2}\left(n\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\boldsymbol{A}(\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}\right) \tag{A.1.7}$$

Second Derivative: The Hessian matrix of the likelihood is:

$$H_l = \begin{pmatrix} \frac{\partial^2 l}{\partial \boldsymbol{\mu}^2}\Big|_{p\times p} & \frac{\partial^2 l}{\partial \boldsymbol{\mu}\partial \boldsymbol{\Sigma}}\Big|_{p\times p^2} \\ \frac{\partial^2 l}{\partial \boldsymbol{\Sigma}\partial \boldsymbol{\mu}}\Big|_{p^2\times p} & \frac{\partial^2 l}{\partial \boldsymbol{\Sigma}^2}\Big|_{p^2\times p^2} \end{pmatrix}_{p(p+1)\times p(p+1)}.$$

Next, we calculate the four submatrices.

(a) The first submatrix is

$$\frac{\partial^2 l}{\partial \boldsymbol{\mu}^2} = -n\boldsymbol{\Sigma}^{-1} \tag{A.1.8}$$

(b) Since, $\frac{\partial \boldsymbol{\Sigma}^{-1}}{\partial \boldsymbol{\Sigma}} = -\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1}$ and $\frac{\partial \boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu})}{\partial \boldsymbol{\Sigma}^{-1}} = \boldsymbol{I} \otimes (\bar{\boldsymbol{x}} - \boldsymbol{\mu})$ we obtain the following:

$$\frac{\partial^2 l}{\partial \boldsymbol{\Sigma}\partial \boldsymbol{\mu}}\Big|_{p^2\times p} = -\left[\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1}\right][\boldsymbol{I} \otimes (\bar{\boldsymbol{x}} - \boldsymbol{\mu})] \tag{A.1.9}$$

(c) The third submatrix is

$$\begin{aligned}
\left.\frac{\partial^2 l}{\partial \boldsymbol{\mu}\partial \boldsymbol{\Sigma}}\right|_{p\times p^2} &= -\frac{1}{2}\frac{\partial}{\partial \boldsymbol{\mu}}\text{vec}\left[n\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\boldsymbol{A}(\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}\right] \\
&= \frac{1}{2}\frac{\partial}{\partial \boldsymbol{\mu}}\text{vec}\left[\boldsymbol{\Sigma}^{-1}\boldsymbol{A}(\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}\right] \\
&= \frac{1}{2}\frac{\partial \text{vec}\left(\boldsymbol{\Sigma}^{-1}\left[-2n\bar{\boldsymbol{x}}\boldsymbol{\mu}^\top + n\boldsymbol{\mu}\boldsymbol{\mu}^\top\right]\boldsymbol{\Sigma}^{-1}\right)}{\partial \boldsymbol{\mu}} \\
&= \frac{1}{2}\left[-2n\frac{\partial \text{vec}\left[(\boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{x}})(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})^\top\right]}{\partial \boldsymbol{\mu}} + n\frac{\partial \text{vec}\left[(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})^\top\right]}{\partial \boldsymbol{\mu}}\right] \\
&= \frac{1}{2}\left[-2n\frac{\partial (\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})^\top \otimes (\boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{x}})^\top}{\partial \boldsymbol{\mu}} + n\frac{\partial (\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})^\top \otimes (\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})^\top}{\partial \boldsymbol{\mu}}\right] \\
&= \frac{1}{2}\left[-2n\boldsymbol{\Sigma}^{-1}\otimes(\boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{x}})^\top + n\boldsymbol{\Sigma}^{-1}\otimes(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})^\top + n(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})^\top\otimes\boldsymbol{\Sigma}^{-1}\right]
\end{aligned} \tag{A.1.10}$$

(d) Following the calculations of Chaudhuri et al. (2007), we obtain

$$\left.\frac{\partial^2 l}{\partial \boldsymbol{\Sigma}^2}\right|_{p^2\times p^2} = \frac{1}{2}\left[n\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1} - \left(\boldsymbol{\Sigma}^{-1}\boldsymbol{A}(\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}\right)\otimes\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\otimes\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{A}(\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}\right)\right] \tag{A.1.11}$$

We have shown the existence of the second derivative and from the above quantities it is evident the the third derivative exists too. Since multivariate normal has all the moments (Chacón and Duong, 2015), so we have essentially verified conditions $\mathscr{F}1-\mathscr{F}4$. Next, we verify $\mathscr{H}1-\mathscr{H}3$ for the constraint. The simplest way is to express it as $h(\boldsymbol{\mu},\boldsymbol{\Sigma}) = \boldsymbol{\Sigma}\boldsymbol{\mu}-\boldsymbol{\mu} = \mathbf{0}$. Here also we need to check the Hessian matrix of the constraint and its corresponding bound. We can establish the coordinate wise bound to be 1. The details are as follows:

First Derivative

$$\left.\frac{\partial h}{\partial \boldsymbol{\mu}}\right|_{p\times p} = \boldsymbol{\Sigma} - \boldsymbol{I}, \qquad \left.\frac{\partial h}{\partial \boldsymbol{\Sigma}}\right|_{p^2\times p} = \boldsymbol{\mu}\otimes\boldsymbol{I} \tag{A.1.12}$$

We denote the first derivative to be $\left(\boldsymbol{H}_{\boldsymbol{\theta}}^{1}\right)_{p+p^2\times p}$ with $rank(\boldsymbol{H}_{\boldsymbol{\theta}}^{1}) = p$

Second Derivative:

$$\left.\frac{\partial^2 h}{\partial \boldsymbol{\mu}^2}\right|_{p^2\times p} = \mathbf{0}, \qquad \left.\frac{\partial^2 h}{\partial \boldsymbol{\Sigma}\partial \boldsymbol{\mu}}\right|_{p^2\times p^2} = \boldsymbol{I}_{p^2}$$
$$\left.\frac{\partial^2 h}{\partial \boldsymbol{\mu}\partial \boldsymbol{\Sigma}}\right|_{p^3\times p} = \boldsymbol{I}\otimes \mathrm{vec}(\boldsymbol{I}_p), \qquad \left.\frac{\partial^2 h}{\partial \boldsymbol{\Sigma}^2}\right|_{p^3\times p^2} = \mathbf{0} \tag{A.1.13}$$

This gives us

$$\left.\boldsymbol{H}_{\boldsymbol{\theta}}^{2}\right|_{(p^2+p^3)\times(p+p^2)} = \left(\left(\frac{\partial^2 h}{\partial \theta_i \partial \theta_j}\right)\right)$$

with $rank(\boldsymbol{H}_{\boldsymbol{\theta}}^{2}) = p + p^2$

4. **Proof of Lemma 2:** We assume that $\boldsymbol{\mu}$ is fixed and are interested in calculating the directional derivative of the Lagrangian in (3.2.1) as a function of $\boldsymbol{\Sigma}$ only in the direction of a symmetric matrix $\boldsymbol{D}$. Let us set $L(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \boldsymbol{X}) = f(\boldsymbol{\Sigma})$. By the definition of directional derivative,

$$\nabla_{\boldsymbol{D}} f = \lim_{h\to 0} \frac{f(\boldsymbol{\Sigma}+h\boldsymbol{D}) - f(\boldsymbol{\Sigma})}{h}$$

where h is a scalar. Now since the terms in the Lagrangian are additive, we calculate the directional derivative of each term separately.

(a) First we focus on the first term ignoring the constant $f^1 = \log \mid \boldsymbol{\Sigma} \mid$.

$$\begin{aligned}
\nabla_{\boldsymbol{D}} f^1 &= \lim_{h\to 0} \frac{1}{h} \log\left[\frac{\mid \boldsymbol{\Sigma}+h\boldsymbol{D} \mid}{\mid \boldsymbol{\Sigma} \mid}\right] = \lim_{h\to 0} \frac{1}{h} \log\left[\frac{\mid \boldsymbol{\Sigma} \mid\mid \boldsymbol{I}+h\boldsymbol{D}\boldsymbol{\Sigma}^{-1} \mid}{\mid \boldsymbol{\Sigma} \mid}\right] \\
&= \lim_{h\to 0} \frac{1}{h} \log\left[\mid \boldsymbol{I}+h\boldsymbol{D}\boldsymbol{\Sigma}^{-1} \mid\right] \\
&= \lim_{h\to 0} \frac{1}{h} \sum_{i=1}^{p} \log\left[1+h\lambda_i(\boldsymbol{D}\boldsymbol{\Sigma}^{-1})\right] \quad \text{(Here } \lambda_i(\boldsymbol{D}) \text{ denote the ith eigenvalue of } \boldsymbol{D})
\end{aligned}$$

$$= \lim_{h\to 0} \frac{1}{h}\left[\sum_{i=1}^{p} h\lambda_i(\boldsymbol{D}\boldsymbol{\Sigma}^{-1}) + \sum_{i=1}^{p}\left\{\log\left[1+h\lambda_i(\boldsymbol{D}\boldsymbol{\Sigma}^{-1})\right] - h\lambda_i(\boldsymbol{D}\boldsymbol{\Sigma}^{-1})\right\}\right] = Tr\left[\boldsymbol{D}\boldsymbol{\Sigma}^{-1}\right]$$

(b) The second term ignoring the constant $f^2 = Tr\left[\boldsymbol{S}\boldsymbol{\Sigma}^{-1}\right]$.

$$\begin{aligned}
\nabla_D f^2 &= \lim_{h\to 0} \frac{1}{h} Tr\left[\boldsymbol{S}\left\{(\boldsymbol{\Sigma}+h\boldsymbol{D})^{-1} - \boldsymbol{\Sigma}^{-1}\right\}\right] \\
&= \lim_{h\to 0} \frac{1}{h} Tr\left[\boldsymbol{S}\left\{\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}(h^{-1}\boldsymbol{D}^{-1} - \boldsymbol{\Sigma}^{-1})^{-1}\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\right\}\right] \\
&= \lim_{h\to 0} \frac{1}{h} Tr\left[\boldsymbol{S}\left\{-\boldsymbol{\Sigma}^{-1}(h^{-1}\boldsymbol{D}^{-1} + \boldsymbol{\Sigma}^{-1})^{-1}\boldsymbol{\Sigma}^{-1}\right\}\right] \\
&= \lim_{h\to 0} \frac{1}{h} Tr\left[\boldsymbol{S}\left\{-h\boldsymbol{\Sigma}^{-1}(\boldsymbol{D}^{-1} + h\boldsymbol{\Sigma}^{-1})^{-1}\boldsymbol{\Sigma}^{-1}\right\}\right] \\
&= \lim_{h\to 0} Tr\left[-\boldsymbol{S}\boldsymbol{\Sigma}^{-1}(\boldsymbol{D}^{-1} + h\boldsymbol{\Sigma}^{-1})^{-1}\boldsymbol{\Sigma}^{-1}\right] \\
&= Tr\left[-\boldsymbol{S}\boldsymbol{\Sigma}^{-1}\boldsymbol{D}\boldsymbol{\Sigma}^{-1}\right]
\end{aligned}$$

(c) The third term without the constant is $f^3 = Tr\left[(\bar{\boldsymbol{x}}-\boldsymbol{\mu})(\bar{\boldsymbol{x}}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}\right]$. The same calculation shows that $\nabla_D f^3 = Tr\left[-(\bar{\boldsymbol{x}}-\boldsymbol{\mu})(\bar{\boldsymbol{x}}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{D}\boldsymbol{\Sigma}^{-1}\right]$

(d) The fourth term $f^4 = \boldsymbol{\alpha}_2^\top(\boldsymbol{\Sigma}\boldsymbol{\mu}-\boldsymbol{\mu})$

$$\begin{aligned}
\nabla_D f^4 &= \lim_{h\to 0} \frac{\left\{\boldsymbol{\alpha}_2^\top(\boldsymbol{\Sigma}+h\boldsymbol{D})\boldsymbol{\mu} - \boldsymbol{\alpha}_2^\top\boldsymbol{\mu} - \boldsymbol{\alpha}_2^\top\boldsymbol{\Sigma}\boldsymbol{\mu} + \boldsymbol{\alpha}_2^\top\boldsymbol{\mu}\right\}}{h} = \lim_{h\to 0} \frac{\left\{\boldsymbol{\alpha}_2^\top(\boldsymbol{\Sigma}+h\boldsymbol{D})\boldsymbol{\mu} - \boldsymbol{\alpha}_2^\top\boldsymbol{\Sigma}\boldsymbol{\mu}\right\}}{h} \\
&= \lim_{h\to 0} \frac{\boldsymbol{\alpha}_2^\top(\boldsymbol{\Sigma}+h\boldsymbol{D}-\boldsymbol{\Sigma})\boldsymbol{\mu}}{h} = \boldsymbol{\alpha}_2^\top\boldsymbol{D}\boldsymbol{\mu}
\end{aligned}$$

With these calculations the final directional derivative of the Lagrangian function is:

$$\nabla_D f = -\frac{n}{2} Tr\left[\boldsymbol{D}\boldsymbol{\Sigma}^{-1}\right] + \frac{n}{2} Tr\left[\boldsymbol{S}\boldsymbol{\Sigma}^{-1}\boldsymbol{D}\boldsymbol{\Sigma}^{-1}\right] + \frac{n}{2} Tr\left[\boldsymbol{B}\boldsymbol{\Sigma}^{-1}\boldsymbol{D}\boldsymbol{\Sigma}^{-1}\right] + \boldsymbol{\alpha}_2^\top\boldsymbol{D}\boldsymbol{\mu}$$

$$= -\frac{n}{2} Tr\left[\boldsymbol{D}\boldsymbol{\Sigma}^{-1}\right] + \frac{n}{2} Tr\left[(\boldsymbol{S}+\boldsymbol{B})\boldsymbol{\Sigma}^{-1}\boldsymbol{D}\boldsymbol{\Sigma}^{-1}\right] + \boldsymbol{\alpha}_2^{\top}\boldsymbol{D}\boldsymbol{\mu} \tag{A.1.14}$$

where $\boldsymbol{B} = (\bar{\boldsymbol{x}} - \boldsymbol{\mu})(\bar{\boldsymbol{x}} - \boldsymbol{\mu})^{\top}$

Now, we calculate the second directional derivative in the direction $\boldsymbol{C}$ which is also a symmetric matrix. The first derivative denoted by $\nabla_{\boldsymbol{D}} f$ has two terms as a function of $\boldsymbol{\Sigma}$. The corresponding notation for second directional derivative is

$$\nabla_{\boldsymbol{C}} \nabla_{\boldsymbol{D}} f = \lim_{h \to 0} \frac{\nabla_{\boldsymbol{D}} f(\boldsymbol{\Sigma} + h\boldsymbol{C}) - \nabla_{\boldsymbol{D}} f(\boldsymbol{\Sigma})}{h}.$$

We will calculate the directional derivative of each of the two terms.

(a) The first term ignoring the constant is $\nabla_{\boldsymbol{D}} f^1 = Tr\left[\boldsymbol{D}\boldsymbol{\Sigma}^{-1}\right]$. This is same as the second term of the original likelihood function. So by applying the same formula we obtain: $\nabla_{\boldsymbol{C}} \nabla_{\boldsymbol{D}} f^1 = -Tr\left[\boldsymbol{D}\boldsymbol{\Sigma}^{-1}\boldsymbol{C}\boldsymbol{\Sigma}^{-1}\right] = -Tr\left[\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\boldsymbol{D}\boldsymbol{\Sigma}^{-1}\boldsymbol{C}\boldsymbol{\Sigma}^{-1}\right]$

(b) The second term ignoring the constant is $\nabla_{\boldsymbol{D}} f^2 = Tr\left[(\boldsymbol{S}+\boldsymbol{B})\boldsymbol{\Sigma}^{-1}\boldsymbol{D}\boldsymbol{\Sigma}^{-1}\right]$. By Woodbury-Sherman matrix formula:

$$\begin{aligned}
&(\boldsymbol{\Sigma} + h\boldsymbol{C})^{-1}\boldsymbol{D}(\boldsymbol{\Sigma} + h\boldsymbol{C})^{-1} - \boldsymbol{\Sigma}^{-1}\boldsymbol{D}\boldsymbol{\Sigma}^{-1} \\
&= \left[\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}(h^{-1}\boldsymbol{C}^{-1} + \boldsymbol{\Sigma}^{-1})^{-1}\boldsymbol{\Sigma}^{-1}\right]\boldsymbol{D}\left[\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}(h^{-1}\boldsymbol{C}^{-1} + \boldsymbol{\Sigma}^{-1})^{-1}\boldsymbol{\Sigma}^{-1}\right] - \boldsymbol{\Sigma}^{-1}\boldsymbol{D}\boldsymbol{\Sigma}^{-1} \\
&= \left[\boldsymbol{\Sigma}^{-1} - h\boldsymbol{\Sigma}^{-1}(\boldsymbol{C}^{-1} + h\boldsymbol{\Sigma}^{-1})^{-1}\boldsymbol{\Sigma}^{-1}\right]\boldsymbol{D}\left[\boldsymbol{\Sigma}^{-1} - h\boldsymbol{\Sigma}^{-1}(\boldsymbol{C}^{-1} + h\boldsymbol{\Sigma}^{-1})^{-1}\boldsymbol{\Sigma}^{-1}\right] - \boldsymbol{\Sigma}^{-1}\boldsymbol{D}\boldsymbol{\Sigma}^{-1} \\
&= -h\boldsymbol{\Sigma}^{-1}(\boldsymbol{C}^{-1} + h\boldsymbol{\Sigma}^{-1})^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{D}\boldsymbol{\Sigma}^{-1} - h\boldsymbol{\Sigma}^{-1}\boldsymbol{D}\boldsymbol{\Sigma}^{-1}(\boldsymbol{C}^{-1} + h\boldsymbol{\Sigma}^{-1})^{-1}\boldsymbol{\Sigma}^{-1} + \mathcal{O}(h^2)
\end{aligned}$$

Using this result, we get:

$$\nabla_{\boldsymbol{C}} \nabla_{\boldsymbol{D}} f^2 = \lim_{h \to 0} \frac{Tr\left[(\boldsymbol{S}+\boldsymbol{B})(\boldsymbol{\Sigma} + h\boldsymbol{C})^{-1}\boldsymbol{D}(\boldsymbol{\Sigma} + h\boldsymbol{C})^{-1}\right] - Tr\left[(\boldsymbol{S}+\boldsymbol{B})\boldsymbol{\Sigma}^{-1}\boldsymbol{D}\boldsymbol{\Sigma}^{-1}\right]}{h}$$

$$
\begin{aligned}
&= \lim_{h\to 0} \frac{Tr\left[(\boldsymbol{S}+\boldsymbol{B})\left\{(\boldsymbol{\Sigma}+h\boldsymbol{C})^{-1}\boldsymbol{D}(\boldsymbol{\Sigma}+h\boldsymbol{C})^{-1}-\boldsymbol{\Sigma}^{-1}\boldsymbol{D}\boldsymbol{\Sigma}^{-1}\right\}\right]}{h} \\
&= -\lim_{h\to 0} Tr\left[(\boldsymbol{S}+\boldsymbol{B})\boldsymbol{\Sigma}^{-1}(\boldsymbol{C}^{-1}+h\boldsymbol{\Sigma}^{-1})^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{D}\boldsymbol{\Sigma}^{-1}\right] \\
&\qquad -\lim_{h\to 0} Tr\left[(\boldsymbol{S}+\boldsymbol{B})\boldsymbol{\Sigma}^{-1}\boldsymbol{D}\boldsymbol{\Sigma}^{-1}(\boldsymbol{C}^{-1}+h\boldsymbol{\Sigma}^{-1})^{-1}\boldsymbol{\Sigma}^{-1}\right] \\
&= -Tr\left[(\boldsymbol{S}+\boldsymbol{B})\boldsymbol{\Sigma}^{-1}\boldsymbol{C}\boldsymbol{\Sigma}^{-1}\boldsymbol{D}\boldsymbol{\Sigma}^{-1}\right] - Tr\left[(\boldsymbol{S}+\boldsymbol{B})\boldsymbol{\Sigma}^{-1}\boldsymbol{D}\boldsymbol{\Sigma}^{-1}\boldsymbol{C}\boldsymbol{\Sigma}^{-1}\right] \\
&= -Tr\left[(\boldsymbol{S}+\boldsymbol{B})\boldsymbol{\Sigma}^{-1}\boldsymbol{C}\boldsymbol{\Sigma}^{-1}\boldsymbol{D}\boldsymbol{\Sigma}^{-1}\right] - Tr\left[(\boldsymbol{S}+\boldsymbol{B})\boldsymbol{\Sigma}^{-1}(\boldsymbol{D}\boldsymbol{\Sigma}^{-1}\boldsymbol{C})^{\top}\boldsymbol{\Sigma}^{-1}\right] \\
&= -Tr\left[2(\boldsymbol{S}+\boldsymbol{B})\boldsymbol{\Sigma}^{-1}\boldsymbol{C}\boldsymbol{\Sigma}^{-1}\boldsymbol{D}\boldsymbol{\Sigma}^{-1}\right]
\end{aligned}
$$

Adding these two we obtain the final directional derivative to be:

$$
\begin{aligned}
\nabla_{\boldsymbol{C}}\nabla_{\boldsymbol{D}} f &= \frac{n}{2} Tr\left[\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\boldsymbol{D}\boldsymbol{\Sigma}^{-1}\boldsymbol{C}\boldsymbol{\Sigma}^{-1}\right] - \frac{n}{2} Tr\left[2(\boldsymbol{S}+\boldsymbol{B})\boldsymbol{\Sigma}^{-1}\boldsymbol{C}\boldsymbol{\Sigma}^{-1}\boldsymbol{D}\boldsymbol{\Sigma}^{-1}\right] \\
&= -\frac{n}{2} Tr\left[\{2(\boldsymbol{S}+\boldsymbol{B})-\boldsymbol{\Sigma}\}\boldsymbol{\Sigma}^{-1}\boldsymbol{C}\boldsymbol{\Sigma}^{-1}\boldsymbol{D}\boldsymbol{\Sigma}^{-1}\right] \qquad \text{(A.1.15)}
\end{aligned}
$$

Now if we assume that the mean vector $\boldsymbol{\mu}$ is estimated by $\bar{\boldsymbol{x}}$, then $\boldsymbol{B}=\boldsymbol{0}$. If we further assume that the estimate of the covariance matrix lies within the set $D_{2S} = \{\boldsymbol{\Sigma} \text{ is pd}, 0 \prec \boldsymbol{\Sigma} \prec 2\boldsymbol{S}\}$ and $\boldsymbol{C}=\boldsymbol{D}$, then

$$
\begin{aligned}
\nabla_{\boldsymbol{D}}\nabla_{\boldsymbol{D}} f &= -\frac{n}{2} Tr\left[\boldsymbol{\Sigma}^{-1/2}\{2\boldsymbol{S}-\boldsymbol{\Sigma}\}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{D}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{D}\boldsymbol{\Sigma}^{-1/2}\right] \\
&= -\frac{n}{2} Tr\left[\boldsymbol{\Sigma}^{-1/2}\boldsymbol{D}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{-1/2}\{2\boldsymbol{S}-\boldsymbol{\Sigma}\}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{D}\boldsymbol{\Sigma}^{-1/2}\right] \leq 0 \qquad \text{(A.1.16)}
\end{aligned}
$$

Therefore, within D_{2S} the constrained likelihood is strictly concave, but outside this set that is not the case as shown by the following counter-example:

If $\boldsymbol{\Sigma} \notin D_{2S}$, then there exists a $\boldsymbol{u}$ such that $\boldsymbol{u}^{\top}(2\boldsymbol{S}-\boldsymbol{\Sigma})\boldsymbol{u} \leq 0$. Choosing $\boldsymbol{D} = \boldsymbol{\Sigma}\boldsymbol{u}\boldsymbol{u}^{\top}\boldsymbol{\Sigma}$ then

$$\nabla_{\boldsymbol{D}}\nabla_{\boldsymbol{D}}f = -\frac{n}{2}Tr\left[\boldsymbol{\Sigma}^{1/2}\boldsymbol{u}\boldsymbol{u}^{\top}(2\boldsymbol{S}-\boldsymbol{\Sigma})\boldsymbol{u}\boldsymbol{u}^{\top}\boldsymbol{\Sigma}^{1/2}\right] = -\frac{n}{2}\boldsymbol{u}^{\top}\boldsymbol{\Sigma}\boldsymbol{u}\boldsymbol{u}^{\top}(2\boldsymbol{S}-\boldsymbol{\Sigma})\boldsymbol{u} \geq 0, \quad \text{(A.1.17)}$$

which completes the proof.

5. **Covariance error bound in Lemma 5:** The error of a new estimate can be bounded in the following way:

$$\left\|\boldsymbol{\Sigma}-\widehat{\boldsymbol{\Sigma}}^{*}\right\|_{\mathscr{F}} \leq \left\|\boldsymbol{\Sigma}-\widehat{\boldsymbol{\Sigma}}_{map}\right\|_{\mathscr{F}} + \left\|\widehat{\boldsymbol{\Sigma}}_{map}-\widehat{\boldsymbol{\Sigma}}^{*}\right\|_{\mathscr{F}}$$

$$\left\|\widehat{\boldsymbol{\Sigma}}_{map}-\widehat{\boldsymbol{\Sigma}}^{*}\right\|_{\mathscr{F}} = \left\|\left(1-\frac{1}{\lambda_P}\right)\sum_{\substack{i=1\\ i\neq i_0}}^{d}\lambda_i\boldsymbol{P}_i\boldsymbol{P}_i^{\top} + \left(\frac{\lambda_{i_0}}{c_{0i_0}^2}-1\right)\widehat{\boldsymbol{\mu}}^{*}\widehat{\boldsymbol{\mu}}^{*\top}\right\|_{\mathscr{F}}$$

Applying triangle inequality :

$$\leq \left\|\left(1-\frac{1}{\lambda_P}\right)\sum_{\substack{i=1\\ i\neq i_0}}^{d}\lambda_i\boldsymbol{P}_i\boldsymbol{P}_i^{\top}\right\|_{\mathscr{F}} + \left\|\left(\frac{\lambda_{i_0}}{c_{0i_0}^2}-1\right)\widehat{\boldsymbol{\mu}}^{*}\widehat{\boldsymbol{\mu}}^{*\top}\right\|_{\mathscr{F}}$$

$$= \left|\left(1-\frac{1}{\lambda_P}\right)\right|\left\|\sum_{\substack{i=1\\ i\neq i_0}}^{d}\lambda_i\boldsymbol{P}_i\boldsymbol{P}_i^{\top}\right\|_{\mathscr{F}} + \left|\left(\frac{\lambda_{i_0}}{c_{0i_0}^2}-1\right)\right|\left\|\widehat{\boldsymbol{\mu}}^{*}\widehat{\boldsymbol{\mu}}^{*\top}\right\|_{\mathscr{F}} \quad \text{(A.1.18)}$$

6. **Proof of Equation 3.3.4:** This follows from standard regression OLS estimate.

7. **Proof of Equation 3.3.5:**

$$f\left(\lambda_{\mathscr{S}}^{'}\right) = Tr\left[\left(\boldsymbol{\Sigma}-\widehat{\boldsymbol{\Sigma}}^{*}\right)^{2}\right]$$

$$\frac{\partial f\left(\lambda_{\mathscr{S}}^{'}\right)}{\partial \lambda_{i_k}} = -2.\mathrm{Tr}\left[\left(\boldsymbol{\Sigma}-\widehat{\boldsymbol{\Sigma}}^{*}\right)\left(\boldsymbol{b}_k\boldsymbol{b}_k^{\top}\right)\right] = 0$$

$$\implies \mathrm{Tr}\left[\boldsymbol{b}_k^{\top}\left(\boldsymbol{\Sigma}-\widehat{\boldsymbol{\Sigma}}^{*}\right)\boldsymbol{b}_k\right] = 0$$

$$\implies \boldsymbol{b}_k^{\top}\boldsymbol{\Sigma}\boldsymbol{b}_k = \boldsymbol{b}_k^{\top}\widehat{\boldsymbol{\Sigma}}^{*}\boldsymbol{b}_k$$

$$\implies \widehat{\lambda'}_{i_k} = \boldsymbol{b}_k^\top \boldsymbol{\Sigma} \boldsymbol{b}_k \tag{A.1.19}$$

A.2 Explicit Calculation of the Lagrange Multiplier

We consider finding the MLE under constraints for an exponential family of distributions:

$$f(\boldsymbol{X};\boldsymbol{\theta}) = \exp\left[b_0(\boldsymbol{X}) + \sum_{i=1}^{q} \theta_i T_i(\boldsymbol{X}) - a(\boldsymbol{\theta}) \right]$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_q)$ is the vector of natural parameters and $\boldsymbol{T}(\boldsymbol{X}) = (T_1(\boldsymbol{X}), T_2(\boldsymbol{X}), \ldots, T_q(\boldsymbol{X}))$ is their complete and sufficient statistics with the following (Lehmann and Casella, 2006)

$$\mathbb{E}[T(\boldsymbol{X})] = \frac{\partial a(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \boldsymbol{m} \quad \text{and} \quad \text{Cov}[T(\boldsymbol{X})] = \frac{\partial^2 a(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} = \boldsymbol{V}.$$

Let the constraint on the parameters be expressed as a function $h(\boldsymbol{m}) = 0$ where $h(\boldsymbol{m}) : \mathbb{R}^q \to \mathbb{R}$ and both $\boldsymbol{m}$ and $\boldsymbol{V}$ are functions of $\boldsymbol{\theta}$. Differentiating the Lagrangian function

$$w(\boldsymbol{X};\boldsymbol{\theta},\alpha_2) = \log f(\boldsymbol{X};\boldsymbol{\theta}) + \alpha_2 h(\boldsymbol{m}) \propto \boldsymbol{\theta}^\top T(\boldsymbol{X}) - a(\boldsymbol{\theta}) + \alpha_2 h(\boldsymbol{m})$$

with respect to $\boldsymbol{\theta}$ and equating it to zero, we obtain

$$\begin{aligned} \frac{\partial w(\boldsymbol{X};\boldsymbol{\theta},\alpha_2)}{\partial \boldsymbol{\theta}} &= T(\boldsymbol{X}) - \frac{\partial a(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} + \alpha_2 \frac{\partial h(\boldsymbol{m})}{\partial \boldsymbol{\theta}} = 0 \\ \boldsymbol{m} &= T(\boldsymbol{X}) + \alpha_2 \nabla \boldsymbol{m}(\boldsymbol{\theta}) \nabla h(\boldsymbol{m}) \end{aligned} \tag{A.2.1}$$

where $\nabla \boldsymbol{m}(\boldsymbol{\theta})$ is a $q \times q$ gradient matrix and $\nabla h(\boldsymbol{m})$ denotes a $q \times 1$ gradient vector. The algorithms proposed in Section 3.1.1 approximates the estimate of Lagrange multiplier within the iterations so that the iterations are free of the nuisance Lagrange parameters, see Matthews and Crowther (1995), Strydom and Crowther (2012). The Taylor series expansion of $h(\boldsymbol{m})$ around $T(\boldsymbol{X})$ and the approximation of unknown γ is performed as follows:

$$0 = h(\boldsymbol{m}) = h(T) + \alpha_2 \nabla h(\boldsymbol{m})^\top \nabla \boldsymbol{m}(\boldsymbol{\theta}) \nabla h(T) + o(||\boldsymbol{m} - T||)$$

$$\alpha_2 = -\left[\nabla h(\boldsymbol{m})^\top \nabla \boldsymbol{m}(\boldsymbol{\theta}) \nabla h(T)\right]^{-1} h(T)$$

Substituting this value in (A.2.1) we obtain the final approximations for $\boldsymbol{m}$ to be the equation (3.1.11).

Example: We focus on the constraint $\boldsymbol{\Sigma\mu} = \boldsymbol{\mu}$ under normal distribution as an example of the general set up described above. We can rewrite the constraint in terms of a suitable differentiable h, and as a function of the expectation of the sufficient statistic. The log-likelihood of normal distribution as in (3.1.1), can also be expressed in terms of natural parameters in the following way:

$$\begin{aligned} l(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \boldsymbol{X}) &\propto n\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \bar{\boldsymbol{x}} - \frac{n}{2} \mathrm{Tr}\left[\boldsymbol{\Sigma}^{-1}\left(\frac{1}{n}\sum_{i=1}^{p} \boldsymbol{x}_i \boldsymbol{x}_i^\top\right)\right] - \frac{n}{2}\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \frac{n}{2}\log[\det(2\pi\boldsymbol{\Sigma})] \\ &= \boldsymbol{\theta}^\top T - a(\boldsymbol{\theta}) \end{aligned}$$

Here T, the sufficient statistic and $\boldsymbol{\theta}$, the natural parameter are:

$$T(\boldsymbol{X}) = \begin{pmatrix} \bar{\boldsymbol{x}} \\ \mathrm{vec}\left(\frac{1}{n}\sum_{i=1}^{p} \boldsymbol{x}_i \boldsymbol{x}_i^\top\right) \end{pmatrix}, \qquad \boldsymbol{\theta} = \begin{pmatrix} n\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\ -\frac{n}{2}\mathrm{vec}(\boldsymbol{\Sigma}^{-1}) \end{pmatrix}$$

with

$$\mathbb{E}(T) = \boldsymbol{m} = \begin{pmatrix} \boldsymbol{\mu} \\ \mathrm{vec}(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top) \end{pmatrix} = \begin{pmatrix} \boldsymbol{m}_1 \\ \boldsymbol{m}_2 \end{pmatrix}, \quad \mathrm{cov}(T) = V = \begin{pmatrix} \boldsymbol{V}_{11} & \boldsymbol{V}_{12} \\ \boldsymbol{V}_{21} & \boldsymbol{V}_{22} \end{pmatrix}$$

where

$$\boldsymbol{V}_{11} = \frac{1}{n}\boldsymbol{\Sigma}, \qquad \boldsymbol{V}_{12} = \frac{1}{n}(\boldsymbol{\Sigma} \otimes \boldsymbol{\mu} + \boldsymbol{\mu} \otimes \boldsymbol{\Sigma})$$

$$V_{21} = V_{21}^{\top}, \qquad V_{22} = \frac{1}{n}\left(I_{p^2} + K\right)\left[\Sigma \otimes \Sigma + \Sigma \otimes \mu\mu^{\top} + \mu\mu^{\top} \otimes \Sigma\right]$$

The matrix K is given by $K = \sum_{i,j=1}^{p} H_{ij} \otimes H_{ij}$, where H_{ij} : zero matrix except (i, j)-th element, $h_{ij} = 1$.

Proof of the Form of h in (3.1.9): The condition $Tr[(\Sigma - I_p)R_\mu] = \sum_{i=1}^{p}(\Sigma - I_p)_{i.}\mu = 0$, where $R_\mu = \mu \otimes \mathbb{1}^{\top}$ and $\mathbb{1} = [1, 1, \ldots, 1]^{\top}$ will be written in the form $h(m) = Tr[(\Sigma - I_p)R_\mu]$. We know that $\text{vec}(R_\mu) = \mathbb{1} \otimes \mu$.

$$\begin{aligned} m_2 &= \text{vec}(\Sigma + \mu\mu^{\top}) = \text{vec}(\Sigma) + \text{vec}(\mu\mu^{\top}) = \text{vec}(\Sigma) + m_1 \otimes m_1 \\ h(m) &= \left[m_2 - m_1 \otimes m_1 - \text{vec}(I_p)\right]^{\top}(\mathbb{1} \otimes m_1) = \left[\text{vec}(\Sigma) - \text{vec}(I_p)\right]^{\top}(\mathbb{1} \otimes \mu) \\ &= \text{vec}\left(\Sigma - I_p\right)^{\top} \text{vec}(R_\mu) = Tr\left[\left(\Sigma - I_p\right)R_\mu\right] \end{aligned}$$

The iteration in (3.1.11) requires ∇h and ∇m. Note that $\nabla m = V$ and ∇h is calculated as follows:

$$\nabla h = \left[\frac{\partial h}{\partial m}\right]_{(p^2+p)\times 1} = \begin{pmatrix} \frac{\partial h}{\partial m_1} \\ \frac{\partial h}{\partial m_2} \end{pmatrix}$$

where

$$\begin{aligned} \frac{\partial h}{\partial m_1} &= \frac{\partial\left(m_2 - \text{vec}(I_p)\right)^{\top}(\mathbb{1} \otimes m_1)}{\partial m_1} - \frac{(m_1 \otimes m_1)^{\top}(\mathbb{1} \otimes m_1)}{\partial m_1} \\ &= \left(\mathbb{1} \otimes I_p\right)^{\top}\left(m_2 - \text{vec}(I_p)\right) - \left(\mathbb{1} \otimes I_p\right)^{\top}(m_1 \otimes m_1) \\ &\qquad - \left(m_1 \otimes I_p + I_p \otimes m_1\right)^{\top}(\mathbb{1} \otimes m_1) \\ \frac{\partial h}{\partial m_2} &= \mathbb{1} \otimes m_1. \end{aligned}$$

Algorithm 2: The detailed steps for finding the constrained mle in an exponential family is:

Step 1. Start with an initial value T_0, the vector of observed canonical statistics.

Step 2. Set $T = T_0$

Step 3A. For the $l-$th iteration of $\boldsymbol{m}$: $\boldsymbol{m}^{(l)} = T$ and calculate $\nabla h\left(\boldsymbol{m}^{(l)}\right)$ and $\boldsymbol{V}^{(l)}$ as a function of $\boldsymbol{m}^{(l)}$ and T.

Step 3B. For the $k-$th iteration of T:

1. calculate $h\left(T^{(k)}\right)$, $\nabla h\left(T^{(k)}\right)$.
2. use (3.1.11) to update:

$$T^{(k+1)} = T^{(k)} - \boldsymbol{V}^{(l)} \nabla h(\boldsymbol{m}^{(l)}) \frac{h\left(T^{(k)}\right)}{\left[\nabla h\left(\boldsymbol{m}^{(l)}\right)^{\top} \boldsymbol{V}^{(l)} \nabla h\left(T^{(k)}\right)\right]}.$$

3. If $\left\| T^{(k+1)} - T^{(k)} \right\| \leq \epsilon$ for some fixed number ϵ determining accuracy, set $T = T^{(k+1)}$ break the loop and go to Step 3A, else repeat the loop.

Step 4. If $\left\| T - \boldsymbol{m}^{(l)} \right\| \leq \epsilon$ the convergence is attained and the constrained MLE estimate is $\boldsymbol{m} = \boldsymbol{m}^{(l)}$.

APPENDIX B

SECOND APPENDIX

B.1 Proofs of Results:

1. In case of normal-inverse Wishart prior the posterior density with parameter $(\boldsymbol{\mu}_n, \kappa_n, \nu_n, \boldsymbol{\Lambda}_n)$ is the following:

$$p(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \boldsymbol{\mu}_n, \kappa_n, \nu_n, \boldsymbol{\Lambda}_n) \propto \kappa_n^{-\frac{p}{2}} \mid \boldsymbol{\Sigma} \mid^{-\left(\frac{\nu_n+p}{2}+1\right)} \exp\left[-\frac{1}{2}\left\{\mathrm{Tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}_n\right) + \kappa_n (\boldsymbol{\mu}-\boldsymbol{\mu}_n)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}-\boldsymbol{\mu}_n)\right\}\right]$$

After taking logarithm and differentiating we get :

$$\frac{\partial \log p(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \boldsymbol{\mu}_n, \kappa_n, \nu_n, \boldsymbol{\Lambda}_n)}{\partial \boldsymbol{\Sigma}^{-1}} = \left(\frac{\nu_n + p}{2} + 1\right)\boldsymbol{\Sigma} - \frac{1}{2}\boldsymbol{\Lambda}_n - \frac{\kappa_n}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}_n)(\boldsymbol{\mu}-\boldsymbol{\mu}_n)^\top = 0$$

$$\widehat{\boldsymbol{\Sigma}}_{map} = \frac{\boldsymbol{\Lambda}_n + \kappa_n (\boldsymbol{\mu}-\boldsymbol{\mu}_n)(\boldsymbol{\mu}-\boldsymbol{\mu}_n)^\top}{\nu_n + p + 2}$$

By applying Theorem 4.2.1 of Bibby et al. (1979) we can say that this is indeed the MAP estimator.

2. Let $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and the prior on $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is normal-shrinkage inverse Wishart prior i.e. $\boldsymbol{\mu} \mid \boldsymbol{\Sigma} \sim N_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}/\kappa_0)$ and $\boldsymbol{\Sigma} \sim \pi_{SIW}(\nu_0, b, \boldsymbol{\Lambda}_0^{-1})$. Here w are interested with $b = 1$ as stated earlier. The posterior distribution is calculated as follows:

$$\pi(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{\mu}_0, \kappa_0, \nu_0, \boldsymbol{\Lambda}_0) \propto \prod_{i=1}^{n} N_p(\boldsymbol{x}_i; \boldsymbol{\mu}, \boldsymbol{\Sigma}) N_p(\boldsymbol{\mu}; \boldsymbol{\mu}_0, \boldsymbol{\Sigma}/\kappa_0) \pi_{SIW}(\boldsymbol{\Sigma}; \nu_0, 1, \boldsymbol{\Lambda}_0^{-1})$$

$$\propto \prod_{i=1}^{n} N_p(\boldsymbol{x}_i; \boldsymbol{\mu}, \boldsymbol{\Sigma}) N_p(\boldsymbol{\mu}; \boldsymbol{\mu}_0, \boldsymbol{\Sigma}/\kappa_0) \frac{\pi_{IW}(\boldsymbol{\Sigma};\, \nu_0, \boldsymbol{\Lambda}_0^{-1})}{\prod_{i<j}(\lambda_i - \lambda_j)}$$

This is because shrinkage-inverse-Wishart can be obtained from inverse-Wishart density by dividing it with $\prod_{i<j}(\lambda_i - \lambda_j)$. We know that the normal-inverse Wishart prior is conjugate with posterior parameters described in (2.6.8), $\kappa_n = \kappa_0 + n$ and $\nu_n = \nu_0 + n$ respectively. The numerator is the posterior distribution corresponding to the normal-inverse Wishart prior and gives us the following:

$$\begin{aligned}
\pi(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{\mu}_0, \kappa_0, \nu_0, \boldsymbol{\Lambda}_0) &\propto \frac{\pi_{NIW}(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \boldsymbol{\mu}_n, \kappa_n, \nu_n, \boldsymbol{\Lambda}_n^{-1})}{\prod_{i<j}(\lambda_i - \lambda_j)} \\
&\propto N_p(\boldsymbol{\mu} \mid \boldsymbol{\Sigma}, \boldsymbol{\mu}_n, \kappa_n) \frac{\pi_{IW}(\boldsymbol{\Sigma} \mid \nu_n, \boldsymbol{\Lambda}_n^{-1})}{\prod_{i<j}(\lambda_i - \lambda_j)} \\
&\propto N_p(\boldsymbol{\mu} \mid \boldsymbol{\Sigma}, \boldsymbol{\mu}_n, \kappa_n) \pi_{SIW}(\boldsymbol{\Sigma} \mid \nu_n, 1, \boldsymbol{\Lambda}_n^{-1})
\end{aligned}$$

From the above calculation we can see that the posterior is normal-shrinkage inverse Wishart with the same parameters as in normal-inverse Wishart proving our hypothesis.

3. **Calculation of the Posterior Distribution:**

$$\begin{aligned}
p(\boldsymbol{\mu}, \boldsymbol{\lambda} \mid \boldsymbol{X}) &\propto \left(\prod_{i=1}^{p-1} \lambda_i\right)^{-\frac{n}{2}} \exp\left[-\frac{1}{2} Tr\left\{\boldsymbol{\Lambda}^{-1} B(\boldsymbol{X}, \boldsymbol{\mu})\right\}\right] \left(\prod_{i=1}^{p-1} \lambda_i\right)^{-\frac{1}{2}} \\
&\qquad \exp\left[-\frac{1}{2} Tr\left\{\boldsymbol{\Lambda}^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_0)(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^\top\right\}\right] \left(\prod_{i=1}^{p-1} \lambda_i\right)^{-a} \exp\left[-\frac{1}{2} Tr\left\{\boldsymbol{\Lambda}^{-1} \boldsymbol{H}_0\right\}\right] \\
&\propto \left(\prod_{i=1}^{p-1} \lambda_i\right)^{-\frac{n+2(a+1)}{2}} \exp\left[-\frac{1}{2} Tr\left\{\boldsymbol{\Lambda}^{-1}\left(B(\boldsymbol{X}, \boldsymbol{\mu}) + (\boldsymbol{\mu} - \boldsymbol{\mu}_0)(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^\top + \boldsymbol{H}_0\right)\right\}\right] \\
&\propto \left(\prod_{i=1}^{p-1} \lambda_i\right)^{-\frac{n+2(a+1)}{2}} \exp\left[-\frac{1}{2} Tr\left\{\boldsymbol{\Lambda}^{-1} \boldsymbol{H}_N\right\}\right]
\end{aligned}$$

where $B(\boldsymbol{X}, \boldsymbol{\mu}) = \boldsymbol{P}(\boldsymbol{\mu})^\top \boldsymbol{A}(\boldsymbol{\mu}) \boldsymbol{P}(\boldsymbol{\mu})$.

B.2 Normal-Shrinkage Inverse Wishart Gibbs Sampling Method:

Here we give a concise description of this 'new method' which uses the Gibbs sampler:

1. From (2.6.10) we get:

$$-\frac{1}{2}Tr(\boldsymbol{\Lambda}_0\boldsymbol{P}\boldsymbol{D}^{-1}\boldsymbol{P}^\top) = -\frac{1}{2}Tr(\boldsymbol{P}^\top\boldsymbol{\Lambda}_0\boldsymbol{P}\boldsymbol{D}^{-1}) = \sum_{i=1}^{p}\frac{a_i}{\lambda_i}$$

where a_i's are the diagonal elements of the matrix $\boldsymbol{P}^\top\boldsymbol{\Lambda}_0\boldsymbol{P}$. This implies that

$$\pi(\boldsymbol{D} \mid P, \boldsymbol{\Lambda}_0) \propto \prod_{i=1}^{p}\lambda_i^{-\frac{\nu_0+p+1}{2}}e^{-\frac{a_i}{\lambda_i}} \tag{B.2.1}$$

i.e. eigenvalues can be generated from inverse gamma $\left(\frac{\nu_0+p-1}{2}, a_i\right)$ and ordered.

2. The next step is to generate from $\pi(\boldsymbol{P} \mid \boldsymbol{D}, \boldsymbol{\Lambda}_0)$. For this step without loss of generality we assume that $\boldsymbol{\Lambda}_0$ is diagonal i.e. $\boldsymbol{\Lambda}_0 = \text{diag}(\Lambda_{01}, \Lambda_{02}, \ldots, \Lambda_{0p})$.The principle is to update each pair of rows in $\boldsymbol{P}$. The k-th update of the row number i and j, we write $P^{(k)} = \text{diag}(\boldsymbol{G}, \boldsymbol{I}_{p-2})\left(P_{ij}^{(k-1)\top}, P_{-ij}^{(k-1)\top}\right)^\top$ where $P_{ij}^{(k-1)}$ denotes the $2 \times p$ matrix with the i-th and j-th rows of $P^{(k-1)}$ and $P_{-ij}^{(k-1)}$ denotes the rest of the matrix of dimension $(p-2) \times p$ in the $(k-1)$ th step. Let us define $\boldsymbol{D}_\epsilon = \text{diag}(\epsilon_1, \epsilon_2)$ with $\epsilon_i = \pm 1$, $\boldsymbol{R}_{rot}(\theta)$ a 2×2 rotation matrix and $\boldsymbol{G} = \boldsymbol{D}_\epsilon\boldsymbol{R}_{rot}(\theta)$. Let us denote $\boldsymbol{\Lambda}_1 = \text{diag}(\Lambda_{0i}, \Lambda_{0j})$ and $\boldsymbol{\Lambda}_2$ diagonal with the rest of the values. The distribution of θ given $\boldsymbol{P}_{ij}^{(k-1)}$, $\boldsymbol{P}_{-ij}^{(k-1)}$,$\boldsymbol{D}^{(k-1)}$ and $\boldsymbol{\Lambda}_0$ is:

$$\pi(\theta \mid \boldsymbol{P}_{ij}, \boldsymbol{P}_{-ij}, \boldsymbol{D}; \boldsymbol{\Lambda}_0) \propto \exp\left[c_0\cos^2(\theta+\omega)\right] \tag{B.2.2}$$

where $\boldsymbol{P}_{ij}\boldsymbol{D}^{-1}\boldsymbol{P}_{ij}^\top = \boldsymbol{R}_{rot}(\omega)\text{diag}(s_1, s_2)\boldsymbol{R}_{rot}^\top(\omega)$ and $c_0 = -\frac{1}{2}(s_1 - s_2)(\Lambda_{0i} - \Lambda_{0j})$. If we do a change of variable by $\alpha = \cos^2(\theta+\omega)$ then

$$\pi(\alpha \mid \boldsymbol{P}_{ij}^{(k-1)}, \boldsymbol{P}_{-ij}^{(k-1)}, \boldsymbol{D}; \boldsymbol{\Lambda}_0) \propto \exp[c_0\alpha]\, \alpha^{-\frac{1}{2}}(1-\alpha)^{-\frac{1}{2}}.$$

Since the normalizing constant in this set up is not known, it is not easy to simulate from this distribution. We have to resort to rejection sampling from Beta($\frac{1}{2}, \frac{1}{2}$). For the rejection sampling, we need to find out a constant M_0 such that

$$\pi(\alpha \mid \boldsymbol{P}_{ij}^{(k-1)}, \boldsymbol{P}_{-ij}^{(k-1)}, \boldsymbol{D}; \boldsymbol{\Lambda}_0) \leq M_0 \text{Beta}\left(\frac{1}{2}, \frac{1}{2}\right).$$

Let M be the normalizing constant. We know that

$$M = \int_0^1 \exp[c_0\alpha]\alpha^{-\frac{1}{2}}(1-\alpha)^{-\frac{1}{2}} d\alpha \geq \int_0^1 2\exp[c_0\alpha] d\alpha = \frac{2(e^{c_0}-1)}{c_0}$$

Since c_0 is negative,

$$\begin{aligned}\pi(\alpha \mid \boldsymbol{P}_{ij}^{(k)}, \boldsymbol{P}_{-ij}^{(k)}, \boldsymbol{D}; \boldsymbol{\Lambda}_0) &= \frac{\exp[c_0\alpha]\alpha^{-\frac{1}{2}}(1-\alpha)^{-\frac{1}{2}}}{M} \leq \frac{\beta\left(\frac{1}{2}, \frac{1}{2}\right)}{M} \frac{\alpha^{-\frac{1}{2}}(1-\alpha)^{-\frac{1}{2}}}{\beta\left(\frac{1}{2}, \frac{1}{2}\right)} \\ &\leq \frac{c_0\beta\left(\frac{1}{2}, \frac{1}{2}\right)}{2(e^{c_0}-1)} \frac{\alpha^{-\frac{1}{2}}(1-\alpha)^{-\frac{1}{2}}}{\beta\left(\frac{1}{2}, \frac{1}{2}\right)}\end{aligned}$$

This gives us the value of M_0 to be $\frac{c_0\beta\left(\frac{1}{2}, \frac{1}{2}\right)}{2(e^{c0}-1)}$.

Another crucial point is the choice of $\theta, \omega \in [-\pi/2, \pi/2]$. ω can be solved from the equation constructed from the off diagonal elements of the matrix:

$$\begin{aligned}\boldsymbol{P}_{ij}\boldsymbol{D}^{-1}\boldsymbol{P}_{ij}^{\top} &= \boldsymbol{R}_{rot}(\omega)\text{diag}(s_1, s_2)\boldsymbol{R}_{rot}^{\top}(\omega) \\ (P_{ij})_{1.}D^{-1}(P_{ij})_{2.}^{\top} &= \frac{s_1 - s_2}{2}\sin(2\omega)\end{aligned}$$

Once we get ω we can use that to compute θ from α giving us the joint update for two rows of $\boldsymbol{P}$. We repeat this process for each pair to update the eigenvector matrix.

APPENDIX C

THIRD APPENDIX

C.1 Simulation tables using Frobenius norm for MLE approximation, Gibbs sampling and MAP approximation

Table C.1: Risk ratio of MLE and MAP approximation (through a lower bound) relative to the MAP of normal-inverse Wishart (using Frobenius norm)

		MLE Approx. Risk		**MAP Approx. Risk**		**Time (Sec)**
n	**p**	**Mean**	**Sigma**	**Mean**	**Sigma**	
50	3	0.4253	1.1331	0.4253	1.1331	3.25
50	5	0.6625	1.1528	0.6625	1.1528	2.59
50	10	1.5009	1.195	1.5009	1.195	5.22
100	3	0.3481	1.4383	0.3481	1.4383	2.52
100	5	0.5342	1.5065	0.5342	1.5065	2.69
100	10	1.401	1.63	1.401	1.63	2.84
300	3	0.3159	2.3753	0.3159	2.3753	2.7
300	5	0.5797	2.6429	0.5797	2.6429	2.95
300	10	1.303	2.6493	1.303	2.6493	3.41

Table C.2: Risk ratio of MAP approximation (from MH within Gibbs sampling) relative to the MAP of normal-inverse Wishart (using Frobenius norm)

		Normal-Inverse Gamma Risk		**Acc Rate (in MH)**	**Time (Sec.)**
n	**p**	**Mean**	**Sigma**		
50	3	1.1173	0.8928	0.4135	980.52
50	5	1.0824	1.0963	0.2822	2029.15
50	10	1.2423	1.2798	0.1218	13574.59
100	3	1.0542	0.9872	0.4352	910.92
100	5	1.1677	1.3662	0.2908	1901.41
100	10	1.1963	1.6229	0.1245	12183.56
300	3	1.1661	1.2546	0.4287	1256.81
300	5	1.5729	2.0163	0.2895	3215.97
300	10	1.5533	2.4053	0.1294	9385

C.2 Plot of Constrained and Unconstrained Normal

The graphical representation of normal distribution is only possible for two dimensional case only. It is shown in Figure C.1. Let us select $\boldsymbol{\mu} = (1,1)^\top$ then the corresponding covariance matrix will be compound symmetric. The following is the form of the covariance matrix in stuctured model proposed in (1.3).

$$\boldsymbol{D} = \begin{bmatrix} 1 & 0 \\ 0 & \frac{1-\rho}{1+\rho} \end{bmatrix}, \quad \boldsymbol{P} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}, \quad \boldsymbol{\Sigma} = \frac{1}{1+\rho} \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

However for unconstrained case we can choose the covariance matrix arbitrarily. The following is a contour plot of constrained and unconstrained normal density. The unconstrained normal density has the same mean and the covariance matrix as $\Sigma = \begin{bmatrix} 2 & 0.7 \\ 0.7 & 3 \end{bmatrix}$

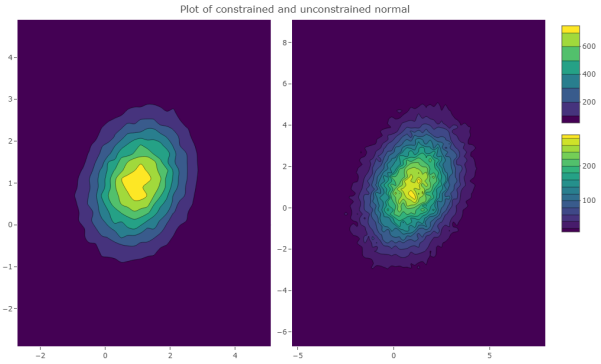

Figure C.1: Plot on the left is constrained normal with $\rho = 0.2$, on the right we have the unconstrained normal

APPENDIX D

FOURTH APPENDIX

D.1 Variable Selection

1. **A Brief Review of Standard Variable Selection (Test-based):**

   Due to the infeasibility of model uncertainty (i.e. looking at all subsets) for large values of p, Breaux (1967) proposed the idea of stepwise regression which either starts from null model and add variables into the model or begins with a full model and proceed by deleting some variables one by one based on some criteria like, lowest p-value, highest adjusted R^2, lowest Mallow's C_p, lowest AIC, AICc, BIC or HQIC (Hannan and Quinn, 1979), lowest prediction error, leave-one-out cross validation, etc. One interesting criticism (Doornik, 2009) is that stepwise proceeds without backtesting and is proved to be biased and inconsistent (Hurvich and Tsai (1990),Flom and Cassell (2007) etc.)

2. **Bayesian Variable Selection and Lasso-Type Model Fitting (Penalty-based):**

   Penalty based variable selection methods like LASSO (Tibshirani, 1996), Ridge regression etc. can be interpreted as Bayesian posterior mode estimate (e.g. independent Laplace prior on parameters in the case of LASSO, see Park and Casella (2008)). Various priors are proposed in the literature in this context e.g. horseshoe prior (Carvalho et al., 2010), Laplace prior (Park and Casella, 2008), a family of priors like g-prior (Fernandez et al. (2001), Berger et al. (2001)) etc. O'Hara et al. (2009) classifies the priors into four categories: indicator model selection (e.g. Kuo and Mallick (1998), Dellaportas et al. (2002)), stochastic search variable selection (SSVS) see George and McCulloch (1993), adaptive shrinkage and model space approach.

3. **Modern Variable Selection Methods: Clustering Variables (Screening):**

   These are ranking methods that rely on some association measure between the dependent

variable and the regressors. The most common and first of it kind is Sure Independence Screening (SIS Fan and Lv (2008)). SIS needs a selection procedure in the end to obtain consistent results (e.g. LASSO). Other screening methods include Covariate Assisted Screening Estimates (CASE Ke et al. (2014)).