MLE of Jointly Constrained Mean-Covariance of Multivariate Normal Distributions

Anupam Kundu and Mohsen Pourahmadi
Texas A&M University, College Station, USA

Abstract

Estimating the unconstrained mean and covariance matrix is a popular topic in statistics. However, estimation of the parameters of $N_p(\mu, \Sigma)$ under joint constraints such as $\Sigma\mu = \mu$ has not received much attention. 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. We derive the non-linear likelihood equations for the constrained maximum likelihood estimator of $(\mu, \Sigma)$ 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. A simulation study illustrates our methodology and shows that the modified estimators perform better than the initial estimators from the iterative methods.

AMS (2000) subject classification. 62H12, 62F10, 62F30, 65H17.
Keywords and phrases. maximum likelihood estimation, iterative methods, lagrange multiplier, positive-definite matrices, covariance matrix.

1 Introduction

Mean and covariance estimation are of central importance in almost every area of multivariate statistics. However, estimation under joint constraints on the mean vector and covariance matrix of data from a $N_p(\mu, \Sigma)$
distribution is relatively uncommon in multivariate statistics (Bibby et al., 1979). Our goal is to study and resolve some new challenges which appear when one attempts to jointly estimate the mean vector and the covariance matrix of a multivariate normal distribution under the following two constraints:

\[ \Sigma \mu = \mu, \quad |\Sigma| = 1. \]  

(1.1)

It is interesting to note that the first constraint forces the mean vector \( \mu \) to be an eigenvector of \( \Sigma \) 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 \( \mu \) 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

\[ \Sigma = PDP^\top = \sum_{i=1}^{p} \lambda_i P_i P_i^\top = \sum_{i=1}^{p-1} \lambda_i P_i P_i^\top + \frac{\mu \mu^\top}{\|\mu\|^2}, \]  

(1.2)

where \( D = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_{p-1}, 1) \) is the diagonal matrix of ordered eigenvalues other than 1 and \( P = [P_1, P_2, \ldots, P_p] \) 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.

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 \( \mu \mu^\top \) in (1.2) 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 and et al, 1975) where the parameters \( (\mu, \Sigma) \) satisfy the constraints in (1.1). As a potential relaxation of the first constraint which is the source of most complications, and for the sake of demonstration, an intermediate constraint \( \Sigma b = \mu \) for some possibly known vector \( b \) is also considered in Theorem 1, hoping that it will shed more light on the nature of the constraints and the possible challenges in standard MLE computation through iteration. We interpret the constraints in the context of factor
and error-in-variable models in multivariate regression (Molstad et al., 2020) where the error covariance matrix and the regression coefficient matrix are parameterically connected.

Though an explicit formula for the MLE of $\theta$ 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). In this paper, focusing on multivariate normal distributions we incorporate the constraints in computing the MLE of the mean-covariance parameters, and derive the (constrained) likelihood equations. In the absence of closed-form MLE, three iterative methods for computing the MLE and the Lagrange multiplier are provided and we study their statistical/computational properties. Unfortunately, computing the Lagrange multiplier in our setup is not straightforward, perhaps due to implicit nonlinearity in the first constraint, and requires special attention. We compute the Lagrange multiplier using either an iterative or explicit methods.

It turns out that the presumed MLEs obtained from the iterative methods invariably do not satisfy the constraint in Eq. 1.1, 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 (1.1), in addition to being symmetric and positive definite. A novel algorithm is developed where starting with any pair of mean-covariance estimators, they are modified so as to satisfy the conditions in (1.1). 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.

The paper is organized as follows: Section 2 provides statistical interpretation of the model with a few examples. Section 3 describes three iterative methods of computing the MLE and their modifications, two of the methods employ explicit calculation of the Lagrange multiplier (Aitchison and Silvey, 1958). Section 4 studies concavity of the Lagrangian function and provides further theoretical justification for using the iterative methods. Section 5 gives the details of developing algorithms to modify estimators satisfying both constraints, and Section 6 illustrates our methods through simulations. Section 7 is the conclusion.
2 Examples of Non-linear Dependency of $\Sigma$ on $\mu$

1. Earth’s Magnetic Pole Data: The mean direction estimate data of Earth’s historic magnetic pole from 33 different sites in Tasmania (Schmidt, 1976) has been studied by Paine et al. (2018). It is shown that the angular Gaussian distribution i.e. the distribution of $Y = \frac{X}{\|X\|} \in S^{p-1}$ where $X \sim N_p(\mu, \Sigma)$, simplifies significantly under the constraint (1.1). The ensuing subfamily is called the elliptically symmetric angular Gaussian (ESAG) distribution (Paine et al. 2018, Section 2.5) has provided strong evidence in favor of the ESAG distribution over the isotropic angular Gaussian (IAG) distribution using a large sample likelihood ratio test.

2. Factor Model: Consider a factor model with a single factor of the form (Rao, 1973, §8f.4)

$$X_i = \mu + \mu w_i + \epsilon_i$$  \hspace{1cm} (2.1)

where $w_i \overset{iid}{\sim} N(0, 1)$ and $\epsilon_i \overset{iid}{\sim} N_p(0, \Sigma_\epsilon)$ are uncorrelated. Note that the mean vector $\mu$ appears as the loading matrix and $w_i$ is the common factor. The covariance matrix of $X_i$ is as in (1.2):

$$\Sigma = \Sigma_\epsilon + \mu \mu^T.$$  

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.

3. Invariant Distribution: When one deals with discrete Markov chain with the transition matrix $P$ (Hoel and Port, 1986) its stationary distribution $\pi$ satisfies $\pi P = \pi$ which looks similar to our first constraint. The Perron-Frobenius theorem shows that $P$ has a unique equilibrium with only one eigenvalue 1 and the existence of a $\pi$ when all entries of the transition matrix are positive (Lemmens and Nussbaum, 2012). This result has many applications: for example in chaos theory, search engine for Google’s PageRank algorithm (Knill, 2011, Lecture 34).

3 Constrained Maximum Likelihood Estimation

The Lagrange multiplier method (Aitchison and Silvey, 1958) is used to incorporate the constraints for finding the MLE of the parameters of a multivariate distribution. We derive the likelihood equations, present three
MLE of Jointly Constrained...  

iterative methods and study some of their computational and statistical 
properties. 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 iter-
ative 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 \( x_1, x_2, \ldots, x_n \) be a sample of size \( n \) from \( N_p(\mu, \Sigma) \) where \( \Sigma \) is a 
positive-definite matrix. If \( X \) is the \( n \times p \) data matrix, then the log-likelihood 
of the multivariate normal distribution is proportional to

\[
\ell(\mu, \Sigma | X) \propto -\frac{n}{2} \log |\Sigma| - \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)\top \Sigma^{-1} (x_i - \mu).
\]  

(3.1)

Let \( y = x - \frac{1}{n} \mu \) and \( \Sigma = \frac{A(\bar{x}) + 2\alpha_2 \mu\top}{n + 2\alpha_1} \), where \( \alpha_1 \) and \( \alpha_2 \) are the Lagrange multipliers.

(a) Under the lone constraint \( \Sigma = 1 \) (\( \alpha_2 = 0 \)), the MLE \( \hat{\mu}_{mle} = \bar{x} \) is the 

sample mean and \( \hat{\Sigma}_{mle} = \frac{A(\bar{x})}{|A(\bar{x})|^{1/p}} \) is a shape matrix.

(b) If \( \Sigma = 1 \) and \( \Sigma b = \mu \) as in (3.2), then the constrained MLE satisfies

\[
\mu = \left( \bar{x} - \frac{1}{n} \alpha_2 \right), \quad \Sigma = \frac{[A(\mu) + 2\alpha_2 \mu\top]}{n + 2\alpha_1},
\]  

(3.3)

\[
| \Sigma | = 1.
\]  

(3.4)

(c) Under both constraints in (1.1), the MLE satisfies

\[
\mu = \left( \bar{x} - \frac{1}{n} (I - \Sigma) \alpha_2 \right), \quad \Sigma = \frac{[A(\mu) + 2\alpha_2 \mu\top]}{n + 2\alpha_1},
\]  

(3.5)

\[
| \Sigma | = 1.
\]
The proof is provided in the Appendix A.1. Unlike the closed-form solution in (a), computing the MLE in (b) and (c) is more challenging and involves both $\alpha_1$ and $\alpha_2$. 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 1.1, every parameter can be expressed in terms of $\alpha_2$ due to the intermediate constraint $\Sigma b = \mu$, so that the iterations will be over $\alpha_2$ only, see Section 3.1.1 for details. By contrast, the case in Theorem 1.1 under $\Sigma \mu = \mu$ is much more challenging, at least, due to the presence of $\Sigma$ in $\mu$. These observations serve as strong motivations for considering the alternative method of explicit calculation of the Lagrange multiplier $\alpha_2$ in Section 3.2. In view of Theorem 1 (a), from here on we focus mostly on the first constraint and deemphasize the second constraint $|\Sigma| = 1$ which is achievable through a scale change.

3.1. Algorithms for Computing Constrained MLE: In spite of the apparent closed forms in (3.3) and (3.5), these can not be implemented or viewed as bona fide estimators because of their dependence on the Lagrange multipliers $\alpha_1$ and $\alpha_2$. 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. Solving (3.3) for $\alpha_2$. Knowing $\alpha_2$ in (3.3), determines all the other unknown quantities. To emphasize dependence on $\alpha_2$, we set $\mu = \mu(\alpha_2)$ and denote the numerator of $\Sigma$ by

$$U(\alpha_2) = A[\mu(\alpha_2)] + 2\alpha_2\mu^T(\alpha_2).$$

From the second constraint in (3.4) it follows that $|U(\alpha_2)|^{1/p} = n + 2\alpha_1$. Replacing the numerator by $U(\alpha_2)$ and the denominator by $|U(\alpha_2)|^{1/p}$ in the right hand side of the second identity of (3.3) leads to

$$\Sigma(\alpha_2) = \frac{U(\alpha_2)}{|U(\alpha_2)|^{1/p}},$$

which is a function of $\alpha_2$. Substituting $\Sigma(\alpha_2)$ in the first expression of (3.4), we obtain

$$\mu(\alpha_2) = \frac{\Sigma(\alpha_2)b}{|\Sigma(\alpha_2)|^{1/p}}, \tag{3.6}$$
where further replacing \( \mu(\alpha_2), \Sigma(\alpha_2), U(\alpha_2) \) and \( n + 2\alpha_1 \) in terms of \( \alpha_2 \) one obtains the following after some algebraic manipulation:

\[
\alpha_2 = \frac{|\Sigma(\alpha_2)|^{1/p} \bar{x} - (n-1)Sb - (1/n^2)\alpha_2 \bar{\alpha}_2 b}{2(\bar{x}^\top b - \bar{\alpha}_2 b/n)} = f(\alpha_2). \tag{3.7}
\]

This being nonlinear in \( \alpha_2 \) suggests using the iterations:

\[
\alpha_2^{(k+1)} = f(\alpha^{(k)}), k = 0, 1, 2 \ldots, \text{ with } \alpha_2^{(0)} = \bar{x},
\]

for solving it.

Although the intermediate constraint seems similar to (1.1), in the next section it is demonstrated that the latter is much harder to work with in that one needs to iterate over the \( \alpha_1 \) as well.

3.1.2. Solving (3.5) for \( \alpha_1 \) and \( \alpha_2 \). After replacing \( \mu \) from the first identity, which involves \( \Sigma \), the second equation in (3.5) reveals that \( \Sigma \) is a nonlinear function of \( \alpha_2 \). This is different from Theorem 1.1 in that not all parameters can be expressed in terms of a single parameter (like \( \alpha_2 \)). Thus, one may resort to iterative methods involving the four parameters \((\mu, \Sigma, \alpha_1, \alpha_2)\) where the updates for the \((k+1)\)-th iteration is done in the following order:

\[
\begin{align*}
\alpha_1^{(k+1)} &= \frac{1}{2} \left( |A(\mu^{(k)}) + 2\alpha_2^{(k)} \mu^{(k)\top}|^{1/p} - n \right), \\
\Sigma^{(k+1)} &= \left[ A(\mu^{(k)}) + 2\alpha_2^{(k)} \mu^{(k)\top} \right]/n = 2\alpha_1^{(k)} \\
\alpha_2^{(k+1)} &= \frac{1}{2} \left( \left(n + 2\alpha_1^{(k)}\right) \Sigma^{(k)} - A(\mu^{(k)}) \right) \mu^{(k)}, \quad \mu^{(k+1)} = \Sigma^{(k)} \left( \bar{x} - \frac{1}{n} (I - \Sigma^{(k)}) \alpha_2^{(k)} \right).
\end{align*} \tag{3.8}
\]

Our suggested initial values are \((\Sigma^{(0)}, \alpha_2^{(0)}, \mu^{(0)}) = (S, \bar{x}, \bar{x})\), and for \( k = 0 \) we compute \( \alpha_1^{(1)} \) using \((\mu^{(0)}, \alpha_2^{(0)})\) from the first equation above. But for the updates \( \Sigma^{(1)} \) and \( \alpha_2^{(1)} \), 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 \((\alpha_1^{(1)}, \alpha_2^{(1)}, \Sigma^{(1)}, \mu^{(1)})\) and repeat the process.

3.1.3. Common Challenges with Iterative Methods for Computing MLE of \( \Sigma \). 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 \( \frac{1}{2} (\Sigma + \Sigma^\top) \), producing an estimator of the form \( A + ab^\top + ba^\top \) where \( a, b \in \mathbb{R}^p \) and a positive definite matrix \( A \). 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 \( a, b \in \mathbb{R}^p \) and \( A \) be a positive definite matrix. Then,

(a) The non-zero eigenvalues of \((ab^\top + ba^\top)\) are \( a^\top b \pm \|a\|\|b\| \)
The matrix $M = A + (ab^\top + ba^\top)$ has at most one negative eigenvalue.

To ensure positive-definiteness, Lemma 1.1 suggests replacing the smallest eigenvalue of $M$ by $(\prod_{j=1}^{p-1} \lambda_j)^{-1}$ where $\lambda_j$’s are the ordered eigenvalues of $M$. This is justified by noting that according to Weyl’s inequality (Bhatia, 2007)

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

In addition, there are a number of existence and convergence problems related to Theorem 1.1. These are dealt with partially in the next two subsections by relying on more explicit calculations of the Lagrange multipliers under the first constraint only.

3.2. 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) (denoted by S&C method) 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 $\Sigma \mu = \mu$ as the scalar function $h : \mathbb{R}^{p^2 + p} \rightarrow \mathbb{R}$ of the natural parameter vector $m$ of a multivariate normal distribution:

$$h(m) = [m_2 - m_1 \otimes m_1 - \text{vec}(I_p)]^\top (1 \otimes m_1) = 0, \quad (3.9)$$

where $m^\top = [\mu^\top, \text{vec}(\Sigma + \mu \mu^\top)]^\top = [m_1^\top, m_2^\top]^\top$.

Using the Taylor’s expansion of $h(m)$ 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:

$$\alpha_2 = -[\nabla h(m)^\top \nabla m(\theta) \nabla h(T)]^{-1} h(T), \quad (3.10)$$

where $\theta$ is the canonical parameter for the multivariate normal distribution, see Appendix B. Substituting this in (B.1) leads to the identity

$$m = T(X) - V \nabla h(m) \frac{h(T)}{[\nabla h(m)^\top V \nabla h(T)]} \quad \text{with} \quad \nabla m(\theta) = V. \quad (3.11)$$

It can be used iteratively via a “double iteration” over $T$ and $m$, see Section 2 of Strydom and Crowther (2012), with the initial values chosen as the observed canonical statistics for both $T$ and $m$, 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. This can be attributed to the explicit calculation of Lagrange multiplier.

3.3. 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 \( U_\epsilon = \{ \theta : \| \theta - \theta_0 \| < \epsilon \} \) where \( \theta_0 \) is the true parameter value for the parameter vector \( \theta = (\mu^T, \text{vec}(\Sigma)^T)^T \) of a multivariate normal distribution.

For the vector-valued constraint function

\[
h(\mu, \Sigma) = \Sigma \mu - \mu,
\]

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

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

The notations \( H^1_\theta \) and \( H^1_{\theta_0} \), with obvious interpretation, are used as needed next. The partitioned matrix \( E = \begin{bmatrix}
B_{\theta_0} & -H^1_{\theta_0} \\
-H^1_{\theta_0} & 0
\end{bmatrix} \) is non singular (Aitchison and Silvey, 1958, Lemma 3) where \( B_{\theta_0} = \begin{pmatrix}
\Sigma^{-1} & 0 \\
0 & \Sigma^{-1} \otimes \Sigma^{-1}
\end{pmatrix}, \) and its inverse is given by

\[
E^{-1} = \begin{bmatrix}
P_\theta & Q_\theta \\
Q_\theta^T & R_\theta
\end{bmatrix}
\]

where

\[
R_\theta = -\left( H^1_{\theta} B^{-1}_{\theta} H^1_{\theta} \right)^{-1} = -\left[ (\Sigma - I)\Sigma(\Sigma - I) + (\mu^T \Sigma \mu) \Sigma \right]^{-1}
\]

\[
Q_\theta = -B_{\theta} H^1_{\theta} R, \quad P_\theta = B^{-1}_{\theta} [I - H^1_{\theta} Q^T].
\]

It follows from Lemmas 1 and 2 (Aitchison and Silvey, 1958) that, under some regularity conditions on the density and the constraint function, the solution to the equation \( \frac{\partial L}{\partial \theta} = 0 \) (first derivative of the Lagrangian function) exists within the set \( U_\epsilon \) almost surely and it maximizes the likelihood function subject to the constraint \( h(\theta) = 0 \). We denote the constrained maximum likelihood estimator by \( \hat{\theta}_n(x) \) and \( \hat{\alpha}_{2n}(x) \) 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:

\[
\sqrt{n} \left( \hat{\theta}_n - \theta_0 \right) \rightarrow N \left( 0, \left[ \frac{P}{\theta_0} 0 \right] \right), \tag{3.12}
\]

Some of the requisite regularity conditions for the above results are verified in the Appendix A 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:

\[
\begin{bmatrix}
\hat{\theta}^{(j+1)} \\
-\frac{1}{2n} \hat{\alpha}_2^{(j+1)}
\end{bmatrix} = \begin{bmatrix}
\hat{\theta}^{(j)} \\
-\frac{1}{2n} \hat{\alpha}_2^{(j)}
\end{bmatrix} + \begin{bmatrix}
P_{\theta} & Q_{\theta} \\
Q_{\theta}^T & R_{\theta}
\end{bmatrix} \begin{bmatrix}
\frac{1}{n} \frac{\partial l(\theta|X)}{\partial \theta} \bigg|_{\theta=\hat{\theta}^{(j)}} + H_{\hat{\theta}^{(j)}} \frac{1}{n} \hat{\alpha}_2^{(j)} \\
0
\end{bmatrix} \begin{bmatrix}
h(\hat{\theta}^{(j)})
\end{bmatrix}
\] \tag{3.13}

where \( \begin{bmatrix} P_{\theta} & Q_{\theta} \\ Q_{\theta}^T & R_{\theta} \end{bmatrix} \) is the inverse of

\[
\begin{bmatrix}
B_{\hat{\theta}^{(j)}} & -H_{\hat{\theta}^{(j)}}^T \\
-H_{\hat{\theta}^{(j)}} & 0
\end{bmatrix}
\]

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.

### 4 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 iterations restricted to this set will move closer to the true parameter.

Recall that with the constraint \( \Sigma \mu = \mu \), the Lagrangian function is

\[
L(\mu, \Sigma \mid X) = l(\mu, \Sigma \mid X) + \alpha_2^T (\Sigma \mu - \mu)
\]

\[
= -\frac{n}{2} \log |\Sigma| - \frac{1}{2} \text{Tr} \left[ A(\mu) \Sigma^{-1} \right] + \alpha_2^T (\Sigma \mu - \mu)
\]

\[
= -\frac{n}{2} \log |\Sigma| - \frac{n}{2} \text{Tr} \left[ S\Sigma^{-1} \right] - \frac{n}{2} \text{Tr} \left[ (\bar{x} - \mu)(\bar{x} - \mu)^T \Sigma^{-1} \right]
\]
MLE of Jointly Constrained...

\[ + \alpha_2^\top (\Sigma \mu - \mu) \tag{4.1} \]

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 \( \Delta A = \{ \Sigma : 0 < \Sigma < A \} \), see (Zwiernik et al., 2017). Let \( S_p \) denote the set of all \( p \times p \) real symmetric matrices as a subset of \( \mathbb{R}^{p(p+1)/2} \) and \( S_{>0,p} \) denotes the open convex cone in \( S_p \) 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 \( \alpha_2^2 \).

**Lemma 2.** For a given value of \( \alpha_2 \) and the mean vector \( \mu \) estimated by \( \bar{x} \), the Lagrangian function \( L : S_p \to \mathbb{R} \) in (4.1) is strictly concave in \( \Sigma \) in the region \( \Delta_{2S} \).

The proof is given in the Appendix (A.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 \( \Sigma_{\text{max}} = \arg\max_{\Sigma \in \Delta_{2S}} L(\bar{x}, \Sigma) \), then \( \Sigma_{\text{max}} \) is unique in \( \Delta_{2S} \).

See Appendix (A.5) for the proof.

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

\[
P \{ \Sigma \in \Delta_{2S} \} = P \left[ 2S - \Sigma \succ 0 \right] = P \left[ 2\Sigma^{-1/2}S\Sigma^{-1/2} - I_p \succ 0 \right] = P \left[ \frac{2}{n}W_{n-1} \succ I_p \right] = P \left[ W_{n-1} \succ \frac{n}{2}I_p \right] = P \left[ \lambda_p(W_{n-1}) > \frac{n}{2} \right].
\]

Interestingly, the probability that the true parameter \( \Sigma \) lies within the set \( \Delta_{2S} \) is independent of \( \Sigma \) and is equal to the probability that \( \lambda_p(W_{n-1}) > \frac{n}{2} \) where \( W_{n-1} \sim \pi_W(n-1, I_p) \). 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.
5 An Algorithm for Enforcing the Constraints

Most estimators presented so far do not necessarily satisfy the constraints. In this section, starting with any reasonable estimators for \((\mu, \Sigma)\) (like those in Sections 3), we present an algorithm for modifying them so as to satisfy both constraints in (1.1). The notation \(M_0 = (\hat{\mu}, \hat{\Sigma})\) is used from here on to denote any such pre-estimate of \((\mu, \Sigma)\) and \(M_i, i = 1, 2, 3\) for its gradual modifications.

5.1. 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 \(\hat{\mu} \in \mathbb{R}^p\) and \(\hat{\Sigma}\) any \(p \times p\) positive-definite covariance matrix with the spectral decomposition \(PDP^\top\) as in (1.2). Set \(P_p^* = \frac{\hat{\mu}}{\|\hat{\mu}\|}\) and apply the Gram - Schmidt orthonormalization process to the set of vectors \(\{P_p^*, P_{p-1}, \ldots, P_2, P_1\}\) to obtain \(\{P_p^*, \ldots, P_2^*, P_1^*\}\). Then, the modified covariance matrix

\[
\hat{\Sigma}^* = \sum_{j=1}^{p-1} \frac{\lambda_j}{\lambda_{pr}} P_j^* P_j^{*\top} + P_p^* P_p^{*\top}
\]

where \(\lambda_{pr} = \left(\prod_{k=1}^{p-1} \lambda_k\right)^{\frac{1}{p-1}}\) (5.1)

satisfies the conditions in (1.1).

We denote this estimator by \(M_1\). In Lemma 4, \(\hat{\mu}\) is effectively forced to become an eigenvector corresponding to the eigenvalue 1 of a modified covariance matrix estimator, i.e. \(\hat{\Sigma}^* \hat{\mu} = \hat{\mu}\). 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 constraints 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. \( \tilde{\mu} = cP_i \) for some eigenvector \( P_i \).

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

\[
c_{0i} = \arg\min_{c \in \mathbb{R}} \| \tilde{\mu} - cP_i \|^2 = \frac{\langle P_i, \tilde{\mu} \rangle}{\|P_i\|^2} \quad \text{and} \quad i_0 = \arg\min_i \left( 1 - \frac{\lambda_i}{\lambda_{0i}} \right)^2. \tag{5.2}
\]

Then, the modified mean-covariance estimators

\[
\hat{\mu}^* = c_{0i_0}P_{i_0}, \quad \hat{\Sigma}^* = \sum_{j \neq i_0} \frac{\lambda_j}{\lambda_{pr}} P_iP_i^\top + \hat{\mu}^*\hat{\mu}^{*\top} \quad \text{where} \quad \lambda_{pr} = \left( \prod_{j \neq i_0} \frac{\lambda_j}{\lambda_{0j}} \right)^{\frac{1}{p-1}} \tag{5.3}
\]

satisfy (1.1).

We refer to the estimator \((\hat{\mu}^*, \hat{\Sigma}^*)\) as \( M_2 \) in the sequel. The intuition behind the method for selecting \( i_0 \) is that from \( \hat{\mu}^*\hat{\mu}^* = c_{0i_0}^2P_iP_i^\top \) it is desirable to have the eigenvalue corresponding to \( \hat{\mu}^* \) to be as close as possible to one of the \( \lambda_i \)'s. Thus, it is reasonable that \( \lambda_i/c_{0i_0}^2 \) should be as close to 1 as possible. More details about such selection can be found in Appendix (6).

Modifying the initial estimator jointly using (5.3) we obtain \((\hat{\mu}^*, \hat{\Sigma}^*)\). 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.

5.2. *The Modification Algorithm: Multiple Regression* In this section we consider a full-fledged multiple linear modeling of \( \hat{\mu} \) on \( P_i \)'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 \( M_3 \).
1. Start with a given \((\hat{\mu}, \hat{\Sigma})\) and its spectral decomposition as in (1.2).

2. **Variable (Basis) Selection:** Write \(\hat{\mu} = \sum_{j=1}^{p} c_j P_j = PC\) where \(c = (c_1, c_2, \ldots, c_p)\).
   - Simple Clustering: Viewing the \(c_i\)’s as weights, select those \(P_i\)’s which have 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 \(S = \{i_1, i_2, \ldots, i_{j_0}\}\) 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 \(\hat{\mu}\) on the span of columns of \(P_{j_0} = [P_{i_1}, P_{i_2}, \ldots, P_{i_{j_0}}]:\)
   \[
   \hat{\beta} = \arg \min_{\beta} \|\hat{\mu} - P_{j_0}\beta\|_2, \quad \hat{\mu}^* = P_{j_0}\hat{\beta}
   \]  
   (5.4)

4. **Orthogonalization to accommodate \(\hat{\mu}^*\):** Let \(g = \arg \max \{|\hat{\beta}_{k^*}| : k^* = i_1, i_2, \ldots, i_{j_0}\}\). Apply the Gram-Schmidt process on \(\{P_{i_1}, \ldots, P_{i_{j_0}-1}, \hat{\mu}^*, P_{i_{j_0}+1}, \ldots, P_{i_{j_0}}\}\) to obtain \(\{\hat{\mu}^*, b_1, b_2, \ldots, b_{j_0-1}\}\) with \(\hat{\mu}^*\) as the starting vector.

5. Set,
   \[
   \hat{\Sigma}^* = \hat{\mu}^*\hat{\mu}^* + \sum_{j \neq \{i_1, \ldots, i_{j_0}\}}^{p} \lambda_j P_j P_j^T + \sum_{k=1}^{j_0-1} \lambda_{i_k} b_k b_k^T
   \]

   estimate \(\lambda_{i_k}'\) by
   \[
   \hat{\lambda}_{i_k} = b_k^T \Sigma b_k, \quad k = 1, 2, \ldots, j_0 - 1.
   \]  
   (5.5)

   (The proof of this step is presented in A.8).

6. Let \(\lambda_{pr} = \left(\prod_{j=1}^{p} \lambda_j \cdot \prod_{k=1}^{j_0} \lambda_{i_k}\right)^{1/p-1}\). The modified estimator \((\hat{\mu}^*, \hat{\Sigma}^*)\) is given by
   \[
   \hat{\mu}^* = P_{j_0}\hat{\beta}
   \]
   \[
   \hat{\Sigma}^* = \sum_{j \neq \{i_1, \ldots, i_{j_0}+1\}}^{p} \lambda_{pr} P_j P_j^T + \sum_{k=1}^{j_0} \frac{\hat{\lambda}_{i_k} b_k b_k^T + \hat{\mu}^* \hat{\mu}^*}{\lambda_{pr}}.
   \]  
   (5.6)

---

Algorithm 1: Modifying an Estimator to Satisfy Eq. (1.1)
6 Simulation Experiments

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 (or S&C method), 3. The (Aitchison and Silvey, 1958) iterative (or A&S) method.

6.1. The Simulation Set up: We have taken sample size and dimension to be \((n,p) = (50,5), (50,10), (300,5), (300,10)\). 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 \(L_2\) loss for the mean vector and the (Förstner and Moonen, 2003) loss (denoted by MF in Table 2) for the covariance matrix to assess the performance of our estimators along with the Kullback-Leibler divergence.

For the parameters of the Gaussian distributions used for data generation we take the entries of the mean vector \(\mu\) to be values of independent standard Gaussian variables. For the covariance matrix, we start with \(\Psi = LL^\top\) where \(L\) 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 \(L\) ensure positive-definiteness of \(\Sigma\). Since such \((\mu, \Psi)\) do not necessarily satisfy conditions (1.1), the covariance matrix is modified first by applying Algorithm (5.1) to \((\mu, \Psi)\) to obtain \((\mu, \Sigma)\). Samples \((X_i)'s\) generated from \(N_p(\mu, \Sigma)\) follow the ESAG distribution of (Paine et al., 2018) when normalized (i.e. \(X_i/\|X_i\|\)) since \((\mu, \Sigma)\) satisfies Eq. (1.1).

6.2. Simulation Results from the Three Iterative Methods:

6.2.1. The Standard MLE: The iterative method for computing the maximum likelihood estimator described in Section 3.1 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 \(\mu\) and \(\alpha_1\) converge, but the rate of decrease of risk for estimating \(\Sigma\) 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 in Table 2), involves iterative updating of the Lagrange multipliers. In contrast, the next two iterative methods involve exact calculation of the Lagrange multiplier.
6.2.2. The S&C Method: The S&C method is described in Section 3.2. 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 1).

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 $\epsilon$ to be 0.1. From the Table 2 we can see that the S&C method is losing very little while achieving the satisfaction of the joint constraint (1.1). This method performs better than the their competitor in estimating the covariance matrix (see Table 2).

6.2.3. The A&S Method: The iterative method for calculation of the constrained MLE described in Section 4 operates inside a closed ball of radius $\delta = \| \theta^{(0)} \|$ around the true parameter. Hence choosing a good initial value for the iterations to run is essential and here we chose $\theta^{(0)} = (\bar{x}^\top, \text{vec}(S)^\top)^\top$.

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

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 performance is even worse than the standard MLE.

6.3. An Example: Estimates of the Historic Position of Earth’s 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 list of locations are close to Tasmania and can

| n  | p  | No of cases with positive definite covariance estimate |
|----|----|-----------------------------------------------------|
| 50 | 5  | 92                                                  |
| 50 | 10 | 97                                                  |
| 300| 5  | 87                                                  |
| 300| 10 | 99                                                  |
Table 2: Risks for the three iterative methods of finding constrained MLE, modified by Algorithm 1 (M3) with K-Means

| Method      | n  | p   | μ (MF) | Σ (MF) | MLE Approx with M3 | MLE Approx | KL Div. | KL Div with M3 |
|-------------|----|-----|--------|--------|--------------------|------------|---------|----------------|
| SMLE        | 50 | 5   | 0.5411 | 1.597  | 0.5545             | 0.8174     | 1.1371  | 1.3813         |
| S&C         | 50 | 5   | 0.4187 | 1.667  | 0.5711             | 0.6335     | 1.2228  | 1.3443         |
| A&S         | 50 | 10  | 0.4692 | 1.654  | 0.6788             | 0.715      | 1.5949  | 1.798          |
| SMLE        | 300| 5   | 0.5392 | 1.417  | 0.6725             | 0.6708     | 2.3977  | 3.1173         |
| S&C         | 300| 5   | 0.4907 | 1.417  | 0.6992             | 0.5314     | 2.5974  | 2.9252         |
| A&S         | 300| 10  | 0.6663 | 1.6173 | 0.6861             | 0.6719     | 2.3325  | 2.8416         |

SMLE: standard MLE; S&C is the method of (Strydom and Crowther, 2012), and A&S denotes the method of (Aitchison and Silvey, 1958).
be viewed as points in $\mathbb{R}^3$ instead of a sphere when the size of the earth is taken into account barring some outliers. This feature will be translated to an unit sphere when the longitude and latitudes are transformed to $X_1, X_2, \ldots, X_{33}$ on unit sphere (Preston and Paine, 2017). Paine et al. (2018, Section 2.5) provided strong evidence in favor of ESAG distribution over isotropic angular gaussian distribution while analysing this dataset using a large sample likelihood ratio test. This inspires us to make normality assumption under the constraint similar to ESAG distribution on the transformed dataset. The constrained maximum likelihood estimate calculated using the numerical method by Strydom and Crowther (2012) with 1000 iterations is:

$$\frac{\mu}{\|\mu\|} = [-0.91, 0.099, 0.392], \quad \Sigma = \begin{bmatrix}
1.114 & -0.012 & 0.269 \\
-0.012 & 0.586 & 0.076 \\
0.269 & 0.076 & 1.607
\end{bmatrix}$$

which is comparable to the maximum likelihood estimate calculated using eliptically symmetric angular Gaussian distribution with a specific parametrization in three dimension (Paine et al., 2018):

$$\frac{\mu}{\|\mu\|} = [-0.553, 0.262, 0.791], \quad \Sigma = \begin{bmatrix}
1.076 & 0.633 & -0.156 \\
0.633 & 1.666 & 0.222 \\
-0.156 & 0.222 & 0.817
\end{bmatrix}$$

One main advantage is that our calculation is not restricted to three dimension.

7 Conclusions

We address construction of a joint estimator for the mean-covariance of a normal distribution under the constraints (1.1). Three iterative methods are presented where the end results do not necessarily satisfy the constrains or the basic requirements of being a covariance matrix. Our novel algorithm modifies any joint estimator of the mean-covariance to satisfy the constraints. Comparison of the three methods for finding constrained maximum likelihood estimator shows an advantage for explicit computation of the Lagrange multiplier, possibly because the corresponding iterative methods are variants of the Newton- Raphson algorithm.

Funding. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.
MLE of Jointly Constrained...

References

AITCHISON, J. and SILVEY, S. (1958). Maximum-likelihood estimation of parameters subject to restraints. Ann. Math. Stat. pp. 813–828.

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

BIBBY, J., KENT, J. and MARDIA, K. (1979). Multivariate analysis. Academic Press, London.

CHACÓN, J. E. and DUONG, T. (2015). Efficient recursive algorithms for functionals based on higher order derivatives of the multivariate gaussian density. Stat. Comput. 25, 5, 959–974.

CHAUDHURI, S., DRTON, M. and RICHARDSON, T.S. (2007). Estimation of a covariance matrix with zeros. Biometrika 94, 1, 199–216.

EFRON, B. and ET AL (1975). Defining the curvature of a statistical problem (with applications to second order efficiency). Ann. Stat. 3, 6, 1189–1242.

FÖRSTNER, W and MOONEN, B. (2003). A metric for covariance matrices. In geodesy-the challenge of the 3rd millennium, pp. 299–309. Springer.

HARTIGAN, J. A. and WONG, M.A. (1979). Algorithm as 136: a k-means clustering algorithm. J. Royal Stat. Soc. Series C (Appl. Stat.) 28, 1, 100–108.

HOEL, P. G. and PORT, S. C. (1986). And Stone. Introduction to stochastic processes. Waveland Press, C. J.

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

KHAN, R. A. (1968). A note on estimating the mean of a normal distribution with known coefficient of variation. Journal of the American Statistical Association 63, 323, 1039–1041.

KNILL, O. (2011). Math 19b: Linear Algebra with Probability. https://people.math.harvard.edu/knill/teaching/math19b_2011/handouts/math19b_2011.pdf. Accessed 7 July 2022.

Fig. 1: This pictorial representation shows how we update when the iteration goes outside the ball in A&S method

\[ \theta^{(i)} = a \]

\[ \delta \]

\[ \theta^{(i+1)} = b \]
Appendix A: Proofs of Results:

A.1 Proof of Theorem 1:

(a) By differentiating the Lagrangian in 1.(a)
\[
L(X; \mu, \Sigma) = l(X; \mu, \Sigma) + \alpha_1 (| \Sigma^{-1} | - 1)
\]
\[
= c + \frac{n}{2} \log | \Sigma^{-1} | - \frac{1}{2} Tr [A(\mu) \Sigma^{-1}] + \alpha_1 (| \Sigma^{-1} | - 1)
\]
(A.1)

with respect to \( \mu \) and setting to zero leads to \( \hat{\mu}_{mle} = \bar{x} \). Differentiation with respect to \( \alpha_1 \) gives us the condition \( | \Sigma^{-1} | = 1 \). Using this and setting derivative with respect to \( \Sigma^{-1} \) to 0, leads to

\[
\frac{\partial L(X; \mu, \Sigma)}{\partial \Sigma^{-1}} = \frac{1}{2} \left[ n \Sigma^{-1} - A(\mu) + 2\alpha_1 | \Sigma^{-1} | \Sigma \right]
\]
\[
\hat{\Sigma} = \frac{A(\mu)}{|A(\mu)|^{1/p}}
\]  
(A.2)

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

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

\[
L(X; \mu, \Sigma) = c + \frac{n}{2} \log |\Sigma^{-1}| - \frac{1}{2} Tr \left[ A(\mu) \Sigma^{-1} \right] + \alpha_1(|\Sigma^{-1}| - 1) - \alpha_2^\top \left( \Sigma^{-1} \mu - b \right)
\]  
(A.3)

Rewriting the condition as \(\Sigma^{-1} \mu = b\) and \(|\Sigma^{-1}| = 1\) in the Lagrangian is necessary for taking the derivative with respect to \(\Sigma^{-1}\) in accordance with the standard practice in the unconstrained case (Bibby et al., 1979, §4.2.2). Differentiating with respect to \(\mu\) and \(\Sigma^{-1}\) we have:

\[
\frac{\partial L(X; \mu, \Sigma)}{\partial \mu} = n \Sigma^{-1}(\bar{x} - \mu) - \Sigma^{-1} \alpha_2
\]

\[
\frac{\partial L(X; \mu, \Sigma)}{\partial \Sigma^{-1}} = \frac{1}{2} \left[ (n + 2\alpha_1) \Sigma - A(\mu) - 2\alpha_2 \mu^\top \right]  
\]  
(A.4)

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

\[
\mu = \bar{x} - \frac{1}{n} \alpha_2, \quad \Sigma = \frac{A(\mu) + 2\alpha_2 \mu^\top}{n + 2\alpha_1}
\]

\[
|\Sigma^{-1}| = 1, \quad \Sigma^{-1} \mu = b
\]

(c) The Lagrangian

\[
L(X; \mu, \Sigma) = c + \frac{n}{2} \log |\Sigma^{-1}| - \frac{1}{2} Tr \left[ A(\mu) \Sigma^{-1} \right] + \alpha_1(|\Sigma^{-1}| - 1) - \alpha_2^\top (I_p - \Sigma^{-1}) \mu
\]  
(A.5)

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

\[
\frac{\partial L(X; \mu, \Sigma)}{\partial \mu} = n \Sigma^{-1}(\bar{x} - \mu) - (I_p - \Sigma^{-1}) \alpha_2
\]

\[
\frac{\partial L(X; \mu, \Sigma)}{\partial \Sigma^{-1}} = \frac{1}{2} \left[ (n + 2\alpha_1) \Sigma - A(\mu) - 2\alpha_2 \mu^\top \right]  
\]  
(A.6)
A. Kundu and M. Pourahmadi

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

\[ \mu = \bar{x} - \frac{1}{n} (I_p - \Sigma) \alpha_2, \]
\[ \Sigma = \frac{A(\mu) + 2\alpha_2 \mu^\top}{n + 2\alpha_1} \]
\[ \left| \Sigma^{-1} \right| = 1, \]
\[ \Sigma^{-1} \mu = b. \]

A.2: Proof of Lemma 1

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

\[ (ab^\top + ba^\top) \left( \frac{a}{\|a\|} + \frac{b}{\|b\|} \right) = (a^\top b + \|a\||\|b\|) \left( \frac{a}{\|a\|} + \frac{b}{\|b\|} \right) \]
\[ (ab^\top + ba^\top) \left( \frac{a}{\|a\|} - \frac{b}{\|b\|} \right) = (a^\top b - \|a\||\|b\|) \left( \frac{a}{\|a\|} - \frac{b}{\|b\|} \right) \]

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

\[ \lambda_j(M) = \lambda_j(B + A) \geq \lambda_j(B) + \lambda_p(A) \]
\[ \geq \lambda_p(B) + \lambda_p(A) \]
\[ = (a^\top b - \|a\||\|b\|) + \lambda_p(A), \quad (\text{By applying the first part.}) \]

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

\[ \lambda_j(M) \geq \lambda_j(B) + \lambda_p(A) = \lambda_p(A) > 0 \]

and \( \lambda_1(M) > 0 \) trivially.

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

3 Verification of Conditions

\( \mathcal{F}1 - \mathcal{F}4 \) and \( \mathcal{H}1 - \mathcal{H}3 \) (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

\[
\frac{\partial l}{\partial \mu} = n\Sigma^{-1}(\bar{x} - \mu), \quad \frac{\partial l}{\partial \Sigma} = -\frac{1}{2} (n\Sigma^{-1} - \Sigma^{-1} A(\mu)\Sigma^{-1}) \tag{A.7}
\]

Second Derivative The Hessian matrix of the likelihood is:

\[
H_l = \begin{pmatrix}
\frac{\partial^2 l}{\partial \mu^2} & \frac{\partial^2 l}{\partial \mu \partial \Sigma} \\
\frac{\partial^2 l}{\partial \Sigma \partial \mu} & \frac{\partial^2 l}{\partial \Sigma^2}
\end{pmatrix}_{p \times p, p \times p^2, p^2 \times p^2, p(p+1) \times p(p+1)}
\]

Next, we calculate the four submatrices.

(a) The first submatrix is

\[
\frac{\partial^2 l}{\partial \mu^2} = -n\Sigma^{-1} \tag{A.8}
\]

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

\[
\frac{\partial^2 l}{\partial \Sigma \partial \mu} = -\left[\Sigma^{-1} \otimes \Sigma^{-1}\right] [I \otimes (\bar{x} - \mu)] \tag{A.9}
\]

(c) The third submatrix is

\[
\frac{\partial^2 l}{\partial \mu \partial \Sigma} = \frac{1}{2} \frac{\partial \text{vec} \left[ n\Sigma^{-1} - \Sigma^{-1} A(\mu)\Sigma^{-1} \right]}{\partial \mu} = \frac{1}{2} \text{vec} \left[ \Sigma^{-1} A(\mu)\Sigma^{-1} \right] = \frac{1}{2} \text{vec} \left[ \Sigma^{-1} \left[ -2n\bar{x}\mu^\top + n\mu\mu^\top \right] \right] = \frac{1}{2} \left[ -2n \frac{\partial\text{vec} \left( \Sigma^{-1} \bar{x}(\Sigma^{-1}\mu)^\top \right)}{\partial \mu} + n \frac{\partial\text{vec} \left( (\Sigma^{-1}\mu)(\Sigma^{-1}\mu)^\top \right)}{\partial \mu} \right] = \frac{1}{2} \left[ -2n \frac{\partial \Sigma^{-1}}{\partial \mu} \otimes (\Sigma^{-1}\bar{x})^\top + n \frac{\partial \Sigma^{-1}}{\partial \mu} \otimes (\Sigma^{-1}\mu)^\top \right] = \frac{1}{2} \left[ -2n \Sigma^{-1} \otimes (\Sigma^{-1}\bar{x})^\top + n \Sigma^{-1} \otimes (\Sigma^{-1}\mu)^\top + n(\Sigma^{-1}\mu)^\top \otimes \Sigma^{-1} \right] \tag{A.10}
\]

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

\[
\frac{\partial^2 l}{\partial \Sigma^2} = \frac{1}{2} \left[ n\Sigma^{-1} \otimes \Sigma^{-1} - (\Sigma^{-1} A(\mu)\Sigma^{-1}) \otimes \Sigma^{-1} - \Sigma^{-1} \otimes (\Sigma^{-1} A(\mu)\Sigma^{-1}) \right] \tag{A.11}
\]
We have shown the existence of the second derivative and from the above quantities it is evident the third derivative exists too. Since multivariate normal has all the moments (Chacón and Duong, 2015), so we have essentially verified conditions $F_1 - F_4$. Next, we verify $H_1 - H_3$ for the constraint. The simplest way is to express it as $h(\mu, \Sigma) = \Sigma \mu - \mu = 0$. 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**

$$ \frac{\partial h}{\partial \mu} \bigg|_{p \times p} = \Sigma - I, \quad \frac{\partial h}{\partial \Sigma} \bigg|_{p^2 \times p} = \mu \otimes I \quad (A.12) $$

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

**Second Derivative:**

$$ \frac{\partial^2 h}{\partial \mu^2} \bigg|_{p^2 \times p} = 0, \quad \frac{\partial^2 h}{\partial \Sigma \partial \mu} \bigg|_{p^2 \times p^2} = I_p^2 $$

$$ \frac{\partial^2 h}{\partial \mu \partial \Sigma} \bigg|_{p^3 \times p} = I \otimes \text{vec}(I_p), \quad \frac{\partial^2 h}{\partial \Sigma^2} \bigg|_{p^3 \times p^2} = 0 \quad (A.13) $$

This gives us

$$ H^2_\theta \bigg|_{(p^2+p^3) \times (p+p^2)} = \left( \begin{array}{c} \left( \frac{\partial^2 h}{\partial \theta_i \partial \theta_j} \right) \end{array} \right) $$

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

**A.4: Proof of Lemma 2:**

We assume that $\mu$ is fixed and are interested in calculating the directional derivative of the Lagrangian in (4.1) as a function of $\Sigma$ only in the direction of a symmetric matrix $D$. Let us set $L(\mu, \Sigma \mid X) = f(\Sigma)$. By the definition of directional derivative,

$$ \nabla_D f = \lim_{h \to 0} \frac{f(\Sigma + hD) - f(\Sigma)}{h} $$

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 $f^1 = \log |\Sigma|$. 

\[
\nabla_D f^1 = \lim_{h \to 0} \frac{1}{h} \log \left[ \frac{|\Sigma + hD|}{|\Sigma|} \right] = \lim_{h \to 0} \frac{1}{h} \log \left[ \frac{||I + hD\Sigma^{-1}||}{|\Sigma|} \right] \\
= \lim_{h \to 0} \frac{1}{h} \log \left[ |I + hD\Sigma^{-1}| \right] \\
= \lim_{h \to 0} \frac{1}{h} \sum_{i=1}^{p} \log \left[ 1 + \lambda_i(D\Sigma^{-1}) \right] \\
= \lim_{h \to 0} \frac{1}{h} \left[ \sum_{i=1}^{p} \lambda_i(D\Sigma^{-1}) + \sum_{i=1}^{p} \left\{ \log \left[ 1 + \lambda_i(D\Sigma^{-1}) \right] - \lambda_i(D\Sigma^{-1}) \right\} \right] \\
= \text{Tr} \left[ D\Sigma^{-1} \right]
\]

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

\[
\nabla_D f^2 = \lim_{h \to 0} \frac{1}{h} \text{Tr} \left[ S \left( (\Sigma + hD)^{-1} - \Sigma^{-1} \right) \right] \\
= \lim_{h \to 0} \frac{1}{h} \text{Tr} \left[ S \left( \Sigma^{-1} - \Sigma^{-1}(h^{-1}D^{-1} - \Sigma^{-1})^{-1}\Sigma^{-1} - \Sigma^{-1} \right) \right] \\
= \lim_{h \to 0} \frac{1}{h} \text{Tr} \left[ S \left( -\Sigma^{-1}(h^{-1}D^{-1} + \Sigma^{-1})^{-1}\Sigma^{-1} \right) \right] \\
= \lim_{h \to 0} \frac{1}{h} \text{Tr} \left[ S \left( -h\Sigma^{-1}(D^{-1} + h\Sigma^{-1})^{-1}\Sigma^{-1} \right) \right] \\
= \lim_{h \to 0} \text{Tr} \left[ -S\Sigma^{-1}(D^{-1} + h\Sigma^{-1})^{-1}\Sigma^{-1} \right] \\
= \text{Tr} \left[ -S\Sigma^{-1}D\Sigma^{-1} \right]
\]

(c) The third term without the constant is $f^3 = \text{Tr} \left[ (\bar{x} - \mu)(\bar{x} - \mu)^\top \Sigma^{-1} \right]$. 

The same calculation shows that $\nabla_D f^3 = \text{Tr} \left[ -(\bar{x} - \mu)(\bar{x} - \mu)^\top \Sigma^{-1}D\Sigma^{-1} \right]$

(d) The fourth term $f^4 = \alpha^\top_2 (\Sigma\mu - \mu)$ 

\[
\nabla_D f^4 = \lim_{h \to 0} \frac{\left\{ \alpha^\top_2 (\Sigma + hD) \mu - \alpha^\top_2 \Sigma \mu + \alpha^\top_2 \mu \right\}}{h} \\
= \lim_{h \to 0} \frac{\left\{ \alpha^\top_2 (\Sigma + hD) \mu - \alpha^\top_2 \Sigma \mu \right\}}{h} \\
= \lim_{h \to 0} \frac{\alpha^\top_2 (\Sigma + hD - \Sigma) \mu}{h} = \alpha^\top_2 D\mu
\]

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

\[
\nabla_D f = \frac{n}{2} \text{Tr} \left[ D\Sigma^{-1} \right] + \frac{n}{2} \text{Tr} \left[ S\Sigma^{-1}D\Sigma^{-1} \right] + \frac{n}{2} \text{Tr} \left[ B\Sigma^{-1}D\Sigma^{-1} \right] + \alpha^\top_2 D\mu
\]
\[ -\frac{n}{2} \text{Tr} \left[ D\Sigma^{-1} \right] + \frac{n}{2} \text{Tr} \left[ (S + B)\Sigma^{-1} D\Sigma^{-1} \right] + \alpha_2^\top D\mu \]

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

Now, we calculate the second directional derivative in the direction \( C \) which is also a symmetric matrix. The first derivative denoted by \( \nabla_D f \) has two terms as a function of \( \Sigma \). The corresponding notation for second directional derivative is

\[ \nabla_C \nabla_D f = \lim_{h \to 0} \frac{\nabla_D f(\Sigma + hC) - \nabla_D f(\Sigma)}{h}. \]

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

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

\[ \nabla_C \nabla_D f^1 = -\text{Tr} \left[ D\Sigma^{-1} C\Sigma^{-1} \right] = -\text{Tr} \left[ \Sigma\Sigma^{-1} D\Sigma^{-1} C\Sigma^{-1} \right] \]

(b) The second term ignoring the constant is \( \nabla_D f^2 = \text{Tr} \left[ (S + B)\Sigma^{-1} D\Sigma^{-1} \right] \).

By Woodbury-Sherman matrix formula:

\[
\begin{align*}
(\Sigma + hC)^{-1} D(\Sigma + hC)^{-1} & - \Sigma^{-1} D\Sigma^{-1} \\
&= \left[ \Sigma^{-1} - \Sigma^{-1}(h^{-1}C^{-1} + \Sigma^{-1})^{-1}\Sigma^{-1} \right] D \left[ \Sigma^{-1} - \Sigma^{-1}(h^{-1}C^{-1} + \Sigma^{-1})^{-1}\Sigma^{-1} \right] \\
&\quad - \Sigma^{-1} D\Sigma^{-1} \\
&= \left[ \Sigma^{-1} - h\Sigma^{-1}(C^{-1} + h\Sigma^{-1})^{-1}\Sigma^{-1} \right] D \left[ \Sigma^{-1} - h\Sigma^{-1}(C^{-1} + h\Sigma^{-1})^{-1}\Sigma^{-1} \right] \\
&\quad - h\Sigma^{-1}(C^{-1} + h\Sigma^{-1})^{-1}\Sigma^{-1} D\Sigma^{-1} \\
&\quad - h\Sigma^{-1}(C^{-1} + h\Sigma^{-1})^{-1}\Sigma^{-1} D\Sigma^{-1} \\
&\quad - h\Sigma^{-1}(C^{-1} + h\Sigma^{-1})^{-1}\Sigma^{-1} D\Sigma^{-1} \\
&\quad + O(h^2)
\end{align*}
\]

Using this result, we get:

\[
\begin{align*}
\nabla_C \nabla_D f^2 &= \lim_{h \to 0} \frac{\text{Tr} \left[ (S + B)(\Sigma + hC)^{-1} D(\Sigma + hC)^{-1} \right] - \text{Tr} \left[ (S + B)\Sigma^{-1} D\Sigma^{-1} \right]}{h} \\
&= \lim_{h \to 0} \frac{\text{Tr} \left[ (S + B) \left( (\Sigma + hC)^{-1} D(\Sigma + hC)^{-1} - \Sigma^{-1} D\Sigma^{-1} \right) \right]}{h} \\
&= -\lim_{h \to 0} \text{Tr} \left[ (S + B) \Sigma^{-1}(C^{-1} + h\Sigma^{-1})^{-1}\Sigma^{-1} D\Sigma^{-1} \right] \\
&\quad - \lim_{h \to 0} \text{Tr} \left[ (S + B) \Sigma^{-1} D\Sigma^{-1}(C^{-1} + h\Sigma^{-1})^{-1}\Sigma^{-1} \right] \\
&= -\text{Tr} \left[ (S + B) \Sigma^{-1} C\Sigma^{-1} D\Sigma^{-1} \right] - \text{Tr} \left[ (S + B) \Sigma^{-1} D\Sigma^{-1} C\Sigma^{-1} \right] \\
&= -\text{Tr} \left[ (S + B) \Sigma^{-1} C\Sigma^{-1} D\Sigma^{-1} \right] - \text{Tr} \left[ (S + B) \Sigma^{-1} (D\Sigma^{-1} C)^\top \Sigma^{-1} \right] \\
&= -\text{Tr} \left[ 2(S + B)\Sigma^{-1} C\Sigma^{-1} D\Sigma^{-1} \right]
\end{align*}
\]

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

\[
\nabla_C \nabla_D f = \frac{n}{2} \text{Tr} \left[ \Sigma\Sigma^{-1} D\Sigma^{-1} C\Sigma^{-1} \right] - \frac{n}{2} \text{Tr} \left[ 2(S + B)\Sigma^{-1} C\Sigma^{-1} D\Sigma^{-1} \right]
\]
MLE of Jointly Constrained…

\[ = -\frac{n}{2} Tr \left[ \{2(S + B) - \Sigma\} \Sigma^{-1} C \Sigma^{-1} D \Sigma^{-1} \right] \tag{A.15} \]

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

\[ \nabla_D \nabla_D f = -\frac{n}{2} Tr \left[ \Sigma^{-1/2} \{2S - \Sigma\} \Sigma^{-1/2} D \Sigma^{-1/2} \Sigma^{-1/2} D \Sigma^{-1/2} \right] \]
\[ = -\frac{n}{2} Tr \left[ \Sigma^{-1/2} D \Sigma^{-1/2} \Sigma^{-1/2} \{2S - \Sigma\} \Sigma^{-1/2} D \Sigma^{-1/2} \right] \leq 0 \tag{A.16} \]

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

\[ \nabla_D \nabla_D f = -\frac{n}{2} Tr \left[ \Sigma^{-1/2} uu^T (2S - \Sigma) uu^T \Sigma^{-1/2} \right] = -\frac{n}{2} u^T \Sigma uu^T (2S - \Sigma) u \geq 0, \tag{A.17} \]

which completes the proof.

### A.5: Proof of Lemma 3:

Suppose there are two matrices \( \Sigma_1 \) and \( \Sigma_2 \) in \( \Delta_{2S} \), which maximize the Lagrangian function for a given \( \alpha_2 \). Then, for the matrix \( \Sigma(t) = (1 - t) \Sigma_1 + t \Sigma_2, t \in [0, 1] \), we have

\[ L(\bar{x}, \Sigma(t)) \geq (1 - t)L(\bar{x}, \Sigma_1) + tL(\bar{x}, \Sigma_2) = L(\bar{x}, \Sigma_1) \]

so that \( \Sigma(t) \)'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 \( \Delta_{2S} \), it is unique.

1.1. * A.6: Covariance error bound in Lemma 5: The error of a new estimate can be bounded in the following way:

\[ \| \Sigma - \hat{\Sigma}^* \|_\mathcal{F} \leq \| \Sigma - \hat{\Sigma}_{map} \|_\mathcal{F} + \| \hat{\Sigma}_{map} - \hat{\Sigma}^* \|_\mathcal{F} \]

\[ \| \hat{\Sigma}_{map} - \hat{\Sigma}^* \|_\mathcal{F} = \left\| \left(1 - \frac{1}{\lambda_P} \right) \sum_{i=1}^{d} \lambda_i P_i P_i^T + \left( \frac{\lambda_{in}}{\sigma_{in}^2} - 1 \right) \hat{\mu}^* \hat{\mu}^*^T \right\|_\mathcal{F} \]

Applying triangle inequality:

\[ \leq \left\| \left(1 - \frac{1}{\lambda_P} \right) \sum_{i=1}^{d} \lambda_i P_i P_i^T \right\|_\mathcal{F} + \left\| \left( \frac{\lambda_{in}}{\sigma_{in}^2} - 1 \right) \hat{\mu}^* \hat{\mu}^*^T \right\|_\mathcal{F} \]
\[ = \left| \left( 1 - \frac{1}{\lambda P} \right) \sum_{i=1}^{d} \lambda_i P_i P_i^\top \right| + \left| \left( \frac{\lambda_i}{\lambda_{i0}} - 1 \right) \mu_i \mu_\star^\top \right| \]

(A.18)

**A.7: Proof of Equation 5.4:**

This follows from standard regression OLS estimate.

**Appendix A.8: Proof of Equation 5.5:**

\[ f (\lambda') = \text{Tr} \left[ (\Sigma - \hat{\Sigma}^*)^2 \right] \]

\[ \frac{\partial f (\lambda')}{\partial \lambda_{ik}} = -2 \text{Tr} \left[ (\Sigma - \hat{\Sigma}^*) (b_k b_k^\top) \right] = 0 \]

\[ \Rightarrow \text{Tr} \left[ b_k^\top (\Sigma - \hat{\Sigma}^*) b_k \right] = 0 \]

\[ \Rightarrow b_k^\top \Sigma b_k = b_k^\top \hat{\Sigma}^* b_k \]

\[ \Rightarrow \hat{\lambda}_{ik} = b_k^\top \Sigma b_k \]  

(A.19)

**Appendix B: Explicit Calculation of the Lagrange Multiplier**

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

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

where \( \theta = (\theta_1, \ldots, \theta_q) \) is the vector of natural parameters and \( T(X) = (T_1(X), T_2(X), \ldots, T_q(X)) \) is their complete and sufficient statistics with the following (Lehmann and Casella, 2006)

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

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

\[ w(X; \theta, \alpha_2) = \log f(X; \theta) + \alpha_2 h(m) \propto \theta^\top T(X) - a(\theta) + \alpha_2 h(m) \]
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 $m$: $m^{(l)} = T$ and calculate $\nabla h(m^{(l)})$ and $V^{(l)}$ as a function of $m^{(l)}$ and $T$.

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

1. calculate $h(T^{(k)})$, $\nabla h(T^{(k)})$.
2. use Eq. (3.11) to update:

$$T^{(k+1)} = T^{(k)} - V^{(l)} \nabla h(m^{(l)}) \frac{h(T^{(k)})}{|\nabla h(m^{(l)})|}$$

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

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

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

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

$$\frac{\partial w(X; \theta, \alpha_2)}{\partial \theta} = T(X) - \frac{\partial a(\theta)}{\partial \theta} + \alpha_2 \frac{\partial h(m)}{\partial \theta} = 0$$

$$m = T(X) + \alpha_2 \nabla m(\theta) \nabla h(m)$$ \hspace{1cm} (B.1)

where $\nabla m(\theta)$ is a $q \times q$ gradient matrix and $\nabla h(m)$ denotes a $q \times 1$ gradient vector. The algorithms proposed in Section 3.1 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 $h(m)$ around $T(X)$ and the approximation of unknown $\gamma$ is performed as follows:

$$0 = h(m) = h(T) + \alpha_2 \nabla h(m) \nabla m(\theta) \nabla h(T) + o(||m - T||)$$

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

Substituting this value in (B.1) we obtain the final approximations for $m$ to be the Eq. (3.11).
Example We focus on the constraint $\Sigma \mu = \mu$ 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), can also be expressed in terms of natural parameters in the following way:

$$l(\mu, \Sigma \mid X) \propto n \mu^\top \Sigma^{-1} \bar{x} - \frac{n}{2} \mathrm{Tr} \left[ \Sigma^{-1} \left( \frac{1}{n} \sum_{i=1}^{p} x_i x_i^\top \right) \right] - \frac{n}{2} \mu^\top \Sigma^{-1} \mu - \frac{n}{2} \log[\det(2\pi \Sigma)]$$

$$= \theta^\top T - a(\theta)$$

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

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

with

$$\mathbb{E}(T) = \mathbf{m} = \begin{pmatrix} \mu \\ \text{vec}(\Sigma + \mu \mu^\top) \end{pmatrix} = \begin{pmatrix} m_1 \\ m_2 \end{pmatrix}, \quad \text{cov}(T) = V = \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix}$$

where

$$V_{11} = \frac{1}{n} \Sigma, \quad V_{12} = \frac{1}{n} (\Sigma \otimes \mu + \mu \otimes \Sigma)$$

$$V_{21} = V_{21}^\top, \quad V_{22} = \frac{1}{n} (I_p^2 + K) \left[ \Sigma \otimes \Sigma + \Sigma \otimes \mu \mu^\top + \mu \mu^\top \otimes \Sigma \right]$$
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 (3.9)** The condition $Tr[(\Sigma - I_p)R_\mu] = \sum_{i=1}^{p}(\Sigma - I_p)\mu = 0$, 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 $\text{vec}(R_\mu) = \mathbb{1} \otimes \mu$.

The iteration in (3.11) requires $\nabla h$ and $\nabla m$. Note that $\nabla m = V$ and $\nabla h$ is calculated as follows:

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

where

$$
\frac{\partial h}{\partial m_1} = \frac{\partial (m_2 - \text{vec}(I_p))^\top (\mathbb{1} \otimes m_1)}{\partial m_1} - \frac{\partial (m_1 \otimes m_1)}{\partial m_1} (\mathbb{1} \otimes m_1)
$$

$$
= (\mathbb{1} \otimes I_p)^\top (m_2 - \text{vec}(I_p)) - (\mathbb{1} \otimes I_p)^\top (m_1 \otimes m_1)
$$

$$
- (m_1 \otimes I_p + I_p \otimes m_1)^\top (\mathbb{1} \otimes m_1)
$$

$$
\frac{\partial h}{\partial m_2} = \mathbb{1} \otimes m_1.
$$

**Appendix C: A Non-Example: Multinomial Distribution**

Here we point out that the mean-covariance of the multinomial distributions do not satisfy the constraints. 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.

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

$$
\mu = nq, \quad \Sigma(Y) = \text{diag}(q) - qq^\top
$$

where $q = (q_1,q_2,\ldots,q_p)^\top$. The covariance matrix is positive semi-definite with one eigenvalue 0 corresponding to the eigenvector $\mathbb{1} = (1,\ldots,1)^\top$. We
note that $\Sigma \mu \neq \mu$. More generally, the class of Dirichlet distributions is another example of this kind which do not satisfy the constraints.

Anupam Kundu
Mohsen Pourahmadi
Department of Statistics,
Texas A&M University,
College Station, USA
E-mail: anupam.kundu@tamu.edu
pourahm@stat.tamu.edu

Paper received: 1 November 2021; accepted 27 July 2022.