NUMERICAL METHODS FOR ESTIMATING CORRELATION COEFFICIENT OF TRIVARIATE GAUSSIANS

WERONIKA J. ŚWIĘCHOWICZ
YUANFANG XIANG
ILLINOIS INSTITUTE OF TECHNOLOGY
DEPARTMENT OF APPLIED MATHEMATICS

RESEARCH ADVISOR
PROFESSOR SONJA PETROVIĆ

.

Abstract. Given observed data, the fundamental task of statistical inference is to understand the underlying data-generating mechanism. This task usually entails several steps, including determining a good family of probability distributions that could have given rise to the observed data, and identifying the specific distribution from that family that best fits the data. The second step is usually called parameter estimation, where the parameters are what determines the specific distribution. In many instances, however, estimating parameters of a statistical model poses a significant challenge for statistical inference. Currently, there are many standard optimization methods used for estimating parameters, including numerical approximations such as the Newton-Raphson method. However, they may fail to find a correct set of maximum values of the function and draw incorrect conclusions, since their performance depends on both the geometry of the function and location of the starting point for the approximation. An alternative approach, used in the field of algebraic statistics, involves numerical approximations of the roots of the critical equations by the method of numerical algebraic geometry. This method is used to find all critical points of a function, before choosing the maximum value(s). In this paper, we focus on estimating correlation coefficients for multivariate normal random vectors when the mean is known. The bivariate case was solved in 2000 by Small, Wang and Yang, who emphasize the problem of multiple critical points of the likelihood function. The goal of this paper is to consider the first generalization of their work to the trivariate case, and offer a computational study using both numerical approaches to find the global maximum value of the likelihood function.

1. Introduction

The multivariate Gaussian distribution, also known as the multivariate normal distribution, is a probability distribution whose shape depends on two quantities: the mean vector and the correlation matrix. The entries of the correlation matrix, called the correlation coefficients, measure a dependence between the random variables. The mean of each random variable and the correlation coefficients are called parameters, and a specific choice of parameter values identifies one particular Gaussian distribution. Allowing the parameter values to vary produces an entire family of probability distributions, jointly called the multivariate Gaussian model. Thus, a statistical model is a set of probability distributions of similar form, and each distribution from the family is specified by a particular value of model parameters.

In statistics, one observes some data and then asks for the 'explanation' of those data. To even attempt to answer this question, one assumes that data form a random sample, that is, that they were produced by a data-generating mechanism from some (fixed) statistical model. In this context, finding a 'best explanation' of the data means finding the parameter values so that the resulting distribution makes the data most likely to have been observed, among all the distributions from that model. Formally, we write this as follows: the observed data

$x_1, \ldots, x_n$ are assumed to be independent and identically distributed from some probability density function $f(x|\theta)$; here, θ is the fixed but unknown parameter vector, while the notation $x|\theta$ indicates that f is a function of the sample x given the fixed value θ. To determine the 'best' value of the parameters, statisticians write the likelihood function $L(\theta|x)$, which essentially equals this probability function of the given data sample, but it is considered as a function of the parameters. In other words, $L(\theta|x) = f(x|\theta)$. For simplicity, $L(\theta|x)$ is often just written as $L(\theta)$. The goal of statistical inference is then to solve the optimization problem of finding the vector(s) θ that maximize the likelihood function $L(\theta)$ of the observed sample. Any solution of this optimization problem $argmaxL(\theta)$ is called a maximum likelihood estimator (or an MLE) for the given model. A closed-form expression for the MLE can be easily obtained for simple examples, as in calculus, by solving the likelihood equations $\frac{\partial}{\partial\theta} \log L(\theta) = 0$ symbolically.

In this paper, the statistical model in question is the family of multivariate Gaussian distributions for which the mean vector is assumed to be known. The case when mean is unknown is a standard, well-known case; but fixing the mean vector results in a constrained optimization problem that has only been solved in the two-dimensional case [SWY]. Thus we study the problem $argmaxL(\theta)$, or, equivalently, $argmax \log L(\theta)$, for the likelihood function arising from the restricted model of multivariate Gaussians where the entries of the parameters vector θ are correlation coefficients. Unfortunately, this likelihood function is challenging to work with and explicit (symbolic) solutions of the likelihood equations are difficult to obtain. Therefore, we rely on other approaches and use two numerical methods, Newton-Raphson and the polynomial homotopy continuation methods, as optimization strategies. Specifically, we show that the work of [SWY] in two dimensions does *not* directly generalize to the three-dimensional case.

The organization of this article is as follows. In section 2, we introduce several fundamental concepts including the multivariate normal distribution and correlation coefficients. In section 3, we present two numerical approaches used in determination of critical points of the log-likelihood function $\log L(\theta)$. In addition, we describe algorithms specified for Newton-Raphson and `PHCpack`, the execution of which gives approximations for critical points of the given function. In section 4, we apply the two methods to analyze simulated data, while in section 5, we bring up another possibility: using sample correlation. Previous work in estimating correlation coefficients by using sample correlation coefficients in the two-dimensional case was done by Shevlykov and Smirnov [SS]. We test the validity of the use of the sample correlation coefficients in both two- and three-dimensional cases, and compare these sample estimators with the critical points computed by `PHCpack` [V99]. In addition, we discuss the behavior of the likelihood-function on the boundary for both two- and three-dimensional cases. We finish the analysis of the log-likelihood $l(\theta)$ with a conclusion section 6, comparing the solutions of the two methods.

2. Problem Set-up

In statistics, the multivariate Gaussian distribution (or the multivariate normal distribution) generalizes the one-dimensional (univariate) normal distribution to several random variables, or, as is customary to say, a random vector. The multivariate normal distribution of a k-dimensional random vector is usually denoted by the following notation: $X \sim \mathcal{N}_k(\mu, \Sigma)$, where μ represents the k-dimensional mean vector and Σ stands for the $k \times k$ covariance matrix. We focus on the three-dimensional case, that is $k = 3$, and use the correlation matrix instead, noting that correlation is just a scaled version of covariance. Thus we let $X \sim \mathcal{N}_3(0, R)$, where R is the correlation matrix with the variances restricted to 1. By definition, the matrix R is a symmetric

positive definite matrix with 1's on the diagonal:

$$R := \begin{bmatrix} 1 & r_{12} & r_{13} \\ r_{12} & 1 & r_{23} \\ r_{13} & r_{23} & 1 \end{bmatrix}. \tag{1}$$

Here the data are three-dimensional column vectors $X = [x_1, x_2, x_3]$. The entries r_{12}, r_{13} and r_{23} are correlation coefficients of x_1 and x_2, x_1 and x_3, x_2 and x_3, respectively. These are the three parameters that we are going to estimate.

For a random sample $X_1, ..., X_n$ from X, the likelihood function for the Gaussian model is

$$\begin{aligned} L(R) = \prod_{i=1}^{n} f(X_i) &= \prod_{i=1}^{n} (2\pi)^{-p/2} det(R)^{-1/2} exp(-\frac{1}{2} X_i^T R^{-1} X_i) \\ &= (2\pi)^{-np/2} det(R)^{-n/2} exp(-\frac{1}{2} \sum_{i=1}^{n} X_i^T R^{-1} X_i). \end{aligned} \tag{2}$$

Observe that

$$\sum_{i=1}^{n} X_i^T R^{-1} X_i = \sum_{i=1}^{n} tr(X_i X_i^T R^{-1}) = tr(\sum_{i=1}^{n} X_i X_i^T R^{-1}),$$

thus the expression (2) can be rewritten using the simple matrix trace formulas, as in [E]:

$$L(R) = (2\pi)^{-np/2} det(R)^{-n/2} exp(-\frac{1}{2} trAR^{-1}), \tag{3}$$

where

$$A := \sum_{i=1}^{n} X_i X_i^T. \tag{4}$$

As outlined in the Introduction, to estimate the values of the correlation coefficients r_{12}, r_{13} and r_{23}, one can - as in calculus - maximize the likelihood function $L(R)$ of the observed sample by finding all critical points of the three likelihood equations $\frac{\partial}{\partial r_{ij}} L(r_{12}, r_{13}, r_{23}) = 0$. In our example, however, the closed-form expression for the maximum likelihood estimator is difficult to compute explicitly. Next, we introduce two numerical techniques called the Newton-Raphson and the polynomial homotopy continuation methods. We will be using these two as the optimization strategies to numerically solve the optimization problem of computing the MLE of the vector (r_{12}, r_{13}, r_{23}).

3. Numerical Approximation of Critical Points

This section is an introduction to two numerical approximations of critical points of a multivariate function. The first method, Newton-Raphson, is used to find one critical point, while the second method, polynomial homotopy continuation, finds all critical points of the function.

3.1. **Newton-Raphson Method.**

Newton-Raphson iteration is a frequently used numerical technique that reckons successively better approximation to a root of a nonlinear real-valued function f. The method involves linearizing the function, that is, replacing f with the first two terms in its Taylor series.

To illustrate the application of the Newton-Rapson for finding one of the critical points of a function $f(x, y, z)$, consider the following example ([KC]). Let $f_1(x, y, z), f_2(x, y, z)$, and

$f_3(x, y, z)$ be a system of partial derivatives of the function f with respect to x, y, and z respectively, such that

$$\begin{cases} 0 = f_1(x, y, z) \\ 0 = f_2(x, y, z) \\ 0 = f_3(x, y, z) \end{cases}. \tag{5}$$

Now suppose that the point (x, y, z) is an estimated solution of (5) with computed correction h, k, and l such that $(x + h, y + k, z + l)$ is a successively better approximation of the solution. Then by using only the linear terms (the first two terms) of the Taylor series expansion of f_1, f_2, and f_3 in three variables, we get that the system

$$\begin{cases} 0 &= f_1(x + h, y + k, z + l) \\ 0 &= f_2(x + h, y + k, z + l) \\ 0 &= f_3(x + h, y + k, z + l) \end{cases} \tag{6}$$

is approximated by

$$\begin{cases} 0 \approx f_1(x, y, z) + h\frac{\partial f_1}{\partial x}\Big|_{x,y,z} + k\frac{\partial f_1}{\partial y}\Big|_{x,y,z} + l\frac{\partial f_1}{\partial z}\Big|_{x,y,z} \\ 0 \approx f_2(x, y, z) + h\frac{\partial f_2}{\partial x}\Big|_{x,y,z} + k\frac{\partial f_2}{\partial y}\Big|_{x,y,z} + l\frac{\partial f_2}{\partial z}\Big|_{x,y,z} \\ 0 \approx f_3(x, y, z) + h\frac{\partial f_3}{\partial x}\Big|_{x,y,z} + k\frac{\partial f_3}{\partial y}\Big|_{x,y,z} + l\frac{\partial f_3}{\partial z}\Big|_{x,y,z} \end{cases}. \tag{7}$$

We try to solve equation (7) for h, k, and l. It is a system of three linear equations. The coefficient matrix is the Jacobian matrix of f_1, f_2, and f_3:

$$J = \begin{bmatrix} \frac{\partial f_1}{\partial x} & \frac{\partial f_1}{\partial y} & \frac{\partial f_1}{\partial z} \\ \frac{\partial f_2}{\partial x} & \frac{\partial f_2}{\partial y} & \frac{\partial f_2}{\partial z} \\ \frac{\partial f_3}{\partial x} & \frac{\partial f_3}{\partial y} & \frac{\partial f_3}{\partial z} \end{bmatrix}.$$

If the matrix J is not singular the solution to (7) exists and is as follows

$$\begin{bmatrix} h \\ k \\ l \end{bmatrix} = -J^{-1} \begin{bmatrix} f_1(x, y, z) \\ f_2(x, y, z) \\ f_3(x, y, z) \end{bmatrix}.$$

Therefore, the Newton-Raphson method uses the following approximation to solutions of the three nonlinear equations (5):

$$\begin{bmatrix} x^{(j+1)} \\ y^{(j+1)} \\ z^{(j+1)} \end{bmatrix} = \begin{bmatrix} x^{(j)} \\ y^{(j)} \\ z^{(j)} \end{bmatrix} + \begin{bmatrix} h^{(j)} \\ k^{(j)} \\ l^{(j)} \end{bmatrix},$$

where $[h^{(j)}, k^{(j)}, l^{(j)}]$ is the solution to the jth Jacobian linear system

$$J \begin{bmatrix} h^{(j)} \\ k^{(j)} \\ l^{(j)} \end{bmatrix} = - \begin{bmatrix} f_1(x^{(j)}, y^{(j)}, z^{(j)}) \\ f_2(x^{(j)}, y^{(j)}, z^{(j)}) \\ f_3(x^{(j)}, y^{(j)}, z^{(j)}) \end{bmatrix}.$$

The following algorithm shows how Taylor's approximation is used to search for a (local) maximum of a multivariate function.

Algorithm: K-dimensional Newton-Raphson
Input: A function $f(\mathbf{x})$ and a starting point $\mathbf{x_0}$
Output: A value $\hat{\mathbf{x}}$ of $\mathbf{x}$ that maximizes $f(\mathbf{x})$

$i \leftarrow 0$
while $||\nabla f'(\mathbf{x_i})|| >$ **tolerance do**

$\mathbf{i} \leftarrow \mathbf{i} + \mathbf{1}$
$\mathbf{x_i} \leftarrow \mathbf{x_{i-1}} - (\mathbf{D^2 f(x_{i-1})})^{-1} \nabla \mathbf{f}(\mathbf{x}_{i-1})$
end while
$\hat{\mathbf{x}} \leftarrow \mathbf{x_i}$.

Remark 1. We use a simple function $f(x) = x^3 - \frac{5}{2}x^2 - 6x + 3$ in one variable to demonstrate how the Newton-Raphson method obtains one of its critical points. Choosing the starting point close to $x = 5$, the algorithm will return the desired value as the final approximation. However, if the initial point is far from the global maximum, say $x = -2$, the outcome of the algorithm will return a different result than that for $x = 5$. From this simple example, it should be evident that the success of Newton-Raphson method strongly depends on the initial point chosen in the analysis.

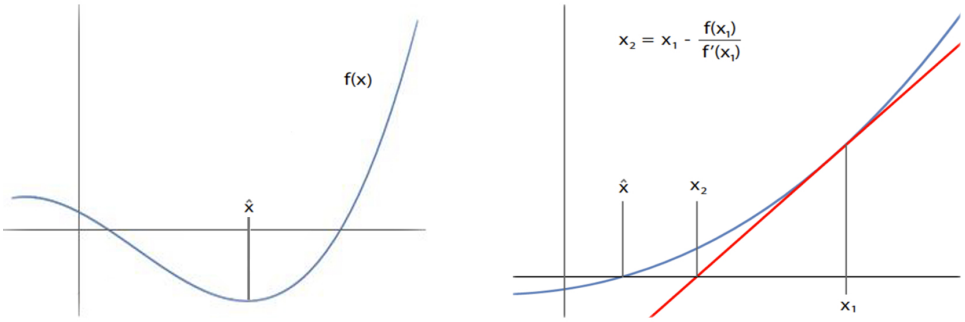

FIGURE 1. Example of a Function f(x) (left) and the Newton-Raphson Method Estimating Critical Points of the Derivative of the given Function (right).

3.2. **Polynomial Homotopy Continuation.** Polynomial homotopy continuation is a numerical method that approximates solutions to a system of equations $G(x)$. The method first establishes a start system $F(x)$ that is easy to solve and has at least as many solutions as the target system $G(x)$. Such a start system is generated based on the structure and the upper bound on the number of solutions of the target system $G(x)$. The method then traces a finite number of solution paths from the start polynomial system to the target system, whose solutions we want to estimate. A subsequent step in the process is to define a homotopy space $H(x, t)$ with embedded paths between the two systems. For this purpose, we use a step function defined on the interval $[0, 1]$ that traces paths from function $F(x)$ to $G(x)$. This path-tracking is illustrated in Figure 2.

Definition 1. Let X, Y be topological spaces, and f, g: X $\to$ Y continuos maps. A homotopy from f to g is a continuos function H: X $\times$ [0,1] $\to$ Y satisfying

$$H(x, 0) = f(x) \text{ and } H(x, 1) = g(x), \text{ for all } x \in X. \tag{8}$$

If such a homotopy exists, we say that f is homotopic to g, and denote this by f $\simeq$ g. A standard reference is [W].

Remark 2. Definition 1 states that a homotopy is a deformation space containing paths from known solutions of an easy function to the unknown solution set of the target function. In a homotopy example illustrated in Figure 2, paths traced from a solution set of function $F(x)$ lead to an approximation for the solutions of $G(x)$; see [L] for a more detailed background on this numerical procedure.

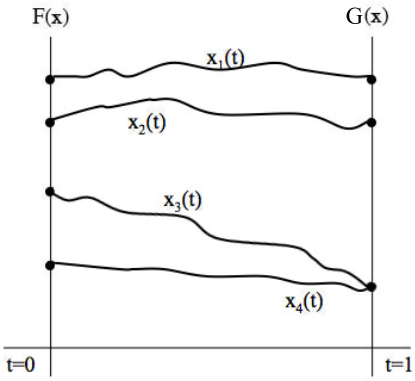

FIGURE 2. Determination of zeros (roots) of a target system G(x). In our application, G(x) is the system of partial derivatives of the log-likelihood function.

Suppose the polynomial system $H(x,t)$ denoting the homotopy is defined as:

$$H(x,t) = cG(x)t + (1-t)F(x) \qquad t \in [0,1], \quad c \in \mathbb{C} \tag{9}$$

where $x = \{x_1, x_2, ..., x_n\}$ is a complex n-tuple. Based on Eq.(9), it should be clear that $H(x,0) = F(x)$ and $H(x,1) = cG(x)$.

The final step in the process involves estimating solutions for the target system by employing a predictor-corrector method [V99] that is built into the software `PHCpack`. This approach uses a step size control that eliminates all divergent and path-crossing solutions when tracing the paths between two systems. It then follows independent solutions paths by small increments t and approximates solutions to the critical points of the target function (for more details, see [V99], [LM]).

4. Comparison of the Two Numerical Methods

The log-likelihood function for the trivariate normal model defined in Section 2 is written in the expanded form as follows:

$$\begin{aligned} l(r_{12}, r_{23}, r_{13}) = &-\frac{3}{2}n\log(2\pi) - \frac{1}{2}n\log(1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2r_{12}r_{13}r_{23}) \\ &- \frac{1}{2(1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2r_{12}r_{13}r_{23})}\Big(a_{11}(1 - r_{23}^2) + a_{22}(1 - r_{13}^2) + a_{33}(1 - r_{12}^2) \\ &+ 2a_{23}(r_{12}r_{13} - r_{23}) + 2a_{13}(r_{12}r_{23} - r_{13}) + 2a_{12}(r_{23}r_{13} - r_{12})\Big), \end{aligned} \tag{10}$$

where r_{12}, r_{13} and r_{23} are correlation coefficients that we are trying to estimate and a_{11}, a_{12}, a_{13}, a_{22}, a_{23}, and a_{33} are entries of the matrix A computed from the data sample.

The approach for testing the two numerical methods from Section 3 is outlined as follows. First, we simulate a random set of data from $X_1, \ldots, X_n$; here, n is the sample size. That sample is then used to generate a positive definite symmetric matrix A satisfying restrictions defined in Equation (4), i.e. $A := \sum_{i=1}^{n} X_i X_i^T$. (`R` code for generating these matrices can be found in Appendix 7.1.) Having generated A, we can now define the log-likelihood function in Equation (10) as a function of three parameters, r_{12}, r_{13}, and r_{23}. Note that the domain of the function is the open cube $(-1,1)^3$, ensuring that there are no zero denominators. Finally, we estimate r_{12}, r_{13}, and r_{23} by maximizing the log-likelihood function, and compare the estimated values to the true values of the parameters.

In the remainder of this section, we explain this procedure in detail and illustrate what happens when the two numerical methods are used to estimate the parameters r_{12}, r_{13}, and r_{23}.

4.1. **Steps needed to estimate a local maximum using the Newton-Raphson method.** To compute one of the critical points for the given log-likelihood function using the Newton-Raphson method, we use a `maxLik` package included in the statistical software `R`. The input provided to `R` consists of the entries of a sample-based positive definite symmetric matrix and the equation of the log-likelihood function l. We then call the `maxLik` solver specifying the function for which the roots need to be found, the initial point of the iteration, and the method that we want to use, in our case the Newton-Raphson. The command to do this in `R` is:

```
mle<-maxLik(l,start=c(-0.993,-0.993,0.993),method="NR")
```

and the output is stored in the variable `mle`. In this example, we used the point $(-0.993, -0.993, 0.993)$ as the initial point of the iteration.

We ran the algorithm for various starting points and several data sets. The algorithm above found only a single root for each one of the sample-based positive definite matrices we simulated. The results are both consistent with the analysis of the function's concavity illustrated in Figure 3, as well as the conjectured geometry of the log-likelihood generated function based on the 2×2 correlation matrix R from [SWY]. However, at this point, we have not verified whether the unique critical point discovered by `maxLik` corresponds to the global maximum of the function. In order do that, we need to find all critical points.

4.2. **Steps needed to approximate all critical points using the polynomial homotopy continuation method.** To find all critical points for the log-likelihood function, we use the `PHCpack` interface with the `Macaulay2` package [GPV] as a black-box polynomial system solver. The input provided to `PHCpack` is the sample size n, the entries of a sample-based positive definite symmetric matrix, and the equations of the partial derivatives of the log-likelihood function l. In order to manipulate polynomials, we first define a polynomial ring of the critical equations by using the standard `Macaulay2` command $R = CC[r_{12}, r_{23}, r_{13}]$. This allows us to define the equations of the first partial derivatives for the log-likelihood l as polynomials in the three variables r_{ij}. We name these three critical equations (i.e., partial derivatives) $f1$, $f2$, and $f3$. In the following, we list and solve a system of three derivative equations $f1 = 0$, $f2 = 0$ and $f3 = 0$ to obtain approximations for the critical points of the given log-likelihood. This is done using `solveSystem` command which calls the black-box `PHCpack` solver. The solutions are stored in list called `sols`.

```
list={f1,f2,f3}
sols=solveSystem (list)
```

By default, the software computes complex solutions of the input system. Thus we need to use the `realPoints` command to extract only those critical points whose entries r_{12}, r_{23} and r_{13} are all real. This is done with a simple command:

```
realpts=realPoints(sols)
```

The algorithm above generated roughly 22 real solutions for each one of sample-based positive define symmetric matrices we simulated; there were 22 most of the time, and sometimes a few less. However, in each case, there was only one critical point that was statistically relevant, meaning that it lies in the open cube $(-1, 1)^3$. The results are consistent with the analysis of the function's concavity illustrated in Figure 3.

4.3. **Performance analysis and comparison of the results.** The ultimate goal is to test how well the two methods actually estimate the true r_{12}, r_{13}, r_{23}. For this, we use the standard statistics trick: we *fix* a correlation matrix R, which we will call the ground truth in our simulations. Then, we sample data from the distribution $X \sim \mathcal{N}_3(0, R)$. Finally, we use the

data X to estimate the entries of the matrix R, as if we did not know what they were. To fix the ground truth R, we use the function `rcorrmatrix` in package `clusterGeneration` to generate the random correlation matrix for each sample size n. The entries of this matrix R are the model parameters, thus a good approximation of the MLE should be close to these parameter values.

The code needed to generate the sample-based matrices A is shown in Section 7.2. We used three different ground-truth matrices R, and for each, we generated data samples of sizes $n = 100$, $n = 1000$, and $n = 10000$. The following three tables summarize the output of both methods on these 9 data sets. The tables show the following information: the ground truth r_{12}, r_{23}, r_{13}, the corresponding A matrices generated using code form Section 7.2, the statistically relevant critical points out of all critical points computed by `PHCpack` with code from Section 4.2, and the estimated values using `maxLik` with code from Section 4.1.

TABLE 1. Computational results for $(r_{12}, r_{23}, r_{13}) = (-0.04, 0.17, 0.54)$

n	The Matrix A	Critical Point (`PHCpack`)	Estimated Max (`maxLik`)
100	a_{11}=83.99, a_{12}=39.91, a_{13}=2.66, a_{22}=70.46, a_{23}=-28.02, a_{33}=84.40	(0.37, -0.08, 0.75)	(0.96, 0.90, -0.01)
1000	a_{11}=1064.31, a_{12}=664.34, a_{13}=-56.65, a_{22}=1046.01, a_{23}=-374.37, a_{33}=1014.02	(0.32, -0.15, 0.69)	(1.00, 0.59, 0.74)
10000	a_{11}=10211.43, a_{12}=6450.77, a_{13}=-639.90, a_{22}=10207.43, a_{23}=-3581.70, a_{33}=10037.0	(0.31, -0.16, 0.68)	(0.60, -0.97, 1.00)

TABLE 2. Computational results for $(r_{12}, r_{23}, r_{13}) = (-0.28, 0.39, -0.58)$

n	The Matrix A	Critical Point (`PHCpack`)	Estimated Max (`maxLik`)
100	a_{11}=121.06, a_{12}=77.21, a_{13}=-18.45, a_{22}=104.54, a_{23}=-37.06, a_{33}=87.23	(0.32, -0.19, 0.62)	(1.00, 0.51, 0.60)
1000	a_{11}=1107.01, a_{12}=729.04, a_{13}=-23.55, a_{22}=1058.10, a_{23}=-277.14, a_{33}=924.09	(0.25, -0.21, 0.68)	(1.0, 0.49, 0.70)
10000	a_{11}=10097.48, a_{12}=6332.98, a_{13}=-657.10, a_{22}=10323.37, a_{23}=-3703.83, a_{33}=10272.4	(0.31, -0.15, 0.68)	(1.0, 0.59, 0.73)

TABLE 3. Computational results for $(r_{12}, r_{23}, r_{13}) = (-0.47, -0.48, -0.14)$

n	The Matrix A	Critical Point `PHCpack`	Estimated Max (`maxLik`)
100	a_{11}=106.45, a_{12}=52.56, a_{13}=19.55, a_{22}=74.70, a_{23}=-22.25, a_{33}=108.49	(0.27, -0.21, 0.71)	(1.00, 0.83, 1.00)
1000	a_{11}=990.01, a_{12}=645.61, a_{13}=-62.66, a_{22}=645.61, a_{23}=1022.90, a_{33}=1011.59	(0.31, -0.16, 0.70)	(1.00, 0.54, 0.74)
10000	a_{11}=10133.35, a_{12}=6273.80, a_{13}=-623.43, a_{22}=10219.34, a_{23}=-3655.68, a_{33}=10036.5	(0.32, -0.14, 0.69)	(0.33, -0.16, 0.97)

The above results reveal a significant discrepancy in the final approximation of critical points generated by `maxLik`. Simply, the maxima estimated using `maxLik` do not converge to a certain point, unlike the critical points approximated using `PHCpack`. That indicates that the Newton-Raphson method is not reliable for computing the maxima or minima in the cases when there is a possibility of multiple critical points. Due to this inconsistency, we study the geometry of

the problem more closely. In particular. we examine the behavior of the log-likelihood function by plotting it using `Mathematica`.

One way of plotting the function in three variables in `Mathematica` is to use `Manipulate` and `plot3D`. In this way, we create a 3-dimensional plot treating one variable, say, r_{13}, as a parameter. The `Mathematica` code generating the three-dimensional plot of the four-dimensional trivariate log-likelihood function $l(r_{12}, r_{13}, r_{23})$ is:

```
Manipulate[Plot3D[l[r12, r13, r23], {r12, -0.99999, 0.99999},
{r13, -0.99999, 0.99999}], {r23, -0.99999, 0.99999}]
```

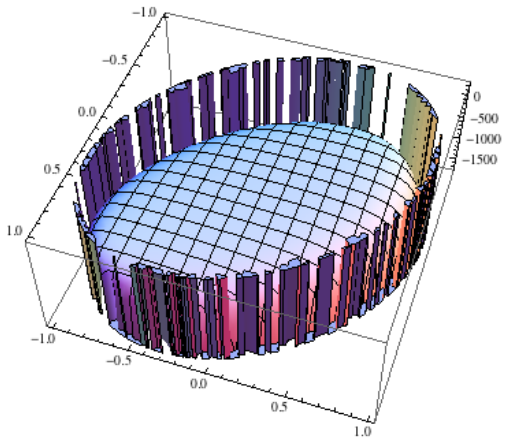

FIGURE 3. Graph of the Log-Likelihood Function $l(r_{12}, r_{13}, r_{23})$

The correlation coefficients r_{12}, r_{23} and r_{13} are restricted by the model to lie inside the open cube $(-1, 1)^3$. Therefore, for the purpose of the graphical representation in `Mathematica`, we replace the actual endpoints -1 and 1 by close approximations: -0.99999 and 0.99999. That let us observe the change in the behavior of the function that depended on the change in the value of r_{13}. Since the value of r_{13} was allowed to vary between -1 and 1, the function's value changed accordingly. Figure 3, which was generated for a fixed value of $r_{13} = -.23$, shows that the function is bounded above with a vertical asymptote around the entire boundary of the function. That indicates a concave down behavior, implying that the only critical point found by `PHCpack` inside $(-1, 1)$ is the global maximum. Perhaps one of the reasons why the Newton-Raphson method fails to produce reasonable outputs is that it cannot converge on the boundary.

Remark 3. A discussion on the computational approach is in order. Namely, one of the difficulties with this example stems from the fact that the closed-form formula for critical points is difficult to obtain. We discovered that there exists another parameterization of the log-likelihood, provided in the Appendix, Section 7.3, that simplifies the computation of the partial derivatives. However, as a result, the function $\ell(r_{12}, r_{23}, r_{13})$ becomes unbounded for all sample-based matrices A! This poses a significant challenge for the Newton-Raphson method in that it fails to find any critical points.

Thus, in our problem, we used the straightforward parameterization given in Equation (7). However, even with the optimal parameterization, the Newton-Raphson method fails to produce results close to those computed by `PHCpack`. One of the reasons why the method fails is because r_{12}, r_{23} and r_{13} all belong to $(-1, 1)$, which is an open interval. The function may increase towards infinity on boundary, and it is well known that the Newton-Raphson method cannot converge on the boundary in such cases.

5. Can Correlation Coefficients be Estimated Well?

In the previous section we described the shortcomings of the Newton-Raphson and the PHC methods in estimating critical points of the log-likelihood function $l(r_{12}, r_{13}, r_{23})$. While the Newton-Raphson was simply failing to converge, PHC was able to find the only critical point, however, this method also failed to approximate the critical point correctly. This could be due to the model symmetry; i.e., changing the sign of a critical point computed may result in the same value of the likelihood function (this is a common issue related to parameter non-identifiability, but it is beyond the scope of this paper).

With the failure of both methods, we search for another solution. A natural question arises: why not use the sample correlation coefficient method to estimate the correlation coefficients? After all, sample correlations can be used to estimate correlation coefficient in the well-studied unconstrained optimization problem case (i.e., when the means are not fixed to zero). In this section, several simulations show that the sample correlation coefficient cannot be used as an MLE because it does not correspond to a solution of the likelihood equations. Nevertheless, practitioners may wish to still use it, because it has other good properties such as small mean square error.

To illustrate our point, it is sufficient to reconsider the bivariate case. In a bivariate normal distribution, there is one off-diagonal entry in the correlation matrix, namely $\rho = r_{12}$. A classical estimator of the correlation coefficient ρ is given by the sample correlation coefficient r: [SS]

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}[(x_i - \bar{x_1})^2(y_i - \bar{y})^2]^{1/2}}, \tag{11}$$

where x_i and y_i are the entires of the ith data vector $X_i = [x_i, y_i]$. It is also the maximum likelihood estimator of ρ for the bivariate normal distribution density. However, it is only true when the mean and variances are unknown because in the event that variances are known, the likelihood function is conditioned on the knowledge that $\mu = 0$ and the diagonal entires of the matrix Σ are equal to 1. Indeed [FR] say we lose crucial information by using the sample correlation coefficient. In fact, let us verify that the sample correlation coefficient is not a solution of the likelihood equation $\frac{\partial lnL}{\partial \rho} = 0$. When means of both x and y are zero and the diagonal entries of the correlation matrix are both 1, the likelihood equation $\frac{\partial lnL}{\partial \rho} = 0$ becomes ([SWY])

$$\rho(1 - \rho^2) + (1 + \rho^2)\frac{\sum xy}{n} - \rho[\frac{\sum(x^2 + y^2)}{n}] = 0. \tag{12}$$

By the Fundamental Theorem of Algebra, it follows that Equation (12) has three solutions in the set of complex numbers, where at least one of the roots is real. We, however, are only interested in the solutions of ρ on the open interval $(-1, 1)$. To check if the sample correlation coefficient r is one of the three roots of (12), and thus appropriate to use as an MLE, we simulate data $X_1, \ldots, X_n$ and compute. Solving Equation (11) and (12) shows that the sample correlation coefficient is 0.186 and the roots of (12) are $\rho = 1, \frac{3}{2}, 5$. So, since none of the roots of ρ are even close to 0.186 and all of them are outside of the interval of interest, we conclude that the sample correlation coefficient should not be used as an MLE.

In the three dimensional case, we do not have one equation such as (12), instead, we have a system of three critical equations. Having lost hope in using the sample correlation coefficient as an MLE, we test if it is a suitable estimator at all. We simulate several data samples and compute their sample correlation coefficients. Below is the summary of the comparison of sample.

correlation coefficient and the true correlation coefficient used in the simulations. We compared them using the standard method of the mean square error.

TABLE 4. Comparison of the Simulated and Approximated Sample Correlation Coefficients using the Mean Square Error Method.

n	True Correlation Coefficients	Sample Correlation Coefficients	Mean Square Error
10	(0.61, -0.35, -0.05)	(0.76, -0.44, -0.24)	0.08
20	(-0.85, 0.45, -0.51)	(-0.90, 0.45, -0.47)	0.02
50	(-0.31, -0.02, -0.44)	(-0.34 ,0.15, -0.47)	0.02
100	(0.32 ,0.68, 0.53)	(0.19, 0.60, 0.46)	0.01
200	(0.10, 0.50, -0.77)	(0.05, 0.48, -0.81)	0.0037

From the above table we see that the mean square error gets smaller as the sample size increases. It indicates that the sample correlation might be a good estimator, because it has a decreasing mean-square error (MSE). However, low MSE may not be the only good property one desires an estimator to have; for example, efficiency or asymptotic normality [CB], which are properties that a maximum likelihood estimator (MLE) is guaranteed to have. Even though sample correlation is not the MLE for this constrained optimization problem (when means and variances are known), we can nevertheless use it as a relatively good simple estimator.

6. Conclusion

The motivation for this paper is maximization of the log-likelihood function for a standardized trivariate normal distribution with means 0 and variances 1. In the bivariate case, [SWY] show that the sole critical equation (i.e., the partial derivative of the log-likelihood function) reduces to a degree-three polynomial equation in one variable, which has three roots, and at least one real root in the interval $(-1, 1)$. [SWY] also suggest that the multiple root problem (i.e., multiple critical point problem) can be 'eliminated' in higher dimensions as well.

In this paper, we tested this claim computationally, and found out several interesting things. First, even in three dimensions, it is already not at all obvious that there will be only one statistically relevant critical point of the likelihood function. For example, the choice of the parametrization of the likelihood function can change the geometry of the problem completely. In the Appendix Section 7.3, we provide an alternative parametrization of the log-likelihood function found in the literature. For this parametrization, there are several statistically significant critical points; the numerical homotopy method can detect them, but the Newton-Rhapson method fails to converge for every data sample we generated. Plotting this alternate parametrization, we discovered that the likelihood function was sometimes bounded and sometimes not; in particular, it was unbounded whenever the positive semidefinite matrices A used as input to the function were generated using real data samples. In contrast, [ZUR] recently (concurrently with our analysis) proved a general result which implies that the log-likelihood is to be bounded for our problem. Therefore, we realized that the parametrization did not preserve convexity of the function. Thus, we chose to work with the natural parametrization proposed by Small, Wang and Yang, because, even though it is more complicated, it captures the correct geometry of the problem.

Second, we tested two of the commonly used numerical methods to estimate correlation coefficients, and compared their performance against the ground truth we used to generated our data. Both of the methods failed to locate the true correlations; however, the numerical homotopy continuation method was at least able to detect that there exists exactly one critical point in the statistical domain of the function, namely, the open cube $(-1, 1)^3$, and it converged

to some value that we may declare the MLE of the parameters. The Newton-Raphson method failed to converge altogether, and provided, as usual, no guarantee that the value it output is anywhere close to the critical point we are looking for. Given this situation, we also considered simply using the sample correlation coefficients as estimators. We showed that this simple point estimator is not a maximum likelihood estimator, however, it does have another nice property which is that its mean square error decreases with sample size. We thus conclude it is a very reasonable estimator after all.

References

[FR] B. K. Fosdick, A. E. Raftery, Estimating the Correlation in Bivariate Normal Data With Known Variances and Small Sample Sizes, Amer. Statist., 66 (2012), pp. 34-41.

[CB] G. Casella, R. Berger, Statistical Inference, 2nd Ed., Brooks/Cole, 2001.

[E] B. S. Everitt, Introduction to Optimization Methods and Their Application in Statistics, Chapman & Hall, Ltd., London, UK, UK, 1987.

[GPV] E. Gross, S. Petrović, J. Verschelde, PHCpack in Macalay2, J. Softw. Algebra Geom., 5 (2013), pp. 20-25.

[KC] D. Kincaid, W. Cheney, Numerical Analysis: Mathematics of Scientific Computing, 3rd Edition, American Mathematical Society, Providence, RI, 2002.

[LM] E. Lee, C. Mavroidis, Solving the Geometric Design Problem of Spatial 3R Robot Manipulators Using Polynomial Homotopy Continuation, ASME. J. Mech. Des., 124 (2002), pp.652-661.

[L] T. Y. Li, Numerical solution of polynomial systems by homotopy continuation methods, Handb. Numer. Anal., 11 (2003), pp. 209-304.

[SS] G. Shevlyakov, P. Smirnov, Robust Estimation of the Correlation Coefficient: An Attempt of Survey, Austrian Journal of Statistics, 40 (2011), pp. 147-156.

[SWY] C. G. Small, J. Wang, Z. Yang, Eliminating multiple root problems in estimation, Statist. Sci., 15 (2000), pp. 313-341.

[SGPR] D. Stasi, E. Gross, E., S. Petrović, D. Richards, Multiple Roots Problems: ML Degree for Estimation of Correlation Matrices, technical report, Pennsylvania State University, 2014.

[V99] J. Verschelde, Algorithm 795: PHCpack: A general-purpose solver for polynomial systems by homotopy continuation, ACM Trans. Math. Softw., 25 (1999), pp. 251-276.

[W] G. W. Whitehead, Elements of homotopy theory, Springer Grad. Texts in Math. 61, 1978.

[ZUR] P. Zwiernik, C. Uhler, D. Richards, Maximum Likelihood Estimation for Linear Gaussian Covariance Models, preprint, available at arXiv:1408.5604v1 [math.ST].

7. Appendix

7.1. **Computing sample-based positive definite matrix A from the data.** Here we provide statistical software `R` code that can be used to 1) generate a random sample $x_1, \ldots, x_n$ from the trivariate normal distribution, and 2) compute the matrices A using that data sample.

```
> # initialize a 3x3 zero matrix:
   A <-matrix(c(rep(0,9)),nrow=3,ncol=3)
> x<-list()
> xtrans <-list()
> prod <-list()
> for(i in 1:100){
+  #Generate one sample matrix from multivariate normal distribution:
+  x[[i]]<-mvrnorm(n=1,rep(0,3),Sigma)
+  #Calculate the transpose of the above matrix:
```

```
+  xtrans[[i]] <-t(x[[i]])
+   #Calculate the product of these two matrices:
+  prod[[i]] <- x[[i]] %*% xtrans[[i]]
+  #Sum up these products according to equation (4)"
+  A <- A + prod[[i]]  }
```

7.2. **Obtaining a correlation matrix.** In order to test the methods, as explained in Section 4, we need to obtain a correlation matrix that we will use as 'ground truth'. We can use the following code in statistical software R to obtain one such matrix R:

```
> library("MASS")
> library("clusterGeneration")
> R<-rcorrmatrix(3)
```

Now, instead of generating the data samples as in Section 7.1 above, we would like to use the ground-truth matrix R. From standard statistical theory (see, e.g., [CB]), we know that matrices A follow a distribution called Wishart. Thus we can generate a sample-based positive definite matrix A from the distribution $N_3(0, R)$ as follows:

```
> n=100
> k=30
> A<-rWishart(k,n,R)
```

Here, R denotes a ground-truth correlation matrix, n the sample size, and k the number of matrices A that we want to generate.

7.3. **An alternative parametrization of the log-likelihood function for bivariate Gaussians.** Here we provide the parametrization of the log-likelihood function for our problem borrowed from [SGPR].

$$
\begin{aligned}
l(r_{12}, r_{23}, r_{123}) = & -3\log(2\pi)n - n(\log(1-r_{12}^2) + \log(1-r_{23}^2) + \log(1-r_{132}^2)) \\
& - \frac{a_{11}}{(1-r_{12}^2)\times(1-r_{132}^2)} \\
& - a_{22}(1 + \frac{(r_{12}-r_{23})^2}{(1-r_{12}^2)(1-r_{23}^2)(1-r_{132}^2)} - \frac{2r_{12}r_{23}r_{132}}{\sqrt{(1-r_{12}^2)(1-r_{23}^2)}(1-r_{132}^2)}) \\
& - \frac{a_{33}}{(1-r_{23}^2)(1-r_{132}^2)} \\
& - 2a_{12}(\frac{-r_{12}}{(1-r_{12}^2)\times(1-r_{132}^2)} + \frac{r_{23}r_{132}}{\sqrt{(1-r_{12}^2)(1-r_{23}^2)}(1-r_{132}^2)}) \\
& - 2a_{23}(\frac{-r_{23}}{(1-r_{23}^2)(1-r_{132}^2)} + \frac{r_{132}}{\sqrt{(1-r_{12}^2)(1-r_{23}^2)}(1-r_{132}^2)}) \\
& + \frac{2a_{13}r_{132}}{\sqrt{(1-r_{12}^2)(1-r_{23}^2)}(1-r_{132}^2)}
\end{aligned}
\tag{13}
$$

This parametrization is nice in the sense that one can derive the partial derivatives explicitly and use them for computing an MLE. Unfortunately, it produces a function that is unbounded above; this means in particular the parameter substations used in deriving the function destroyed the geometry of the model.