A calibration method for non-positive definite covariance matrix in multivariate data analysis

DOI:
10.1016/j.jmva.2017.03.001

Document Version
Accepted author manuscript

A calibration method for non-positive definite covariance matrix in multivariate data analysis

Chao Huang [1], Daniel Farewell [2] and Jianxin Pan [3*]

[1] *South East Wales Trials Unit, Cardiff University*
Heath Park, Cardiff, CF14 4YS, UK
[2] *Division of Population Medicine, School of Medicine, Cardiff University*
Heath Park, Cardiff, CF14 4YS, UK
[3] *School of Mathematics, University of Manchester,*
Manchester M13 9PL, UK

Abstract

Covariance matrices that fail to be positive definite arise often in covariance estimation. Approaches addressing this problem exist, but are not well supported theoretically. In this paper, we propose a unified statistical and numerical matrix calibration, finding the optimal positive definite surrogate in the sense of Frobenius norm. The proposed algorithm can be directly applied to any estimated covariance matrix. Numerical results show that the calibrated matrix is typically closer to the true covariance, while making only limited changes to the original covariance structure.

Keywords:
Covariance matrix calibration, Nearness problem, Non-positive definiteness, Spectral decomposition

1. Introduction

The estimation of covariance matrices plays an essential role in multivariable data analysis. Covariances are required by many statistical modelling approaches, including multivariate regression and the analysis of spatial data.

*Corresponding author. Tel: +44 161 2755864; Fax: +44 161 275 5819.
Email Addresses: HuangC12@cardiff.ac.uk (C. Huang); farewelld@cf.ac.uk (D. Farewell); Jianxin.Pan@manchester.ac.uk (J. Pan)

Often, well-estimated covariance matrices improve efficiency in estimating parameters in a mean function [22]. In some circumstances, the covariance matrix may itself be of direct scientific interest: for instance, in spatial variation analysis for geographical data, and in volatility analysis for financial data.

However, it is not uncommon that estimators of covariance matrices fail to be positive definite. A typical example is the sample covariance matrix, which is often singular when the sample size is close to, or less than, the dimension of the random samples [3]. If singularity is caused by collinearity, conventional ridge regression [18] or modern variable selection [6, 21] approaches may solve the problem by excluding redundant variables. Dimension reduction approaches such as Principle Component Analysis [19] can also help to exclude eigenvalues with ignorable contributions.

However, these resolutions only apply in cases where such redundance truly exists. More often, non-positive definiteness may be put down to the generic difficulty of maintaining positive definiteness in covariance estimation; resulting estimators may not even be positive semidefinite. Even for elaborately designed statistical approaches, the estimators of covariance matrices can be ill-conditioned [5, 14]. A number of approaches have been proposed to resolve this issue. However, these are either limited to special circumstances or lack theoretical support. For instance, one alternative is to use the Moore-Penrose inverse of a non-positive definite matrix to replace the regular inverse typically used in statistical inferences [20]. However, this does not directly resolve the non-positive definiteness, and is lack of statistical interpretation. Alternatively, a smoothing approach exists [23] in which non-positive eigenvalues of the covariance matrix estimator are replaced by certain positive values. However, justification for the selection of these positive values was scant.

Based on the fundamental work of Halmos [7], Higham [9] proposed a solution for finding the nearest (in the sense of Frobenius norm) positive semidefinite matrix to an arbitrary input matrix. However, this surrogate positive semidefinite matrix is still singular [9, 10], so difficulty persists in using the surrogate matrix in statistical practice. Rebonto and Jäckel [17] considered a correlation matrix calibration using the hyperspherical decomposition and eigenvalue correction, which again leads to positive semidefinite correlation matrices. Hendrikse et al.[8] proposed an eigenvalue correction method using bootstrap resampling in order to reduce the bias arising in sample eigenvalues. Their work focused on the correction of the sample

covariance, where the performance of the correction method relies on the assumed distribution of the covariance matrix eigenvalues in the population.

In this paper, we propose a unified approach to calibrate a non-positive definite covariance matrix to ensure positive definiteness. The calibrated covariance matrix is usually closer to the true covariance matrix than the original covariance matrix estimator. Our proposed approach is implemented through a straightforward screening algorithm. In Section 2, we briefly review the matrix nearness problem, before proposing our novel calibration method together with its integrated criterion and algorithm. In Section 3 we conduct two simulation studies, and in Section 4 we discuss two case studies, including a calibration of the non-positive definite covariance matrix obtained by nonparametric regression in Diggle and Verbyla [5]. Conclusions are presented in Section 5.

2. Calibration method

2.1. The matrix nearness problem

In numerical analysis, a nearness problem involves finding, for a given matrix and a particular matrix norm, the nearest matrix that has certain important properties. Examples include finding the nearest covariance matrix [9] or correlation matrix [2, 16] in the sense of the Frobenius norm (or 2-norm).

Given an arbitrary square matrix X of order n, we denote its Frobenius norm by $\|X\| = \text{trace}(X^\top X)^{1/2}$. The nearness problem involves finding the nearest symmetric positive semidefinite matrix $P_0(X)$:

$$P_0(X) = \arg\min_{A \geq 0} \|X - A\| \tag{1}$$

Throughout, we shall assume that $A \geq 0$ denotes both non-negative definiteness and symmetry $A = A^\top$. Higham [9] used a polar decomposition to show that the solution to (1) has the explicit form $P_0(X) = (B + H)/2$, where $B = S(X) = (X + X^\top)/2$ is the symmetric matrix version of X, and H is the symmetric polar factor of B, satisfying $B = UH$ with U a unitary matrix and $H \geq 0$. This solution has been compiled in a MATLAB file named `poldex.m`, which can be found in the Matrix Computation Toolbox [11]. Clearly, if X is symmetric then the solution becomes $P_0(X) = (X + H)/2$. If, further, we are given the spectral decomposition of a symmetric $X = X^\top$ (that is, $X = Q\Lambda Q^\top$ for $Q^\top Q = I$ and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$), we have

$P_0(X) = Q\text{diag}\{\max(\lambda_1, 0), \ldots, \max(\lambda_n, 0)\}Q^\top$. In other words, the nearest positive semidefinite matrix $P_0(X)$ can be obtained by replacing by zero any negative eigenvalues of a symmetric X [10], eliminating the corresponding columns of Q (and causing some information loss). A immediate alternative is to instead replace negative eigenvalues by positive values, so that a positive definite correction of X is formed without this loss of information about Q. However, the theory of this idea need to be justified, particularly on how to choose appropriate replacement positive values, for which we will address in this paper.

2.2. A new calibration approach

We now aim to find a positive definite matrix surrogate for a generic X. First, we formulate this question as a nearness problem. For $c \geq 0$, let $\mathcal{D}_c = \{A : A - cI \geq 0\}$ be the set of positive definite matrices with no eigenvalue smaller than c. Given X, finding the nearest matrix $P_c(X) \in \mathcal{D}_c$ to X in terms of the Frobenius norm amounts to defining

$$P_c(X) = \arg\min_{A \in \mathcal{D}_c} \|X - A\|. \tag{2}$$

An explicit expression for $P_c(X)$ is given in Theorem 1.

Theorem 1. *Given X and a constant $c \geq 0$, the nearest (in the sense of Frobenius norm) matrix $P_c(X) \in \mathcal{D}_c$ to X is of the form*

$$P_c(X) = P_0(X - cI) + cI \tag{3}$$

where (as before) $P_0(X - cI) = (B + H)/2$ for $B = S(X - cI)$ and H the polar factor of B. Furthermore, if X is symmetric with spectral decomposition $X = Q\text{diag}(\lambda_1, \ldots, \lambda_n)Q^\top$ then $P_c(X)$ has the simplified form

$$P_c(X) = Q\text{diag}\{\max(\lambda_1, c), \ldots, \max(\lambda_n, c)\}Q^\top. \tag{4}$$

Proof: The details of the proof are deferred to the Appendix. □

Maintaining symmetry in covariance estimation is typically not difficult, so direct use of (4) will often be sufficient in practice. For non-symmetric X, one may directly symmetrize X before calibration. Note that $P_c(X)$ only depends on X via its symmetric version $S(X)$, so (4) can equivalently be applied to $S(X)$.

2.3. Selection criterion

Clearly $P_c(X)$ varies with c and, as c decreases, the domain $\mathcal{D}_c$ of A expands. At $c = 0$, $P_c(X) = P_0(X)$ becomes positive semidefinite (unless, of course, all eigenvalues of X are already positive). Consequently, we require a criterion for selecting an appropriate positive value, $c = c_*$, say. Let $\lambda^+_{\min}$ be the smallest positive eigenvalue of an estimated covariance matrix X. In order to maintain, as far as possible, the covariance structure of X, it is reasonable to constrain $0 \leq c_* \leq \lambda^+_{\min}$. Rather than make simple choices such as $c_* = \lambda^+_{\min}/2$, here we propose a tuning approach, balancing proximity to X with proximity to singularity. Writing $c_\alpha = 10^{-\alpha}\lambda^+_{\min}$, where $\alpha \geq 0$ is a tuning parameter, with $c_0 = \lambda^+_{\min}$ and $c_\infty = 0$ (i.e., $c \to 0$ as $\alpha \to \infty$), we choose c_* as follows:

Definition 1. *Define* $c_* = c_{\alpha_*}$ *via*

$$\alpha_* = \arg\min_{\alpha} \|X - P_{c_\alpha}(X)\| + \alpha, \tag{5}$$

where (as before) $c_\alpha = 10^{-\alpha}\lambda^+_{\min}$.

Rather than simply minimize the quantity $\|X - A\|$, in (5) we also add a penalty (namely, α) that penalises small values of c. Such penalty terms are widely used in a variety of statistical contexts, such as the AIC/BIC and penalty functions [1, 6, 21]. Reassuringly, positive definite covariance matrices remain unchanged after calibration. To see this, note that $P_{\lambda^+_{\min}}(X) = X$ if X is itself positive definite. In this case, choosing $\alpha = 0$ (so $c = \lambda^+_{\min}$) thus makes both $\|X - P_{c_\alpha}(X)\|$ and α vanish, so $c_* = \lambda^+_{\min}$ and the solution $P_*(X) = P_{c_*}(X)$ completely reduces to X.

The tuning parameter α can also be interpreted in terms of the condition number of the matrix $P_{c_\alpha}(X)$ [24, p146]. For a positive semi-definite matrix, the condition number is the ratio of its biggest to smallest eigenvalues. The condition number can warn us the numerical inaccuracy in calculating the inverse of a given matrix. In our case, let $\lambda^+_{\max}$ be the biggest positive eigenvalues of X and $d = \lambda^+_{\max}/\lambda^+_{\min}$. Then the condition number of the calibrated matrix $P_{c_\alpha}(X)$ is $\kappa(P_{c_\alpha}(X)) = 10^\alpha d$. Therefore, the penalty α approximates the number of digits of accuracy we are prepared to sacrifice in the inversion of $P_{c_\alpha}(X)$ in order to reduce $\|X - P_{c_\alpha}(X)\|$.

2.4. Algorithm

In practice, we implement a screening-search strategy for the tuning parameter α. Rather than let $\alpha \in [0, +\infty)$, we constrain the screening to a feasible region. This strategy is employed in the following algorithm.

Step 1. Given a feasible region of α, say $[0, \alpha_N]$, create a partition $0 = \alpha_0 < \alpha_1 < \alpha_2 < \ldots < \alpha_N$. For $\alpha \in \{\alpha_0, \ldots, \alpha_N\}$, compute the corresponding c_α and use (4) to calculate the resulting solution matrix $P_{c_\alpha}(X)$. We choose

$$\alpha_* = \underset{\alpha \in \{\alpha_0, \ldots, \alpha_N\}}{\arg\min} \|X - P_{c_\alpha}(X)\| + \alpha. \tag{6}$$

Step 2. Set $c_* = c_{\alpha_*}$, and return $P_*(X) = P_{c_*}(X)$ as the final calibrated covariance matrix.

In terms of the screening region $[0, \alpha_N]$ and its partition, we make the following recommendations. In most applications, $10^{-\alpha}$ become negligible when $\alpha > 10$, so we take our default option to be $\alpha_N = 10$. Options of larger α_N are possible when the original $\lambda^+_{\min}$ is in large scale. However, we would not recommend a too large α_N, as it corresponds to a large condition number of $P_*(X)$. When screening $\alpha \in [0, \alpha_N]$, we suggest a uniform partition of the region: for example, given $\alpha_N = 10$, we could use the partition $0, 1, \ldots, 10$ or $0, 0.5, \ldots, 10$. More refined partition would be preferable when extra accuracy in calibration is demanded.

3. Simulation studies

In this section, we carry out two simulation studies to assess the performance of our proposed calibration method. In Simulation 1, we consider three commonly used covariance structures: compound symmetry, first-order autoregressive (AR(1)) and tri-diagonal. In Simulation 2, a more general covariance structure formed by the modified Cholesky decomposition [15] is investigated. Covariance matrices are fitted via the nonparametric covariance estimation approach of Diggle and Verblya [5]. The advantage of that approach lies in that the variogram considered therein has a clear statistical interpretation, being useful in describing spatial correlation in geo-statistics [4], and in judging if the covariance structure is stationary. Here we focus on the non-positive definite covariance matrices obtained by that approach to assess the performance of the proposed calibration method.

The longitudinal data we consider here are described by $(y_{ij}, t_{ij}), i = 1, \ldots, n,\ j = 1, \ldots, n_i$, where y_{ij} represents the measurement j (out of n_i) on subject i and t_{ij} is its measurement time. Let $\mu_{ij} = \mu_i(t_{ij})$ be the mean of y_{ij} and $\mu_i = (\mu_{i1}, \ldots, \mu_{in_i})^\top$ be the vector of the means of the responses $y_i = (y_{i1}, \ldots, y_{in_i})^\top$. Assume Σ_i is the covariance matrix of the responses y_i, where the elements of Σ_i are defined by a generic covariance function $(\Sigma_i)_{j,k} = \Sigma(t_{ij}, t_{ik})$. Following [5], a multivariate normal distribution is assumed, i.e., $y_i \sim \mathcal{N}(\mu_i, \Sigma_i)$. The main covariance fitting process of [5] is briefly summarized as follows. Firstly, a local polynomial smoothing technique is used to estimate the variances in Σ_i, $\widehat{(\Sigma_i)}_{j,j}$, using the sample variances $(y_{ij} - \hat{\mu}_{ij})^2, j = 1, \ldots, n_i$, where $\hat{\mu}_{ij}$ are the fitted means through certain nonparametric regression estimation methods, such as [25]. Secondly, a bivariate local polynomial smoothing method is used to model the variograms $\hat{v}_{ijk}$ through the sample variograms $\{(y_{ij} - \hat{\mu}_{ij}) - (y_{ik} - \hat{\mu}_{ik})\}^2/2$. Finally, the off-diagonal elements of Σ_i, $\widehat{(\Sigma_i)}_{j,k}, j \neq k$, are estimated through $\widehat{(\Sigma_i)}_{j,k} = \{\widehat{(\Sigma_i)}_{j,j} + \widehat{(\Sigma_i)}_{k,k}\}/2 - \hat{v}_{ijk}$.

3.1. Simulation 1

We generate 100 datasets based on the Gaussian process mechanism described above. In each dataset there are $n = 50$ subjects and $n_i = m = 10$ or 20 repeated measurements for each subject. The means μ_{ij} are formulated as $\mu_{ij} = t_{ij} + \sin(t_{ij})$ for all i, j, with measurement times $t_{ij} = j$ for all j. Given a common variance σ^2 and a correlation parameter ρ, the covariance structure Σ_i of subject i is assumed to have the particular structure described below:

1. Compound symmetry. Within-subject correlation is assumed equal for any disjoint pair of observations. In other words, $\Sigma_i = \sigma^2\{(1-\rho)I + \rho J\}$, where $\rho \in (-1/(m-1), 1)$, I is an identity matrix and J is a matrix of ones with order m.
2. AR(1). Within-subject correlation decreases with the time separation as $\Sigma_i = \sigma^2(\rho^{|j-k|})$ $(j, k = 1, \ldots, m,\ \rho \in (-1, 1))$.
3. Tri-diagonal. Within-subject correlation vanishes except for adjacent observations, i.e., the (j, k)-th element of Σ_i is given by

$$(\Sigma_i)_{j,k} = \begin{cases} \sigma^2, & j = k \\ \sigma^2\rho, & |j-k| = 1 \\ 0, & |j-k| \geq 2 \end{cases}$$

where $\rho \in (-\{\cos(\pi/(m+1))\}^{-1}/2, \{\cos(\pi/(m+1))\}^{-1}/2)$.

Table 1: Calibration for conventional covariance structures with $m = 10$

	ρ	0	0.2	0.5	0.8
Compound symmetry	$\|\widehat{\Sigma} - \Sigma\|$	13.8075	13.8161	13.8507	13.8815
	non-P.D.frequency (%)	94	96	100	100
	$\|P_*(\widehat{\Sigma}) - \Sigma\|$	13.7601	13.7694	13.7956	13.8280
	$\|\widehat{\Sigma} - P_*(\widehat{\Sigma})\|$	0.6648	0.7194	1.0337	1.3536
AR(1)	$\|\widehat{\Sigma} - \Sigma\|$	13.8075	13.8715	13.9560	13.8937
	non-P.D.frequency (%)	94	67	20	9
	$\|P_*(\widehat{\Sigma}) - \Sigma\|$	13.7601	13.8371	13.9207	13.8611
	$\|\widehat{\Sigma} - P_*(\widehat{\Sigma})\|$	0.6648	0.4311	0.5305	0.5815
Tri-diagonal	$\|\widehat{\Sigma} - \Sigma\|$	13.8075	13.8687	14.0129	/
	non-P.D.frequency (%)	94	69	27	/
	$\|P_*(\widehat{\Sigma}) - \Sigma\|$	13.7601	13.8331	13.9826	/
	$\|\widehat{\Sigma} - P_*(\widehat{\Sigma})\|$	0.6648	0.4425	0.4304	/

We explore with four different values of $\rho \in \{0, 0.2, 0.5, 0.8\}$, representing different degrees of within-subject correlation. An exception exists in the tri-diagonal case, where ρ can only take three different values $\rho = 0, 0.2, 0.5$, as $\rho = 0.8$ leads to a non-positive definite matrix Σ_i. Table 1 and Table 2 display the simulation results with $m = 10$ and $m = 20$. The labels $\|\widehat{\Sigma} - \Sigma\|$, $\|\widehat{\Sigma} - P_*(\widehat{\Sigma})\|$ and $\|P_*(\widehat{\Sigma}) - \Sigma\|$ denote the average Frobenius norm between the true covariance matrix Σ, the estimated covariance matrix $\widehat{\Sigma}$ and the calibrated covariance matrix $P_*(\widehat{\Sigma})$. The row headed "non-P.D.frequency" records the proportion of non-positive definite $\widehat{\Sigma}$ arising in the covariance estimation procedure. The very high rate of non-positive definite $\widehat{\Sigma}$, particularly when $m = 20$, demonstrates the need for calibration. Also, the $\|P_*(\widehat{\Sigma}) - \Sigma\|$ are smaller than $\|\widehat{\Sigma} - \Sigma\|$, indicating that, on average, $P_*(\widehat{\Sigma})$ is closer to the true covariance matrix than $\widehat{\Sigma}$. Furthermore, the reported $\|\widehat{\Sigma} - P_*(\widehat{\Sigma})\|$ are relatively small, meaning that the calibration method makes no essential changes to the original covariance matrix, other than in correcting its non-positive definiteness. Comparing Table 2 with Table 1, the averaged Frobenius norms $\|P_*(\widehat{\Sigma}) - \Sigma\|$, $\|\widehat{\Sigma} - \Sigma\|$ and $\|\widehat{\Sigma} - P_*(\widehat{\Sigma})\|$ increase with m, indicating the dimension dependency of the Frobenious norm. Another potential reason is that $\widehat{\Sigma}$ with higher dimension were more frequently to be non-positive definiteness, so that more actual calibrations were manipulated. When $\rho = 0$ all three structures reduce to the identity structure and the results become identical.

Table 2: Calibration for conventional covariance structures with $m = 20$

	ρ	0	0.2	0.5	0.8
Compound symmetry	$\|\widehat{\Sigma} - \Sigma\|$	29.0470	29.2198	29.2109	29.3859
	non-P.D.frequency (%)	100	100	100	100
	$\|P_*(\widehat{\Sigma}) - \Sigma\|$	28.7750	29.0383	29.1101	29.3128
	$\|\widehat{\Sigma} - P_*(\widehat{\Sigma})\|$	3.2369	2.0760	1.6720	1.6933
AR(1)	$\|\widehat{\Sigma} - \Sigma\|$	29.0470	29.1571	29.3189	29.4197
	non-P.D.frequency (%)	100	100	100	74
	$\|P_*(\widehat{\Sigma}) - \Sigma\|$	28.7750	28.8966	29.1032	29.3403
	$\|\widehat{\Sigma} - P_*(\widehat{\Sigma})\|$	3.2369	2.9150	2.1133	0.7128
Tri-diagonal	$\|\widehat{\Sigma} - \Sigma\|$	29.0470	29.0928	29.2723	/
	non-P.D.frequency (%)	100	100	100	/
	$\|P_*(\widehat{\Sigma}) - \Sigma\|$	28.7750	28.8416	29.0362	/
	$\|\widehat{\Sigma} - P_*(\widehat{\Sigma})\|$	3.2369	2.8607	2.5349	/

3.2. Simulation 2

We now consider a more general covariance structure via the modified Cholesky decomposition [15]. With a covariance matrix Σ_i of order m, the modified Cholesky decomposition of Σ_i is specified by $T_i \Sigma_i T_i^\top = D_i$, where

$$T_i = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ -\phi_{i21} & 1 & 0 & \cdots & 0 \\ -\phi_{i31} & -\phi_{i32} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -\phi_{im1} & -\phi_{im2} & -\phi_{im3} & \cdots & 1 \end{pmatrix}, D_i = \begin{pmatrix} \sigma_{i1}^2 & 0 & \cdots & 0 \\ 0 & \sigma_{i2}^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{im}^2 \end{pmatrix},$$

and where ϕ_{ijk} and σ_{ij}^2 are the generalized autoregressive parameters and innovation variances, respectively. We then parameterize ϕ_{ijk} and σ_{ij}^2 as functions of their corresponding measurement times, $\phi_{ijk} = g(t_{ij}, t_{ik})$ and $\ln \sigma_{ij}^2 = q(t_{ij})$, where $g(.,.)$ and $q(.)$ are two- and one-dimensional smoothing functions, respectively. With different specifications for $g(.,.)$ and $q(.)$, the covariance matrix Σ_i encompases a wide range of covariance structures. Here we assume $g(t_{ij}, t_{ik}) = m^{-2}(t_{ij}^2 + t_{ik}^2) \exp\{-(t_{ij} - t_{ik})/4\}$ and $q(t_{ik}) = 2 \ln[\ln\{t_{ik}/(m+2)\}]$, with $m = 10, 20$. With the same mean function of Simulation 1, 100 simulated datasets are generated. The numerical results are presented in Table 3. Again, we see that our proposed method provides, on average, a closer-to-true surrogate covariance matrix. In the case of $m = 20$, the calibrated covariance matrix $P_*(\widehat{\Sigma})$ substantially improves $\widehat{\Sigma}$ in the sense of the Frobenius norm ($\|P_*(\widehat{\Sigma}) - \Sigma\| = 36.5162$ while $\|\widehat{\Sigma} - \Sigma\| = 45.5746$). Comparing the case of $m = 10$ to $m = 20$, $\|\widehat{\Sigma} - P_*(\widehat{\Sigma})\|$ substantially increases

from 1.1041 to 26.5426. This is partly because of the dimension dependency of the F-norm. It may also imply that the covariance estimate $\widehat{\Sigma}$ with $m = 20$ were more ill-conditioned and deeper calibrations were made.

Table 3: Calibration of a more general covariance structure with $m = 10$ and $m = 20$

m	10	20
$\|\widehat{\Sigma} - \Sigma\|$	14.4988	45.5746
non-P.D.frequency (%)	89	100
$\|P_*(\widehat{\Sigma}) - \Sigma\|$	14.4718	36.5162
$\|\widehat{\Sigma} - P_*(\widehat{\Sigma})\|$	1.1041	26.5426

4. Example analyses

In this section, we revisit two substantive analyses to demonstrate the properties of the proposed calibration method for covariance matrix calibration.

4.1. Cattle data

Kenward's cattle data [13] involves 60 cattle assigned randomly into two treatment groups: half of the cattle received treatment A, and the other half received treatment B. The cattle were weighed 11 times in total over a nineteen-week period. The data are balanced in the sense that the weighing times were the same for every cow. Using Diggle and Verbyla's nonparametric regression approach [5] to model these two treatment groups separately, we calculate the associated covariance matrix estimates for the two groups. In this instance, both groups' estimated covariance matrices are positive definite. Here, we illustrate the need for calibration in missing data settings, by artificially removing different portions of the cattle data. For each subject, the first six repeated measurements are kept observed, but from the seventh repeated measurement onwards the measurement is set to be missing with probability θ. Once a measurement is missing, all subsequent measurements for this subject are also dropped. With this manipulation, non-positive definite covariance estimates emerge as θ increases. The missingness rate is set to be $\theta = 0\%$ (full data), $10\%, 20\%, 30\%$, respectively. These data sets with missingness are then analyzed and calibrated where necessary. Because the true covariance Σ is unknown, the sample covariance based on the full data, Σ_s, is used as a benchmark for Σ.

Table 4: Calibration results of cattle data with/without missingness

		Positive	F-Norm among $\widehat{\Sigma}$, Σ_s and $P_*(\widehat{\Sigma})$		
	θ	definiteness of $\widehat{\Sigma}$	$\|\widehat{\Sigma} - \Sigma_s\|$	$\|P_*(\widehat{\Sigma}) - \Sigma_s\|$	$\|\widehat{\Sigma} - P_*(\widehat{\Sigma})\|$
A	0%	Yes	116.8018	116.8018	0
	10%	Yes	197.2300	197.2300	0
	20%	No	215.1713	213.5823	12.7821
	30%	No	347.9251	344.2363	24.2091
B	0%	Yes	109.7911	109.7911	0
	10%	Yes	307.9060	307.9060	0
	20%	Yes	348.7275	348.7275	0
	30%	No	511.1006	498.6707	30.1888

Table 4 shows that for the full data set or cases with relatively low missingness rates (treatment A with missing rate up to 10%, treatment B with missing rate up to 20%), the $\widehat{\Sigma}$ are positive definite. In these cases, the calibrated matrices $P_*(\widehat{\Sigma})$ are identical to $\widehat{\Sigma}$, the calibration keeping $\widehat{\Sigma}$ unchanged. When the missing rate increases to 20% for treatment A and 30% for treatment B, $\widehat{\Sigma}$ become non-positive definite. In these circumstances, the proposed calibration method yields surrogate matrices $P_*(\widehat{\Sigma})$ that are positive-definite and whose Frobenius distances to Σ_s are shorter than those from $\widehat{\Sigma}$.

4.2. CD4+ data

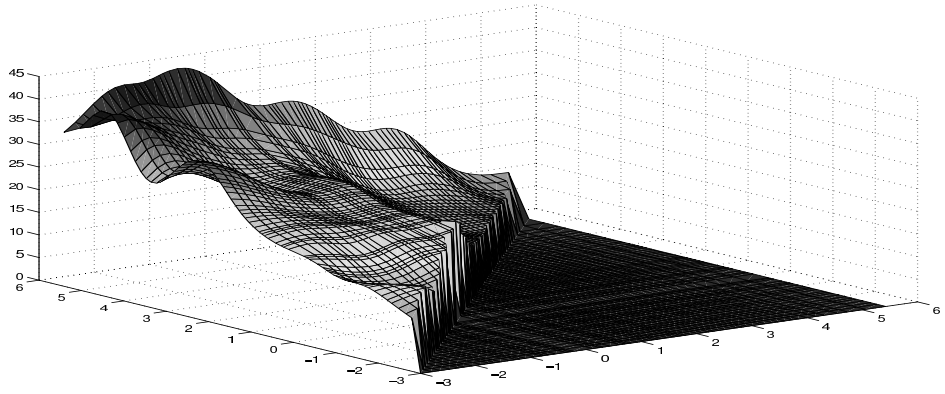

Figure 1: Variograms of CD4+ data

The CD4+ data comes from an AIDS cohort study [12] comprising 369 infected patients. In total, 2376 repeated measurements of CD4+ cell counts were taken over a period of eight and half years. The data are highly unbalanced, with measurement times varying from subject to subject. Diggle and Verbyla [5] analyzed the CD4+ data using their proposed nonparametric covariance structure estimation method. Their estimated covariance matrix turns to be non-positive definite, however. We reanalyze the CD4+ data and then use our proposed calibration method to calibrate the original covariance matrix estimate.

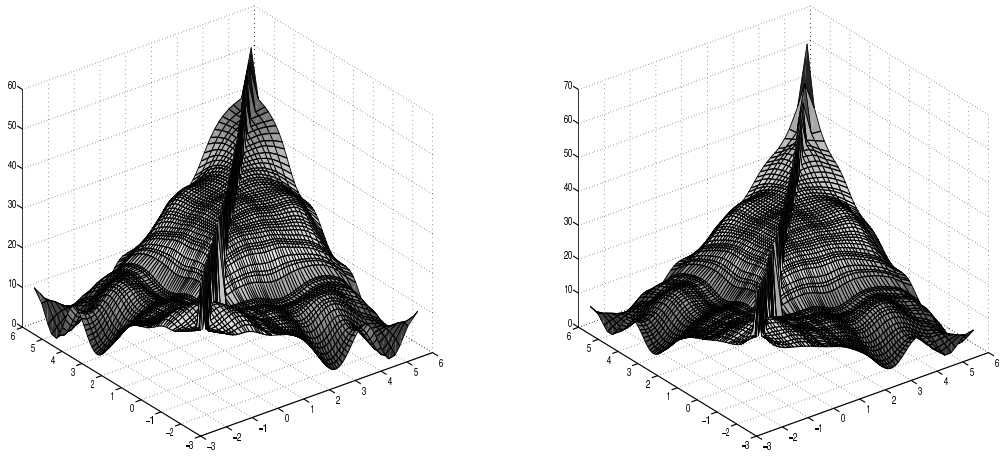

Figure 2: Original covariance estimate $\widehat{\Sigma}$ (left) and its calibration matrix $P_*(\widehat{\Sigma})$ (right) in the CD4+ data

The estimated variogram surface is presented in Figure 1, corresponding to Figure 8 of [5]. The variogram varies for time pairs with equal lags, implying that the underlying longitudinal process for the CD4+ cell counts may be non-stationary. In Figure 2, we plot the original covariance matrix estimate $\widehat{\Sigma}$ and its calibrated covariance matrix $P_*(\widehat{\Sigma})$, where $\widehat{\Sigma}$ is found to be non-positive definite, as mentioned by [5]. From Figure 2 we can see that $\widehat{\Sigma}$ and $P_*(\widehat{\Sigma})$ are very similar in terms of shape and structure, indicating that the calibration approach maintains the major characteristics of the original covariance matrix.

5. Discussion

We have proposed a calibration approach that provides a positive definite surrogate for any given non-positive definite matrix. The calibrated covariance matrix preserves the major characteristics of the original matrix, while being closer to the true covariance than the original matrix, in the sense of the Frobenius norm. Figure 3 illustrates the idea behind our approach, where the dashed circle represents the domain $\mathcal{D}_0$ of all positive semidefinite matrices whilst the solid circle is the domain $\mathcal{D}_c$ of all positive definite matrices with eigenvalues no smaller than c. Given a positive constant c, a non-positive definite matrix $\widehat{\Sigma}$'s nearest positive definite matrix $P_c(\widehat{\Sigma})$ will be closer to the true covariance matrix, provided $\Sigma \in \mathcal{D}_c$. We might therefore expect that using our positive definite surrogate will improve efficiency and accuracy in mean estimation.

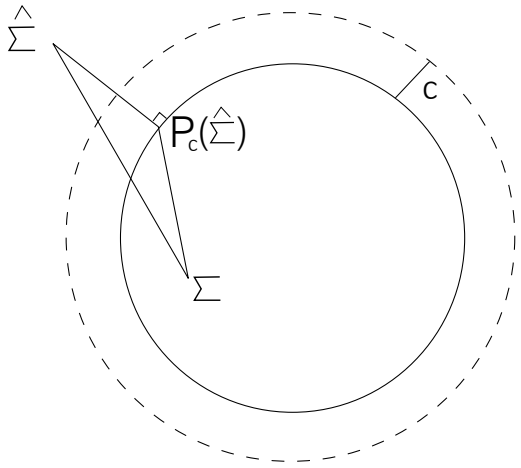

Figure 3: Illustration plot on Σ, $\widehat{\Sigma}$, $P_c(\widehat{\Sigma})$ and c

One potential extension is to replace the domain $\mathcal{D}_c$ by a more general set $\{A : A - \text{diag}(c_1, \ldots, c_n) \geq 0\}$ where the c_i are all positive. However, this extension implies different restrictions imposed on the eigenvalues at different positions, which may be difficult to justify. It also uses the same number of parameters as the dimension of the covariance matrix, requiring

intensive computational efforts, particularly for high-dimensional data. We feel that this particular extension is unlikely to offer substantial benefit.

Our proposed method is not constrained by model assumptions, and hence can be used in both likelihood-based methods (such as generalised linear models) and moment-based approaches (such as generalized estimating equations). Neither is it limited by data structures, indicating it can be applied into any multivariate data setting. In principle, it is applicable to any field of multivariate data analysis where non-positive definiteness of a covariance matrix estimator is a concern. Since the proposed approach is a calibration approach, rather than a covariance estimation approach itself, it can be directly incorporated in any existing covariance estimation process, and offers a routine check and calibration of covariance matrix estimators.

Appendix: Proof of Theorem 1

We seek

$$P_c(X) = \arg\min_{A \in \mathcal{D}_c} \|X - A\|.$$

Rewrite this as

$$\arg\min_{A:(A-cI)\in\mathcal{D}_0} \|(X - cI) - (A - cI)\| = \arg\min_{A'\in\mathcal{D}_0} \|(X - cI) - A'\| + cI.$$

From [9], this latter is just $P_0(X - cI) + cI$, as required. If, further, X is symmetric, it has spectral decomposition $X = Q\Lambda Q^\top$ (say) for orthogonal Q and diagonal Λ. Therefore $X - cI = Q(\Lambda - cI)Q^\top$, and

$$P_c(X) = Q\text{diag}\{\max(\lambda_1 - c, 0), \ldots, \max(\lambda_n - c, 0)\}Q^\top + cI.$$

But $cI = QcQ^\top$, so $P_c(X) = Q\text{diag}\{\max(\lambda_1, c), \ldots, \max(\lambda_n, c)\}Q^\top$, as required. □

Acknowledgements

We gratefully acknowledge very constructive comments and suggestions made by the Editor, AE and one anonymous reviewer, which leads to significant improvements to the paper.

References

[1] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716-723.

[2] Borsdorf, R. and Higham, N.J. (2010). A Preconditioned Newton Algorithm for the Nearest Correlation Matrix. *IMA Journal of Numerical Analysis*, **30**, 94-107.

[3] Chtelat, D. and Wells, M.T. (2016). Improved second order estimation in the singular multivariate normal model. *Journal of Multivariate Analysis*, **147**, 1-19.

[4] Diggle, P. J. (1988). An approach to the analysis of repeated measures. *Biometrics*, **44**, 959-971.

[5] Diggle, P.J. and Verbyla, A. P. (1998). Nonparametric estimation of covariance structure in longitudinal data. *Biometrics*, **54**, 401-415.

[6] Fan, J.Q. and Li, R.Z. (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of Royal Statistical Association*, **96**, 1348-1360.

[7] Halmos, P.R. (1972). Positive approximants of operators. *Indiana Univ. of Math. J.*, **21**, 951-960.

[8] Hendrikse, A.J., Spreeuwers, L.J. and Veldhuis, R.N.J. (2009). A Bootstrap Approach to Eigenvalue Correction. *Ninth IEEE International Conference on Data Mining*, 818-823.

[9] Higham, N.J. (1988). Computing a nearest symmetric positive semidefinite matrix. *Linear Algebra and Appl.*, **103**, 103-118.

[10] Higham, N.J. (2008). Functions of Matrices: Theory and Computation *SIAM.*

[11] Higham, N.J. The matrix Computation Toolbox. `http://www.maths.manchester.ac.uk/~higham/mctoolbox/.`

[12] Kaslow, R. A., Ostrow, D. G., Detels, R., et al. (1987). The Multicenter AIDS Cohort Study: Rationale, organization and selected characteristics of the participants. *American Journal of Epidemiology*, **126**, 310-318.

[13] Kenward, M.G. (1987). A method for comparing profiles of repeated measurements. *A. ppl. Statist.*, **36**, 296-308.

[14] Li, Y. (2011). Efficient semiparametric regression for longitudinal data with nonparametric covariance estimation. *Biometrika*, **98(2)**, 355-370.

[15] Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika*, **86**, 677-90.

[16] Qi, H.D. and Sun, D. (2006). A quadratically convergent Newton method for computing the nearest correlation matrix. *SIAM J. Matrix Anal. Appl.*, **28**, 360-385.

[17] Rebonato, R. and Jäckel P. (2000). The most general methodology to create a valid correlation matrix for risk management and option pricing purposes. *J. of Risk*, **2**, 17-27.

[18] Saleh, A.K.Md.E and Shalabh (2014). A ridge regression estimation approach to the measurement error model. *Journal of Multivariate Analysis*, **123**, 68-84.

[19] Sheena, Y. (2013). Modified estimators of the contribution rates of population eigenvalues. *Journal of Multivariate Analysis*, **115**, 301-316.

[20] Srivastava, M.S. and Yanagihara, H. (2010). Testing the equality of several covariance matrices with fewer observations than the dimension. *Journal of Multivariate Analysis*, **101(6)**, 1319-1329.

[21] Tibshirani, R.J. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society , Series B*, **58**, 267-288.

[22] Wang, Y.G. and Carey, V. (2003). Working correlation structure misspecification, estimation and covariate design: Implications for generalized estimating equations performance. *Biometrika*, **90**, 29-41.

[23] Wothke, W. (1993). Nonpositive definite matrices in structural modeling. In K. A. Bollen, K.A. and Long, J.S. (Editors), Testing structural equation models. *Newbury Park, CA: Sage*, 256-293.

[24] Zarowski., Christopher J. (2004). An Introduction to Numerical Analysis for Electrical and Computer Engineers. *Wiley.*

[25] Zeger, S.L. and Diggle, P.J. (1994). Semiparametric Models for Longitudinal Data with Application to CD4 Cell Numbers in HIV Seroconverters. *Biometrics*, **50**, 689-699.