Computational Statistics and Data Analysis 92 (2015) 13–25

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Covariance matrix estimation for left-censored data[☆]

Maju Pesonen [a,*], Henri Pesonen [a], Jaakko Nevalainen [b]

[a] *Department of Mathematics and Statistics, University of Turku, Finland*

[b] *School of Health Sciences, University of Tampere, Finland*

HIGHLIGHTS

- ML based covariance matrix estimator for left-censored data is introduced.
- Computation times are decreased considerably with parallelized pairwise estimation.
- The proposed estimators produce unbiased estimates with reasonable variation.

ARTICLE INFO

Article history:
Received 12 December 2014
Received in revised form 23 June 2015
Accepted 24 June 2015
Available online 2 July 2015

Keywords:
Maximum likelihood estimation
Covariance matrix
Left-censoring
Non-detects

ABSTRACT

Multivariate methods often rely on a sample covariance matrix. The conventional estimators of a covariance matrix require complete data vectors on all subjects—an assumption that can frequently not be met. For example, in many fields of life sciences that are utilizing modern measuring technology, such as mass spectrometry, left-censored values caused by denoising the data are a commonplace phenomena. Left-censored values are low-level concentrations that are considered too imprecise to be reported as a single number but known to exist somewhere between zero and the laboratory's lower limit of detection. Maximum likelihood-based covariance matrix estimators that allow the presence of the left-censored values without substituting them with a constant or ignoring them completely are considered. The presented estimators efficiently use all the information available and thus, based on simulation studies, produce the least biased estimates compared to often used competing estimators. As the genuine maximum likelihood estimate can be solved fast only in low dimensions, it is suggested to estimate the covariance matrix element-wise and then adjust the resulting covariance matrix to achieve positive semi-definiteness. It is shown that the new approach succeeds in decreasing the computation times substantially and still produces accurate estimates. Finally, as an example, a left-censored data set of toxic chemicals is explored.

1. Introduction

Multivariate methods often rely on the sample covariance matrix. For example, principal component analysis, which is used to describe high dimensional data in lower dimensions, uses the eigenvalue decomposition of the covariance matrix. In canonical correlation analysis, blocks of the covariance matrix are used to find the maximal correlation between linear combinations of variables belonging to two different data sets. The conventional estimators of the covariance matrix require

[☆] A data set used in Section 5 is available online at http://dx.doi.org/10.1016/j.csda.2015.06.005.

[*] Correspondence to: Department of Mathematics and Statistics, Assistentinkatu 7, Publicum 4th floor, 20014 University of Turku, Finland. Tel.: +358 2 3335444.

E-mail address: mekuja@utu.fi (M. Pesonen).

http://dx.doi.org/10.1016/j.csda.2015.06.005

complete data vectors on all subjects, which is an assumption that can frequently not be met. Many fields of life sciences are constrained by the measurement accuracy of modern measurement technology. Often, it is impossible to quantify the exact concentration or the complete absence of a compound, especially at the low end of the detectable concentration range. The lowest concentration that can be reliably detected with the given analytical method is referred to as the lower limit of detection (LLOD) (Browne and Whitcomb, 2010). Measurements falling below the LLOD are referred to as left-censored values or, in some contexts, non-detects.

For data that contain left-censored values, if the analysis is carried out using only the completely observed data, the means of the concentrations would be overestimated and the standard deviations would be underestimated. Consequently, any related test statistic or estimate would be biased. Thus, neglecting these informative censored values can lead to a severe bias and a loss of precision due to the decrease of the effective sample size.

In the presence of left-censored values, a common and simple way to deal with the estimation of the covariance matrix is to delete any subject containing at least one censored value. To avoid completely ignoring these subjects, Little and Rubin (2002) propose to build the estimate of the covariance matrix one element at a time by using all observations for which both values are present. However, the resulting covariance matrix estimate is not necessarily positive semi-definite. According to Mehrotra (1995), the efficient use of all observed data is more important than the possible lack of positive semi-definiteness. If the positive semi-definiteness is lost, he recommends the element-wise estimation of the variances and covariances combined with a possible adjustment.

Another commonly used approach is to substitute the left-censored values with a suitable constant and then to compute the sample covariance from the resulting complete data. The potential substitution value can be the sample mean of the uncensored values for the corresponding variable, zero, LLOD/2, or the minimum of the observed values. These alternative approaches have been investigated in previous studies (Farnham et al., 2002; El-Shaarawi and Esterby, 1992; Succop et al., 2004). All of them are more or less biased, but they are still used despite the criticism (Helsel, 2005, 2006).

In addition to the mentioned quick fixes, there exists more eloquent methods to overcome the challenges set by censored data. Instead of single and simple substitutions, one approach is to multiply impute the censored values. The key idea is to use the conditional distribution of the observed data to generate a set of plausible imputations for the censored data (Rubin, 2004, 1996; Carpenter and Kenward, 2013). Imputations are repeated several times, creating multiple data sets, which are then analyzed individually as if they were complete. Thus, if the main interest lies in estimating the covariance matrix for the left-censored data, the multiple imputation would be followed by computing the standard sample covariances for each of the imputed data sets. Finally, the results are combined across all multiply imputed data sets by so-called "Rubin's rules", which incorporate the imputation-related uncertainty into the analysis (Rubin, 2004). The multiple imputation methods for left-censored data may be appealing due to their relatively simple computational algorithms. The literature includes applications in univariate (Baccarelli et al., 2005; Huybrechts et al., 2002), bivariate (Chen et al., 2011) and multivariate settings (Hopke et al., 2001; Chen et al., 2013).

Hewett and Ganser (2007) divide censored data analysis methods into four categories: substitution methods, log-probit regression, maximum likelihood (ML) estimation methods, and non-parametric methods. None of the methods has been recommended for all different scenarios. The recommendation depends on the sample size, the divergence from log-normal distribution or the degree of censoring. However, due to its many desirable statistical properties, ML estimation is often considered the gold standard provided the data is well-described by some parametric probability distribution (Helsel, 1990; Koo et al., 2002; Zhao and Frey, 2006).

Extensive literature exists regarding univariate and bivariate ML-based methods for estimating the measures of centrality and variability in the presence of the left-censored values, such as Lyles et al. (2001), Lynn (2001) and Williams and Ebel (2014). The majority of these works are based on the normality assumption. Song et al. (2004) propose an alternative, more robust approach based on generalized estimating equations to estimate the correlation between two continuous variables with left-censored values. Their method also has an advantage of not requiring time consuming optimization routines.

The extensions of ML-based estimation techniques to the multivariate setting are fewer. If it could be assumed that each variable is identically distributed (normal, gamma, or Weibull), then the results proposed by Gupta (1952) and Harter and Moore (1967) could be applied. In practice, however, this is not often a realistic assumption. Chung (1993) proposes a covariance matrix estimation method that uses marginal ML estimation. Despite its good practical properties, theoretically it is not a ML estimator but rather a good-enough approximation.

Building on the work of Lyles et al. (2001), Perkins et al. (2013) propose a ML estimator for the mean vector and the covariance matrix based on the multivariate normality assumption in the presence of left-censored values. They formulate the likelihood function as a product of the marginal multivariate distribution for the observed variables and the conditional multivariate distribution for the left-censored variables. They also develop an approximation for the Fisher information and covariance matrix for the estimated parameters and give an example based on a three-parameter multivariate normal distribution.

To overcome the complexity of estimating the parameters of multivariate normal and log-normal models by maximizing the logarithm of the likelihood equations, Hoffmann and Johnson (2014) proposes a pseudo-likelihood approach. The pseudo-likelihood estimate maximizes a computationally simpler approximation of the log-likelihood function but is not equal to it.

Against this background, we address the challenge of dealing with the left-censoring in the analysis of multivariate data and introduce an efficient way to use the values that are below the set LLOD. We formulate the likelihood maximization