Estimating High-dimensional Covariance and Precision Matrices under General Missing Dependence

Seongoh Park[1], Xinlei Wang[2], and Johan Lim*[3]

[1]Department of Statistics, Sungshin Women's University, Seoul, Korea

[2]Department of Statistical Science, Southern Methodist University, Dallas, TX, USA

[3]Department of Statistics, Seoul National University, Seoul, Korea

Abstract

A sample covariance matrix $\boldsymbol{S}$ of completely observed data is the key statistic in a large variety of multivariate statistical procedures, such as structured covariance/precision matrix estimation, principal component analysis, and testing of equality of mean vectors. However, when the data are partially observed, the sample covariance matrix from the available data is biased and does not provide valid multivariate procedures. To correct the bias, a simple adjustment method called inverse probability weighting (IPW) has been used in previous research, yielding the IPW estimator. The estimator can play the role of $\boldsymbol{S}$ in the missing data context, thus replacing $\boldsymbol{S}$ in off-the-shelf multivariate procedures such as the graphical lasso algorithm. However, theoretical properties

*To whom all correspondence should be addressed. Email: johanlim@snu.ac.kr

(e.g. concentration) of the IPW estimator have been only established in earlier work under very simple missing structures; every variable of each sample is independently subject to missingness with equal probability. We investigate the deviation of the IPW estimator when observations are partially observed under general missing dependency. We prove the optimal convergence rate $O_p(\sqrt{\log p/n})$ of the IPW estimator based on the element-wise maximum norm, even when two unrealistic assumptions (known mean and/or missing probability) frequently assumed to be known in the past work are relaxed. The optimal rate is especially crucial in estimating a precision matrix, because of the "meta-theorem" that claims the rate of the IPW estimator governs that of the resulting precision matrix estimator. In the simulation study, we discuss one of practically important issues, non-positive semi-definiteness of the IPW estimator, and compare the estimator with imputation methods.

Keywords: Convergence rate; covariance matrix; dependent missing structure; element-wise maximum norm; inverse probability weighting.