Automatic sparse estimation of high-dimensional cross-covariance matrix

Tetsuya Umino[a], **Kazuyoshi Yata**[b], **and Makoto Aoshima**[b]

[a]Graduate School of Science and Technology, University of Tsukuba
[b]Institute of Mathematics, University of Tsukuba

A common feature of high-dimensional data is that the data dimension is high, however, the sample size is relatively small. This is the so-called " HDLSS " or " large p, small n " data situation where $p/n \to \infty$; here p is the data dimension and n is the sample size. Such data situations occur in many areas of modern science such as genomics, medical imaging, text recognition, finance, chemometrics, and so on.

Suppose we take samples, $\boldsymbol{x}_j,\ j = 1, \ldots, n$, of size $n\ (\geq 4)$, which are independent and identically distributed (i.i.d.) as a p-variate distribution. Here, we consider situations where the data dimension p is very high compared to the sample size n. Let $\boldsymbol{x}_j = (\boldsymbol{x}_{1j}^T, \boldsymbol{x}_{2j}^T)^T$ and assume $\boldsymbol{x}_{ij} \in \mathbf{R}^{p_i}$, $i = 1, 2$, with $p_1 \in [1, p-1]$ and $p_2 = p - p_1$. We assume that $\boldsymbol{x}_j$ has an unknown mean vector, $\boldsymbol{\mu} = (\boldsymbol{\mu}_1^T, \boldsymbol{\mu}_2^T)^T$, and unknown covariance matrix,

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_1 & \boldsymbol{\Sigma}_* \\ \boldsymbol{\Sigma}_*^T & \boldsymbol{\Sigma}_2 \end{pmatrix}\ (\geq \boldsymbol{O}),$$

that is, $\mathrm{E}(\boldsymbol{x}_{ij}) = \boldsymbol{\mu}_i$, $\mathrm{Var}(\boldsymbol{x}_{ij}) = \boldsymbol{\Sigma}_i$, $i = 1, 2$, and $\mathrm{Cov}(\boldsymbol{x}_{1j}, \boldsymbol{x}_{2j}) = \mathrm{E}(\boldsymbol{x}_{1j}\boldsymbol{x}_{2j}^T) - \boldsymbol{\mu}_1\boldsymbol{\mu}_2^T = \boldsymbol{\Sigma}_*$.

Aoshima and Yata [1] and Yata and Aoshima [4, 5] considered testing the cross-covariance matrix by

$$H_0 : \boldsymbol{\Sigma}_* = \boldsymbol{O} \quad \text{vs.} \quad H_1 : \boldsymbol{\Sigma}_* \neq \boldsymbol{O} \tag{1}$$

for high-dimensional settings. When $(p_1, p_2) = (p-1, 1)$ or $(1, p-1)$, (1) implies the test of correlation coefficients. Aoshima and Yata [1] gave a test statistic for the test and Yata and Aoshima [4, 5] improved the test statistic by using a method called the *extended cross-data-matrix (ECDM) methodology*.

In this talk, we consider the problem of estimating the cross-covariance matrix, $\boldsymbol{\Sigma}_*$. There have been several studies on sparse estimation of the entire covariance matrix. For example, Bien and Tibshirani [3] proposed a sparse estimator of the covariance matrix based on L1-penalties, and Bickel and Levina [2] proposed a thresholding estimator of the covariance matrix. However, to our knowledge, sparse estimation of the cross-covariance matrix does not seem to have been studied in high-dimensional settings.

Recently, Yata and Aoshima [6] proposed a new sparse PCA (SPCA) method called the automatic SPCA (A-SPCA). A-SPCA does not depend on any threshold (tuning) values. In this talk, by applying the idea of A-SPCA to the estimation of the cross-covariance matrix,

we propose a new sparse estimator of $\mathbf{\Sigma}_*$. We show that the proposed estimator is consistent without any threshold (tuning) values.

Acknowledgements: Research of the second author was partially supported by Grant-in-Aid for Scientific Research (C), JSPS, under Contract Number 22K03412. Research of the third author was partially supported by Grants-in-Aid for Scientific Research (A) and Challenging Research (Exploratory), JSPS, under Contract Numbers 20H00576 and 22K19769.

References

[1] M. Aoshima and K. Yata. Two-stage procedures for high-dimensional data. *Sequential Analysis (Editor's special invited paper)*, 30: 356–399, 2011.

[2] P. J. Bickel and E. Levina. Covariance regularization by thresholding. *The Annals of Statistics*, 36: 2577 – 2604, 2008.

[3] J. Bien and R. J. Tibshirani. Sparse estimation of a covariance matrix. *Biometrika*, 98: 807–820, 2011.

[4] K. Yata and M. Aoshima. Correlation tests for high-dimensional data using extended cross-data-matrix methodology. *Journal of Multivariate Analysis*, 117:313–331, 2013.

[5] K. Yata and M. Aoshima. High-dimensional inference on covariance structures via the extended cross-data-matrix methodology. *Journal of Multivariate Analysis*, 151:151–166, 2016.

[6] K. Yata and M. Aoshima. Automatic sparse PCA for high-dimensional data. *Statistica Sinica*, 2025 (in press).