

Data Transformation for High-Dimension, Strongly Spiked Eigenvalue Models using Automatic Sparse Estimation

Tetsuya Umino^a, Kazuyoshi Yata^b, and Makoto Aoshima^b

^aGraduate School of Science and Technology, University of Tsukuba

^bInstitute of Mathematics, University of Tsukuba

1 Introduction

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

Suppose we have a sample of n (≥ 4) independent and identically distributed (i.i.d.) p -variate random vectors, $\mathbf{x}_1, \dots, \mathbf{x}_n$. Here, we consider situations where the data dimension p is very high compared to the sample size n . We also assume that $\mathbf{x}_j$ has a mean vector $\mathbf{0}$ and an unknown covariance matrix $\mathbf{\Sigma}$ ($\geq \mathbf{O}$), and the eigen-decomposition of $\mathbf{\Sigma}$ is given by $\sum_{s=1}^p \lambda_s \mathbf{h}_s \mathbf{h}_s^T$, where $\lambda_1 \geq \dots \geq \lambda_p (\geq 0)$ are the eigenvalues of $\mathbf{\Sigma}$ and $\mathbf{h}_1, \dots, \mathbf{h}_p$ are the corresponding eigenvectors. Recently, Aoshima and Yata [1] established two disjoint models for high-dimensional data based on the covariance structure of $\mathbf{\Sigma}$: the strongly spiked eigenvalue (SSE) model and the non-SSE (NSSE) model. The SSE model is defined by

$$\liminf_{p \rightarrow \infty} \frac{\lambda_{\max}(\mathbf{\Sigma})}{\sqrt{\text{tr}(\mathbf{\Sigma}^2)}} > 0, \quad (1)$$

where $\lambda_{\max}(\mathbf{\Sigma})$ is the largest eigenvalue of $\mathbf{\Sigma}$. On the other hand, the NSSE model is defined by

$$\frac{\lambda_{\max}(\mathbf{\Sigma})}{\sqrt{\text{tr}(\mathbf{\Sigma}^2)}} = 0, \quad \text{as } p \rightarrow \infty. \quad (2)$$

High-dimensional statistical methods are often developed under the NSSE model. However, in practice, the NSSE model may not hold due to high correlations between variables or data mixing, meaning that these methods may not provide valid analytical results. In this talk, we propose a data transformation method that transforms high-dimensional data $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]$ generated from the SSE model into a new dataset $\tilde{\mathbf{X}} = [\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_n]$ that approximately follows the NSSE model.

2 Data Transformation Method using Automatic Sparse PCA

Let $\mathbf{S}$ be the sample covariance matrix, and consider its eigen-decomposition as $\mathbf{S} = \sum_{s=1}^p \hat{\lambda}_s \hat{\mathbf{h}}_s \hat{\mathbf{h}}_s^T$, where $\hat{\lambda}_1 \geq \dots \geq \hat{\lambda}_p (\geq 0)$. For $l = 1, \dots, p$, let $\Psi_{(l)} = \sum_{s=l}^p \lambda_s^2$. Under the SSE model (1), we assume the following condition:

(A-i) There exists a positive fixed integer k such that $\lambda_1, \dots, \lambda_k$ are distinct in the sense that $\liminf_{p \rightarrow \infty} (\lambda_s / \lambda_t - 1) > 0$ when $1 \leq s < t \leq k$, and λ_k and λ_{k+1} satisfy

$$\liminf_{p \rightarrow \infty} \frac{\lambda_k^2}{\Psi_{(k)}} > 0 \quad \text{and} \quad \frac{\lambda_{k+1}^2}{\Psi_{(k+1)}} \rightarrow 0 \quad \text{as } p \rightarrow \infty.$$

Condition (A-i) was introduced by Aoshima and Yata [1] as a special case of the SSE model. Under (A-i), we define the matrix $\mathbf{A}$ by

$$\mathbf{A} = \mathbf{I}_p - \sum_{s=1}^k \mathbf{h}_s \mathbf{h}_s^T = \sum_{s=k+1}^p \mathbf{h}_s \mathbf{h}_s^T.$$

Then, transforming $\mathbf{x}_j$ into $\mathbf{A}\mathbf{x}_j$ yields data following the NSSE model (2). However, as pointed out by Aoshima and Yata [1], the sample eigenvectors $\hat{\mathbf{h}}_s$ cannot ensure an accurate estimation of $\mathbf{h}_s$ in high-dimensional settings.

In this talk, focusing on the representation $\mathbf{A}\mathbf{x}_j = \mathbf{x}_j - \sum_{s=1}^k (\mathbf{x}_j^T \mathbf{h}_s) \mathbf{h}_s$, we propose a novel method for estimating $\mathbf{A}\mathbf{x}_j$ by combining the estimators of $(\mathbf{x}_j^T \mathbf{h}_s)$ and $\mathbf{h}_s$. Let $x_{js} = \mathbf{x}_j^T \mathbf{h}_s$ for $s = 1, \dots, k$, and let $\tilde{x}_{js}$ be the estimator of x_{js} proposed by Aoshima and Yata [1]. In addition, let $\tilde{\mathbf{h}}_{s*}$, $s = 1, \dots, k$, be the estimator of $\mathbf{h}_s$ obtained by the automatic sparse PCA proposed by Yata and Aoshima [2]. Based on these estimators, for each $j = 1, \dots, n$, we estimate $\mathbf{A}\mathbf{x}_j$ as

$$\tilde{\mathbf{x}}_j = \mathbf{x}_j - \sum_{s=1}^k \tilde{x}_{js} \tilde{\mathbf{h}}_{s*}.$$

Then, we have the following result.

Theorem 1. Let $\widetilde{\mathbf{X}} = [\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_n]$. Under (A-i) and some regularity conditions, it holds that

$$\frac{1}{n} \|\mathbf{A}\mathbf{X} - \widetilde{\mathbf{X}}\|_F^2 = \frac{1}{n} \sum_{j=1}^n \|\mathbf{A}\mathbf{x}_j - \tilde{\mathbf{x}}_j\|^2 = o_P(\lambda_1)$$

as $p, n \rightarrow \infty$.

References

- [1] Aoshima, M. and Yata, K. (2018). Two-sample tests for high-dimension, strongly spiked eigenvalue models. *Statistica Sinica*, 28:43–62.
- [2] Yata, K. and Aoshima, M. (2025). Automatic sparse PCA for high-dimensional data. *Statistica Sinica*, 35:1069–1090.