A classification procedure for high-dimension, low-sample-size data under the strongly spiked eigenvalue model

Aki Ishii
Department of Information Sciences, Tokyo University of Science

Nowadays, you can see many types of high-dimensional data such as genetic microarrays, medical imaging, text recognition, finance, chemometrics, and so on. A common feature of high-dimensional data is that the data dimension is extremely high, however, the sample size is relatively low. We call such data "HDLSS" or "large p, small n" data, where p is the data dimension and n is the sample size. In this talk, we consider high-dimensional classification based on eigenstructures. Note that one cannot use a typical classification rule for HDLSS data. Suppose we have two classes $\pi_i,\ i=1,2$, and take independent p dimensional samples from each π_i having a mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i\ (\geq \boldsymbol{O})$. We do not assume any distribution functions. Let $\lambda_{1(i)},...,\lambda_{p(i)}$ be eigenvalues of $\boldsymbol{\Sigma}_i$, where $\lambda_{1(i)} \geq \cdots \geq \lambda_{p(i)} (\geq 0)$. Aoshima and Yata (2017) proposed two types of eigenvalue models. One is called the strongly spiked eigenvalue (SSE) model and defined as follows:

$$\liminf_{p\to\infty} \left\{ \frac{\lambda_{1(i)}^2}{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)} \right\} > 0 \quad \text{for } i=1 \text{ or } 2. \tag{1}$$

The other one is called the non-SSE (NSSE) model and defined as follows:

$$\frac{\lambda_{1(i)}^2}{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)} \to 0 \quad \text{as } p\to\infty \text{ for } i=1,2. \tag{2}$$

In this talk, we focus on (1). We often see (1) when we analyze microarray data. Ishii et al. (2016) gave asymptotic properties of the first eigenspace under (1). Aoshima and Yata (2014) gave an effective classifier called the distance-based classifier for (2). In order to develop the distance-based classifier for (1), we consider the data transformation given by Aoshima and Yata (2017). By using the noise-reduction methodology given by Yata and Aoshima (2012), we estimate the first eigenspaces and show that our classifier has some preferable asymptotic properties when p is large theoretically. Also, we give some simulation results and data analysis.

[1] Aoshima, M. and Yata, K. (2017). Two-sample tests for high-dimension, strongly spiked eigenvalue models. *Statist. Sinica*, in press.

[2] Aoshima, M. and Yata, K. (2014). A distance-based, misclassification rate adjusted classifier for multiclass, high-dimensional data. *Ann. Inst. Statist. Math.*, 66, 983-1010.

[3] Ishii, A., Yata, K. and Aoshima, M. (2016). Asymptotic properties of the first principal component and equality tests of covariance matrices in high-dimension, low-sample-size context. *J. Stat. Plan. Inference*, 170, 186-199.

[4] Yata, K. and Aoshima, M. (2012). Effective PCA for high-dimension, low-sample-size data with noise reduction via geometric representations. *J. Multivariate Anal.* 105, 193-215.