A quadratic classifier for high-dimensional data under the strongly spiked eigenvalue model

Aki Ishii[1], **Kazuyoshi Yata**[2] **and Makoto Aoshima**[2]

[1]Department of Information Sciences, Tokyo University of Science
[2]Institute of Mathematics, University of Tsukuba

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 $\boldsymbol{\Sigma}_i\ (\geq \boldsymbol{O})$. 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 (2018a) 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:

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

Let $\bar{\boldsymbol{x}}_{in_i}$ and $\boldsymbol{S}_{in_i}$ be the sample mean vector and the sample covariance matrix for $i = 1,2$. Aoshima and Yata (2011, 2018b) gave an effective quadratic classifier called the geometric classifier. By using the geometric classifier, one classifies the individual $\boldsymbol{x}_0$ into π_1 if

$$\frac{p||\boldsymbol{x}_0-\bar{\boldsymbol{x}}_{1n_1}||^2}{\mathrm{tr}(\boldsymbol{S}_{1n_1})} - \frac{p||\boldsymbol{x}_0-\bar{\boldsymbol{x}}_{2n_2}||^2}{\mathrm{tr}(\boldsymbol{S}_{2n_2})} - p\log\left\{\frac{\mathrm{tr}(\boldsymbol{S}_{2n_2})}{\mathrm{tr}(\boldsymbol{S}_{1n_1})}\right\} - \frac{p}{n_1} + \frac{p}{n_2} < 0$$

and into π_2 otherwise. They also showed the asymptotic normality of the classifier under (2) and discuss the sample size determination to control the misclassification rate. In this talk, we focus on (1) and give a new quadratic classifier. We often see (1) when we analyze microarray data, so it is very important to develop theories and methodologies for (1). We propose a new geometric classifier for (1) by using the data transformation technique given by Aoshima and Yata (2018a). We show that our new classifier has several preferable asymptotic properties for high-dimensional data under (1). Finally, we demonstrate our new classifier by using microarray data sets.

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

[2] Aoshima, M. and Yata, K. (2018a). Two-sample tests for high-dimension, strongly spiked eigenvalue models. *Statist. Sinica*, 28, 43-62.

[3] Aoshima, M. and Yata, K. (2018b). High-dimensional quadratic classifiers in non-sparse settings. *Methodol. Comput. Appl. Probab.*, to appear.