高次元データ解析におけるグラフ表現と漸近的特性

Geometric representation of high-dimensional data and its asymptotic properties

数学専攻　河口　裕
Kawaguchi Yutaka

Abstract

In recent years, high dimension low sample size (HDLSS) data are emerging in various areas of science, which are genetic microarrays, medical image and finance. Such HDLSS data presents a substantial challenge to many methods for classical statistical analysis. Namely, because the covariance matrix for HDLSS data is not of full rank, the inverse for this one does not exist. Accordingly, statistical methods can not be used for HDLSS data.

Consider a random sample of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ from a p-dimensional population. The high dimension low sample size (HDLSS) data can be regarded as n vectors or points in p-dimensional space. We discuss the asymptotic behavior of HDLSS as p tends to infinity. Recently, there is a considerable interest for a high-dimensional data set when the dimension is large. In high-dimensional asymptotic theory, it is assumed that (i) p tends to infinity and n is fixed, or (ii) both p and n tend to infinity. The first high-dimensional framework is used for high-dimensional low sample data (HDLSS). Assuming that $\boldsymbol{x}_i$'s are a sample from $N(0, I_p)$. Hall *et al.* (2005) showed that the three geometric statistics satisfy the following under large-dimension-fixed-sample size;

$$\begin{aligned} \|\boldsymbol{x}_i\| &= \sqrt{p} + \mathrm{O}_p(1), \quad i = 1, \ldots, n, \\ \|\boldsymbol{x}_i - \boldsymbol{x}_j\| &= \sqrt{2p} + \mathrm{O}_p(1), \quad i, j = 1, \ldots n,\ i \neq j, \\ \mathrm{ang}(\boldsymbol{x}_i, \boldsymbol{x}_j) &= \frac{\pi}{2} + \mathrm{O}_p(p^{-1/2}), \quad i, j = 1, \ldots n,\ i \neq j, \end{aligned}$$

where $\|\cdot\|$ is the Euclidean distance and O_p denotes the stochastic order. These results imply that the data converge to the vertices of a deterministic regular simplex. In non-normal case, these properties were extended with some assumptions. They extended these properties to the case that two data sets are drawn from different distributions, and examined the performance of some discrimination rules.

In this paper, we mainly refine their results and study influence of dimension p on these properties in standard normal case. Our results are refined results of Hall *et al*, and may be used to extend the statistical insights based on the asymptotic behaviors to a middle-dimensional case.

We firstly try to refine these results in multivariate standard normal case by asymptotic expansion of distributions of geometric features in Section 2. To get

them, we difined three statistics

$$T_1 = \sqrt{2}(\|\boldsymbol{x}_i\| - \sqrt{p}),$$
$$T_2 = \|\boldsymbol{x}_i - \boldsymbol{x}_j\| - \sqrt{2p},$$
$$T_3 = \sqrt{q}\left(\frac{\pi}{2} - \theta\right),$$

where the variable θ denotes the angle of $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, $q = p - \Delta$ and Δ is the correction term. Then the limiting distributions of these statistics are the standard normal distributions. The distribution of $T_1 = \sqrt{2}(\|\boldsymbol{x}_i\| - \sqrt{p})$ is expanded as

$$\Phi(x) - \phi(x)\left[\frac{1}{\sqrt{p}}\ell_1(x) + \frac{1}{p}\ell_2(x)\right] + \mathrm{o}(p^{-1}).$$

Here $\ell_1(x)$ and $\ell_2(x)$ are defined as follows,

$$\ell_1(x) = \frac{\sqrt{2}}{12}h_2(x) - \frac{\sqrt{2}}{4}h_0(x),$$
$$\ell_2(x) = \frac{1}{144}[-15h_5(x) - 6h_3(x) + 16h_2(x) - 81h_1(x) + 72h_0(x)],$$

where $h_i(x)$ denotes the Hermite polynomial. In addition, asymptotic expansion of distribution of T_3 is

$$\Phi(x) + \frac{1}{12q}\left[h_3(u) + 6(2\Delta - 1)h_1(x)\right]\phi(x) + \mathrm{o}(q^{-1}).$$

In Section 3, we obtain computable error bounds for limiting distributions of the length and the one of distance i.e

$$|\mathrm{P}(T_i \leq x) - \Phi(x)| \leq B(p) = \mathrm{O}(p^{-1/2}), \quad (i = 1, 2)$$

where

$$B(p) = \min_{\lambda} D(\lambda, p) + \frac{2}{\mathrm{e}\sqrt{p\pi}}. \tag{1}$$

The idea to get the error bounds is based on Ulyanov et al (2006). They obtain some computable error bounds of $\mathrm{O}(n^{-1})$ for the chi-squared approximation of transformed chi-squared random variables with n degrees of freedom. In expression (1), $\min_{\lambda} D(\lambda, p)$ denotes the following error bound;

$$\sup_x \left|\mathrm{P}\left(\frac{\chi_p^2 - p}{\sqrt{2p}} < x\right) - \Phi(x)\right| \leq \min_{\lambda} D(\lambda, p) = \mathrm{O}(p^{-1/2}).$$

By the centeral limit theorem, $P((\chi_p^2 - p)/\sqrt{2p} < x)$ converges the normal distribution $\Phi(x)$. In Section 3.2, we modify the result of Ulyanov et al (2006) to get this error bound by two approaches and compare these bounds .

In Section 4, we briefly introduce the extension, which is led by Hall(2005), of properties in non-normal case. A single sample case is treated in Section 4.1. Then the following three conditions are assumed to examine the limiting behavior of a sample $\mathcal{X}(p) = (\boldsymbol{x}_1, \boldsymbol{x}_2 \ldots, \boldsymbol{x}_n)$.

1. The fourth moments of the entries of the data vectors are uniformly bounded.

2. For a constant σ^2,

$$\frac{1}{p}\sum_{j=1}^{p}\mathrm{Var}(x_{ij}) \to \sigma^2 \tag{2}$$

3. The infinite data vector $\boldsymbol{x}_i$ is ρ mixing for functions that are dominated by quadratics, where ρ mixing condition is accurately defined in Appendix;

To be brief, 3rd assumption implies that the correlation between component i and $j = i + r$ gets weak as r increases. In Section 4.2, we extend these properties in two data sets from different distributions. Properties in Section 4.2 are applied for the analysis of discrimination methods. In non-normal case, we need a ρ mixing condition to satisfy properties. This condition is somewhat too strict because the condition is equivalent to have a strong collinearity among variables and the condition also depends on the order of entries, which can be arbitrary.

To research asymptotic properties of the sample covariance matrix in a normal case, Jeongyoun (2007) shows that the same geometric representation hold under a mild assumption on the population eigenvalues. Note that Jeongyoun (2007) considers dual sample covariance $S_D = X^T X/n$ instead of primal sample covariance $S_P = XX^T/n$, where X is $p \times n$ data matrix. it has the same positive eigenvalues as S_P. To show geometric representation for HDLSS data, the following condtions are assumed;

1. The fourth moments of the entries of the data vectors are uniformly bounded.

2. The eigenvalues of Σ_p are sufficiently diffused, in the sense that

$$\frac{\sum_{j=1}^{p}\lambda_j^2}{(\sum_{j=1}^{p}\lambda_j)^2} \to 0 \quad as \quad p \to \infty, \tag{3}$$

   where $\lambda_1 \geq \cdots \geq \lambda_p$ is eigenvalues of a nonnegative definite covariance matrices Σ.

assumption (3) is uesd at a population version of the locally most powerful invariant test statistic for sphericity. In multivariate normal distiributions, the empirical version is the locally most powerful invariant test statistic for sphericity. In Section 5, in addition to a new assumption about cumulants, we extend the idea to non-normal case.

In Secton 6, this new geometric representation is used to analyse the HDLSS performance of support vector machine (SVM). SVM is a new discrimination method proposed by Vapnik, and so on. The origin of SVM is Optimal Separating Hyperplane proposed by Vapnik in the 1960's ,and then in the 1990's, the method was extended to nonlinear discrimination by a kernel and soft margin. SVM is the notable method at the present time. From the point of view of VC-dimension, which was introduced by Vapnik and Chervonenkis, good generalization performance is

guaranteed for SVM in case that the sample size is finite. Here, VC-dimension denotes the one of measures of complex for a function set. And it is known that the idea such that the margin between two groups become maximum is most suitable in the sense that the risk become minimum, and the performance does not depend on the dimension of data. And the performance for HDLSS data is researched by Hall *et,al* (2005). They paid attention to the distance of new data from centroid of simplex. Their result is introduced in Section 6.1.

In Section 6.2, we consider the case of two multivariate standard normal populations $\Pi_1 : N(\boldsymbol{\mu}^{(1)}, I_p)$ and $\Pi_2 : N(\boldsymbol{\mu}^{(2)}, I_p)$ where $\boldsymbol{\mu}^{(i)} = (\mu_1^{(i)}, \ldots, \mu_p^{(i)})$ $(i = 1, 2)$ is the vector of means of the ith population, $i = 1, 2$. $\boldsymbol{\mu}^{(1)}$ and $\boldsymbol{\mu}^{(2)}$ satisfy the condition that

$$\frac{1}{p}\sum_{k=1}^{p}\left\{\mu_k^{(1)} - \mu_k^{(2)}\right\}^2 = \mu \quad (\mu : \text{constant}).$$

Let D_1 and D_2 be difined as the distances of new data X_0 from m-simplex and n-simplex respectively. Then the new data X_0 is classified to Π_1 or Π_2 according as

$$D < 0 \Rightarrow X_0 \in \Pi_1,$$
$$D > 0 \Rightarrow X_0 \in \Pi_2.$$

Here, $D = D_1 - D_2$. The probability of misclassification, if the new data is from Π_1, is approximated following;

$$\Pr\left(\frac{D}{\sqrt{2}} > 0 \middle| X_0 \in \Pi_1\right) \simeq \Phi\left(-\sqrt{\frac{p}{2}}\left(\sqrt{2+\mu^2} - \sqrt{2}\right)\right).$$

The probability of misclassification, if the new data is from Π_1, is also. The interesting consequence of this result is that the probability of misclassification is decreasing, as p is increasing. And if $\mu = 0$, i.e, each population has the same mean, the probability of misclassification is $1/2$, which denotes that discrimination method (SVM) is meaningless.

References

[1] Ahn Jeongyoun, Marron J. S., Muller Keith E., Chi Yueh-Yun. (2007). The high dimension, low sample size geometric representation holds under conditoins.

[2] Hall, P., Marron, J. S., & Neeman, A. (2005). Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society. Series B* **67**, 427-444.

[3] Ulyanov, V. V., Christoph G. and Fujikoshi, Y. (2006). On approximations of transformed chi-squared distributions in statistical applications. *Siberian Mathematical Journal* **47 No.6**, 1154-1166.