MR4488251 62G10 62G07 62G20
Ghosh, Santu [**Ghosh, Santu**[2]] (1-AUGU); **Polansky, Alan M.** (1-NIL-NDM)
Large-scale simultaneous testing using kernel density estimation. (English summary)
Sankhya A **84** (2022), *no. 2,* 808–843.

The main contribution of this article is to demonstrate that when suitably implemented, the two-sample t-test statistic using kernel density estimators (KDEs) can improve the error rate of a normal approximation to its p-value by an order of magnitude. Let $\{\mathbf{X}_j\}_{j=1}^n \in \mathbb{R}^d$ and $\{\mathbf{Y}_j\}_{j=1}^m \in \mathbb{R}^d$ be i.i.d. random samples drawn independently from two populations $\mathbf{X}$ and $\mathbf{Y}$. For notational simplicity, the i-th component of $\mathbf{W} \in \{\mathbf{X}, \mathbf{Y}\}$ is denoted as $\mathbf{W}(i)$ hereinafter. To identify different component-wise features, consider d sets of null and alternative hypotheses

$$H_{0i} : \mu_{\mathbf{X}(i)} = \mu_{\mathbf{Y}(i)} \quad \text{vs} \quad H_{1i} : \mu_{\mathbf{X}(i)} \neq \mu_{\mathbf{Y}(i)}, \quad i = 1, \dots, d,$$

where $\mu_{\mathbf{W}(i)} = \mathbb{E}(\mathbf{W}(i))$.

To construct the KDE-based two-sample t-statistic, for $i = 1, \dots, d$, let

$$\widehat{f}_{\mathbf{X}(i)}(\cdot) := \frac{1}{nh_i} \sum_{j=1}^{n} K\left(\frac{\mathbf{X}_j(i) - \cdot}{h_i}\right) \quad \text{and} \quad \widehat{f}_{\mathbf{Y}(i)}(\cdot) := \frac{1}{mh_i} \sum_{j=1}^{m} K\left(\frac{\mathbf{Y}_j(i) - \cdot}{h_i}\right)$$

be the KDEs of $\mathbf{X}(i)$ and $\mathbf{Y}(i)$, where $K(u) = e^{-u^2/2}/\sqrt{2\pi}$ is the univariate Gaussian kernel, and $h_i > 0$ is the common bandwidth that shrinks toward zero at a certain rate. Also, let $\widehat{F}_{\mathbf{W}(i)}$ be the cumulative distribution function (CDF) implied by the KDE $\widehat{f}_{\mathbf{W}(i)}$. The test statistic for the testing of H_{0i} against H_{1i} is defined as

$$\widetilde{T}_i := \frac{\tilde{\mu}_{\mathbf{X}(i)} - \tilde{\mu}_{\mathbf{Y}(i)}}{\sqrt{\widetilde{S}^2_{\mathbf{X}(i)}/n + \widetilde{S}^2_{\mathbf{Y}(i)}/m}},$$

where $\tilde{\mu}_{\mathbf{W}(i)}$ and $\widetilde{S}^2_{\mathbf{W}(i)}$ are plug-in estimators of $\mu_{\mathbf{W}(i)}$ and $\sigma^2_{\mathbf{W}(i)} = \mathrm{Var}(\mathbf{W}(i))$ based on $\widehat{F}_{\mathbf{W}(i)}$, respectively.

A benefit of employing the Gaussian kernel is that only the second-order cumulant of $\widehat{F}_{\mathbf{W}(i)}$ differs from that of the corresponding empirical CDF $\widehat{G}_{\mathbf{W}(i)}$ (by a margin of h_i^2). It follows that $\widetilde{T}_i$ can be rewritten as

$$\widetilde{T}_i = \frac{\bar{\mu}_{\mathbf{X}(i)} - \bar{\mu}_{\mathbf{Y}(i)}}{\sqrt{S^2_{\mathbf{X}(i)}/n + S^2_{\mathbf{Y}(i)}/m + h_i^2(1/n + 1/m)}},$$

where $\bar{\mu}_{\mathbf{W}(i)}$ and $S^2_{\mathbf{W}(i)}$ are the sample mean and variance of $\mathbf{W}(i)$, respectively. Observe that when $h_i = 0$, $\widetilde{T}_i$ collapses to the conventional two-sample t-statistic

$$T_i := \frac{\bar{\mu}_{\mathbf{X}(i)} - \bar{\mu}_{\mathbf{Y}(i)}}{\sqrt{S^2_{\mathbf{X}(i)}/n + S^2_{\mathbf{Y}(i)}/m}}.$$

Suppose that $\widetilde{T}_i$ takes some value $\tilde{t}_i$. Then, an Edgeworth expansion of the p-value $\mathbb{P}(|\widetilde{T}_i| \geq |\tilde{t}_i|)$ yields

$$\mathbb{P}\big(|\widetilde{T}_i| \geq |\tilde{t}_i|\big) = \mathbb{P}(|Z| \geq |\tilde{t}_i|) - C_1 h_i^2 - C_2 N^{-1} + O(N^{-\tau}), \quad N = n + m \to \infty,$$

where Z is a standard normal random variable, $C_1 > 0$ and C_2 are constants that depend on population moments of $\mathbf{W}(i)$, the standard normal density $\phi(\cdot)$ and CDF $\Phi(\cdot)$, and $\tau = \min\{2k + 1/2, 3/2\}$ for some $k > 0$ satisfying $h_i^2 = O(N^{-k})$. It follows that if $C_2 < 0$, then putting $h_i^2 = -(C_2/C_1)N^{-1}$ leads to

$$\mathbb{P}\big(|\widetilde{T}_i| \geq |\tilde{t}_i|\big) = 2\big(1 - \Phi(|\tilde{t}_i|)\big) + O(N^{-3/2}), \quad N \to \infty.$$

On the other hand, by an Edgeworth expansion of the p-value for the conventional two-sample t-statistic,

$$\mathbb{P}\big(|T_i| \geq |\tilde{t}_i|\big) = 2\big(1 - \Phi(|\tilde{t}_i|)\big) + O(N^{-1}), \quad N \to \infty.$$

Therefore, the two-sample t-statistic constructed by a smoothed version of the empirical CDF can reduce the normal approximation error by an order of magnitude.

This article establishes that the $O(N^{-3/2})$ approximation error can be attained even after all population moments are replaced by their sample counterparts. It is also demonstrated that when $C_2 > 0$, the KDE-based t-statistic using properly rescaled observations can retrieve the $O(N^{-3/2})$ rate. Furthermore, it is shown that as long as the total number of tests d diverges but $d = o(N^{3/2})$, the KDE-based test can control the false discovery rate.

Masayuki Hirukawa