Journal of Multivariate Analysis 162 (2017) 122–133

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Comparison of a large number of regression curves

Guanghui Wang, Zhaojun Wang, Changliang Zou *

Institute of Statistics and LPMC, Nankai University, China

ARTICLE INFO

Article history:
Received 29 March 2017
Available online 28 September 2017

AMS 2010 subject classifications:
62H15
62G10

Keywords:
Asymptotic normality
Comparison of regression curves
High-dimensional data analysis
Nonparametric inference
U-statistics

ABSTRACT

We revisit the classical statistical inference problem of comparing regression curves. Traditional methods assume that the number of curves is small and fixed, while the sample size on which each curve is based tends to infinity. In contrast, we consider the case where the number of curves tends to infinity and the sample sizes are bounded by a common value. Our test is motivated by the fact that two Borel measurable functions are equivalent if and only if their Fourier transforms are identical (Bierens, 1994). An unbiased statistic is then proposed to avoid noise accumulation in a high-dimensional context. The asymptotic null distribution of the test statistic is derived and its power is studied via simulation. An illustration involving cholesterol data is provided.

1. Introduction

High-dimensional data analysis has gained in popularity over the last few years, as observations involving a large number of variables now arise in a wide range of applications, e.g., in genomics, signal processing, and finance. Traditional statistical inference methods may not work in the high-dimensional regime, mainly due to noise accumulation and spurious correlations. This challenge calls for new statistical tests to deal with high-dimensional data sets. See [1,4,5,10,25] for inferences on high-dimensional mean vectors, [6,14,21,31] for tests for high-dimensional covariance structures, and [11,30] for high-dimensional regression coefficient tests.

In this work, we consider the problem of testing the equality of p regression curves of unknown functional form. In the classical literature, p is assumed to be small and fixed, and asymptotic analysis proceeds by letting the sample size on which each curve is based tend to infinity. In contrast, we consider a setting where the number of curves tends to infinity. This testing problem arises in the framework of high-dimensional statistical inference mentioned above. The related issue which consists of testing the equality of a large number of densities was recently discussed in [29].

Suppose p regression curves are observed independently and, for $i \in \{1, \ldots, p\}$, the ith curve consists of n_i random observations $(X_{i1}, Y_{i1}), \ldots, (X_{in_i}, Y_{in_i})$. Further assume that the response Y_{ij} is related to the d-dimensional covariate X_{ij} according to a nonparametric regression model, i.e.,

$$\forall_{i\in\{1,\ldots,p\}}\ \forall_{j\in\{1,\ldots,n_i\}} \quad Y_{ij} = m_i(X_{ij}) + \epsilon_{ij},$$

where $m_1, \ldots, m_p$ are unknown regression functions and the ϵ_{ij}s are random errors satisfying $\mathrm{E}(\epsilon_{ij} \mid X_{ij}) = 0$. We are interested in justifying the equality of these p curves under the assumption that all the X_{ij}s are independent and identically distributed (iid) as X. The null hypothesis to be tested is thus

$$\mathcal{H}_0 : \Pr\{m_1(X) = \cdots = m_p(X)\} = 1. \tag{1}$$

* Corresponding author.
E-mail address: nk.chlzou@gmail.com (C. Zou).

https://doi.org/10.1016/j.jmva.2017.09.003
0047-259X/

We focus on the case where many curves are based on relatively small samples.

The comparison of two or more regression curves has been the object of much work; see Section 7 of [12] for a recent review. For nonparametric models, most approaches rely on estimators of all regression curves using smoothing techniques. Generally, there are two ways to construct a test statistic. The first approach relies on such estimators directly, either by contrasting all individual estimators with the pooled one, or by performing pairwise comparisons; see, e.g., [8,28]. The second approach is based on the distribution of the residuals or some of their characteristics. For instance, Dette and Neumeyer [8] based their test statistic on a comparison of the variance of a pooled estimator with those of its individual components. Pardo-Fernández et al. [18] compared empirical cumulative distribution functions of the residuals via Kolmogorov–Smirnov and Cramér–von Mises type statistics while Pardo-Fernández et al. [17] considered empirical estimators of characteristic functions of the residuals. Neumeyer and Dette [16] also suggested a test based on the comparison of marked empirical processes of the residuals. Note that the above mentioned tests all depend on nonparametric estimators of regression curves. In contrast, Delgado [7] used a weighted empirical process that avoids fitting regression curves in specific cases.

It is widely acknowledged that most existing test statistics converge to their weak limit at a very slow rate. Furthermore, the asymptotic distribution of some tests involves features of the underlying data generating process and thus can only be used in conjunction with re-sampling calibration procedures such as the bootstrap. In high-dimensional situations where $p \to \infty$, classical asymptotic results no longer hold. An obvious limitation is that it is often assumed that each ratio $n_i/(n_1 + \cdots + n_p)$ converges to a bounded positive constant. Besides, high-dimensionality also brings non-negligible bias accumulation when smoothing-based estimators are used. Moreover, re-sampling procedures are undesirable in that they require extensive computational efforts.

In this work, we propose a new test tailored for high-dimensional settings. The test statistic avoids explicitly fitting the regression curves and overcomes the noise accumulation problem. Its null distribution is asymptotically normal and free of any unknown quantities. Hence, the asymptotic test is easy to use. Simulation studies show that it performs well for configurations of dimension and sample sizes commonly encountered in practice.

The remainder of this article is organized as follows. In Section 2, we present our test statistic and its theoretical properties. Section 3 provides some simulation results and a real data example as an illustration. Section 4 concludes with some remarks, and proofs are detailed in Appendix.

2. Methodology

2.1. Comparison of two regression curves

To motivate our test statistic, we first consider the comparison of two regression curves, i.e., $p = 2$. Our main idea stems from a result due to Bierens (see Theorem 3.1.1 of [3]), which plays a critical role in the construction of our test statistic. This result is recalled below.

Lemma 1. *Let m_1 and m_2 be Borel measurable real-valued functions defined on $\mathbb{R}^d$. Let X be a random vector in $\mathbb{R}^d$ such that $\mathrm{E}|m_1(X)| < \infty$ and $\mathrm{E}|m_2(X)| < \infty$. For arbitrary $t \in \mathbb{R}^d$, let $\phi_1(t) = \mathrm{E}\{m_1(X)\exp(it^\top X)\}$ and $\phi_2(t) = \mathrm{E}\{m_2(X)\exp(it^\top X)\}$. Then $\Pr\{m_1(X) = m_2(X)\} = 1$ if and only if $\phi_1(t) = \phi_2(t)$ for all $t \in \mathbb{R}^d$.*

For a complex-valued function f defined on $\mathbb{R}^d$, the complex conjugate of f is denoted by $\bar{f}$ and $\|f\|^2 = f\bar{f}$. We define the $\|\cdot\|_w$-norm as $\|f\|_w^2 = \int_{\mathbb{R}^d} \|f(t)\|^2 w(t)\mathrm{d}t$, where w is a positive weight function for which the integral exists. By Lemma 1, the equality of two regression functions, in the sense that $\Pr\{m_1(X) = m_2(X)\} = 1$, is equivalent to $\Delta = \|\phi_1 - \phi_2\|_w^2 = 0$.

Under the assumption that $\mathrm{E}|m_i(X)| < \infty$ for $i \in \{1, 2\}$, it follows from Fubini's Theorem that

$$\Delta = \mathrm{E}\{m_1(X)m_1(X')K(X - X')\} + \mathrm{E}\{m_2(X)m_2(X')K(X - X')\} - 2\mathrm{E}\{m_1(X)m_2(X')K(X - X')\},$$

where X' is a copy of X and $K(u) = \int_{\mathbb{R}^d} \cos(t^\top u)w(t)\mathrm{d}t$. Therefore, if we further assume that $\mathrm{E}(\varepsilon_{ij}\varepsilon_{i\ell} \mid X_{ij}, X_{i\ell}) = 0$ for every $i \in \{1, \ldots, p\}$ and $j, \ell \in \{1, \ldots, n_i\}$ (or assume (A2) below holds), then we can use an unbiased estimate of Δ to formulate a test statistic, viz.

$$\hat{\Delta} = \frac{1}{P_{n_1}^2}\sum_{j=1}^{n_1}\sum_{\ell=1,\ell\neq j}^{n_1} Y_{1j}Y_{1\ell}K(X_{1j} - X_{1\ell}) + \frac{1}{P_{n_2}^2}\sum_{j=1}^{n_2}\sum_{\ell=1,\ell\neq j}^{n_2} Y_{2j}Y_{2\ell}K(X_{2j} - X_{2\ell}) - \frac{2}{n_1 n_2}\sum_{j=1}^{n_1}\sum_{\ell=1}^{n_2} Y_{1j}Y_{2\ell}K(X_{1j} - X_{2\ell}),$$

with an appropriate choice of weight function w, where $P_n^r = n(n-1)\cdots(n-r+1)$.

2.2. Comparison of a large number of curves

In our asymptotic scheme, we assume that $p \to \infty$ while all sample sizes are kept fixed. We propose to base a test of the null hypothesis (1) on the statistic

$$T_p = \frac{1}{p(p-1)}\sum_{i=1}^{p}\sum_{k=1,k\neq i}^{p} \hat{\Delta}_{ik},$$