Statistica Sinica **33** (2023), 2787-2808
doi:https://doi.org/10.5705/ss.202021.0285

MEASURING, TESTING, AND IDENTIFYING HETEROGENEITY OF LARGE PARALLEL DATASETS

Liuhua Peng, Guanghui Wang and Changliang Zou

The University of Melbourne, East China Normal University and Nankai University

Abstract: When working with large parallel data sets, it is necessary to check whether they are collected from different regression models before conducting further modeling, estimation, and inference. We propose a novel metric for such heterogeneity based on a projection strategy. We then use this metric to a new fully data-driven test for the equivalence of a large number of unknown regression models. We also construct the asymptotic normality of the proposed test, and apply the test to identify outlying data sets with regression models that deviate from the majority. Extensive numerical studies demonstrate that our methods perform satisfactorily.

Key words and phrases: Heterogeneity, outlier detection, parallel data sets, projections, U-statistics

1. Introduction

Large parallel data sets are becoming increasingly common in scientific fields such as bioinformatics, computer science, mechanical engineering, and economics, owing to the advancement of data collection techniques and devices. Thus, it is necessary to measure and test the homogeneity of such data sets before further data processing can occur. For example, in experimental studies, before integrating the data, it is necessary to check the extend to which the underlying distributions or models of the parallel data sets collected under different conditions or treatments differ (Borgwardt et al. (2006); Tang and Song (2016)). Moreover, even in the same treatment group, one needs to determine whether individuals share the same model before group-specific modeling (Ke, Li and Zhang (2016); Vogt and Linton (2017)). Therefore, we require a formal test to provide uncertainty quantification on data homogeneity. In such scenarios, we also need to accurately estimate the overall pattern using a post-test diagnostic that can identify outlying groups or individuals.

Corresponding author: Changliang Zou, School of Statistics and Data Science, LPMC, LEBPS, and KLMDASR, Nankai University, Tianjin 300071, China. E-mail:nk.chlzou@gmail.com.

Heterogeneity in data sets may come from various forms of variability across parallel studies. In this study, we focus on data sets of paired measurements of responses and covariates. Thus, testing the heterogeneity across multiple data sets essentially means checking whether the data sets share a common regression function. Suppose we have p parallel data sets, and the kth data set consists of n_k paired members $\{(Y_{ki}, X_{ki}),\ i = 1, \ldots, n_k\}$, for $k = 1, \ldots, p$, where Y_{ki} is a scalar response and X_{ki} denotes the associated d-dimensional covariates. We model the data as follows:

$$Y_{ki} = m_k(X_{ki}) + \varepsilon_{ki}, \quad i = 1, \ldots, n_k;\ k = 1, \ldots, p, \tag{1.1}$$

where m_k is a regression function and ε_{ki} denotes random noise satisfying $E(\varepsilon_{ki} \mid X_{ki}) = 0$, almost surely (a.s.). For $k = 1, \ldots, p$, let $\mathcal{Z}_k = \{Z_{k1}, \ldots, Z_{kn_k}\}$, with $Z_{ki} = (Y_{ki}, X_{ki})$, for $i = 1, \ldots, n_k$. We assume that the p data sets are collected independently, that is, $\mathcal{Z}_k$ is independent of $\mathcal{Z}_\ell$, for any (k, ℓ), such that $1 \le k \ne \ell \le p$, and the paired covariates and noise $(X_{ki}, \varepsilon_{ki})$ are independent and identically distributed (i.i.d.) as (X, ε), for $i = 1, \ldots, n_k$ and $k = 1, \ldots, p$. Testing the homogeneity of the p data sets can thus be formulated as testing the following null hypothesis:

$$H_0 :\ \mathrm{P}\{m_1(X) = \cdots = m_p(X)\} = 1. \tag{1.2}$$

Here, we require no knowledge of the structured functional forms of $\{m_k\}_{k=1}^p$ and treat them as nonparametric. We consider a large-scale setup in the sense that the number of parallel data sets $p \to \infty$. Once we have rejected the null hypothesis H_0 in (1.2), we need to identify the outlying data sets that possess different regression functions from the majority.

In fact, testing for heterogeneity among several (usually two) regression functions has been widely researched. It is natural to perform a classical analysis of variance if all regression functions are restricted to some parametric forms, such as linear models. More recently, testing procedures with no restrictions on the parametric structures have been proposed; see, for instance, Neumeyer and Dette (2001), Neumeyer and Dette (2003), Pardo-Fernández, Van Keilegom and González-Manteiga (2007), Srihera and Stute (2010), Koul and Li (2020), and Cai and Wang (2021), among others. See also Section 7 in González-Manteiga and Crujeiras (2013) for a brief overview. Our context differs from those of past studies, because the number of regression functions can be very large, that is, asymptotically speaking, $p \to \infty$. For example, treatments in experimental studies may be indicated by a categorical variable that takes a value from a large

collection of candidates, which naturally results in large parallel data sets. As another example, econometricians investigating panel data may consider whether the data is poolable over time (Baltagi, Hidalgo and Li (1996); Barras, Scaillet and Wermers (2010)), where the number of periods could be very large. Furthermore, in studies of longitudinal data in which we need to estimate the overall pattern from a large number of subjects (Chiou and Li (2007); Qiu and Xiang (2014)), we need to know whether there are significant differences between individual subjects. Such applications motivate the need for heterogeneity tests for large parallel data sets.

In this paper, we first propose a model-free metric for the departure of two regression functions, based on a projection approach. We then use this metric to create a test statistic for testing for heterogeneity in large parallel data sets. The proposed procedure makes no parametric assumptions on the regression functions, and does not require direct estimations of the nonparametric models. Compared with prior works, our approach is free of nuisance parameters, making it particularly useful for the case of large p. We construct the asymptotic properties of the test statistic when the sample sizes of all data sets diverge. We also propose a bootstrap remedy to mimic the null distribution in cases of conservative sizes in finite-sample performance, and establish its asymptotic validity and consistency. In addition, we apply the proposed heterogeneity testing procedure to identify outlying data sets. We offer a new perspective on outlier detection by performing a sequence of heterogeneity tests in a large-scale manner. We show that the proposed method performs satisfactorily in terms of correctly detecting outlying data sets, while controlling the false positive rate well.

A closely related work is that of Wang, Wang and Zou (2017), who study the testing aspect. Our proposed test statistic is similar to theirs. For example, both get rid of nonparametric estimations of the underlying regression models, and both are related to U-statistics. Nevertheless, our study contributes to the literature in three ways. First, the proposed projection-based metric of heterogeneity is new. Second, our test statistic is free of any tuning parameters. In contrast, the test statistic of Wang, Wang and Zou (2017) involves an additional nuisance parameter that needs to be specified in an elaborate manner. Moreover, they treat the nuisance parameter as fixed, and provide no theoretical guarantees if a data-dependent estimate is plugged-in. Third, our numerical studies reveal that their procedure is sometimes conservative for large sample sizes. This issue is mitigated by our method by our use of the proposed bootstrap calibrations. Furthermore, the calibrations are based on an elaborate analysis of the asymptotic behavior of the proposed test statistics, which makes the theoretical derivations

much more involved. In addition, we present a novel outlier detection scheme based on the proposed testing method.

The remainder of the paper is organized as follows. In Section 2, we develop the heterogeneity measure and derive our test statistic. In Section 3, we investigate the proposed method from a theoretical viewpoint, and in Section 4, we apply the proposed measure and testing procedure to identify outlying data sets. In Section 5, we present several examples based on simulated and real data to evaluate the numerical performance of the proposed method. We conclude the paper in Section 6. Proofs of the theoretical results and additional numerical results are deferred to the Supplementary Material.

2. A New Test Statistic for Heterogeneity Checking

In this section, we introduce a novel measure for heterogeneity that quantifies the difference between two regression functions, and then propose our test statistic based on the heterogeneity measure.

2.1. A novel measure for heterogeneity

Our testing procedure for (1.2) is motivated by a novel measure that characterizes the equivalence or departure of two regression functions based on projections. It induces a testing procedure that avoids directly estimating the regression functions by, for example, using kernel smoothing methods, and can be applied to covariates with moderate or even large dimension. The key observation is provided in Lemma 1.

Lemma 1. *Suppose $E|m_k(X)| < \infty$, for $k = 1, 2$. A necessary and sufficient condition for $m_1(X) = m_2(X)$ a.s. to hold is that*

$$E\{m_1(X)\mathbb{I}(\beta^{\mathrm{T}}X \le u)\} = E\{m_2(X)\mathbb{I}(\beta^{\mathrm{T}}X \le u)\}$$

holds almost everywhere $(\beta, u) \in \mathbb{S}^{d-1} \times \mathbb{R}$, where $\mathbb{S}^{d-1} = \{\beta \in \mathbb{R}^d : \|\beta\| = 1\}$ is the $(d-1)$-dimensional unit sphere.

Projection-based characterizations are often used to avoid the curse of dimensionality in the literature on goodness-of-fit testing (Escanciano (2006); Lavergne and Patilea (2008); Xia (2009); Patilea, Sánchez-Sellero and Saumard (2016); Cuesta-Albertos et al. (2019)). Lemma 1 offers a two-sample version. To aggregate all information over $(\beta, u) \in \mathbb{S}^{d-1} \times \mathbb{R}$, we propose the following projection-averaging (PA)-based measure for the equivalence or departure of two regression functions:

$$\mathrm{PA}(m_1, m_2) = \int_{\beta \in \mathbb{S}^{d-1}} \int_{u \in \mathbb{R}} \left[E\left\{ m_1(X) \mathbb{I}(\beta^{\mathrm{T}} X \le u) \right\} - E\left\{ m_2(X) \mathbb{I}(\beta^{\mathrm{T}} X \le u) \right\} \right]^2 \mathrm{d}F_{\beta^{\mathrm{T}} X}(u) \mathrm{d}\lambda_{\mathbb{S}^{d-1}}(\beta), \tag{2.1}$$

where $F_{\beta^{\mathrm{T}} X}$ is the cumulative distribution function of the projected covariate $\beta^{\mathrm{T}} X$, and $\lambda_{\mathbb{S}^{d-1}}$ represents the uniform probability measure on $\mathbb{S}^{d-1}$. It is obvious that $\mathrm{PA}(m_1, m_2) \ge 0$ and $\mathrm{PA}(m_1, m_2) = 0$ if and only if $\mathrm{P}(m_1(X) = m_2(X)) = 1$. PA techniques similar to (2.1) are used by, among others, Escanciano (2006), Zhu et al. (2017), and Kim, Balakrishnan and Wasserman (2020) for different inferential purposes. One advantage of the PA approach is that it involves a closed-form expression, as shown in Proposition 1.

Proposition 1. *Suppose $E|m_k(X)| < \infty$, for $k = 1, 2$. Let X', X'' be i.i.d. copies of X. Then,*

$$\begin{aligned} \mathrm{PA}(m_1, m_2) =& E\{m_1(X) m_1(X') K(X, X', X'')\} \\ &+ E\{m_2(X) m_2(X') K(X, X', X'')\} \\ &- 2E\{m_1(X) m_2(X') K(X, X', X'')\}, \end{aligned} \tag{2.2}$$

where $K(X, X', X'') = 2^{-1} - (2\pi)^{-1} \arccos[\{(X - X'')^{\mathrm{T}}(X' - X'')\}/(\|X - X''\| \, \|X' - X''\|)]$ if $X \ne X''$ and $X' \ne X''$. If $X = X'' \ne X'$ or $X' = X'' \ne X$, then $K(X, X', X'') = 1/2$, and if $X = X' = X''$, then $K(X, X', X'') = 1$.

Moreover, as we show later, the benefit of using (2.1) or (2.2) appears clearer by observing that $E(Y_{ki} \mid X_{ki}) = m_k(X_{ki})$ a.s., for $i = 1, \ldots, n_k$ and $k = 1, \ldots, p$, which indicates how we construct the test statistic for H_0 without any kernel estimations of the regression functions.

2.2. The test statistic

The idea in the two-sample scenario can be generalized to the context of multiple samples. To this end, we introduce

$$\theta_p = \{p(p-1)\}^{-1} \sum_{1 \le k \ne \ell \le p} \mathrm{PA}(m_k, m_\ell)$$

to serve as the heterogeneity measure of $m_1, \ldots, m_p$. By Lemma 1, the p regression functions m_k are equivalent (i.e., H_0 holds) if and only if $\theta_p = 0$. To form a test statistic, we need an estimate of θ_p or estimates of all pairwise discrepancy measures $\mathrm{PA}(m_k, m_\ell)$.

For any $x, x' \in \mathbb{R}^d$, define

$$\mathcal{K}(x, x') = E\{K(X, X', X'') \mid X = x, X' = x'\},$$

where K is defined in Proposition 1. Suppose, for the moment, that $\mathcal{K}$ is known. Recall that $E(Y_{ki} \mid X_{ki}) = m_k(X_{ki})$, for $i = 1, \ldots, n_k$ and $k = 1, \ldots, p$. For each pair (k, ℓ) such that $1 \le k \ne \ell \le p$, we propose an unbiased estimate of $\mathrm{PA}(m_k, m_\ell)$ as

$$\begin{aligned}\widetilde{\mathrm{PA}}_{k,\ell} =& \{n_k(n_k-1)\}^{-1} \sum_{1\le i_1\ne i_2\le n_k} Y_{ki_1}Y_{ki_2}\mathcal{K}(X_{ki_1}, X_{ki_2}) \\ &+ \{n_\ell(n_\ell-1)\}^{-1} \sum_{1\le j_1\ne j_2\le n_\ell} Y_{\ell j_1}Y_{\ell j_2}\mathcal{K}(X_{\ell j_1}, X_{\ell j_2}) \\ &- 2n_k^{-1}n_\ell^{-1}\sum_{i=1}^{n_k}\sum_{j=1}^{n_\ell} Y_{ki}Y_{\ell j}\mathcal{K}(X_{ki}, X_{\ell j}).\end{aligned}$$

Then, an unbiased estimate of θ_p is

$$U_p = \{p(p-1)\}^{-1} \sum_{1\le k\ne\ell\le p} \widetilde{\mathrm{PA}}_{k,\ell}. \tag{2.3}$$

However, $\mathcal{K}$ is difficult to specify in a closed form, and may even be unknown. Hence, we need good approximations of $\mathcal{K}$. To this end, we propose using the following moment estimates. For each pair (k, ℓ) such that $1 \le k < \ell \le p$, define

$$\widehat{\mathcal{K}}_{-k\ell}(x, x') = n_r^{-1}\sum_{s=1}^{n_r} K(x, x', X_{rs}),$$

where $r = \ell + 1$ if $\ell < p$, and $r = 1 + \mathbb{I}(k = 1)$ if $\ell = p$. For $1 \le \ell < k \le p$, we define $\widehat{\mathcal{K}}_{-k\ell}(x, x') = \widehat{\mathcal{K}}_{-\ell k}(x, x')$. We estimate $\mathrm{PA}(m_k, m_\ell)$, again without bias, by

$$\begin{aligned}\widehat{\mathrm{PA}}_{k,\ell} =& \{n_k(n_k-1)\}^{-1} \sum_{1\le i_1\ne i_2\le n_k} Y_{ki_1}Y_{ki_2}\widehat{\mathcal{K}}_{-k\ell}(X_{ki_1}, X_{ki_2}) \\ &+ \{n_\ell(n_\ell-1)\}^{-1} \sum_{1\le j_1\ne j_2\le n_\ell} Y_{\ell j_1}Y_{\ell j_2}\widehat{\mathcal{K}}_{-k\ell}(X_{\ell j_1}, X_{\ell j_2}) \\ &- 2n_k^{-1}n_\ell^{-1}\sum_{i=1}^{n_k}\sum_{j=1}^{n_\ell} Y_{ki}Y_{\ell j}\widehat{\mathcal{K}}_{-k\ell}(X_{ki}, X_{\ell j}).\end{aligned}$$

This motivates the test statistic

$$T_p = \{p(p-1)\}^{-1} \sum_{1\le k\neq \ell\le p} \widehat{\mathrm{PA}}_{k,\ell}. \tag{2.4}$$

Note that T_p is free of any tuning parameters. Because T_p is also unbiased for θ_p, large values of T_p indicate that we should rejeect H_0 that these p data sets are homogeneous.

3. Theoretical Properties

In this section, we establish the asymptotic null distribution of the test statistic T_p, and use the jackknife method to estimate its asymptotic variance in order to implement the test. In addition, we propose a bootstrap procedure to calibrate the critical value of the test. Finally, we study the asymptotic power of the test under a finite-component mixture model.

3.1. Asymptotic null distribution

Our discussion is under a large-scale setup in the sense that the number of parallel data sets $p \to \infty$. Recall that $\mathcal{Z}_k = \{(Y_{k1}, X_{k1}), \ldots, (Y_{kn_k}, X_{kn_k})\}$, for $k = 1, \ldots, p$. We first make the following assumptions.

Assumption 1. (Model). *$\mathcal{Z}_k$ is independent of $\mathcal{Z}_\ell$, for any $1 \le k \neq \ell \le p$, and $(X_{ki}, \varepsilon_{ki})$ are i.i.d. as (X, ε), for $i = 1, \ldots, n_k$ and $k = 1, \ldots, p$. In addition, there exists a constant $\delta > 0$ such that $E\left\{|m_k(X)|^{2+\delta}\right\} < \infty$, for $k = 1, \ldots, p$, and $E(|\varepsilon|^{2+\delta}) < \infty$.*

Assumption 2. (Number of data sets and sample sizes). *Suppose $p \to \infty$ and $n_k \to \infty$ as $p \to \infty$, for $k = 1, \ldots, p$; in addition, there exist positive constants c_1 and c_2 such that $c_1 \le \inf_{1\le k,\ell\le p} n_k/n_\ell \le \sup_{1\le k,\ell\le p} n_k/n_\ell \le c_2$.*

With $z_i = (y_i, x_i)$, for $i = 1, 2$, define

$$\begin{aligned} h^{(2,0)}(z_1, z_2) = {} & y_1 y_2 \mathcal{K}(x_1, x_2) + E\left\{m_1(X) m_1(X') \mathcal{K}(X, X')\right\} \\ & - y_1 E\left\{m_1(X) \mathcal{K}(x_1, X)\right\} - y_2 E\left\{m_1(X) \mathcal{K}(x_2, X)\right\}. \end{aligned} \tag{3.1}$$

Denote $\sigma_{p,0}^2 = 8p^{-1} \sum_{k=1}^{p} \{n_k(n_k-1)\}^{-1} E\big[\left\{h^{(2,0)}(Z_{k1}, Z_{k2})\right\}^2\big]$. The next theorem gives the asymptotic distribution of T_p under the null hypothesis.

Theorem 1. *Suppose Assumptions 1–2 hold. Under H_0, $p^{1/2} T_p/\sigma_{p,0} \to N(0,1)$ in distribution as $p \to \infty$, and $\sigma_{p,0}^2 = O(n_1^{-2})$.*

By Theorem 1, the convergence rate of T_p to its population counterpart, that is, the heterogeneity measure θ_p, is of the order $p^{-1/2}n_1^{-1}$. The proof of Theorem 1 is given in the Supplementary Material. The key idea is to use Hoeffding's decomposition (Hoeffding (1948)) of U-statistics. In fact, T_p is asymptotically equivalent to U_p, a U-statistic of degree two with a kernel that may depend on p and n_k, under H_0. Moreover, the kernel of U_p is a two-sample U-statistic of degree $(2, 2)$ on its own. Following the proof of Theorem 1 in the Supplementary Material, we have $T_p = 2p^{-1}\sum_{k=1}^{p}\Xi_{p,k} + O_p(p^{-1}n_1^{-1})$ under H_0, where

$$\Xi_{p,k} = \{n_k(n_k-1)\}^{-1}\sum_{1\le i_1\ne i_2\le n_k} h^{(2,0)}(Z_{ki_1}, Z_{ki_2}).$$

In other words, T_p is asymptotically equivalent to an average of a sequence of independent, but not identically distributed, random variables. We can then construct the asymptotic normality of T_p using a central limit theorem for double arrays of random variables.

3.2. A jackknife estimate of variance

It remains to estimate $\sigma_{p,0}^2$ in order to fulfill the testing procedure based on T_p. Instead of directly estimating $E\big[\big\{h^{(2,0)}(Z_{k1}, Z_{k2})\big\}^2\big]$, we use the jackknife estimator of the variance of U-statistics of degree two (Sen (1977)). Denote $\hat{U}_{p,k} = (p-1)^{-1}\sum_{\substack{\ell=1\\ \ell\ne k}}^{p}\widehat{\mathrm{PA}}_{k,\ell}$. We estimate $\sigma_{p,0}^2$ by

$$\hat{\sigma}_{p,0}^2 = 4(p-1)(p-2)^{-2}\sum_{k=1}^{p}\Big(\hat{U}_{p,k} - T_p\Big)^2.$$

In fact, under H_0, $\hat{U}_{p,k}$ can be viewed as an approximation to $E(\widehat{\mathrm{PA}}_{k,\ell} \mid \mathcal{Z}_k)$, or essentially to $\Xi_{p,k}$. Hence, $\hat{\sigma}_{p,0}^2$ is simply the sample variance based on $\{2\hat{U}_{p,1}, \ldots, 2\hat{U}_{p,p}\}$ up to a negligible factor $(p-1)^2(p-2)^{-2}$. The following proposition guarantees the consistency of $\hat{\sigma}_{p,0}^2$.

Proposition 2. *Suppose Assumptions 1–2 hold. Under H_0, $\hat{\sigma}_{p,0}^2/\sigma_{p,0}^2 \to 1$ in probability as $p \to \infty$.*

The null hypothesis is rejected when $p^{1/2}T_p/\hat{\sigma}_{p,0} > z_\alpha$, where z_α is the upper α-quantile of $N(0, 1)$. Slutsky's theorem combined with Theorem 1 and Proposition 2 ensures that $p^{1/2}T_p/\hat{\sigma}_{p,0}$ is asymptotically standard normal, and thus is the size of the test at the significance level α.

3.3. Bootstrap calibrations

Our testing procedure with the jackknife variance estimate $\hat{\sigma}_{p,0}^2$ can be applied directly, free of any nuisance parameters. However, our simulation studies indicate that in some finite-sample situations, this results in conservative sizes. Noting that, under H_0, $T_p = 2p^{-1}\sum_{k=1}^{p}\Xi_{p,k} + O_p(p^{-1}n_1^{-1})$, where $\{\Xi_{p,k}\}_{k=1}^{p}$ are independent, but not identically distributed, we propose using a Studentized bootstrap procedure to calibrate the critical value of the test.

We use $\hat{U}_{p,k} = (p-1)^{-1}\sum_{\substack{\ell=1\\ \ell\neq k}}^{p}\widehat{\mathrm{PA}}_{k,\ell}$ to approximate $\Xi_{p,k}$, for $k = 1, \ldots, p$. Proposition S.1 in the Supplementary Material indicates that $\hat{U}_{p,k}$ is consistent for $\Xi_{p,k}$ under the null hypothesis. Let $F_{p,U}$ be the empirical distribution of $\{\hat{U}_{p,k}\}_{k=1}^{p}$. We randomly draw $\hat{U}_{p,1}^*, \ldots, \hat{U}_{p,p}^*$ from $F_{p,U}$. Because $E(\hat{U}_{p,1}^* \mid F_{p,U}) = p^{-1}\sum_{k=1}^{p}\hat{U}_{p,k} = T_p$, a Studentized bootstrap version of the test statistic is $p^{1/2}(T_p^* - T_p)/S_{p,U}^*$, where $T_p^* = p^{-1}\sum_{k=1}^{p}\hat{U}_{p,k}^*$ and $S_{p,U}^{*2} = p^{-1}\sum_{k=1}^{p}(\hat{U}_{p,k}^* - T_p^*)^2$. Then, the distribution of $p^{1/2}(T_p^* - T_p)/S_{p,U}^*$ conditional on $F_{p,U}$ is used to estimate that of $p^{1/2}T_p/\hat{\sigma}_{p,0}$ under the null hypothesis. The following theorem establishes the theoretical support for the bootstrap method.

Theorem 2. *Suppose Assumptions* 1–2 *hold. Under* H_0, *as* $p \to \infty$,

$$\sup_{x\in\mathbb{R}}\left|\mathrm{P}\left\{\frac{p^{1/2}(T_p^* - T_p)}{S_{p,U}^*} \le x \,\middle|\, F_{p,U}\right\} - \mathrm{P}\left(\frac{p^{1/2}T_p}{\hat{\sigma}_{p,0}} \le x\right)\right| = o_p(1).$$

Let z_α^* be the upper α-quantile of the conditional distribution of $p^{1/2}(T_p^* - T_p)/S_{p,U}^*$. Then, the null hypothesis is rejected when $p^{1/2}T_p/S_{p,U} > z_\alpha^*$. The value of z_α^* can be approximated using a Monte Carlo simulation by repeatedly sampling from $F_{p,U}$ a large number of times. Our numerical studies indicate that the test based on the bootstrap calibration enjoys better performance in term of size.

Remark 1. The computational complexity of the jackknife or bootstrap-based testing procedures lies mainly in the calculation of the test statistic T_p, which is $O(p^2n_1^3)$ if all n_k are $O(n_1)$. Once we have $\{\hat{U}_{p,k}\}_{k=1}^{p}$ in the computation of T_p, both procedures can be performed in an efficient manner, with no additional refitting required. However, computing the test statistic T_p is itself sometimes computationally expensive, especially for very large p and n_k, which is unavoidable at this moment, given how we construct the test statistic.

3.4. Asymptotic power

To determine the asymptotic power of our proposed test with a bootstrap calibration, we introduce the following finite-component mixture model.

Assumption 3. (Clusters). *There are finite (say L) different regression functions $\{m_1^*, \ldots, m_L^*\}$ such that $\mathrm{P}\{m_{\ell_1}^*(X) = m_{\ell_2}^*(X)\} < 1$, for any $1 \le \ell_1 \ne \ell_2 \le L$, and the underlying regression functions $\{m_1, \ldots, m_p\}$ fall into L clusters, such that in the ℓth cluster, the regression functions are identical to m_ℓ^*, for $\ell = 1, \ldots, L$.*

Theorem 3. *Suppose Assumptions* 1–3 *hold. Let p_ℓ be the number of regression functions in the ℓth cluster, for $\ell = 1, \ldots, L$, and $p_{(1)}$ and $p_{(2)}$ be the largest two values of $\{p_\ell\}_{\ell=1}^L$, and assume that $p_{(2)} \to \infty$ and $p_{(2)}/(p^{1/2}n_1^{-1/2}) \to \infty$. Then, the power of the test with a bootstrap calibration converges to one as $p \to \infty$.*

Theorem 3 shows that as long as the number of regression functions in the second largest cluster satisfies $p_{(2)} \to \infty$ and $p_{(2)}/(p^{1/2}n_1^{-1/2}) \to \infty$, our proposed test is consistent against the alternative with a fixed number of clusters. If $p = O(n_1)$, it suffices to require that $p_{(2)} \to \infty$ to achieve consistency. We restrict $p_{(2)} \to \infty$, because our estimated variance $\hat{\sigma}_{p,0}$ may be inconsistent under the alternative models.

Remark 2. To gain some insight into the conditions on $p_{(2)}$, we consider the case of only two clusters for the regression functions. Let p_1 and p_2 denote the number of curves in each cluster, and we assume $p_1 \ge p_2$, without loss of generality. In this case, the signal

$$\theta_p = \{p(p-1)\}^{-1} \sum_{1 \le k \ne \ell \le p} \mathrm{PA}(m_k, m_\ell) = 2p_1p_2\{p(p-1)\}^{-1}\mathrm{PA}(m_1^*, m_2^*),$$

which is of order p_2p^{-1}. In addition, the noise (standard deviation of T_p) is of order $p^{-1/2}n_1^{-1/2}$. Thus, the signal-to-noise ratio is at an order of at least $p_2p^{-1/2}n_1^{1/2}$, which diverges when $p_2/\big(p^{1/2}n_1^{-1/2}\big) \to \infty$.

4. Application: Identifying Heterogeneity

Once we have rejected the null hypothesis H_0 that the p data sets are homogeneous, we need to identify any outlying data sets that depart from the majority of data sets that possess a common regression function. In this setting, we assume that the proportion of outlying data sets is not too large. In the terminology of outlier detection, we refer to an outlying data set that consists of multiple

measurements as an outlier.

Conventionally, outlier detection starts by finding a "clean" subset of the data, from which we can estimate the overall pattern of the underlying data-generating process, or, in our context, the common regression function. We then perform marginal comparisons between the models estimated by each data set with the overall pattern. If the marginal discrepancy measure takes a large value, then the corresponding data set is declared an outlier. This essentially performs a sequence of *two-sample* comparisons between each data set and the set of identified homogeneous data sets. Instead of estimating each marginal model and the common regression models, we first use a screening rule to obtain a "clean" subset of the data consisting of homogeneous data sets. The screening statistics are simply components of our global heterogeneity test statistic,

$$\hat{U}_{p,k} = (p-1)^{-1} \sum_{\substack{\ell=1 \\ \ell \neq k}}^{p} \widehat{\mathrm{PA}}_{k,\ell}, \text{ for } k = 1, \ldots, p. \tag{4.1}$$

If k corresponds to an outlying regression function, then $\hat{U}_{p,k}$ will tend to be large. Denote the order statistic of $\hat{U}_{p,k}$ as $\hat{U}_{p,(1)} \leq \hat{U}_{p,(2)} \leq \cdots \leq \hat{U}_{p,(p)}$, such that $\hat{U}_{p,(j)} = \hat{U}_{p,k_j}$, for $j = 1, \ldots, p$ and $k_j \in \{1, \ldots, p\}$. We can treat the data sets with indices $\mathcal{S} = \{k_1, \ldots, k_{S_p}\}$ as homogeneous or clean, say $\mathcal{Z}_{\mathcal{S}} = \{\mathcal{Z}_k, k \in \mathcal{S}\}$. Our numerical studies suggest that $S_p = \lfloor cp \rfloor$ with $c = 70\%$ exhibits robust performance. Proposition 3 provides a certain theoretical guarantee in the case of a simple two-cluster model.

Proposition 3. *Suppose Assumptions* 1, 2, *and* 3 *hold with* $L = 2$. *Let* $\mathcal{C}_j$ *be the set of indices corresponding to the data sets in the* j*th cluster, with cardinality* p_j, *for* $j = 1, 2$. *Without loss of generality, we assume* $p_1 \geq p_2$. *If* $p^{(3+\delta)/(2+\delta)}/\{(p_1 - p_2)n_1^{1/2}\} \to 0$, *then there exists a positive constant* C *such that as* $p \to \infty$ *and* $n_1 \to \infty$,

$$\mathrm{P}\left(\min_{k_2 \in \mathcal{C}_2} \hat{U}_{p,k_2} - \max_{k_1 \in \mathcal{C}_1} \hat{U}_{p,k_1} \geq C(p_1 - p_2)p^{-1} \right) \to 1.$$

Proposition 3 shows that outlying data sets can be differentiated from the clean data sets with probability approaching one under the two-cluster model. The condition $p^{(3+\delta)/(2+\delta)}/\{(p_1 - p_2)n_1^{1/2}\} \to 0$ requires that p cannot diverge too fast. In addition, p_1 and p_2 cannot be too close, which is reasonable in the context of outlier detection. In the case of $\{p_1 - p_2\}/p \to c_0$, for some positive constant $c_0 \leq 1$ as $p \to \infty$, the condition is satisfied when $p = o(n_1^{1+\delta/2})$. However, this

weakens with a higher moment condition on $m_k(X)$ and ε.

Remark 3. To calculate the screening statistics $\hat{U}_{p,k}$, we use a proportion of the data, say $\{(Y_{ki_j}, X_{ki_j}),\ i_j \in \{1,\ldots,n_k\},\ j = 1,\ldots,m\}_{k=1}^{p}$, to facilitate our computation when n_k is large. The screening rule is still valid if m replacing n_1 satisfies the conditions in Proposition 3.

Once we obtain the clean or normal majority, we can apply the heterogeneity testing procedure to form marginal outlier detection statistics. The key idea is to construct parallel data sets for each marginal comparison. That is, for each $k = 1,\ldots,p$, we first randomly divide the kth data set $\mathcal{Z}_k$ into $q_k = \lfloor {n_k}^{1/2} \rfloor$ disjoint subsets, such that the sample sizes of these q_k parallel data sets are roughly equal. Then, we sample n_k measures from the normal majority apart from $\mathcal{Z}_k$, that is, $\mathcal{Z}_{\mathcal{S}} \backslash \mathcal{Z}_k$. These are split further into q_k new data sets, again each with roughly equal sample sizes. Now, we obtain $2q_k$ parallel data sets, with one half exactly substituting the original $\mathcal{Z}_k$, and the other half sampled from the normal data sets. Finally, we apply our heterogeneity testing procedure in Section 2 to these $2q_k$ data sets, and denote the resulting p-value as $\mathcal{P}_k$. If $\mathcal{Z}_k$ is from the normal majority, we would not expect to have enough evidence to reject the null hypothesis that the $2q_k$ data sets are homogeneous. In contrast, if $\mathcal{Z}_k$ is a true outlier that possesses a regression function that deviates from the common one, then the p-value $\mathcal{P}_k$ should be small to reject the null. Hence, $\mathcal{P}_k$ is an appropriate measure for outlier detection. We denote the set of identified outliers as $\mathcal{O} = \{k : \mathcal{P}_k \leq \alpha\}$, for some nominal significance level α.

Remark 4. Dividing $\mathcal{Z}_k$ into a large number (q_k) of bins is motivated by the nature of the proposed heterogeneity testing procedure. One can trivially treat all screened normal data sets as a whole and perform the heterogeneity test on these $q_k + 1$ data sets. However, this could be computationally inefficient, and may cause unbalanced sample sizes, which goes against the assumption for the validity of our proposed test. Hence, we propose sampling just n_k measurements, which are split further into another q_k data sets, with each of the final $2q_k$ subsets having roughly comparable sample sizes. The detailed algorithm is provided as Algorithm 1 in the Supplementary Material.

Let p_{normal} and p_{outliers} be the numbers of normal and outlying data sets, respectively, among all p data sets. Denote $|\mathcal{O}_{\text{normal}}|$ as the number of data sets declared as outliers, but that are actually normal. By Proposition 3, our proposed detection rule guarantees that the false positive rate is approximately controlled at α, that is, $E\{|\mathcal{O}_{\text{normal}}|\}/p_{\text{normal}} \approx \alpha$.

5. Numerical Studies

We carried out extensive numerical studies, including simulations and real-data analyses, to assess the performance of our proposed test and algorithm.

5.1. Heterogeneity testing

In this section, we evaluate the finite-sample performance of our proposed method by considering the test statistic with a jackknife estimate of the variance (Section 3.2) and the procedure with bootstrap calibrations (Section 3.3), referred to as "Jack" and "Boots," respectively. The method proposed by Wang, Wang and Zou (2017), which is based on Fourier transformations, is considered as a benchmark, and is referred to as "WWZ." The WWZ method involves a nuisance parameter that is specified according to the suggestion given by Wang, Wang and Zou (2017).

To allow different sample sizes, we consider Model (1.1) with $n_k = N_k + 2$, for $k = 1, \ldots, p$, where N_k are independently sampled from a Poisson distribution with mean $n_0 - 2$. We vary p over the values $\{10, 25, 50, 100\}$ and vary n_0 over $\{10, 20, 40\}$. The dimension of all covariates $\{X_{ki},\ i = 1, \ldots, n_k;\ k = 1, \ldots, p\}$ is fixed at $d = 5$, and they are i.i.d. sampled from (i) a standard normal distribution $N(0, 1)$ or (ii) a standardized uniform distribution with zero mean and unit variance. The noise $\{\varepsilon_{ki},\ i = 1, \ldots, n_k,\ k = 1, \ldots, p\}$ is i.i.d. sampled from (1) $N(0, \sigma^2)$ or (2) $\sigma\{\mathrm{Exp}(1) - 1\}$, where $\mathrm{Exp}(1)$ is the exponential distribution with rate one. We consider the following regression functions of $x = (x^{(1)}, \ldots, x^{(d)})^{\mathrm{T}}$:

$$
\begin{aligned}
&m_1^*(x) = d^{-1/2} \sum_{j=1}^{d} x^{(j)},\ m_2^*(x) = \sqrt{2 \log d} \left\{ \max_{j=1,\ldots,d} x^{(j)} - \sqrt{2 \log d} \right\}, \\
&m_3^*(x) = \sum_{j=1}^{d} x^{(j)},\ m_4^*(x) = \sum_{j=1}^{d_1} x^{(j)} + b \sin\left(\pi \prod_{j=d_1+1}^{d} x^{(j)} \right), \\
&m_5^*(x) = \sum_{j=1}^{d_1} x^{(j)} + b \exp\left\{ - \left(\sum_{j=d_1+1}^{d} x^{(j)} \right)^2 \right\}, \\
&m_6^*(x) = \sum_{j=1}^{d_1} x^{(j)} + b \left(\sum_{j=d_1+1}^{d} x^{(j)} \right)^2 .
\end{aligned}
$$

To examine the sizes of the three tests, we consider three scenarios for $\{m_1, \ldots, m_p\}$: (I) $m_k = m_1^*$, for $k = 1, \ldots, p$; (II) $m_k = m_2^*$, for $k = 1, \ldots, p$; and (III) $m_k = m_4^*$, for $k = 1, \ldots, p$, with $d_1 = 3$ and $b = 1$. We set $\sigma = 2$.

Table 1. Observed sizes (in %) of various tests carried out at the 5% nominal level for various (p, n_0)-settings under scenarios (I–III)-(i)(1).

p	10			25			50			100		
n_0	10	20	40	10	20	40	10	20	40	10	20	40
Scenario (I)-(i)(1)												
Boots	5.6	4.0	2.5	4.8	4.2	4.5	4.9	5.6	5.7	4.9	4.8	6.0
Jack	3.8	2.0	1.4	2.6	2.4	2.6	2.4	3.6	3.5	2.7	3.1	4.4
WWZ	3.7	3.0	2.4	3.2	3.0	2.8	2.5	3.3	3.4	3.6	2.5	4.4
Scenario (II)-(i)(1)												
Boots	6.7	3.6	2.2	4.1	4.9	4.0	5.0	4.5	6.1	5.4	4.6	5.5
Jack	3.8	2.4	1.2	2.6	3.2	1.8	2.7	2.8	3.6	3.6	2.9	4.3
WWZ	3.4	3.1	2.0	3.1	3.2	2.9	3.0	3.6	4.6	4.1	3.2	4.1
Scenario (III)-(i)(1)												
Boots	5.3	3.7	3.0	3.9	3.9	4.3	4.5	4.7	5.8	4.5	4.5	5.1
Jack	2.5	2.2	1.9	1.9	1.9	2.3	3.0	2.6	3.3	2.7	2.9	3.5
WWZ	3.0	2.9	2.4	1.4	2.3	2.2	3.3	3.0	3.5	3.5	3.1	3.4

Table 1 presents the observed sizes (in %) of the three tests carried out at the 5% nominal level for various (p, n_0)-settings under scenarios (I–III)-(i)(1). The results show that the sizes of all tests are, in general, close to the nominal level. The WWZ and Jack methods based jackknife estimates seem a little conservative. In contrast, the Boots method that performs bootstrap calibrations yields better performance, overall. Complete numerical results for various combinations of model settings are deferred to Figures S.1–S.3 in Section S3 of the Supplementary Material, which show similar conclusions.

Then, we consider two scenarios to examine the power of the three tests. The first is a two-component mixture model: (IV) Each m_k, for $k = 1, \ldots, p$, is identical to m_1^* with probability $1 - \rho$ and to m_2^* with probability ρ. We consider the following two examples: (IV-a) p varies, $n_0 = 20$, $\rho = 0.5$, and $\sigma = 3$; and (IV-b) $p = 100$, $n_0 = 20$, ρ varies, and $\sigma = 3$. Figure 1 depicts the observed power (in %) of the three tests carried out at the 5% nominal level under Examples (IV-a)-(i/ii)(1) and (IV-b)-(i/ii)(1). The results show that under (i), where the covariates are normally distributed, WWZ performs slightly better than Boots. However, Boots is overwhelmingly better than WWZ under scenario (ii), where the distribution of the covariates is uniform.

The second experiment is conducted using a four-component mixture model: (V) Each m_k, for $k = 1, \ldots, p$, is allocated to clusters of ID 1–4 with probabilities $(\rho_1, \rho_2, \rho_3, \rho_4)$, such that $\sum_{\ell=1}^{4} \rho_\ell = 1$, where in the ℓth cluster, for $\ell = 1, \ldots, 4$, the regression functions are identical to $m_{\ell+2}^*$. Again, we examine two examples:

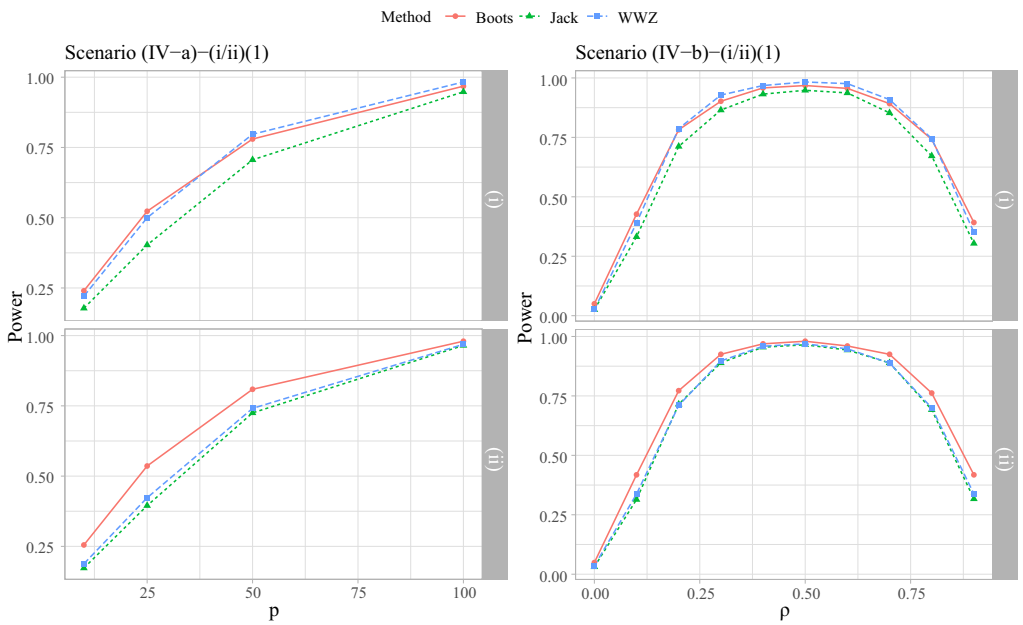

Figure 1. Observed power (in %) of various tests carried out at the 5% nominal level under Examples (IV-a)-(i/ii)(1) and (IV-b)-(i/ii)(1).

(V-a) $p = 100$, n_0 varies, $\rho_\ell = 0.25$, for $\ell = 1, \ldots, 4$, and $\sigma = 6$; and (V-b) $p = 100$, $n_0 = 20$, ρ_1 varies with $\rho_2 = \rho_3 = \rho_4 = (1 - \rho_1)/3$, and $\sigma = 4$. In both examples, we fix $d_1 = 3$ and $b = 1$. Figure 2 shows the observed power (in %) of the three tests carried out at the 5% nominal level under Examples (V-a)-(i/ii)(1) and (V-b)-(i/ii)(1). We can see from this plot that under both scenarios (i) and (ii) for the distribution of the covariates, Boots uniformly outperforms WWZ and Jack methods because of the proposed bootstrap calibrations.

5.2. Identifying outlying data sets

Now, we investigate the finite-sample performance of our proposed in detecting outlying data sets. To facilitate practical use, Algorithm 1 in the Supplementary Material describes the outlier detection process. For illustrative purposes, we consider the numerical setting (IV) in Section 5.1 with $p \in \{100, 200\}$, $n_0 \in \{50, 100, 200\}$, $d = 5$, and $\sigma = 1$, and we range the proportion of outlying data sets ρ over the values $\{5\%, 10\%, 15\%, 20\%\}$. We set the nominal significance level to 5%.

Table 2 reports the average sizes (in %), that is, the proportions of falsely identified outlying data sets among all homogeneous data sets, under different

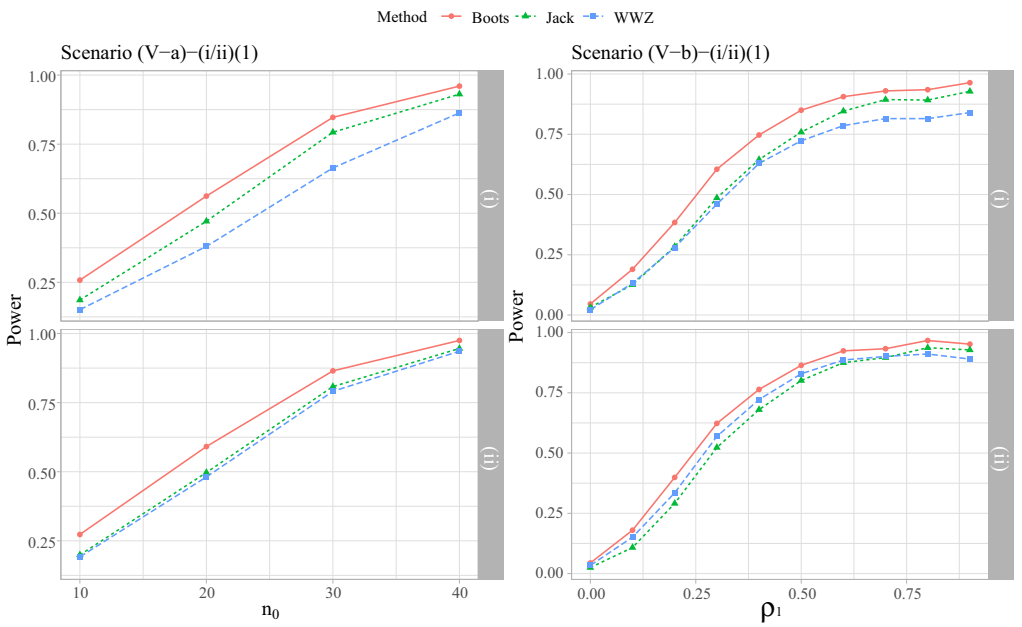

Figure 2. Observed power of various tests carried out at the 5% nominal level under Examples (V-a)-(i/ii)(1) and (V-b)-(i/ii)(1).

Table 2. Average sizes (in %) under different configurations of (p, n_0, ρ) when the covariates and noise are both normally distributed.

	$p = 100$			$p = 200$		
ρ	$n_0 = 50$	$n_0 = 100$	$n_0 = 200$	$n_0 = 50$	$n_0 = 100$	$n_0 = 200$
5%	4.18	4.27	4.46	4.24	4.35	4.46
10%	4.29	4.44	4.68	4.30	4.37	4.57
15%	4.14	4.64	4.92	4.29	4.41	4.74
20%	4.35	4.81	5.76	4.26	4.44	5.08

configurations of (p, n_0, ρ) when the covariates and noise are both normally distributed. Here, our proposed method provides satisfactory observed sizes. In other words, the number of mistakenly declared outliers is well controlled. Figure 3 depicts the average power (in %), that is, the proportions of correctly identified outlying data sets among all truly outlying data sets under the same settings. We observe that most truly outlying data sets are discovered by our method, with the type-I error rate being well controlled. Moreover, the average power increases as more measurements are collected.

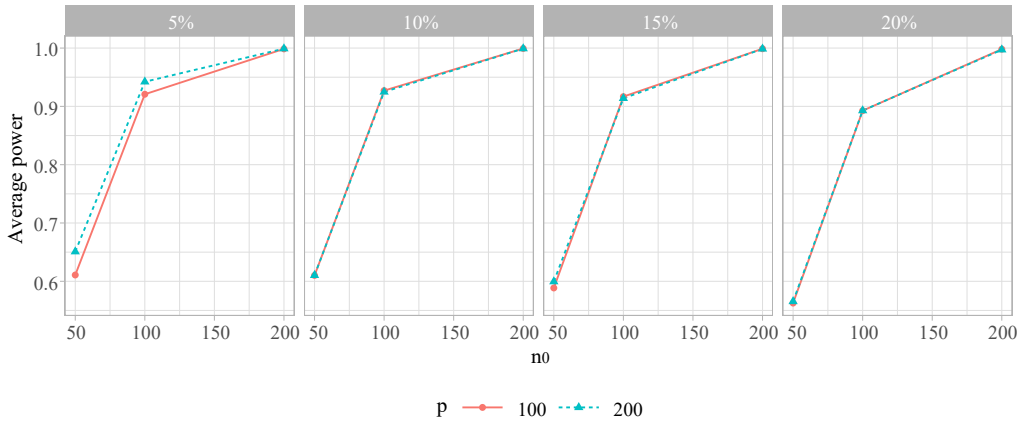

Figure 3. Average power (in %) under different configurations of (p, n_0, ρ) when the covariates and noise are both normally distributed.

5.3. Real data analysis

In this section, we apply our proposed testing procedure and heterogeneity identification algorithm to Melbourne housing data. The data set consists of transaction details of properties in 2016 in Melbourne, Victoria, Australia. We focus on transactions with the property type "House" and suburbs with at least 50 transactions in 2016. After preprocessing, the subset contains 238 suburbs with 33,973 transactions. We are interested in the relative change of housing prices in 2016 compared with those in 2015, as explained by eight covariates: the month of transaction, latitude, longitude, number of bedrooms, number of bathrooms, land size, building area, and year the property was built. For each suburb, we define the growth rate of housing prices as the excess rate of the sold price of the property in each transaction compared with the median price within the suburb in 2015. This variable serves as the response.

By treating each suburb as a group, we obtain 238 parallel data sets. Here, we want to know whether the eight covariates contribute differently to the housing price growth rate across suburbs. After standardizing the eight covariates, we implement our proposed heterogeneity test on the 238 suburbs to test whether they share the same regression function. The p-value of the heterogeneity test is less than 0.0001 when using 10,000 bootstrap iterations, indicating significant evidence against the eight covariates having the same contribution to the relative change in housing prices across suburbs.

Now that we have rejected the homogeneity hypothesis, we wish to identify outlying suburbs. In order to avoid unbalanced sample sizes across suburbs,

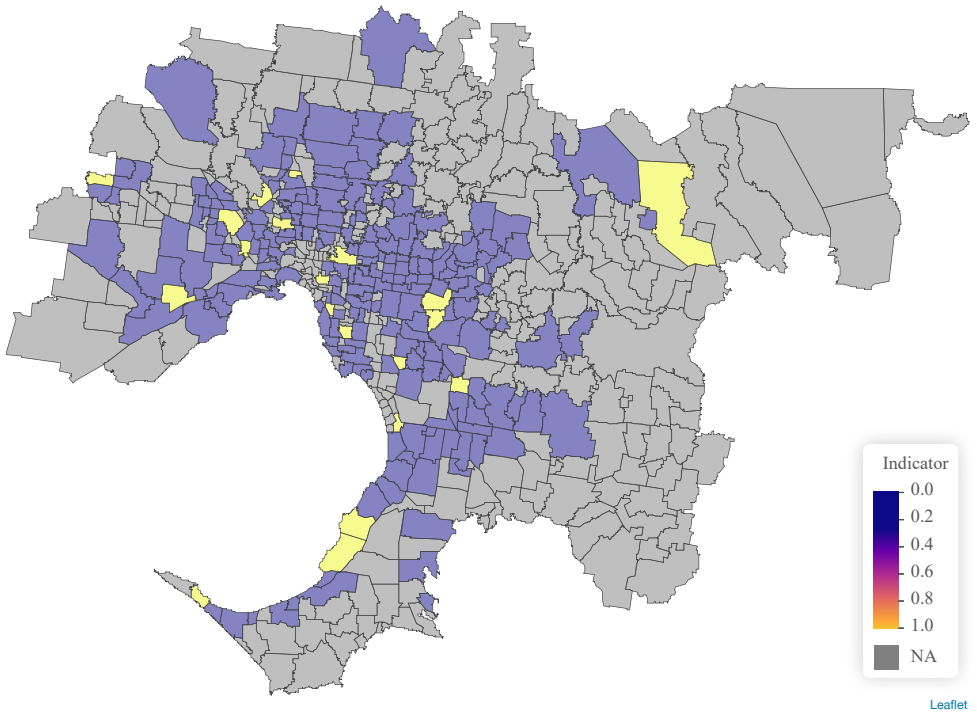

Figure 4. Plot of all suburbs, with detected outliers indicated in yellow and the majority in blue. The suburbs in gray are those with less than 50 transactions in 2016.

we randomly sample 50 transactions from each suburb. The proposed outlier detection algorithm identified 20 out of 238 suburbs as outliers, with significance level $\alpha = 5\%$. Figure 4 shows whether a suburb is detected as an outlier (filled in yellow). The suburbs in gray are those with less than 50 transactions in 2016.

After identifying the 20 outlying suburbs, we need to determine how they differ from the majority (the other 218 suburbs) in terms of modeling and prediction. Thus, we carry out the following simulation to verify the significance of separating the outlying suburbs from the majority. Denote the data set containing all 20 outlying suburbs as O_1, and the data set containing all 218 majority suburbs as M_1. In each iteration, we randomly split M_1 into two subsets, M_2 and M_3, where M_3 has the same size as O_1. The subset M_2 serves as the training set from the majority, and M_3 and O_1 are test sets from the majority and outliers, respectively. A random forest model is trained on M_2, and then used for predictions based on M_3 and O_1. Averaged over 100 iterations, the R^2 of the random forest modeling is 69.5%, which indicates a reasonable fit. The average mean squared error for M_3 is 0.1249, and is 0.1392 for O_1, which is 11.45% higher than

that of M_3. This simulation result suggests the risk of obtaining a larger error when using the model fitted from the majority to predict outlying suburbs. In summary, the outlier detection result provides guidance for further data modeling and analyses of specific outlying suburbs.

6. Conclusion

We have discussed heterogeneity measurement, testing, and identification problems for large parallel data sets collected from multiple regression models. We have proposed a new metric for the equivalence or departure of two regression models, based on the projection approach. Motivated by this, we developed a procedure to test for homogeneity. For non-homogeneous data sets, we have proposed a detection procedure to identify outlying data sets that come from different regression models to that of the majority. The proposed method is model free and data adaptive, which makes it convenient to use in practice.

Our proposed method assumes that the covariates are i.i.d. across data sets. However, domain shifts could happen in practical applications; that is, the distributions of the covariates may differ across data sets. Of primary interest is to test whether domain shifts occur. If the covariates have densities, Zhan and Hart (2014) propose using a kernel smoothing-based procedure to test the equality of a large number of densities. We believe that similar projection averaging-based metrics for the departure of two distributions (e.g., Kim, Balakrishnan and Wasserman (2020)), together with the proposed large-scale testing scheme, can still be applied to achieve this purpose. If domain shifts do occur, the proposed method cannot be used directly to test for the equality of the regression functions (cf., Eq. (1.2)). Model-free and data-adaptive testing procedures that allow for different distributions of the covariates warrant further research. Recently, Xiao, Ke and Li (2021) studied individual regression heterogeneity in panel data and allowed the data to exhibit outliers. On the other hand, regression functions and domain shifts are both sources of heterogeneity in parallel data sets. It is challenging, and possibly infeasible to identify the real cause of heterogeneity in a nonparametric setup, which is left as future work.

Recently, there has been an increase in the number of distance-based approaches, such as the distance correlation (Székely, Rizzo and Bakirov (2007)) and the maximum mean discrepancy (Gretton et al. (2012)). The projection-based method proposed here also seeks to avoid the curse of dimensionality. Thus, it would be interesting to study the relationship between the distance-based method and our proposed projection-based approach. This too is left to future research.

Supplementary Material

The online Supplementary Material contains proofs of the main results presented in the paper, an algorithm for identifying outlying data sets, and additional numerical results.

Acknowledgments

The authors contributed equally to this work, and are listed in alphabetical order. The authors would like to acknowledge the editor, associate editor, and two referees for their constructive comments and suggestions. Wang's research was supported by the NNSF of China Grant 11901314. Zou's research was supported by the NNSF of China Grants 11925106, 11931001, and 11971247, the NSF of Tianjin Grant 18JCJQJC46000, and the 111Project B20016.

References

Baltagi, B. H., Hidalgo, J. and Li, Q. (1996). A nonparametric test for poolability using panel data. *Journal of Econometrics* **75**, 345–367.

Barras, L., Scaillet, O. and Wermers, R. (2010). False discoveries in mutual fund performance: Measuring luck in estimated alphas. *The Journal of Finance* **65**, 179–216.

Borgwardt, K. M., Gretton, A., Rasch, M. J., Kriegel, H.-P., Schölkopf, B. and Smola, A. J. (2006). Integrating structured biological data by Kernel Maximum mean Discrepancy. *Bioinformatics* **22**, e49–e57.

Cai, L. and Wang, S. (2021). Global statistical inference for the difference between two regression mean curves with covariates possibly partially missing. *Statistical Papers* **62**, 2573–2602.

Chiou, J.-M. and Li, P.-L. (2007). Functional clustering and identifying substructures of longitudinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69**, 679–699.

Cuesta-Albertos, J. A., García-Portugués, E., Febrero-Bande, M. and González-Manteiga, W. (2019). Goodness-of-fit tests for the functional linear model based on randomly projected empirical processes. *The Annals of Statistics* **47**, 439–467.

Escanciano, J. C. (2006). A consistent diagnostic test for regression models using projections. *Econometric Theory* **22**, 1030–1051.

González-Manteiga, W. and Crujeiras, R. M. (2013). An updated review of Goodness-of-Fit tests for regression models. *TEST* **22**, 361–411.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B. and Smola, A. (2012). A kernel two-sample test. *The Journal of Machine Learning Research* **13**, 723–773.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics* **19**, 293–325.

Ke, Y., Li, J. and Zhang, W. (2016). Structure identification in panel data analysis. *The Annals of Statistics* **44**, 1193–1233.

Kim, I., Balakrishnan, S. and Wasserman, L. (2020). Robust multivariate nonparametric tests via projection averaging. *The Annals of Statistics* **48**, 3417–3441.

Koul, H. L. and Li, F. (2020). Comparing two nonparametric regression curves in the presence of long memory in covariates and errors. *Metrika* **83**, 499–517.

Lavergne, P. and Patilea, V. (2008). Breaking the curse of dimensionality in nonparametric testing. *Journal of Econometrics* **143**, 103–122.

Neumeyer, N. and Dette, H. (2001). Nonparametric analysis of covariance. *The Annals of Statistics* **29**, 1361–1400.

Neumeyer, N. and Dette, H. (2003). Nonparametric comparison of regression curves: An empirical process approach. *The Annals of Statistics* **31**, 880–920.

Pardo-Fernández, J. C., Van Keilegom, I. and González-Manteiga, W. (2007). Testing for the equality of k regression curves. *Statistica Sinica* **17**, 1115–1137.

Patilea, V., Sánchez-Sellero, C. and Saumard, M. (2016). Testing the predictor effect on a functional response. *Journal of the American Statistical Association* **111**, 1684–1695.

Qiu, P. and Xiang, D. (2014). Univariate dynamic screening system: An approach for identifying individuals with irregular longitudinal behavior. *Technometrics* **56**, 248–260.

Sen, P. K. (1977). Some invariance principles relating to jackknifing and their role in sequential analysis. *The Annals of Statistics* **5**, 316–329.

Srihera, R. and Stute, W. (2010). Nonparametric comparison of regression functions. *Journal of Multivariate Analysis* **101**, 2039–2059.

Székely, G. J., Rizzo, M. L. and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics* **35**, 2769–2794.

Tang, L. and Song, P. X. K. (2016). Fused Lasso approach in regression coefficients clustering – learning parameter heterogeneity in data Integration. *Journal of Machine Learning Research* **17**, 1–23.

Vogt, M. and Linton, O. (2017). Classification of non-parametric regression functions in longitudinal data models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79**, 5–27.

Wang, G., Wang, Z. and Zou, C. (2017). Comparison of a large number of regression curves. *Journal of Multivariate Analysis* **162**, 122–133.

Xia, Y. (2009). Model checking in regression via dimension reduction. *Biometrika* **96**, 133–148.

Xiao, D., Ke, Y. and Li, R. (2021). Homogeneity structure learning in large-scale panel data with heavy-tailed errors. *Journal of Machine Learning Research* **22**.

Zhan, D. and Hart, J. D. (2014). Testing equality of a large number of densities. *Biometrika* **101**, 449–464.

Zhu, L., Xu, K., Li, R. and Zhong, W. (2017). Projection correlation between two random vectors. *Biometrika* **104**, 829–843.

Liuhua Peng

School of Mathematics and Statistics, the University of Melbourne, Victoria 3010, Australia.

E-mail: liuhua.peng@unimelb.edu.au

Guanghui Wang

KLATASDS-MOE, School of Statistics, and Academy of Statistics and Interdisciplinary Sciences, East China Normal University, Shanghai 200062, China.

E-mail: ghwang.nk@gmail.com

Changliang Zou
School of Statistics and Data Science, LPMC, LEBPS, and KLMDASR, Nankai University, Tianjin 300071, China.
E-mail: nk.chlzou@gmail.com

(Received August 2021; accepted April 2022)