Statistica Sinica Preprint No: SS-2021-0285	
Title	Measuring, Testing, and Identifying Heterogeneity of Large Parallel Datasets
Manuscript ID	SS-2021-0285
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202021.0285
Complete List of Authors	Liuhua Peng, Guanghui Wang and Changliang Zou
Corresponding Authors	Changliang Zou
E-mails	nk.chlzou@gmail.com
Notice: Accepted version subject to English editing.	

Statistica Sinica

MEASURING, TESTING, AND IDENTIFYING HETEROGENEITY OF LARGE PARALLEL DATASETS

Liuhua Peng, Guanghui Wang and Changliang Zou

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

Abstract: In applications, large datasets of parallel situations are encountered more and more often. It is necessary to check whether they are collected from different regression models before further modeling, estimation and inference. A novel metric for such heterogeneity is proposed based on the projection strategy, whose strength is then borrowed to form a new test for the equivalence of a large number of unknown regression models that is fully data-driven. Asymptotic normality of the proposed test is constructed. The testing procedure is further applied to identify outlying datasets whose regression models deviate from the majority. Extensive numerical studies demonstrate that our methods have satisfactory performance.

Key words and phrases: Heterogeneity, Parallel datasets, Projections, Outlier detection, U-statistics

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

1. Introduction

Large datasets of parallel situations collected from various sources or platforms are encountered more and more often in many scientific fields such as bioinformatics, computer science, mechanical engineering, and economics, thanks to the advancement of data collection techniques and devices. Measuring and testing homogeneity of such large parallel-structured datasets is a fundamental problem before further data processing. For example, in experimental studies, it is necessary to check whether the underlying distributions or models of parallel datasets collected under different conditions or treatments differ largely before data integration (Borgwardt et al., 2006; Tang and Song, 2016). Moreover, even in the same treatment group, one also needs to figure out whether different individuals in that group indeed share the same model before group-specific modeling (Ke et al., 2016; Vogt and Linton, 2017). A formal test is thereby required to provide uncertainty quantification on data homogeneity. In such scenarios, a post-test diagnostic, which is able to identify outlying groups or individuals, is also desirable in order to accurately estimate the overall pattern.

Heterogeneity of multiple datasets may come from various forms of variability across parallel studies. In this paper, we handle datasets each of which contains paired measurements of responses and covariates, and

thus to test heterogeneity across multiple datasets is essentially to check whether all datasets share the common regression function. Suppose we have collected p datasets of parallel situations and the kth dataset consists of n_k paired members $\{(Y_{ki}, X_{ki}),\ i = 1, \ldots, n_k\}$ for $k = 1, \ldots, p$, where Y_{ki}'s are scalar responses and X_{ki}'s are the associated d-dimensional covariates. We model the data via

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

where m_k's are the regression functions and ε_{ki}'s are random noises 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 these p datasets are collected independently, i.e., $\mathcal{Z}_k$ is independent of $\mathcal{Z}_\ell$ for any (k, ℓ) such that $1 \le k \ne \ell \le p$, and the paired covariates and noises $(X_{ki}, \varepsilon_{ki})$'s are independent and identically distributed (i.i.d.) as (X, ε) for $i = 1, \ldots, n_k;\ k = 1, \ldots, p$. Testing homogeneity of these p datasets can thus be formulated as testing the following null hypothesis

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

Here, we require no knowledge of any structured functional forms of $\{m_k\}_{k=1}^p$

and treat them as nonparametric. In this paper, we consider a large-scale set-up in the sense that the number of parallel datasets $p \to \infty$. Once the null hypothesis H_0 in (1.2) is rejected, another task is to identify the outlying datasets that possess different regression functions from the majority.

In fact, testing heterogeneity among several (usually two) regression functions has been widely discussed in the literature. It is natural to perform a classical analysis of variance if all regression functions are restricted to some parametric forms such as linear models. Recent years have also seen the development of testing procedures when no parametric structures are restricted. This has been the object of much work, see, for instance, Neumeyer and Dette (2001, 2003); Pardo-Fernández et al. (2007); Srihera and Stute (2010); Koul and Li (2020) and Cai and Wang (2021) among many others. We refer to Section 7 in González-Manteiga and Crujeiras (2013) for a brief overview. Our context differs from the traditional ones largely in the amount of parallel cases in the sense that the number of regression functions in comparison can be very large, i.e., asymptotically speaking $p \to \infty$. For example, treatments in experimental studies may be indicated by a categorical variable which takes values among a large collection of candidates, which naturally results in large parallel datasets. As another example, econometricians handling panel data may consider

whether the data is poolable over time (Baltagi et al., 1996; Barras et al., 2010), where the number of time periods could be much large. In a study of longitudinal data where one is interested in estimating the overall pattern from a large number of subjects (Chiou and Li, 2007; Qiu and Xiang, 2014), it requires careful examination of whether there are significant differences among all individual subjects. Such applications have raised the need to carry out heterogeneity tests for large parallel datasets.

In this paper, we first propose a model-free metric for the departure of two regression functions based on a projection approach, whose strength is then borrowed to form a test statistic for heterogeneity testing for large parallel datasets. The proposed testing procedure makes no parametric assumptions on the regression functions and does not require any direct estimation of the nonparametric models. Compared to existing work, our approach is free of any nuisance parameters, making it particularly useful in the situation with large p. We construct asymptotic properties of the test statistic when the sample sizes of all datasets diverge. We also propose a bootstrap remedy to mimic the null distribution in cases of conservative sizes in finite-sample performance. Its asymptotic validity and consistency are established. In addition, we illustrate how to apply the proposed heterogeneity testing procedure to fulfill the task of identification of outlying

datasets. 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 has satisfactory performance in terms of correctly detecting outlying datasets with the false positive rate being well controlled.

A closely related work is Wang et al. (2017) which studied the testing aspect. Our proposed test statistic shares certain similarities to theirs. For example, both get rid of nonparametric estimation of the underlying regression models, and both are related to U-statistics. However, our contribution is three-fold. Firstly, the proposed projection-based metric of heterogeneity is new. Secondly, our test statistic is free of any tuning parameters. In contrast, their test statistic involves an additional nuisance parameter which need to be elaborately specified; moreover, their asymptotic analysis is built by treating the nuisance parameter as being fixed, and no theoretical guarantees are provided if a data-dependent estimate is plugged-in. Thirdly, numerical studies reveal that their procedure could sometimes be conservative for large sample sizes, while this issue is mitigated by our method via the proposed bootstrap calibrations; furthermore, the calibrations are built on an elaborate analysis of the asymptotic behavior of the proposed test statistics, which makes the theoretical derivations much more involved. In addition, a novel outlier detection scheme is proposed based on the proposed

testing method.

The remainder of our paper is organized as follows. In Section 2, we develop the heterogeneity measure and then derive our test statistic. Section 3 provides theoretical investigations on the proposed method. Section 4 presents the application of the proposed measure and testing procedure to identify outlying datasets. A number of simulated and real-data examples are given in Section 5 regarding numerical performance. Some remarks conclude the paper in Section 6. Proofs of theoretical results and some additional numerical results are deferred in 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 would induce a testing procedure that avoids directly estimating the regression functions by, for example, using kernel smoothing

methods, and could be applied to covariates with moderate or even large dimension. The key observation is 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*

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

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 in order to avoid the curse of dimensionality are frequently used in the literature of goodness-of-fit testing (Escanciano, 2006; Lavergne and Patilea, 2008; Xia, 2009; Patilea et al., 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 a Projection-Averaging (PA) based measure for the equivalence or departure of two regression functions

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

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 $\mathrm{PA}(m_1, m_2) \geq 0$, and $\mathrm{PA}(m_1, m_2) = 0$ if and only if $\mathrm{P}(m_1(X) = m_2(X)) = 1$. The projection-averaging techniques similar to that used in (2.1) have become popular recently, see for example Escanciano (2006), Zhu et al. (2017), Kim et al. (2020) for different inferential purposes. One advantage of the PA approach is that it can entail a closed-form expression as shown in Proposition 1 below.

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 $K(X, X', X'') = 2^{-1} - (2\pi)^{-1} \arccos \frac{(X-X'')^{\mathrm{T}}(X'-X'')}{\|X-X''\|\|X'-X''\|}$ *if* $X \neq X''$ *and* $X' \neq X''$. *If* $X = X'' \neq X'$ *or* $X' = X'' \neq X$, *then* $K(X, X', X'') = 1/2$ *and if* $X = X' = X''$, *then* $K(X, X', X'') = 1$.

Moreover, as we will illustrate later, the benefit of using (2.1) or (2.2) appears clearer by observing the key fact that $E(Y_{ki} \mid X_{ki}) = m_k(X_{ki})$ a.s.

for $i = 1, \ldots, n_k;\ k = 1, \ldots, p$, which sheds light on the construction of the test statistic for H_0 free of any kernel estimation of the regression functions.

2.2 The test statistic

The idea in the two-sample scenario can be naturally generalized to the context of multiple comparison. 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's 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 essentially estimates of all pairwise discrepancy measures $\mathrm{PA}(m_k, m_\ell)$'s.

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. To fix the idea, suppose $\mathcal{K}$ is known at the moment. Recall the key fact that $E(Y_{ki} \mid X_{ki}) = m_k(X_{ki})$ for $i = 1, \ldots, n_k;\ k = 1, \ldots, p$. 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 naturally 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 hard to be specified in closed forms or even unknown. Hence we need some "well" approximations to $\mathcal{K}$. To this end, we propose to use 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

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

The T_p is free of any tuning parameters, and thus is convenient for direct usage. As T_p is also unbiased for θ_p, large values of T_p indicate rejecting H_0 that these p datasets are homogeneous.

3. Theoretical properties

In this section, we establish the asymptotic null distribution of the test statistic T_p and use Jackknife to estimate it 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 set-up in the sense that the number of parallel datasets $p \to \infty$. 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 (Model)**.** $\mathcal{Z}_k$ is independent of $\mathcal{Z}_\ell$ for any $1 \le k \ne \ell \le p$, and $(X_{ki}, \varepsilon_{ki})$'s are i.i.d. as (X, ε) for $i = 1, \ldots, n_k$, $k = 1, \ldots, p$. 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 (Number of datasets 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

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

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 3.1–3.1 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 the Hoeffding's decomposition (Hoeffding, 1948) of U-statistics. In fact, T_p is asymptotically equivalent to U_p, a U-statistic of degree 2 with a kernel that may depend on p and n_k's, 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

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

In other words, T_p is asymptotically equivalent to an average of a sequence of independent but not identically distributed random variables. Asymptotic normality of T_p could then be constructed by 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 2 (Sen, 1977). Denote $\hat{U}_{p,k} = (p-1)^{-1}\sum_{\substack{\ell=1\\ \ell\neq k}}^{p} \widehat{\mathrm{PA}}_{k,\ell}$. We can 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 that 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 3.1–3.1 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 the size of the test at significance level α.

3.3 Bootstrap calibrations

Our testing procedure with the jackknife variance estimate $\hat{\sigma}_{p,0}^2$ could be applicable directly free of any nuisance parameters. However, our simulation studies indicate that in some finite-sample situations it would result 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 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 **??** 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 $\{\hat{U}_{p,k}\}_{k=1}^{p}$ and we randomly draw $\hat{U}_{p,1}^*, \ldots, \hat{U}_{p,p}^*$ from $F_{p,U}$. As $E(\hat{U}_{p,1}^* \mid F_{p,U}) = p^{-1}\sum_{k=1}^{p}\hat{U}_{p,k} = T_p$, 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 theoretical support for the bootstrap method.

Theorem 2. *Suppose Assumptions 3.1–3.1 hold. Under H_0, as $p \to \infty$,*

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

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 via Monte Carlo simulation by repeatedly sampling from $F_{p,U}$ a large number of times. Numerical studies indicates that the test based on bootstrap calibration enjoys better performance in term of sizes.

Remark 1. It is remarkable to point out that the computational complexity of either the Jackknife or Bootstrap-based testing procedure is mainly from the calculation of the test statistic T_p, which is of the order $O(p^2 n_1^3)$ if all n_k's are of the order $O(n_1)$. Upon the availability of $\{\hat{U}_{p,k}\}_{k=1}^p$ in the computation of T_p, both the Jackknife and Bootstrap scheme can be preformed in a very efficient manner, and no additional refitting are required. However, computing the test statistic T_p itself is sometimes computationally expensive, especially for very large p and n_k's, which is unavoidable at this moment due to the construction of the test statistic.

3.4 Asymptotic power

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

Assumption (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 3.1–3.4 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 $\{p_\ell\}_{\ell=1}^L$, 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 bootstrap calibration converges to 1 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 that $p_{(2)} \to \infty$ since our estimated variance $\hat{\sigma}_{p,0}$ may be inconsistent under the alternative models.

Remark 2. To gain some insights into the conditions on $p_{(2)}$, we consider the case when there are 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

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

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

4. Application: heterogeneity identification

Once the null hypothesis H_0 that these p datasets are homogeneous is rejected, a natural question is that "could we identify outlying datasets that depart from the majority or normal ones possessing a common regression function?". In this setting, we assume that the proportion of outlying datasets should not be too large. In the terminology of outlier detection, an outlying dataset that consists of multiple measurements is referred to as an outlier here.

Conventionally, outlier detection starts by finding a "clean" subset of the data to estimate the overall pattern of underlying data generating process, to wit, in the current context, the common regression function, and then proceeds by performing marginal comparisons between the model esti-

mated by each dataset with the overall pattern. If the marginal discrepancy measure takes a large value, then the corresponding dataset should be declared as an outlier. This essentially performs a sequence of *two-sample* comparisons between each dataset and the set of identified homogeneous datasets. Instead of estimating each marginal and the common regression models, we first use a screening rule to obtain a "clean" subset of the data consisting of homogeneous datasets. The involving screening statistics are just components of our global heterogeneity test statistic, to be specific,

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

If k corresponds to an outlying regression function, then $\hat{U}_{p,k}$ would tend to be large. Denote the order statistic of $\hat{U}_{p,k}$'s 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 simply treat datasets with indices $\mathcal{S} = \{k_1, \ldots, k_{S_p}\}$ as homogeneous or clean, say $\mathcal{Z}_{\mathcal{S}} = \{\mathcal{Z}_k, k \in \mathcal{S}\}$. Numerical studies suggest that $S_p = \lfloor cp \rfloor$ with $c = 70\%$ has a robust performance. Proposition 3 provides certain theoretical guarantee in case of a simple two-cluster model.

Proposition 3. *Suppose Assumption 3.1, 3.1 and 3.4 holds with* $L = 2$. *Let* $\mathcal{C}_j$ *be the set of indices corresponding to datasets 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 the outlying datasets could be distinguished from the clean one with probability approaching 1 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 $p = o(n_1^{1+\delta/2})$, which weakens with a higher moment condition on $m_k(X)$ and ε.

Remark 3. To calculate the screening statistics $\hat{U}_{p,k}$'s, we could 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's are 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 further apply the heterogeneity testing procedure to form marginal outlier detection statistics.

The key idea is to construct parallel datasets for each marginal comparison. To wit, for each $k = 1, \ldots, p$, we first randomly divide the kth dataset $\mathcal{Z}_k$ into $q_k = \lfloor {n_k}^{1/2} \rfloor$ disjoint subsets such that the sample sizes of these q_k parallel datasets are roughly equal. Then we sample n_k measures form the normal majority apart from $\mathcal{Z}_k$, that is, $\mathcal{Z}_{\mathcal{S}} \backslash \mathcal{Z}_k$, and they are further split into q_k new datasets again each with roughly equal sample sizes. Now we obtain $2q_k$ parallel datasets, with one half exactly substituting the original $\mathcal{Z}_k$ and the other half sampled form the normal datasets. Finally, we apply our heterogeneity testing procedure in Section 2 to these $2q_k$ datasets, and denote the resulted p-value as $\mathcal{P}_k$. It could be expected that if $\mathcal{Z}_k$ is from the normal majority, we would not have enough evidence to reject the null hypothesis that these $2q_k$ datasets are homogeneous. On the contrary, if $\mathcal{Z}_k$ is a true outlier which possesses a regression function deviating from the common one, the p-value $\mathcal{P}_k$ should be small to reject the null, or in other words, $\mathcal{Z}_k$ would be declared as an outlier. 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 significant 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 datasets as a whole and perform the heterogeneity

test on these $q_k + 1$ datasets. However, this could be computationally inefficient and also causes very unbalanced sample sizes that goes against the assumption for the validity of our proposed test. Hence we proposed sampling just n_k measurements, which are further split into another q_k datasets, and each of the final $2q_k$ subsets possesses roughly comparable sample sizes. The detailed algorithm is deferred in Algorithm **??** in the Supplementary Material.

Let p_{normal} and p_{outliers} be the number of normal and outlying datasets among all p datasets. Denote $|\mathcal{O}_{\text{normal}}|$ as the number of declared outliers but actually the normal ones. By Proposition 3, our proposed detection rule would guarantee 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 carry out extensive numerical studies including simulations and real data analysis to assess the performance of our proposed test and outlying datasets identification 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), which are termed as "Jack" and "Boots", respectively. The method proposed by Wang et al. (2017) which is based on the Fourier transformations is considered as a benchmark, and it is termed as "WWZ". The WWZ method involves a nuisance parameter which is specified according to the suggestion given by Wang et al. (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's 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 noises $\{\varepsilon_{ki},\ i = 1, \ldots, n_k,\ k = 1, \ldots, p\}$ are i.i.d. sampled from (1) $N(0, \sigma^2)$ or (2) $\sigma\{\text{Exp}(1)-1\}$, where Exp(1) is the exponential distribution with rate 1. A list of re-

gression functions of $x = (x^{(1)}, \ldots, x^{(d)})^{\mathrm{T}}$ are considered below:

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

To examine the sizes of the three tests, we consider three scenarios for $\{m_1, \ldots, m_p\}$: (I) $m_k = m_1^*, \; k = 1, \ldots, p$, (II) $m_k = m_2^*, \; k = 1, \ldots, p$ and (III) $m_k = m_4^*, \; k = 1, \ldots, p$ with $d_1 = 3$ and $b = 1$. We set $\sigma = 2$. 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 the scenarios (I–III)-(i)(1). It can be seen that the sizes of all tests are generally close to the nominal level. The WWZ and Jack methods that based on the Jackknife estimates seem a little conservative sometimes. In contrast, our Boots method that performs bootstrap calibrations yields better performance overall. A complete numerical result regarding various combinations of the model settings is deferred in Figures **??**–**??** in Section **??** in the Supplementary Material,

Table 1: Observed sizes (in %) of various tests carried out at the 5% nominal level for various (p, n_0)-settings under the 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

from which similar conclusion can be reached.

Then we consider two scenarios to examine the power of the three tests. The first one is a two-component mixture model: (IV) Each m_k, $k = 1, \ldots, p$ 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 $\sigma = 3$; (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)C(1) and (IV-b)-(i/ii)C(1), from which we see that under (i) that the covariates are normally distributed, the WWZ performs slightly better than our Boots method sometimes, while the Boots

is overwhelmingly better than the WWZ under (ii) that the distribution of covariates is uniform.

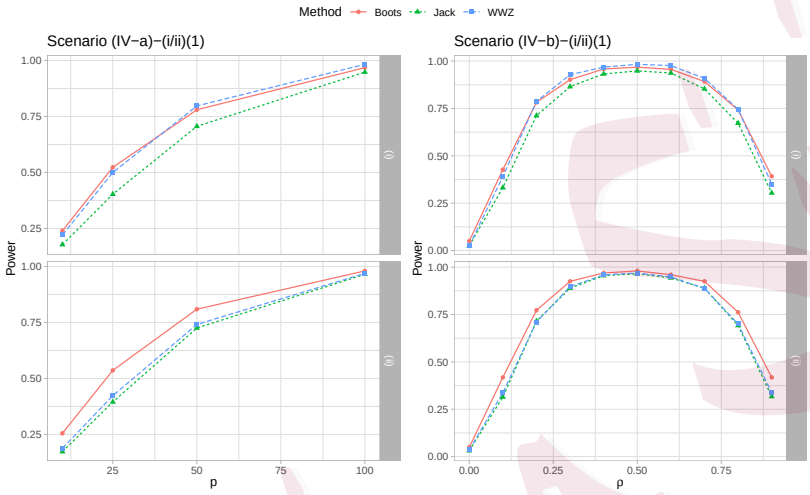

Figure 1: Observed power (in %) of various tests carried out at the 5% nominal level under Examples (IV-a)-(i/ii)C(1) and (IV-b)-(i/ii)C(1).

The second experiment is conducted via a four-component mixture model: (V) Each m_k, $k = 1, \ldots, p$ is allocated to the clusters of ID's 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, $\ell = 1, \ldots, 4$, the regression functions are identical to $m^*_{\ell+2}$. Again, two examples are visited: (V-a) $p = 100$, n_0 varies, $\rho_\ell = 0.25$ for $\ell = 1, \ldots, 4$, and $\sigma = 6$; (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)C(1) and (V-b)-

(i/ii)C(1). We can see from this plot that under both scenarios (i) and (ii) for the distribution of covariates, our Boots method uniformly outperforms the WWZ and Jack methods due to the proposed bootstrap calibrations.

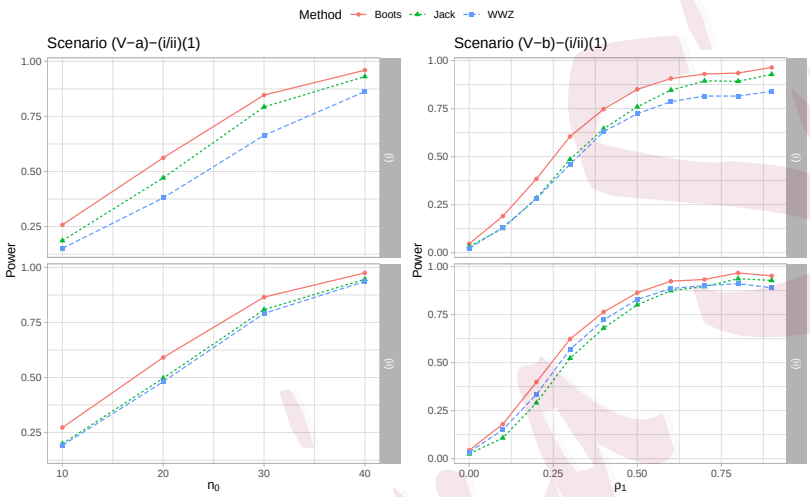

Figure 2: Observed power of various tests carried out at the 5% nominal level under Examples (V-a)-(i/ii)C(1) and (V-b)-(i/ii)C(1).

5.2 Identifying outlying datasets

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

set the nominal significant level as 5%.

Table 2 reports the averaged sizes (in %), i.e., the proportions of falsely identified outlying datasets among all homogeneous datasets, under different configurations of (p, n_0, ρ) when the covariates and noises are both normally distributed. It can be seen that our proposed method has very satisfactory observed sizes. In other words, the number of mistakenly declared outlying datasets could be well controlled. Figure 3 depicts the averaged power (in %), i.e., the proportions of correctly identified outlying datasets among all truly outlying datasets under the same settings. We observe that most truly outlying datasets could be discovered by our method with the type I error rate being well controlled. Moreover, the averaged power increases as more measurements are collected.

Table 2: Averaged sizes (in %) under different configurations of (p, n_0, ρ) when the covariates and noises 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

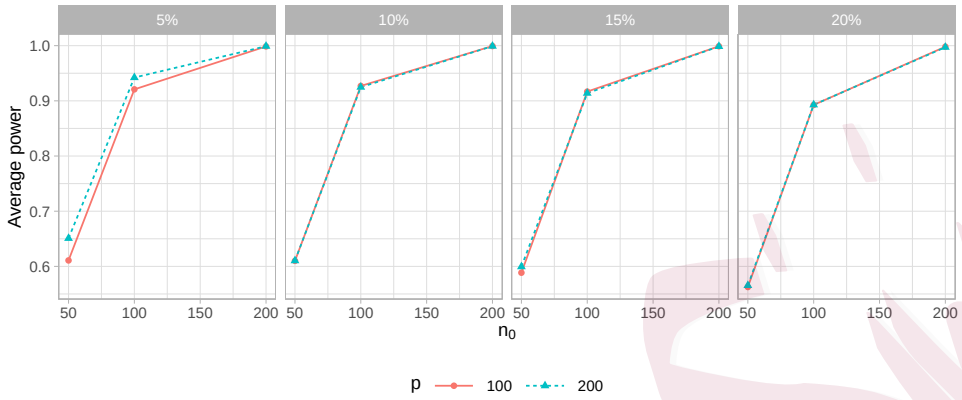

Figure 3: Averaged power (in %) under different configurations of (p, n_0, ρ) when the covariates and noises are both normally distributed.

5.3 Real data analysis

We use the Melbourne Housing data as an example to illustrate the application of our proposed testing procedure and heterogeneity identification algorithm. The dataset consists of transaction details of properties in 2016 in Melbourne, Victoria, Australia. We focus on the transactions with property type "House" and the suburbs with at least 50 transactions in 2016. After data pre-processing, the subset contains 238 suburbs with $33,973$ transactions. We are interested in the relative change of housing price in 2016 compared to 2015 explained by 8 covariates, including the month of transaction, latitude, longitude, number of bedrooms, number of bathrooms, land size, building area and built year of the property. For each suburb, we define the growth rate of housing price as the excess rate of the

sold price of the property in each transaction compared to 2015's median price within this suburb, and this variable serves as the response.

By treating each suburb as a group, we obtain 238 parallel datasets. We are interested in whether the 8 covariates contribute differently on housing price growth rate across different suburbs. After standardizing the 8 covariates, we implement our proposed heterogeneity test on the 238 suburbs to test whether they share the same regression function. It turns out that the p-value of the heterogeneity test is less than 0.0001 when using $10,000$ bootstrap iterations, which indicates that there are significant evidence against that the 8 covariates have same contribution on the relative change of housing price across different suburbs.

With the homogeneity hypothesis being rejected, we are interested in identifying the outlying suburbs. In order to avoid unbalanced sample sizes across different suburbs, we randomly sample 50 transactions from each suburb. By carrying out the outlier detection algorithm proposed in Section 4, there are 20 out of 238 suburbs detected as outliers with significance level $\alpha = 5\%$. Figure 4 shows whether a suburb is detected as an outlier (filled in yellow), and the suburbs in gray are those with less than 50 transactions in 2016.

After the 20 outlying suburbs are detected, it is of interest to explore

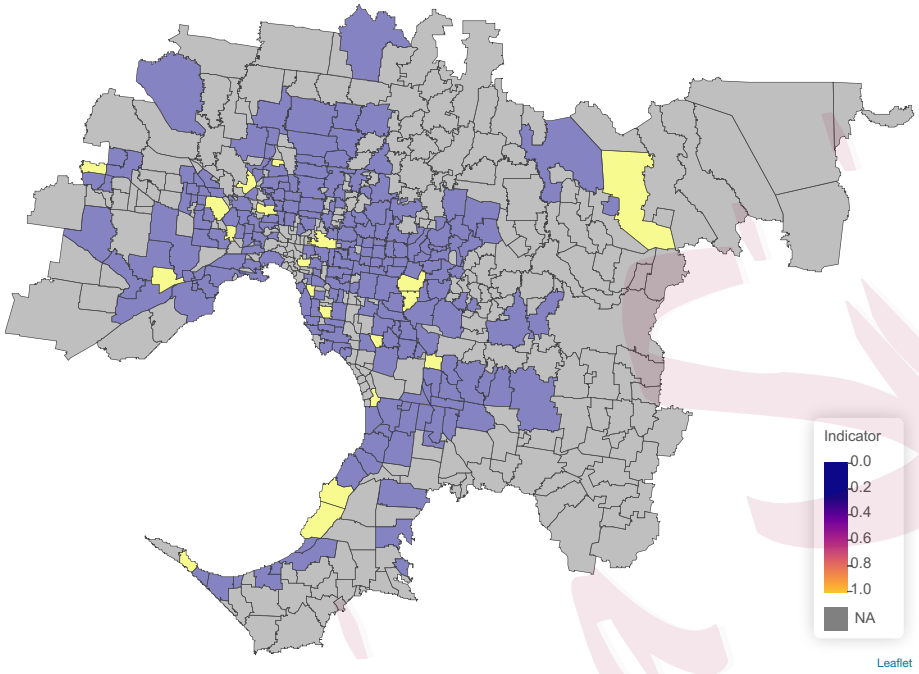

Figure 4: Plot of all suburbs with detected outlying ones being indicated in yellow and the majorities in blue. The suburbs in gray are those with less than 50 transactions in 2016.

how they are different from the majorities (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 dataset including all 20 outlying suburbs as O_1, and the dataset including 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 is served as the training set from the majority, and M_3 and O_1 are two test sets from the majority and outlier, respectively. A random

forest model is trained on M_2 and then used to do prediction on M_3 and O_1. Averaged from 100 iterations, the R^2 of the random forest modeling is 69.5%, which indicates reasonable fitting. The average mean squared error (AMSE) for M_3 is 0.1249, while the AMSE is 0.1392 for O_1, which is 11.45% higher than that of M_3. This simulation result implies the risk of getting a larger error when using the model fitted from the majority to do prediction on the outlying suburbs. In summary, the outlier detection result could provide guidance for further data modeling and the analysis of specific outlying suburbs.

6. Concluding remarks

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

Our development builds on the assumption that the covariates are i.i.d. across datasets. However, domain shifts could happen in practical applications, that is, the distributions of covariates may be different across datasets. Of primary is to test whether domain shifts occur. If the covariates have densities, Zhan and Hart (2014) proposed a kernel smoothing-based procedure to test equality of a large number of densities. We believe that similar projection averaging-based metrics for the departure of two distributions (e.g., Kim et al. (2020)) in conjunction with the proposed large-scale testing scheme is still applicable to achieve this purpose. If domain shifts indeed occur, the proposed method cannot be directly used to test equality of regression functions (cf. Eq. (1.2)). How to develop model-free and data-adaptive testing procedures that allow for different distributions of covariates warrants further research. Recently, Xiao et al. (2021) studied individual regression heterogeneity in panel data and allowed the data to exhibit outliers, which shed some lights to this discussion. On the other hand, both the difference of regression functions and the domain shift are sources of heterogeneity of parallel datasets. It can be challenging if not in-feasible to identify the real cause of the heterogeneity in a nonparametric setup, which is left as future work.

The literature has witnessed the expansion of distance based approaches,

e.g., the distance correlation (Szkely et al., 2007) and the maximum mean discrepancy (Gretton et al., 2012). The projection-based method proposed in this paper, even from a different point of view, share a similar idea of avoiding the curse of dimensionality. Thus, it is of great interest to study the relationship between the distance-based method and our proposed projection-based approach, and we leave it to future investigation.

Supplementary Materials

The supplementary materials consist of the proofs of main results in the paper, an algorithm for identification of outlying datasets, 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, the associate editor, and two referees for their constructive comments and suggestions, which improved the quality of the article greatly. Wang was supported by NNSF of China Grant 11901314. Zou was supported by NNSF of China Grants (11925106, 11931001 and 11971247), NSF of Tianjin Grant 18JCJQJC46000, and the 111Project B20016.

References

Baltagi, B. H., J. Hidalgo, and Q. Li (1996). A nonparametric test for poolability using panel data. *Journal of Econometrics 75*(2), 345–367.

Barras, L., O. Scaillet, and R. Wermers (2010). False discoveries in mutual fund performance: Measuring luck in estimated alphas. *The journal of finance 65*(1), 179–216.

Borgwardt, K. M., A. Gretton, M. J. Rasch, H. P. Kriegel, B. Scholkopf, and A. J. Smola (2006). Integrating structured biological data by Kernel Maximum Mean Discrepancy. *Bioinformatics 22*(14), e49–e57.

Cai, L. and S. Wang (2021). Global statistical inference for the difference between two regression mean curves with covariates possibly partially missing. *Statistical Papers 62*(6), 2573–2602.

Chiou, J.-M. and P.-L. Li (2007). Functional clustering and identifying substructures of longitudinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69*(4), 679–699.

Cuesta-Albertos, J. A., E. García-Portugués, M. Febrero-Bande, and W. González-Manteiga (2019). Goodness-of-fit tests for the functional

linear model based on randomly projected empirical processes. *The Annals of Statistics 47*(1), 439–467.

Escanciano, J. C. (2006). A Consistent Diagnostic Test for Regression Models Using Projections. *Econometric Theory 22*(06), 1030–1051.

González-Manteiga, W. and R. M. Crujeiras (2013). An updated review of Goodness-of-Fit tests for regression models. *TEST 22*(3), 361–411.

Gretton, A., K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola (2012). A kernel two-sample test. *The Journal of Machine Learning Research 13*(1), 723–773.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Ann. Math. Statist. 19*, 293–325.

Ke, Y., J. Li, and W. Zhang (2016). Structure identification in panel data analysis. *Annals of Statistics 44*(3), 1193–1233.

Kim, I., S. Balakrishnan, L. Wasserman, et al. (2020). Robust multivariate nonparametric tests via projection averaging. *Annals of Statistics 48*(6), 3417–3441.

Koul, H. L. and F. Li (2020). Comparing two nonparametric regres-

sion curves in the presence of long memory in covariates and errors. *Metrika 83*(4), 499–517.

Lavergne, P. and V. Patilea (2008). Breaking the curse of dimensionality in nonparametric testing. *Journal of Econometrics 143*(1), 103–122.

Neumeyer, N. and H. Dette (2001). Nonparametric analysis of covariance. *The Annals of Statistics 29*(5), 1361–1400.

Neumeyer, N. and H. Dette (2003). Nonparametric comparison of regression curves: an empirical process approach. *Ann. Statist. 31*(3), 880–920.

Pardo-Fernández, J. C., I. Van Keilegom, and W. González-Manteiga (2007). Testing for the equality of k regression curves. *Statist. Sinica 17*(3), 1115–1137.

Patilea, V., C. Sánchez-Sellero, and M. Saumard (2016). Testing the predictor effect on a functional response. *Journal of the American Statistical Association 111*(516), 1684–1695.

Qiu, P. and D. Xiang (2014). Univariate Dynamic Screening System: An Approach For Identifying Individuals With Irregular Longitudinal Behavior. *Technometrics 56*(2), 248–260.

Sen, P. K. (1977). Some Invariance Principles Relating to Jackknifing and

Their Role in Sequential Analysis. *The Annals of Statistics 5*(2), 316–329.

Srihera, R. and W. Stute (2010). Nonparametric comparison of regression functions. *Journal of Multivariate Analysis 101*(9), 2039–2059.

Szkely, G. J., M. L. Rizzo, and N. K. Bakirov (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics 35*(6), 2769–2794.

Tang, L. and P. X. K. Song (2016). Fused Lasso Approach in Regression Coefficients Clustering – Learning Parameter Heterogeneity in Data Integration. *Journal of Machine Learning Research 17*(113), 1–23.

Vogt, M. and O. Linton (2017). Classification of non-parametric regression functions in longitudinal data models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 79*(1), 5–27.

Wang, G., Z. Wang, and C. Zou (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*(1), 133–148.

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

Zhan, D. and J. D. Hart (2014). Testing equality of a large number of densities. *Biometrika 101*(2), 449–464.

Zhu, L., K. Xu, R. Li, and W. Zhong (2017). Projection correlation between two random vectors. *Biometrika 104*(4), 829–843.

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

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

Academy of Statistics and Interdisciplinary Sciences, East China Normal University, Shanghai 200062, China

E-mail: ghwang.nk@gmail.com

School of Statistics and Data Science, Nankai University, Tianjin 300071, China

E-mail: nk.chlzou@gmail.com