Submitted to the Annals of Statistics

CHANGE-POINT DETECTION IN MULTINOMIAL DATA WITH A LARGE NUMBER OF CATEGORIES

By Guanghui Wang[*], Changliang Zou[*] and Guosheng Yin[†]

Nankai University[*] and The University of Hong Kong[†]

We consider a sequence of multinomial data for which the probabilities associated with the categories are subject to abrupt changes of unknown magnitudes at unknown locations. When the number of categories is comparable to or even larger than the number of subjects allocated to these categories, conventional methods such as the classical Pearson's chi-squared test and the deviance test may not work well. Motivated by high-dimensional homogeneity tests, we propose a novel change-point detection procedure that allows the number of categories to tend to infinity. The null distribution of our test statistic is asymptotically normal and the test performs well with finite samples. The number of change-points is determined by minimizing a penalized objective function based on segmentation, and the locations of the change-points are estimated by minimizing the objective function with the dynamic programming algorithm. Under some mild conditions, the consistency of the estimators of multiple change-points is established. Simulation studies show that the proposed method performs satisfactorily for identifying change-points in terms of power and estimation accuracy, and it is illustrated with an analysis of a real data set.

1. Introduction. Change-point detection plays a critical role in data processing, modeling, estimation, and inference. Although most of the literature focuses on continuous data, in many data generation and collection processes, the observations either are measured on a discrete scale or naturally have some categorical structures. For such categorical data, there are rather limited approaches to change-point detection (Braun, Braun and Müller, 2000). The standard procedure is to apply binary segmentation and perform homogeneity tests on two contiguous samples under multinomial assumptions (Srivastava and Worsley, 1986; Horváth and Serbinowska, 1995). Classical methods, such as Pearson's chi-squared test and the deviance test, work well when each category contains sufficient amount of data. However, in modern applications, it is possible that the number of categories is comparable to or even larger than the number of subjects. For example, in

MSC 2010 subject classifications: Primary 62H15, secondary 62H12

Keywords and phrases: Asymptotic normality, Categorical data, High-dimensional homogeneity test, Multiple change-point detection, Sparse contingency table

the digitized text era, the word composition in different corpuses collected over time often experiences multiple abrupt changes. One may be interested in detecting such change-points that split the text data into segments for gaining more insights. The number of words can be very large, while the count for each word is often small or even zero. As a result, classical test statistics are typically not well-defined due to sparse contingency tables and the asymptotic theory developed for a fixed number of categories is generally not applicable.

In a sequence of multinomial data with the number of categories tending to infinity, we are interested in detecting changes in the probabilities associated with the categories over time. Specifically, we collect T independent observations with p possible outcomes, $\mathbf{X}_t = (X_{t1}, \ldots, X_{tp})^\top, t = 1, \ldots, T$. We assume that $\mathbf{X}_t$ follows a multinomial distribution, $\mathbf{X}_t \sim \text{Multi}(n_t, \mathbf{q}^{(t)})$, where n_t trials are conducted at time point t and the probabilities of outcomes $\mathbf{q}^{(t)} = (q_1^{(t)}, \ldots, q_p^{(t)})^\top$ satisfy $\sum_{j=1}^p q_j^{(t)} = 1$. Following the modern terminology of "large p, small n" problems (Chen and Qin, 2010), we use p to denote the number of outcomes which can be very large, i.e., $p \to \infty$. We consider the change-point model,

$$\mathbf{X}_t \sim \begin{cases} \text{Multi}(n_t, \mathbf{q}_0), & \text{for } t = 1, \ldots, \tau^*, \\ \text{Multi}(n_t, \mathbf{q}_1), & \text{for } t = \tau^* + 1, \ldots, T, \end{cases} \tag{1}$$

where $\tau^* > 0$ is an unknown change-point and $\mathbf{q}_l = (q_{l1}, \ldots, q_{lp})^\top$ for $l = 0, 1$. Our goal is to test whether there exists a change-point, with $H_0 : \tau^* = T$ versus $H_1 : \tau^* < T$, and to further estimate τ^* if H_0 is rejected.

This change-point detection problem is essentially related to a two- or multi-sample comparison with categorical data, for which a homogeneity test is typically used to examine whether all the T ($T \geq 2$) multinomial distributions are the same. Toward this goal, Pearson's chi-squared statistic (Agresti, 2013) can be constructed,

$$K_T = \sum_{t=1}^{T} \sum_{j=1}^{p} \frac{(X_{tj} - n_t \sum_{t=1}^{T} X_{tj}/N)^2}{n_t \sum_{t=1}^{T} X_{tj}/N},$$

where X_{tj} is the j-th component of $\mathbf{X}_t$ and $N = \sum_{t=1}^{T} n_t$. Under the null hypothesis $H_0 : \mathbf{q}_1 = \cdots = \mathbf{q}_T$, K_T follows a $\chi^2_{(p-1)(T-1)}$ distribution for a fixed p as $N \to \infty$. When we allow $p \to \infty$, in the context of one-sample homogeneity test, Holst (1972) and Morris (1975) developed asymptotic theory for Pearson's chi-squared test. Moderate and large deviation theorems for Pearson's chi-squared statistic and the likelihood ratio statistic in multinomial distributions are given in Kallenberg (1985). When all $p, n_1, \ldots, n_T \to \infty$,

K_T is related to the class of multi-dimensional *decomposable statistics* whose asymptotic normality after suitable normalization is also established; for example, see Ivchenko and Levin (1976) and Bykov and Ivanov (1991). More recently, Baranov and Baranov (2005) considered the T-sample homogeneity problem. However, these existing methods are not readily applicable to our change-point problem in (1) as detailed in the following. First, the test statistic K_T is not well-defined if some category does not contain any observation when p is large but N is small. Further, it is difficult to verify the imposed conditions, and the asymptotic result only allows p to grow at a linear rate of n_t. Second, it is required to estimate certain normalizing parameters which involve complicated asymptotic expansions of mixed moments of Poisson distributed variables. The normality property with plugged-in estimators is not guaranteed from asymptotic viewpoints. Third, the theory of decomposable statistics is not directly applicable, when the number of observations T in (1) diverges to infinity.

To overcome these drawbacks, we develop a novel testing procedure that is capable of accommodating large p. Based on the martingale central limit theorem, the proposed test statistic is shown to be asymptotically normal. Our method includes the two- and multi-sample homogeneity tests as special cases. Furthermore, we form an objective function based on segmentation when searching for multiple change-points, and determine the number of change-points by minimizing its penalized version. The locations of the change-points can be estimated via dynamic programming in conjunction with utilization of the intrinsic order structure of the objective function.

The remainder of this article is organized as follows. In Section 2, we present the new test statistic and its theoretical properties. In Section 3, we develop the estimation procedure for multiple change-points. Section 4 provides extensive simulation studies and a real data example as an illustration. Section 5 concludes with some remarks, and all technical proofs and additional numerical studies are delineated in the Supplementary Material.

2. Change-point test and estimation.

2.1. *Test statistic.* We are interested in testing the null hypothesis $H_0 : \mathbf{X}_t \sim \text{Multi}(n_t, \mathbf{q}_0)$, for $t = 1, \ldots, T$, against the alternative in (1) with $\tau^* < T$. We allow $p \to \infty$ and consider the triangular arrays $\mathbf{q}_l = (q_{l1}, \ldots, q_{lp})^\top$ for $l = 0, 1$, where we omit its dependence on p for simplicity. Let $N = \sum_{t=1}^{T} n_t \to \infty$ as $p \to \infty$, while T can either be fixed or diverge to infinity.

If τ is the true change-point ($\tau < T$), it is equivalent to testing whether the two groups, segmented by τ, come from the same multinomial distribution. Let $\mathbf{Z}_{0\tau} = \sum_{t=1}^{\tau} \mathbf{X}_t \sim \text{Multi}(N_{0\tau}, \mathbf{q}_0)$ and $\mathbf{Z}_{1\tau} = \sum_{t=\tau+1}^{T} \mathbf{X}_t \sim$

Multi($N_{1\tau}, \mathbf{q}_1$), where $N_{0\tau} = \sum_{t=1}^{\tau} n_t$ and $N_{1\tau} = \sum_{t=\tau+1}^{T} n_t$. Based on the L_2-norm, an intuitive test statistic can be constructed as

$$L_\tau = \sum_{j=1}^{p} \frac{N_{0\tau}N_{1\tau}}{N} \left(\frac{Z_{0\tau j}}{N_{0\tau}} - \frac{Z_{1\tau j}}{N_{1\tau}} \right)^2, \tag{2}$$

where $Z_{0\tau j}$ and $Z_{1\tau j}$ are the j-th components of $\mathbf{Z}_{0\tau}$ and $\mathbf{Z}_{1\tau}$, respectively. This statistic is similar to Pearson's chi-squared statistic for testing the homogeneity of two multinomial samples $\mathbf{Z}_{0\tau}$ and $\mathbf{Z}_{1\tau}$, which is given by

$$K_{2,\tau} = \sum_{j=1}^{p} \frac{N_{0\tau}N_{1\tau}}{N} \left(\frac{Z_{0\tau j}}{N_{0\tau}} - \frac{Z_{1\tau j}}{N_{1\tau}} \right)^2 \left(\frac{\sum_{t=1}^{T} X_{tj}}{N} \right)^{-1}. \tag{3}$$

In contrast, L_τ in (2) removes the component-wise standardization terms $\hat{q}_j \equiv \sum_{t=1}^{T} X_{tj}/N$, $j = 1, \ldots, p$, to circumvent the cases with $\hat{q}_j = 0$ in the large p but small N situation, such that L_τ is always well-defined. Moreover, by removing such terms, it can further relax the dimensionality and allow p to grow at a faster rate than the sample size. On the other hand, removing the $\hat{q}_j$'s is reasonable when they are of similar order in magnitudes. In the literature, a common assumption is that all the proportions are spread-out and diminishing, i.e., $\max_{1\le j\le p} q_{0j} \to 0$ as $p \to \infty$, where q_{lj} is the j-th component of $\mathbf{q}_l$; see for example, Holst (1972), Morris (1975) and Baranov and Baranov (2005). However, in practice, there may be spikes at certain proportions if a large number of categories are involved, say some of the $\hat{q}_j$'s are large relative to others. For example, in the modern e-commerce, there are always best-sellers among similar products under certain subcategories, whose sales (reflecting buyers' tendency) are much more outstanding than nonpopular ones. In the word composition in a writer's work, function words and certain content words, such as pronouns, could appear more frequently than others in the writing.

To strike a balance between L_τ and $K_{2,\tau}$, we replace the assumption that $\max_{1\le j\le p} q_{0j} \to 0$ by a more relaxed one.

(A1) For $l = 0, 1$, there exists a set $\mathscr{B}_l \subset \{1, \ldots, p\}$ such that $\max_{j\in\mathscr{B}_l} q_{lj}a_p \to 0$ with $a_p^{-1} = O(1)$ as $p \to \infty$. Further let $\mathscr{A}_l = \{1, \ldots, p\}\backslash\mathscr{B}_l$ be the complement of $\mathscr{B}_l$ and assume that $\min_{j\in\mathscr{A}_l} q_{lj}a_p > \varepsilon$ for some $\varepsilon > 0$ as $p \to \infty$.

Assumption (A1) divides p categories into two disjoint subsets $\mathscr{A}_l$ and $\mathscr{B}_l$ according to the magnitudes of their corresponding probabilities, either "significant" or "diminishing", while changes may occur in either subset (more

precisely, on some categories in either set). It requires that these two subsets can be separated by a_p at the population level. When a_p is bounded away from zero, there are finite elements in $\mathscr{A}_l$ and $\max_{j\in\mathscr{B}_l} q_{lj} \to 0$.

Change-point detection using all the proportions in L_τ may cause difficulty in the interpretation and degrade the performance due to the fact that $\sum_{j\in\mathscr{A}_0} \frac{N_{0\tau}N_{1\tau}}{N}\left(\frac{Z_{0\tau j}}{N_{0\tau}} - \frac{Z_{1\tau j}}{N_{1\tau}}\right)^2$ may dominate $\sum_{j\in\mathscr{B}_0} \frac{N_{0\tau}N_{1\tau}}{N}\left(\frac{Z_{0\tau j}}{N_{0\tau}} - \frac{Z_{1\tau j}}{N_{1\tau}}\right)^2$. To improve the detection power in high-dimensional settings, screening methods (Fan and Lv, 2008) can be used to select potential interesting features for further analysis. We first separate out the proportions not less than the order $O(a_p^{-1})$ from $\{1,\dots,p\}$, possibly before and after the change, i.e., $\mathscr{A} = \mathscr{A}_0 \cup \mathscr{A}_1$, by using

$$\hat{\mathscr{A}} = \{j : \hat{q}_j a_p > C\varepsilon\} \text{ for some } C > 0. \tag{4}$$

We then construct two individual test statistics on both $\hat{\mathscr{A}}$ and $\{1,\dots,p\}\backslash\hat{\mathscr{A}} \equiv \hat{\mathscr{B}}$, and finally combine these two parts together.

Let

$$L_{\tau j} = \frac{N_{0\tau}N_{1\tau}}{N}\left(\frac{Z_{0\tau j}}{N_{0\tau}} - \frac{Z_{1\tau j}}{N_{1\tau}}\right)^2 \text{ and } R_{\tau j} = L_{\tau j}/\hat{q}_j.$$

We propose to run over all possible change-points as

$$\begin{aligned} Q_{p,\hat{\mathscr{A}}} &= \sum_{\tau\in\mathcal{T}}\sum_{j\in\hat{\mathscr{B}}}\left(L_{\tau j} - L_{\tau j}^{(0)}\right) + e_p \mathrm{I}\Big(\max_{\tau\in\mathcal{T}}\max_{j\in\hat{\mathscr{A}}} R_{\tau j} > r_p\Big) \\ &\equiv S_{p,\hat{\mathscr{A}}} + E_{p,\hat{\mathscr{A}}}, \end{aligned} \tag{5}$$

where $\mathrm{I}(\cdot)$ is the indicator function, and

$$L_{\tau j}^{(0)} = \frac{N_{0\tau}N_{1\tau}}{N}\left(\frac{Z_{0\tau j}}{N_{0\tau}^2} + \frac{Z_{1\tau j}}{N_{1\tau}^2}\right)$$

is a bias-correction term to make the expectation of $S_{p,\hat{\mathscr{A}}}$ negligible compared to $\sqrt{\mathrm{Var}(S_{p,\hat{\mathscr{A}}})}$. In $E_{p,\hat{\mathscr{A}}}$, the second term of (5), e_p is a large enough constant and r_p is chosen to be slightly larger than the maximum noise level such that $E_{p,\hat{\mathscr{A}}}$ is zero under H_0 with high probability but diverges quickly under H_1 with some $j \in \mathscr{A}$. Note that for $\hat{\mathscr{A}}$ and $\hat{\mathscr{B}}$ we use the max-norm and L_2-norm based test statistics, respectively. It is widely acknowledged that the max-norm test is more suitable for sparse and strong signals, whereas the L_2-norm test is for dense but faint signals (Chen and Qin, 2010; Fan, Liao and Yao, 2015). Similar to the power-enhancement test statistic

proposed by Fan, Liao and Yao (2015), the advantage of using $Q_{p,\hat{\mathscr{A}}}$ would be more transparent by examining its asymptotic behavior in Section 2.2. We use the trimmed summation, say $\mathcal{T} = [\lceil a(T-1) \rceil, \lceil b(T-1) \rceil]$ with fixed constants $0 < a < b < 1$ (for example, $a = 0.1$ and $b = 0.9$), to avoid some technical difficulties when $T \to \infty$, where $\lceil x \rceil$ denotes the smallest integer not less than x; for example, see Perron and Vogelsang (1992).

Remark 1 Conventionally, $M_{p,\hat{\mathscr{A}}} \equiv \max_{\tau \in \mathcal{T}} \sum_{j \in \hat{\mathscr{B}}} L_{\tau j}$ is used as the change-point test statistic rather than $S_{p,\hat{\mathscr{A}}}$ (Csörgö and Horváth, 1997). However, it is recognized the rate of convergence of the maximum statistic is slow; see Section 1.3 of Csörgö and Horváth (1997). Consequently, the asymptotic quantiles do not work well with the typical values of n_t and T in real applications. In contrast, $S_{p,\hat{\mathscr{A}}}$ (also $Q_{p,\hat{\mathscr{A}}}$) is asymptotically normal under some mild conditions and thus can greatly facilitate the construction of the test. Our numerical analysis demonstrates that the power of $S_{p,\hat{\mathscr{A}}}$ is at least comparable to that of $M_{p,\hat{\mathscr{A}}}$. In fact, $M_{p,\hat{\mathscr{A}}}$ and $S_{p,\hat{\mathscr{A}}}$ can be respectively viewed as the CUSUM and Shiryaev–Roberts procedures (Srivastava and Wu, 1993).

When the null hypothesis is rejected, the change-point τ^* can be naturally estimated by

$$\hat{\tau}^* = \begin{cases} \arg\max_{\tau \in \mathcal{T}} \max_{j \in \hat{\mathscr{A}}} R_{\tau j}, & \text{if } E_{p,\hat{\mathscr{A}}} = e_p, \\ \arg\max_{\tau \in \mathcal{T}} \sum_{j \in \hat{\mathscr{B}}} (L_{\tau j} - L_{\tau j}^{(0)}), & \text{otherwise.} \end{cases} \tag{6}$$

Under certain conditions, we can establish the consistency of this estimator.

2.2. *Null distribution of the test statistic.* We begin with the separation consistency of our procedure (4).

THEOREM 1. *Suppose that Assumption (A1) holds, and if* $N a_p^{-2} (\log a_p)^{-1} \to \infty$ *as* $(p, N) \to \infty$, *then*

(i) under H_0, $\Pr(\hat{\mathscr{A}} = \mathscr{A}_0) \to 1$ *for any* $0 < C < 1$.
(ii) under H_1, $\Pr(\hat{\mathscr{A}} = \mathscr{A}_0 \cup \mathscr{A}_1) \to 1$ *for any* $0 < C < \min(\kappa_0, 1 - \kappa_0)/2$, *where* κ_0 *is the limit of* $N_{0\tau^*}/N$, *i.e.,* $N_{0\tau^*}/N \to \kappa_0$ *as* $N \to \infty$.

By Theorem 1, we conclude that $\Pr(Q_{p,\hat{\mathscr{A}}} \le x) = \Pr(Q_{p,\mathscr{A}} \le x) + o(1)$ for any x, and hence it suffices to study the asymptotic behavior of $Q_{p,\mathscr{A}}$. The following assumptions are needed for further theoretical development.

(A2) There exist $0 < \underline{\rho}, \overline{\rho} < \infty$, such that $\underline{\rho} \le N_{0\tau}/N_{1\tau} \le \overline{\rho}$ for any $\tau \in \mathcal{T}$.
(A3) For $l = 0, 1$, $N^{-2} (\sum_{j \in \mathscr{B}} q_{lj}^2)^{-1} \to 0$ and $N^{-1} (\sum_{j \in \mathscr{B}} q_{lj}^3)(\sum_{j \in \mathscr{B}} q_{lj}^2)^{-2} \to 0$, as $(p, N) \to \infty$.

(A4) Assume $(\sum_{j\in\mathscr{B}} q_{0j}^4)(\sum_{j\in\mathscr{B}} q_{0j}^2)^{-2} \to 0$, as $p \to \infty$.
(A5) Assume $Na_p^{-2}\{\log(Ta_p)\}^{-1} \to \infty$, as $(p, N) \to \infty$.

Remark 2 Assumption (A2) is a technical condition that requires the pre- and post-τ sample sizes to be comparable. Assumptions (A3) and (A4) are mild. For example, if $q_{lj} \asymp p^{-1}$ for $j = 1, \ldots, p; l = 0, 1$, i.e., there exist $0 < \underline{C}, \overline{C} < \infty$ such that $\underline{C} \le pq_{lj} \le \overline{C}$, then $\sum_{j=1}^p q_{lj}^r \asymp p^{-r+1}$ for $r = 2, 3, 4$. Consequently, (A3)–(A4) are satisfied if $p/N^2 \to 0$ which is faster than the linear rate, $p/N = O(1)$, as in Baranov and Baranov (2005). Assumption (A5) imposes a condition on a_p, which holds trivially when a_p is bounded away from zero.

THEOREM 2. *Suppose that H_0 and Assumptions (A1)–(A2) hold.*

(i) The expectation and variance of $S_{p,\mathscr{A}}$ are given by

$$\mathrm{E}(S_{p,\mathscr{A}}) = o\Big\{\sqrt{\mathrm{Var}(S_{p,\mathscr{A}})}\Big\},$$
$$\mathrm{Var}(S_{p,\mathscr{A}}) = \Big(2\sum_{\tau<\tau'} \frac{N_{0\tau}N_{1\tau'}}{N_{0\tau'}N_{1\tau}} + \Lambda_T\Big) \cdot 2\sum_{j\in\mathscr{B}} q_{0j}^2\{1 + o(1)\},$$

respectively, as $(p, N) \to \infty$, where $\Lambda_T = \lceil b(T-1)\rceil - \lceil a(T-1)\rceil + 1$.

(ii) Suppose further Assumptions (A3)–(A4) hold, then as $(p, N) \to \infty$,

$$\frac{S_{p,\mathscr{A}}}{\sqrt{\mathrm{Var}(S_{p,\mathscr{A}})}} \xrightarrow{\mathcal{D}} N(0, 1).$$

(iii) Suppose further Assumption (A5) holds, and if $r_p\{\log(Ta_p)\}^{-1} \to \infty$, then as $(p, N) \to \infty$,

$$\frac{Q_{p,\mathscr{A}}}{\sqrt{\mathrm{Var}(S_{p,\mathscr{A}})}} \xrightarrow{\mathcal{D}} N(0, 1).$$

Theorem 2 (ii) establishes the asymptotic normality of $S_{p,\mathscr{A}}$ under the null hypothesis and Theorem 2 (iii) reveals that $S_{p,\mathscr{A}}$ and $Q_{p,\mathscr{A}}$ would have the same asymptotic null behavior given an appropriate sequence of r_p. The proof is outlined in the Supplementary Material with key steps described as follows. In fact, the observations can be decomposed into $X_{tj} = \sum_{i=N_{0,t-1}+1}^{N_{0t}} Y_{ij}$ for $j = 1, \ldots, p,\ t = 1, \ldots, T$, with the convention of $N_{00} = 0$ and $N_{0T} = N$, where $\{(Y_{i1}, \ldots, Y_{ip})^\top\}_{i=1}^N$ are independent and follow the multinomial distribution, $\mathrm{Multi}\big(1, (q_{01}, \ldots, q_{0p})^\top\big)$. It can be shown that $S_{p,\mathscr{A}} - \mathrm{E}(S_{p,\mathscr{A}})$ is asymptotically equivalent to a martingale difference

sequence and consequently the assertion is proved by applying the martingale central limit theorem; e.g., see Corollary 3.1 of Hall and Heyde (1980). Note that $\sum_{j\in\mathscr{B}} L_{\tau j}$ essentially shares a form similar to the high-dimensional two-sample test statistic (Bai and Saranadasa, 1996; Chen and Qin, 2010). However, their results are not directly applicable to $S_{p,\mathscr{A}}$ because the unobservable variables Y_{ij}'s do not satisfy the data structure that the validity of asymptotic normality relies upon. In addition, the treatment on the summation of dependent statistics $L_{\tau j}$ for $\tau \in \mathcal{T}$ is nontrivial.

The variance of $S_{p,\mathscr{A}}$ depends on the unknown quantities $\sum_{j\in\mathscr{B}} q_{0j}^2$. Therefore, we need to find a ratio-consistent estimator of $\sum_{j\in\mathscr{B}} q_{0j}^2$ in order to use the asymptotic normality result in practice. Given $\mathscr{A}$, we propose to use

$$U_{N,\mathscr{A}} = \frac{N}{N-1}\sum_{j\in\mathscr{B}}\Big(\hat{q}_j^2 - \frac{1}{N}\hat{q}_j\Big), \tag{7}$$

for which the ratio-consistency property holds as shown in the following proposition.

PROPOSITION 1. *Suppose that H_0 and Assumption (A3) hold, then as $(p, N) \to \infty$,*

$$U_{N,\mathscr{A}} / \sum_{j\in\mathscr{B}} q_{0j}^2 \xrightarrow{\mathcal{P}} 1.$$

By Slutsky's theorem, we obtain that as $(p, N) \to \infty$,

$$\frac{Q_{p,\mathscr{A}}}{\sqrt{2c_{N,T}U_{N,\mathscr{A}}}} \xrightarrow{\mathcal{D}} N(0,1),$$

where $c_{N,T} = 2\sum_{\tau<\tau'} N_{0\tau}N_{1\tau'}/(N_{1\tau}N_{0\tau'}) + \Lambda_T$. As a result, we reject H_0 at an α level of significance if $Q_{p,\hat{\mathscr{A}}}\Big/\sqrt{2c_{N,T}U_{N,\hat{\mathscr{A}}}}$ exceeds z_α, where z_α is the upper αth quantile of the standard normal distribution.

2.3. *Consistency of the test and estimator.* We investigate the asymptotic behavior of our test under the alternative hypothesis, i.e., the one change-point model in (1). We consider a local alternative hypothesis with $\delta_j = q_{0j} - q_{1j}$ for $j = 1, \ldots, p$, and assume that the true change-point is not at the boundary, i.e., $\tau^* = \lceil \gamma(T-1) \rceil$ for $0 < \gamma < 1$. It is well recognized that change-point tests do not usually work well when the change-point is at the boundary (Chen and Gupta, 2000).

THEOREM 3. *Suppose that Assumptions (A1), (A2), and (A5) hold, $N_{0\tau^*}/N \to \kappa_0$, $r_p\{\log(Ta_p)\}^{-1} \to \infty$ and $e_pT^{-1} \to \infty$, as $(p, N) \to \infty$. If the shift sizes δ_j's satisfy either of the following two conditions:*

(i) $N\sum_{j\in\mathscr{B}}\delta_j^2\Big(\max_{l=0,1}\sum_{j\in\mathscr{B}}q_{lj}^2\Big)^{-1/2}\to\infty$,

(ii) $N\delta_{j'}^2 r_p^{-1}\to\infty$ *for some* $j'\in\mathscr{A}$,

then

$$\frac{Q_{p,\hat{\mathscr{A}}}}{\sqrt{2c_{N,T}U_{N,\hat{\mathscr{A}}}}}\xrightarrow{\mathcal{P}}\infty.$$

This theorem entails the rationale for the combination of $S_{p,\hat{\mathscr{A}}}$ and $E_{p,\hat{\mathscr{A}}}$. Suppose that a_p is bounded away from zero, and r_p is chosen as $\log T\log\log T$. When the signal under the alternative is dense, say the changes occur mainly in $\mathscr{B}$ such that condition (i) is satisfied, $Q_{p,\hat{\mathscr{A}}}$ is as powerful as $S_{p,\hat{\mathscr{A}}}$. For example, if $q_{lj}\asymp p^{-1}$ for $l=0,1$ and $j\in\mathscr{B}$, this result demonstrates our test has non-trivial power under the contiguous alternatives of $O(N^{-1}p^{-1/2})$ in terms of $\sum_{j\in\mathscr{B}}\delta_j^2$. On the other hand, in sparse alternatives where most of the proportions do not change over time but some of δ_j's are particularly large so that $N\delta_{j'}^2/(\log T\log\log T)\to\infty$ for some $j'\in\mathscr{A}$, $Q_{p,\hat{\mathscr{A}}}$ would also be powerful due to the dominance of $E_{p,\hat{\mathscr{A}}}$. In such situations, our proposed test is consistent against the contiguous alternative of order larger than $N^{-1/2}(\log T\log\log T)^{1/2}$ which is a nearly optimal rate for the change detection with a fixed p. The statistic $Q_{p,\hat{\mathscr{A}}}$ gains strength by borrowing information from the pre-separation and thus it is able to balance the detection between the sparse and dense signals.

A by-product of the proof of this theorem is the consistency of our change-point estimator defined in (6).

COROLLARY 1. *Suppose that Assumptions (A1)–(A3), and (A5) hold, and* $N_{0\tau^*}/N\to\kappa_0$ *as* $N\to\infty$.

(i) If there exists some $j'\in\mathscr{A}$ *such that* $N\delta_{j'}^2 r_p^{-1}\to\infty$, *then*

$$\Pr(|\hat{\tau}^*-\tau^*|<\zeta_{\mathscr{A},T})\to 1,$$

where $\zeta_{\mathscr{A},T}>0$ *satisfies that* $(N_{0,\tau^*\pm\zeta_{\mathscr{A},T}}-N_{0\tau^*})\delta_{j'}^2\to\infty$.

(ii) If all $\delta_j=0$ *for* $j\in\mathscr{A}$ *but* $N\sum_{j\in\mathscr{B}}\delta_j^2\Big(\max_{l=0,1}\sum_{j\in\mathscr{B}}q_{lj}^2\Big)^{-1/2}\to\infty$, *then*

$$\Pr(|\hat{\tau}^*-\tau^*|<\zeta_{\mathscr{B},T})\to 1,$$

where $\zeta_{\mathscr{B},T}>0$ *satisfies that*

$$(N_{0,\tau^*\pm\zeta_{\mathscr{B},T}}-N_{0\tau^*})\sum_{j\in\mathscr{B}}\delta_j^2\Big/\sqrt{\max_{l=0,1}\sum_{j\in\mathscr{B}}q_{lj}^2}\to\infty.$$

2.4. *A special case: two-sample homogeneity test.* The proposed $Q_{p,\hat{\mathscr{A}}}$ includes the two-sample homogeneity test as a special case,

$$H_0 : \mathbf{q}_0 = \mathbf{q}_1 \text{ versus } H_1 : \mathbf{q}_0 \neq \mathbf{q}_1,$$

where the two groups $\mathbf{X}_0 \sim \text{Multi}(n_0, \mathbf{q}_0)$ and $\mathbf{X}_1 \sim \text{Multi}(n_1, \mathbf{q}_1)$ are independent. The test statistic can be formulated as

$$Q_{p,\hat{\mathscr{A}}} = \sum_{j \in \mathscr{B}} \Big(L_j - L_j^{(0)} \Big) + e_p \mathrm{I}\Big(\max_{j \in \hat{\mathscr{A}}} \frac{L_j}{\hat{q}_j} > r_p \Big),$$

where $L_j = n_0 n_1 / N (X_{0j}/n_0 - X_{1j}/n_1)^2$, $L_j^{(0)} = X_{0j}/n_0^2 + X_{1j}/n_1^2$, $N = n_0 + n_1$, and $X_{0j}, X_{1j}, \hat{q}_j$ are the j-th component of $\mathbf{X}_0, \mathbf{X}_1$ and $(\mathbf{X}_0 + \mathbf{X}_1)/N$ respectively. A direct application of Theorem 2 yields the following corollary.

COROLLARY 2. *Suppose that H_0 and Assumptions (A1), (A3)–(A4) and (A5) with $T = 1$ hold, and $n_0/N \to \kappa_0 \in (0,1)$ as $N \to \infty$, then $Q_{p,\hat{\mathscr{A}}} / \sqrt{2U_{N,\hat{\mathscr{A}}}} \xrightarrow{\mathcal{D}} N(0,1)$ as $(p, N) \to \infty$, where $U_{N,\hat{\mathscr{A}}} = N/(N-1) \sum_{j \in \mathscr{B}} \big(\hat{q}_j^2 - N^{-1} \hat{q}_j \big)$.*

Chen and Zhang (2013) proposed a graph-based test for two-sample comparison with categorical data when the contingency table is sparsely populated. Their method utilizes similarity information on the sample space and thus may improve power in certain cases. Compared with our proposal, the graph-based test is more computationally intensive, and it requires permutation procedures because the asymptotic null distribution of the test statistic depends on some nuisance parameters that cannot be estimated easily.

3. Multiple change-point estimation. To extend the proposed method to multiple change-points, we assume

$$\mathbf{X}_t \sim \text{Multi}(n_t, \mathbf{q}_l), \ \tau_l^* < t \leq \tau_{l+1}^*, \ l = 0, 1, \ldots, L^*,$$

where L^* is the true number of change-points ($L^* \geq 1$), τ_l^*'s are the locations of these change-points with the convention of $\tau_0^* = 0$ and $\tau_{L^*+1}^* = T$, and $\mathbf{q}_l$ is the vector of probabilities of outcomes for segment $l+1$ satisfying $\mathbf{q}_l \neq \mathbf{q}_{l+1}$.

Intuitively, the binary segmentation for a single change-point discussed earlier can be applied recursively to detect multiple change-points. Although binary segmentation is computationally efficient and roughly linear with sample size, it only provides an approximate solution and may lead to poor estimation of the number and locations of multiple change-points; see

Fryzlewicz (2014) and the references therein for variants of binary segmentation. In contrast, we define an objective function based on segmentation and minimize its penalized version, which can be viewed as a global procedure (Killick, Fearnhead and Eckley, 2012).

We first generalize Assumption (A1) to the multiple change-points setting:

(B1) For any $l = 0, 1, \ldots, L^*$, there exists a nonempty set $\mathscr{B}_l \subset \{1, \ldots, p\}$ such that $\max_{j \in \mathscr{B}_l} q_{lj} \to 0$ as $p \to \infty$, where q_{lj} is the j-th component of $\mathbf{q}_l$. Further let $\mathscr{A}_l = \{1, \ldots, p\} \backslash \mathscr{B}_l$ be the complement of $\mathscr{B}_l$ and assume that $\min_{j \in \mathscr{A}_l} q_{lj} > \varepsilon$ for some $\varepsilon > 0$ as $p \to \infty$.

For ease of discussion, Assumption (B1) simply considers a_p bounded away from zero which may be the case of the most interest. We introduce $\tau^*_{\mathscr{A},1}, \ldots, \tau^*_{\mathscr{A},A^*}$ as all possible change-points at which changes could occur only in set $\mathscr{A}$. Within each range $(\tau^*_{\mathscr{A},a}, \tau^*_{\mathscr{A},a+1}]$, $a = 0, 1, \ldots, A^*$, with the convention of $\tau^*_{\mathscr{A},0} = 0$ and $\tau^*_{\mathscr{A},A^*+1} = T$, we then let $\tau^*_{\mathscr{B},a,1}, \ldots, \tau^*_{\mathscr{B},a,B^*_a}$ be the remaining possible change-points at which changes could occur only in set $\mathscr{B}$. Note that $A^* + \sum_{a=0}^{A^*} B^*_a = L^*$ and we allow that A^* and B^*_a's could be 0. In line with the argument in Section 2, the penalized objective functions for $\mathscr{A}$ and $\mathscr{B}$ should not be the same. Define

$$\hat{\mathscr{A}} = \{j : \hat{q}_j > C\varepsilon \text{ for some } C > 0\} \text{ and } \hat{\mathscr{B}} = \{1, 2, \ldots, p\} \backslash \hat{\mathscr{A}},$$

as (4) in Section 2. We propose a two-step detection procedure as follows.

Step 1: For a candidate set of A change-points, $\tau_1 < \cdots < \tau_A$, we define the objective function,

$$\mathcal{S}_{\hat{\mathscr{A}}}(\tau_1, \ldots, \tau_A) = \sum_{a=0}^{A} \sum_{t=\tau_a+1}^{\tau_{a+1}} \sum_{j \in \hat{\mathscr{A}}} \frac{\{X_{tj} - n_t \bar{X}_j(\tau_a, \tau_{a+1})\}^2}{n_t \hat{q}_j},$$

where $\tau_0 = 0$, $\tau_{A+1} = T$ and $\bar{X}_j(\tau_a, \tau_{a+1}) = \sum_{t=\tau_a+1}^{\tau_{a+1}} X_{tj} / \sum_{t=\tau_a+1}^{\tau_{a+1}} n_t$. The A change-points τ_a's can then be estimated by

$$(\hat{\tau}_{A,1}, \ldots, \hat{\tau}_{A,A}) = \underset{\tau_1 < \cdots < \tau_A}{\operatorname{argmin}} \mathcal{S}_{\hat{\mathscr{A}}}(\tau_1, \ldots, \tau_A).$$

To determine A, we observe that $\mathcal{S}_{\hat{\mathscr{A}}}(\hat{\tau}_{A,1}, \ldots, \hat{\tau}_{A,A})$ is a nonincreasing function in A. Hence we can add a penalty for large A to strike a balance between the value of the objective function and the number of change-points. We determine A by minimizing

$$\mathcal{S}^{\mathrm{Pen}}_A = \mathcal{S}_{\hat{\mathscr{A}}}(\hat{\tau}_{A,1}, \ldots, \hat{\tau}_{A,A}) + A\xi_{p,N}, \tag{8}$$

with respect to $A \geq 0$, where $\xi_{p,N}$ is chosen to be slightly larger than the maximum variation level (no change) so that $\mathcal{S}_{\mathscr{A}}(\hat{\tau}_{A,1}, \ldots, \hat{\tau}_{A,A})$ would be dominated by $A\xi_{p,N}$ under overfitting models with high probability. We denote the resulting estimators as $\hat{A}$ and $(\hat{\tau}_{\hat{A},1}, \ldots, \hat{\tau}_{\hat{A},\hat{A}})$ if $\hat{A} > 0$.

Step 2: The whole sampling range can then be divided into $\hat{A}+1$ subintervals, $(\hat{\tau}_{\hat{A},a}, \hat{\tau}_{\hat{A},a+1}]$, $a = 0, 1, \ldots, \hat{A}$ with the convention of $\hat{\tau}_{\hat{A},0} = 0$ and $\hat{\tau}_{\hat{A},\hat{A}+1} = T$. In the $(a+1)$-th subinterval, we introduce B_a candidate change-points $\tau_1^{(a)} < \cdots < \tau_{B_a}^{(a)}$ and consider the following objective function,

$$\mathcal{S}_{\hat{\mathscr{B}}}\left(\tau_1^{(a)}, \ldots, \tau_{B_a}^{(a)}\right) = \sum_{b=0}^{B_a} \sum_{t=\tau_b^{(a)}+1}^{\tau_{b+1}^{(a)}} \sum_{j \in \hat{\mathscr{B}}} \left\{X_{tj} - n_t \bar{X}_j\left(\tau_b^{(a)}, \tau_{b+1}^{(a)}\right)\right\}^2 \Big/ n_t,$$

where $\tau_0^{(a)} = \hat{\tau}_{\hat{A},a}$ and $\tau_{B_a+1}^{(a)} = \hat{\tau}_{\hat{A},a+1}$. Similarly, we let

$$\left(\hat{\tau}_{B_a,1}^{(a)}, \ldots, \hat{\tau}_{B_a,B_a}^{(a)}\right) = \underset{\tau_1^{(a)} < \cdots < \tau_{B_a}^{(a)}}{\operatorname{argmin}} \mathcal{S}_{\hat{\mathscr{B}}}\left(\tau_1^{(a)}, \ldots, \tau_{B_a}^{(a)}\right)$$

and then minimize

$$\mathcal{S}_{B_a}^{\mathrm{Pen}} = \mathcal{S}_{\hat{\mathscr{B}}}\left(\hat{\tau}_{B_a,1}^{(a)}, \ldots, \hat{\tau}_{B_a,B_a}^{(a)}\right) + B_a\{\hat{Q}_{\hat{\mathscr{B}}}(\hat{\tau}_{\hat{A},a}, \hat{\tau}_{\hat{A},a+1}) + \eta_{p,N}\} \tag{9}$$

with respect to $B_a \geq 0$, where

$$\hat{Q}_{\hat{\mathscr{B}}}(\hat{\tau}_{\hat{A},a}, \hat{\tau}_{\hat{A},a+1}) = \sum_{t=\hat{\tau}_{\hat{A},a}+1}^{\hat{\tau}_{\hat{A},a+1}} \sum_{j \in \hat{\mathscr{B}}} X_{tj} \Big/ \sum_{t=\hat{\tau}_{\hat{A},a}+1}^{\hat{\tau}_{\hat{A},a+1}} n_t$$

together with $\eta_{p,N}$ serve the purpose of penalization. We denote the final estimators as $\hat{B}_a$ and $\left(\hat{\tau}_{\hat{B}_a,1}^{(a)}, \ldots, \hat{\tau}_{\hat{B}_a,\hat{B}_a}^{(a)}\right)$ if $\hat{B}_a > 0$ for $a = 0, 1, \ldots, \hat{A}$.

Remark 3 In the low-dimensional situation such as $\mathcal{S}_{\mathscr{A}}(\hat{\tau}_{A,1}, \ldots, \hat{\tau}_{A,A})$, the total variation reduced due to adding a redundant change-point is of the same order of the maximum noise level and thus can be dominated by $\xi_{p,N}$ in $\mathcal{S}_A^{\mathrm{Pen}}$. However, this is not directly applicable in the high-dimensional setting. The reduction of the objective function $\mathcal{S}_{\hat{\mathscr{B}}}\left(\tau_1^{(a)}, \ldots, \tau_{B_a}^{(a)}\right)$ caused by adding a new point includes two terms, the expectation and the variation, while the latter in fact vanishes compared to the former. In $\mathcal{S}_{B_a}^{\mathrm{Pen}}$, $\hat{Q}_{\hat{\mathscr{B}}}(\hat{\tau}_{\hat{A},a}, \hat{\tau}_{\hat{A},a+1})$ is an approximation to the expectation term and $\eta_{p,N}$ is chosen to be slightly larger than the maximum variation level. However, $\mathcal{S}_{\hat{\mathscr{B}}}\left(\hat{\tau}_{B_a,1}^{(a)}, \ldots, \hat{\tau}_{B_a,B_a}^{(a)}\right) + B_a\hat{Q}_{\hat{\mathscr{B}}}(\hat{\tau}_{\hat{A},a}, \hat{\tau}_{\hat{A},a+1})$ is no longer a nonincreasing

function in B_a, so that standard techniques in establishing consistency, e.g., Yao (1988) and Bai and Perron (1998), become invalid. In fact, our theoretical derivations are almost completely different and highly nontrivial.

The minimization problems in (8) and (9) can be solved via the dynamic programming (DP) algorithm (Hawkins, 2001) or the pruned exact linear time (PELT) algorithm (Killick, Fearnhead and Eckley, 2012). Finding the exact solutions is straightforward and fast; the computational cost is linear in p and could possibly be close to linear in T when using the PELT. In the worst scenario that the "pruned" part is negligible, the total complexity is $O(pT^2)$ which is equivalent to using the standard DP algorithm.

To study the consistency of our two-step detection procedure, we first extend Theorem 1 to the case of $L^* > 1$.

COROLLARY 3. *Suppose that Assumption (B1) holds and, as* $(N,T) \to \infty$, $N_{0\tau_l^*}/N \to \kappa_l$ *for* $l = 0, 1, \ldots, L^*$ *with* $\kappa_0 < \kappa_1 < \cdots < \kappa_{L^*}$. *Then, we have* $\Pr(\hat{\mathscr{A}} = \mathscr{A}) \to 1$ *for any* $0 < C < \min_{l<L^*}\{\kappa_{l+1} - \kappa_l\}/L^*$ *with the convention of* $\kappa_{L^*+1} = T$.

Let $\lambda_{\mathscr{A},T} = \min_{0\le a\le A^*}(\tau^*_{\mathscr{A},a+1} - \tau^*_{\mathscr{A},a})$, $\lambda_{\mathscr{B},T} = \min_{0\le l\le L^*}(\tau^*_{l+1} - \tau^*_l)$, $\Delta_{\mathscr{A}} = \min_{1\le a\le A^*}\sum_{j\in\mathscr{A}}(q_{\tau^*_{\mathscr{A},a},j} - q_{\tau^*_{\mathscr{A},a-1},j})^2/q_j^{(\kappa)}$, $\Delta_{\mathscr{B}} = \min_{1\le l\le L^*}\sum_{j\in\mathscr{B}}(q_{lj} - q_{l-1,j})^2$ and $\underline{n} = \min_{1\le t\le T} n_t$, where $q_j^{(\kappa)} = \sum_{l=0}^{L^*}(\kappa_{l+1} - \kappa_l)q_{lj}$. Two additional assumptions are required for the theoretical development.

(B2) If $A^* > 0$, as $(p,N,T) \to \infty$,

$$\frac{\lambda^2_{\mathscr{A},T}\underline{n}^2 N^{-1}\Delta_{\mathscr{A}}}{\max\{\log T, \underline{n}^{-1}(\log T)^2\}} \to \infty.$$

(B3) If $B_a^* > 0$, as $(p,N,T) \to \infty$,

$$\frac{\lambda^2_{\mathscr{B},T}\underline{n}^2 N^{-1}\Delta_{\mathscr{B}}}{\max\left\{(\max_l \sum_{j\in\mathscr{B}} q_{lj}^2)^{1/2}\log T, \underline{n}^{-1/2}(\log T)^{1/2}, \underline{n}^{-1}(\log T)^2\right\}} \to \infty.$$

Assumptions (B2) and (B3) impose theoretical requirements for the smallest signal strength and distance between two change-points so that the change-points are asymptotically distinguishable. It is intuitive that if two successive distributions are very different, then we do not need a large λ_T to locate the change-point. Theorem 4 firstly establishes the consistency of the estimated change-points for set $\mathscr{A}$, and Theorem 5 then does for set $\mathscr{B}$.

THEOREM 4. *Suppose that Assumptions (B1) and (B2) hold, the upper bound on L^* is bounded, and $N_{0\tau_l^*}/N \to \kappa_l$ as $(N,T) \to \infty$. If $\xi_{p,N}$ is chosen such that $\xi_{p,N}/\max\{\log T, \underline{n}^{-1}(\log T)^2\} \to \infty$, then*

$$\Pr\left(\hat{A} = A^*;\ |\hat{\tau}_{\hat{A},a} - \tau^*_{\mathscr{A},a}| \le \delta_{\mathscr{A},T}, a = 0,1,\ldots,A^*\right) \to 1,$$

as $(p,N,T) \to \infty$, provided that $\delta^2_{\mathscr{A},T}\underline{n}^2 N^{-1}\Delta_{\mathscr{A}}/\xi_{p,N} \to \infty$.

THEOREM 5. *Suppose the conditions in Theorem 4 and Assumption (B3) hold. If $\eta_{p,N}$ is chosen such that*

$$\max\left\{\big(\max_l \sum_{j\in\mathscr{B}} q_{lj}^2\big)^{1/2}\log T, \underline{n}^{-1/2}(\log T)^{1/2}, \underline{n}^{-1}(\log T)^2\right\}\eta_{p,N}^{-1} \to 0,$$

then

$$\Pr\left(\hat{B}_a = B_a^*;\ |\hat{\tau}^{(a)}_{\hat{B}_a,b} - \tau^*_{\mathscr{B},a,b}| \le \delta_{\mathscr{B},T}, b = 0,1,\ldots,B_a^*\right) \to 1,$$

as $(p,N,T) \to \infty$, provided that $\delta^2_{\mathscr{B},T}\underline{n}^2 N^{-1}\Delta_{\mathscr{B}}/\eta_{p,N} \to \infty$ and

(10)
$$\frac{\Delta_{\mathscr{A}}/\max\{\log T, \underline{n}^{-1}(\log T)^2\}}{\Delta_{\mathscr{B}}/\max\{(\max_l \sum_{j\in\mathscr{B}} q_{lj}^2)^{1/2}\log T, \underline{n}^{-1/2}(\log T)^{1/2}, \underline{n}^{-1}(\log T)^2\}} \to \infty.$$

The condition in (10) requires that the signal strength in set $\mathscr{A}$ dominates that in set $\mathscr{B}$. This ensures that the difference between the estimated change-point $\hat{\tau}_{\hat{A},a}$ and the true one, $\tau^*_{\mathscr{A},a}$, would not affect the detection performance in $\mathscr{B}$. Intuitively speaking, this condition can be easily satisfied because the changes in a low-dimensional environment are always more detectable than those in a high-dimensional setting.

Theorems 4 and 5 can be shown using a concentration inequality for degenerate U-statistics on the basis of an independent vector-valued sample, see Section 3.4.3 of Giné and Nickl (2016). As the concentration inequality is sharp, the rate of $\delta_{\cdot,T}$ given in the theorems is "near-optimal" and cannot be improved beyond the degree of $(\log T)^c$ for some $c > 0$.

Choices of $\xi_{p,N}$ and $\eta_{p,N}$ depend on $\underline{n}$ and $\max_l \mathbf{q}_l^\top \mathbf{q}_l$. To guarantee a reasonable detection precision, $\underline{n}$ cannot be too small, and of course the larger the better. For practical use, we suggest to choose $\xi_{p,N}$ and $\eta_{p,N}$ so that the conditions $\xi_{p,N}/\log T \to \infty$ and $\eta_{p,N}/\left\{\left(\max_l \mathbf{q}_l^\top \mathbf{q}_l\right)^{1/2}\log T\right\} \to \infty$

are roughly satisfied. Empirically, we recommend $\xi_{p,N} = c_\xi(\log T)^{1.5}$ and $\eta_{p,N} = c_\eta \bar{U}_t^{1/2}(\log T)^{1.1}$, where $\bar{U}_t = T^{-1}\sum_{t=1}^{T} U_t$ and

$$U_t = \frac{1}{n_t(n_t-1)} \sum_{j \in \hat{\mathscr{B}}} (X_{tj}^2 - X_{tj}).$$

Note that $\bar{U}_t$ can be regarded as an approximation to the lower bound of $\max_l \mathbf{q}_l^\top \mathbf{q}_l$. A slightly conservative choice helps to prevent underfitting, as one is often reluctant to miss any important change-point. Our simulation results indicate that $c_\xi = 2$ and $c_\eta = 1.2$ provide reasonably good performance in most cases.

4. Numerical studies.

4.1. *Two-sample homogeneity test.* To evaluate the performance of our proposed test and the change-point detection procedure, we first consider the two-sample homogeneity problem and compare it with some "off-the-shelf" procedures. A natural benchmark is the classical Pearson's chi-squared test which is modified by removing all the terms with $\hat{q}_j = 0$ to accommodate large p,

$$W_p = \frac{n_0 n_1}{N} \sum_{j=1, \hat{q}_j \neq 0}^{p} \Big(\frac{X_{0j}}{n_0} - \frac{X_{1j}}{n_1}\Big)^2 / \hat{q}_j.$$

The critical value is approximated by $\chi^2_{\alpha,\tilde{p}-1}$, the upper αth quantile of the χ^2-distribution with degrees of freedom $\tilde{p}-1$, where $\tilde{p} = \sum_{j=1}^{p} \mathrm{I}(\hat{q}_j \neq 0)$. Another alternative is the well-known Hellinger test,

$$H_p = \frac{4n_0 n_1}{N} \sum_{j=1}^{p} \Big(\sqrt{\frac{X_{0j}}{n_0}} - \sqrt{\frac{X_{1j}}{n_1}}\Big)^2,$$

which rejects the null hypothesis if $H_p > \chi^2_{\alpha,\tilde{p}-1}$. All simulation results are based on 5,000 replications.

Table 1 presents the empirical sizes at a 5% significance level under the null hypothesis $H_0 : \mathbf{q}_0 = \{\omega/d\mathbf{1}_d^\top, (1-\omega)/(p-d)\mathbf{1}_{p-d}^\top\}^\top$ and different (p, N)-settings, where $0 < \omega < 1$, d is an integer and $\mathbf{1}_d$ stands for the d-dimensional vector with all components being 1. We set $n_0 = n_1 = N/2$. If $d \ll p$, this null model means that $\mathscr{A} \approx \{1, \ldots, d\}$. To obtain a reasonable estimator $\hat{\mathscr{A}}$ in practice, we consider the curve of the cumulative sum of the decreasingly ordered $\hat{q}_j$'s, with an expectation that there would be a

relatively slow growth after d. Thus, we can maximize the angle between the two contiguous slopes of the piecewise linear curve,

$$\hat{d} = \underset{i=1,\ldots,p-1}{\operatorname{argmax}} \frac{-1 - \hat{q}_{(i)}\hat{q}_{(i+1)}}{\sqrt{1+\hat{q}_{(i)}^2}\sqrt{1+\hat{q}_{(i+1)}^2}},$$

where $\hat{q}_{(1)} \geq \cdots \geq \hat{q}_{(p)}$ are the ordered values of $\hat{q}_j$'s. We observe that the sizes of the proposed $Q_{p,\hat{\mathscr{A}}}$ test are generally close to the nominal level under all the scenarios. In contrast, both W_p and H_p work well under relatively small p settings as expected, but encounter serious size distortion under "small N, large p" scenarios.

TABLE 1

Comparison of empirical sizes (%) at a 5% significance level for the two-sample homogeneity test under $H_0 : \mathbf{q}_0 = \{\omega/d\mathbf{1}_d^\top, (1-\omega)/(p-d)\mathbf{1}_{p-d}^\top\}^\top$ *and different* (p, N)*-settings, with* $\omega = 0.5$ *and* $d = 6$.

		p								
N	Test	10	20	50	100	200	500	1000	2000	5000
500	$Q_{p,\hat{\mathscr{A}}}$	5.36	5.40	5.44	5.12	6.02	5.72	5.34	5.42	5.76
	H_p	5.20	6.84	47.50	99.12	100.00	100.00	100.00	100.00	100.00
	W_p	4.44	4.42	3.28	1.18	0.12	0.00	0.00	0.00	0.00
1000	$Q_{p,\hat{\mathscr{A}}}$	5.84	6.50	5.24	5.88	5.78	5.66	5.50	5.10	5.02
	H_p	5.24	5.86	15.48	76.36	100.00	100.00	100.00	100.00	100.00
	W_p	4.98	4.70	4.02	3.16	1.24	0.04	0.00	0.00	0.00

To evaluate power of the three tests, we consider two alternative hypotheses:

(i) dense but faint signals,

$$\mathbf{q}_1 = \{\omega/d\mathbf{1}_d^\top, (1+s)(1-\omega)/(p-d)\mathbf{1}_{\lfloor (p-d)/2 \rfloor}^\top, (1-s)(1-\omega)/(p-d)\mathbf{1}_{\lfloor (p-d)/2 \rfloor}^\top\}^\top,$$

where $\lfloor x \rfloor$ denotes the largest integer not greater than x;

(ii) sparse but strong signals,

$$\mathbf{q}_1 = \{(1+s)\omega/d\mathbf{1}_{\lfloor d/2 \rfloor}^\top, (1-s)\omega/d\mathbf{1}_{\lfloor d/2 \rfloor}^\top, (1-\omega)/(p-d)\mathbf{1}_{p-d}^\top\}^\top.$$

We choose $p = 500, 1000$, and for each p, let $N = 500, 1000$. Figure 1 shows the relationship between empirical power and s. Our $Q_{p,\hat{\mathscr{A}}}$ test is clearly more powerful than W_p. The empirical sizes of H_p deviate far from the nominal level as shown in Table 1, which renders unnecessarily high power for H_p. Overall, our $Q_{p,\hat{\mathscr{A}}}$ test is demonstrated to maintain the test size as well as attaining high power.

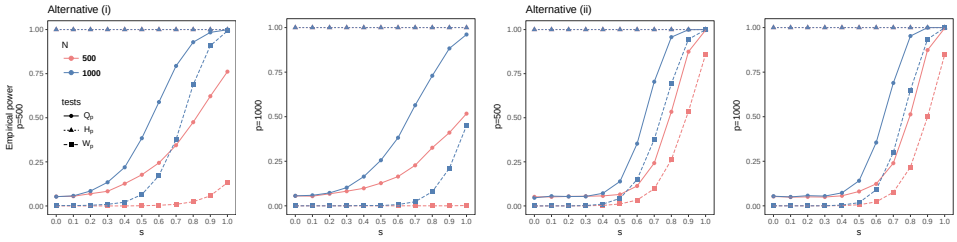

FIG 1. *Comparison of empirical power at a 5% significance level for the two-sample homogeneity test under different* (p, N)*-settings and the alternative hypothesis (i) with* $\omega = 0.3$ *and the alternative hypothesis (ii) with* $\omega = 0.7$*, where* $d = 6$*. Here* Q_p *is short for* $Q_{p,\hat{\mathscr{A}}}$*.*

4.2. *Change-point problem.* When the number of categories, p, is fixed, Srivastava and Worsley (1986) and Horváth and Serbinowska (1995) studied change-point tests with multinomial data based on Pearson's chi-squared statistic. In particular, Srivastava and Worsley (1986) proposed to use $W_p^{(\mathrm{SW})} = \max_\tau K_{2,\tau}$ and gave a conservative approximation of the null distribution based on an improved Bonferroni inequality. Horváth and Serbinowska (1995) developed a weighted version, $W_p^{(\mathrm{HS})} = \max_\tau N_{0\tau}N_{1\tau}/N^2 K_{2,\tau}$, and showed under some conditions $W_p^{(\mathrm{HS})} \xrightarrow{\mathcal{D}} \sup_{0\le t\le 1}\sum_{j=1}^{p-1} B_j^2(t)$, where $\{B_j(t),\ 0 \le t \le 1\}$, $1 \le j \le p-1$, are independent Brownian bridges. To accommodate large p, we replace $K_{2,\tau}$ and p by

$$\tilde{K}_{2,\tau} = \sum_{j=1,\hat{q}_{0j}\neq 0}^{p} \frac{N_{0\tau}N_{1\tau}}{N}\Big(\frac{Z_{0\tau j}}{N_{0\tau}} - \frac{Z_{1\tau j}}{N_{1\tau}}\Big)^2 \big/ \hat{q}_{0j}$$

and $\tilde{p} = \sum_{j=1}^{p} \mathrm{I}(\hat{q}_{0j} \neq 0)$, respectively. As pointed out by Aue et al. (2009), $(W_{\tilde{p}}^{(\mathrm{HS})} - \tilde{p}/4)/\sqrt{\tilde{p}/8} \xrightarrow{\mathcal{D}} N(0,1)$ when p is large. For fairness, we use the same trimmed summation or maximization in these competitors as our $Q_{p,\hat{\mathscr{A}}}$, i.e., $\mathcal{T} = [\lceil a(T-1)\rceil, \lceil b(T-1)\rceil]$.

We again consider $H_0 : \mathbf{q}_0 = \{\omega/d\mathbf{1}_d^\top, (1-\omega)/(p-d)\mathbf{1}_{p-d}^\top\}^\top$. For simplicity, we fix $n_t = n = N/T$ for $t = 1, \ldots, T$, and set $a = 0.1$ and $b = 0.9$ in the proposed test. Table 2 presents the empirical sizes at a 5% significance level under various scenarios with $T = 100$. The results with $T = 10$ and significance levels of 1% and 10% are reported in the Supplementary Material. We observe that the empirical sizes of our test are close to the nominal level, while both $W_{\tilde{p}}^{(\mathrm{SW})}$ and $W_{\tilde{p}}^{(\mathrm{HS})}$ encounter serious size distortion in most cases. Note that it is unnecessary for n to be sufficiently large compared to p.

For power comparison of the three tests, we consider $T = 100, n = 20$

TABLE 2
Comparison of empirical sizes (%) at a 5% significance level for the change-point test under $H_0 : \mathbf{q}_0 = \{\omega/d\mathbf{1}_d^\top, (1-\omega)/(p-d)\mathbf{1}_{p-d}^\top\}^\top$ *and different* (p, N)*-settings when* $T = 100$.

		$\omega = 0.3$			$\omega = 0.5$			$\omega = 0.7$		
p	n	$Q_{p,\hat{\mathscr{A}}}$	$W_{\tilde{p}}^{(\mathrm{SW})}$	$W_{\tilde{p}}^{(\mathrm{HS})}$	$Q_{p,\hat{\mathscr{A}}}$	$W_{\tilde{p}}^{(\mathrm{SW})}$	$W_{\tilde{p}}^{(\mathrm{HS})}$	$Q_{p,\hat{\mathscr{A}}}$	$W_{\tilde{p}}^{(\mathrm{SW})}$	$W_{\tilde{p}}^{(\mathrm{HS})}$
500	10	5.48	1.50	0.54	5.96	3.58	0.12	5.62	6.46	0.02
	20	5.44	1.90	3.46	5.42	3.52	1.70	5.98	6.40	0.44
	50	5.34	2.70	8.84	5.64	2.60	7.34	5.96	4.24	4.44
1000	10	5.80	1.02	0.02	5.76	3.36	0.00	6.14	6.88	0.00
	20	5.12	1.74	0.42	5.52	3.44	0.14	5.76	6.68	0.04
	50	5.34	2.16	4.42	5.40	3.08	2.42	5.68	6.14	0.56

and the locations of change-points at $\tau^* = 20, 50$. We examine the previous two alternatives (i) and (ii) for $\mathbf{q}_1$. Figure 2 depicts the power curves of the three tests, $Q_{p,\mathscr{A}}$, $W_{\tilde{p}}^{(\mathrm{SW})}$ and $W_{\tilde{p}}^{(\mathrm{HS})}$, versus s. As s increases, the power curve of the proposed procedure increases much more sharply than the other two, especially in the sparse signal scenario. We also observe that the power becomes larger when τ^* moves closer to $T/2$, which coincides with Corollary 1. Overall, $Q_{p,\hat{\mathscr{A}}}$ performs better than the other two competitors in terms of attaining high power while maintaining the test size, and the advantage becomes more pronounced for larger p. Such findings are consistent with our theoretical analysis that Pearson's chi-squared test may not work well because the contamination bias in estimating the marginal proportions grows rapidly with p. When p and N are comparable, the inverse of the estimated proportions in the test statistic would no longer bring in benefit.

In Figure 3, we make comparisons with three other approaches: one is the maximum of L_τ, i.e., $M_p = \max_\tau \sum_{j=1}^p L_{\tau j}$; and the other two correspond to the summation and maximum of the Hellinger test statistics, $H_p^{(\mathrm{sum})} = \sum_\tau H_\tau$ and $H_p^{(\mathrm{max})} = \max_\tau H_\tau$, where

$$H_\tau = \frac{4N_{0\tau}N_{1\tau}}{N} \sum_{j=1}^{p} \Big(\sqrt{\frac{Z_{0\tau j}}{N_{0\tau}}} - \sqrt{\frac{Z_{1\tau j}}{N_{1\tau}}} \Big)^2.$$

However, it is difficult to obtain approximate threshold values for these tests. For fairness, we perform a size-corrected power comparison in the sense that the actual threshold values are found through simulations so that these tests approximately maintain a type I error rate of 0.05. Both the M_p and $Q_{p,\hat{\mathscr{A}}}$

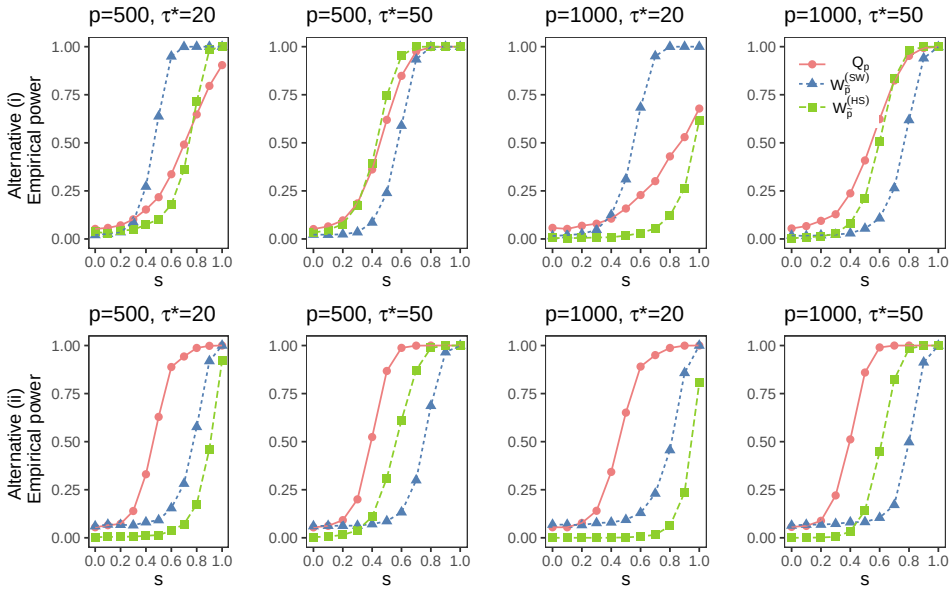

FIG 2. *Comparison of empirical power for the proposed* $Q_{p,\mathscr{A}}$ *test,* $W_{\tilde{p}}^{(\mathrm{SW})}$ *by Srivastava and Worsley (1986) and* $W_{\tilde{p}}^{(\mathrm{HS})}$ *by Horváth and Serbinowska (1995) under the alternative hypothesis (i) with* $\omega = 0.3$*, and the alternative hypothesis (ii) with* $\omega = 0.7$*, where* $d = 6$*.*

tests outperform $H_p^{(\mathrm{sum})}$ in most cases, because $\sum_j L_{\tau j}$ possesses certain advantage over H_τ as conveyed by Figure 1. The performances of the two $\sum_j L_{\tau j}$-based methods are comparable in the dense signal setting, while the M_p test appears to be slightly more powerful when τ^* is small. In the sparse signal setting, the M_p test breaks down even when $\tau^* = \lfloor T/2 \rfloor$, which demonstrates the benefit of the power-enhancement term $E_{p,\hat{\mathscr{A}}}$ in our test statistic $Q_{p,\hat{\mathscr{A}}}$. Besides, the advantages of using $Q_{p,\hat{\mathscr{A}}}$ are obvious: its null distribution is asymptotically normal and the asymptotic test has excellent finite-sample performance as shown by Table 2.

Once the null hypothesis is rejected, we estimate the change-point under the alternatives (i) and (ii). As shown in Table 3, the biases appear to be negligible for all the change-points, and as expected the standard deviations increase as p becomes large and decrease as N becomes large. Overall, the proposed estimators are consistent and work well in most cases.

4.3. *Multiple change-point detection.* To assess our approach for detecting multiple change-points, we consider two different data generation processes. The first one assumes that changes only occur on $\mathscr{B}$ and the change-

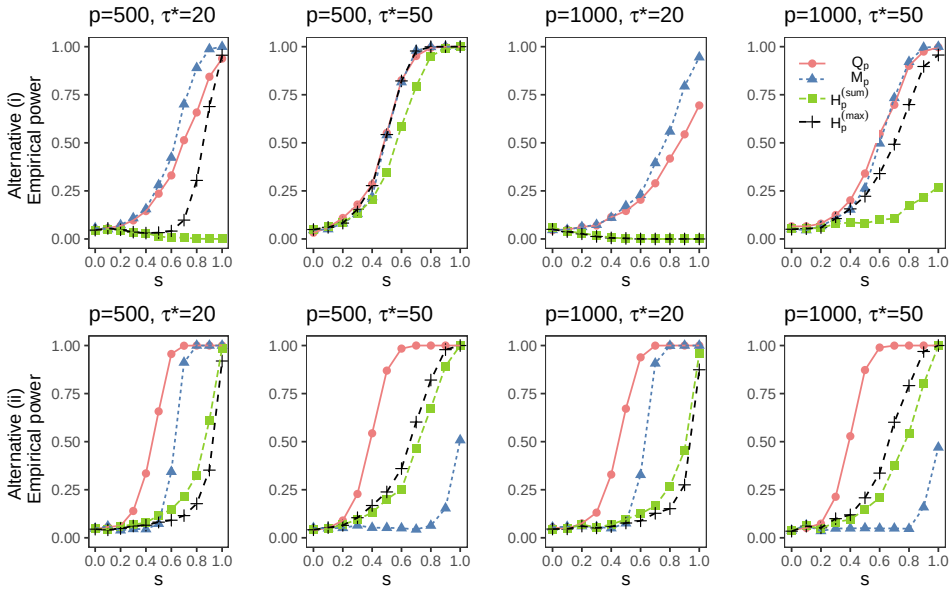

FIG 3. *Comparison of size-corrected empirical power under the alternative hypotheses (i) and (ii).*

points are generated as $\{\tau_1^*, \ldots, \tau_{L^*}^*\}/T = \{0.25, 0.5, 0.75\}$. For each $l = 0, 1, \ldots, L^*$, let $\mathscr{A} = \{1, \ldots, d\}, \mathscr{B} = \{d+1, \ldots, p\}$ and

$$\mathbf{q}_l = \Big\{ \frac{\omega}{d}\mathbf{1}_d^\top, \frac{(1-\omega)s}{p-d}\mathbf{1}_{p-d,\mathscr{B}_l}^\top + \frac{1-\omega-\frac{(1-\omega)sp_0}{p-d}}{p-d-p_0}\mathbf{1}_{p-d,\mathscr{B}\setminus\mathscr{B}_l}^\top \Big\}^\top,$$

where $\mathscr{B}_l$ is a randomly chosen subset of $\mathscr{B}$ with a cardinality of $p_0 = 0.01(p-d)$ and $\mathbf{1}_{p-d,\mathscr{B}_l}$ is a $(p-d)$-dimensional vector with elements taking a value of 1 if belonging to $\mathscr{B}_l$ and 0 otherwise. The second data generation allows changes to occur on both $\mathscr{A}$ and $\mathscr{B}$, and the change-points are designed as $\{\tau_1^*, \ldots, \tau_{L^*}^*\}/T = \{0.2, 0.4, 0.6, 0.8\}$. Further let

$$\begin{aligned}
\mathbf{q}_l =& \Big(\frac{\omega_l}{d}\mathbf{1}_d^\top, \frac{1-\omega_l}{p-d}\mathbf{1}_{p-d}^\top \Big)^\top \quad \text{for } l = 0, 1, 4, \\
\mathbf{q}_2 =& \Big\{ \frac{\omega_2}{d}\mathbf{1}_d^\top, \frac{(1+s_{\mathscr{B}})(1-\omega_2)}{p-d}\mathbf{1}_{\lfloor\frac{p-d}{2}\rfloor}^\top, \frac{(1-s_{\mathscr{B}})(1-\omega_2)}{p-d}\mathbf{1}_{\lfloor\frac{p-d}{2}\rfloor}^\top \Big\}^\top \quad \text{and} \\
\mathbf{q}_3 =& \Big\{ \frac{(1+s_{\mathscr{A}})\omega_3}{d}\mathbf{1}_{\lfloor\frac{d}{2}\rfloor}^\top, \frac{(1-s_{\mathscr{A}})\omega_3}{d}\mathbf{1}_{\lfloor\frac{d}{2}\rfloor}^\top, \frac{(1+s_{\mathscr{B}})(1-\omega_3)}{p-d}\mathbf{1}_{\lfloor\frac{p-d}{2}\rfloor}^\top, \\
& \frac{(1-s_{\mathscr{B}})(1-\omega_3)}{p-d}\mathbf{1}_{\lfloor\frac{p-d}{2}\rfloor}^\top \Big\}^\top,
\end{aligned}$$

TABLE 3

Simulation study on the consistency of our change-point detection procedure with the mean and standard deviation (in parentheses) of $|\hat{\tau}^* - \tau^*|$'s under alternatives (i) and (ii). Note that New, SW, HS and Hel refer to the estimators by our method, $\operatorname{argmax}_\tau K_{2,\tau}$ *(Srivastava and Worsley, 1986),* $\operatorname{argmax}_\tau N_{0\tau}N_{1\tau}K_{2,\tau}$ *(Horváth and Serbinowska, 1995) and Hellinger's* $\operatorname{argmax}_\tau H_\tau$*, respectively.*

	Alternative (i)				Alternative (ii)			
	$p = 500$		$p = 1000$		$p = 500$		$p = 1000$	
	$\tau^* = 20$	$\tau^* = 50$	$\tau^* = 20$	$\tau^* = 50$	$\tau^* = 20$	$\tau^* = 50$	$\tau^* = 20$	$\tau^* = 50$
New	$1.70_{(6.14)}$	$0.75_{(2.23)}$	$5.41_{(12.7)}$	$1.67_{(3.39)}$	$1.70_{(3.30)}$	$0.93_{(1.79)}$	$1.88_{(3.92)}$	$0.97_{(1.85)}$
SW	$2.68_{(3.32)}$	$2.68_{(7.22)}$	$4.24_{(3.71)}$	$13.2_{(14.9)}$	$2.29_{(7.67)}$	$2.20_{(6.14)}$	$3.81_{(12.1)}$	$2.92_{(8.37)}$
HS	$5.21_{(3.78)}$	$9.60_{(1.21)}$	$12.9_{(4.91)}$	$10.0_{(1.69)}$	$16.0_{(5.36)}$	$9.23_{(1.35)}$	$16.8_{(4.21)}$	$9.27_{(1.15)}$
Hel	$1.01_{(1.62)}$	$0.64_{(1.31)}$	$12.6_{(5.57)}$	$1.47_{(2.02)}$	$26.7_{(6.35)}$	$1.34_{(1.90)}$	$27.9_{(4.07)}$	$0.89_{(1.21)}$

where $\omega_1 = \omega_2 = \omega_3$. It covers both cases of sparse signals on $\mathscr{A}$ and dense signals on $\mathscr{B}$. We take $p = 1000$ and $T = 100$ for illustration.

To evaluate the finite-sample performance, we introduce the distance between the estimated change-point set and the true one, representing the over- and under-segmentation error respectively (Zou et al., 2014),

$$\mathrm{OE} = \sup_{r=1,\ldots,L^*} \inf_{l=1,\ldots,\hat{L}} |\hat{\tau}_l - \tau_r^*| \quad \text{and} \quad \mathrm{UE} = \sup_{l=1,\ldots,\hat{L}} \inf_{r=1,\ldots,L^*} |\hat{\tau}_l - \tau_r^*|,$$

for which a desirable estimator should be able to strike a balance. In addition, the estimation error on the number of change-points, $\#\mathrm{E} = |\hat{L} - L^*|$, is also examined. Table 4 presents the mean and standard deviation of #E, OE and UE, based on 2,000 replications. It can be seen that all the three error values are small, and the performances are generally stable. This demonstrates that the proposed global estimator in conjunction with the use of the empirical $\xi_{p,N} = c_\xi(\log T)^{1.5}$ and $\eta_{p,N} = c_\eta \bar{U}_t^{1/2}(\log T)^{1.1}$ (with $c_\xi = 2.0$ and $c_\eta = 1.2$) can deliver satisfactory detection performance in the presence of multiple change-points.

4.4. *Real data application.* We illustrate the proposed method with the *Entree Chicago Recommendation Data* from the University of California at Irvine Machine Learning Repository. This data set contains user interactions with the entree Chicago restaurant recommendation system, which recommended restaurants based on cuisine, price, style, atmosphere etc. to users, from September, 1996 to April, 1999. We focus on the end point of each user interaction, which is represented by the numeric ID of the Chicago restaurant that the user last visited. There are $T = 134$ weekly records, a total

TABLE 4

Performance evaluation on detection of multiple change-points with the standard deviations given in parentheses. We set $\omega = 0.3$ *and* $s = 10$ *in data generation 1 and* $\omega_0 = 0.3, \omega_1 = \omega_2 = \omega_3 = 0.7, \omega_4 = 0.5$ *and* $s_{\mathscr{A}} = s_{\mathscr{B}} = 0.9$ *in data generation 2, and* $d = 6$ *in both settings. Note that* $\#E = |\hat{L} - L^*|$*, and OE and UE represent over- and under-segmentation errors, respectively.*

	Data generation 1			Data generation 2		
n	#E	OE	UE	#E	OE	UE
50	$0.29_{(0.61)}$	$0.77_{(1.22)}$	$2.53_{(5.29)}$	$1.51_{(1.36)}$	$15.7_{(6.40)}$	$6.86_{(7.05)}$
100	$0.13_{(0.40)}$	$0.09_{(0.34)}$	$1.21_{(4.14)}$	$0.82_{(0.64)}$	$16.4_{(6.62)}$	$3.34_{(5.48)}$
200	$0.10_{(0.35)}$	$0.00_{(0.05)}$	$1.08_{(4.29)}$	$0.37_{(0.63)}$	$0.49_{(1.93)}$	$2.23_{(4.42)}$
500	$0.08_{(0.30)}$	$0.00_{(0.00)}$	$1.00_{(4.07)}$	$0.29_{(0.59)}$	$0.03_{(0.25)}$	$1.46_{(3.38)}$

of $N = 43,573$ user interactions and $p = 617$ restaurants. We are interested in testing whether the proportions allotted to all the restaurants based on users' final choices changed over time. Figure 4 (a) depicts the sample size n_t by weeks, and Figure 4 (b) shows the scatter plot of the proportions of two randomly chosen restaurants over time. The heatmaps of the frequencies and proportions of the user interactions in all the restaurants for 134 weeks are given in Figures 4 (c)–(d), respectively.

Figure 5 gives an empirical way to quantify the sparsity pattern $\mathscr{A}$. In particular, Figure 5 (a) shows the estimated proportions $\hat{q}_j$'s, $j = 1, \ldots, p$, for all samples, and Figure 5 (b) exhibits the sorted $\hat{q}_j$'s. The top-ranked $\hat{q}_j$'s are much larger than the average level, $1/p \approx 1.62 \times 10^{-3}$. Further, the zoom-in plot (c) suggests that we may simply select restaurants with the largest 10 $\hat{q}_j$'s as $\hat{\mathscr{A}}$ because those ten proportions are all larger than 0.01 and they occupy 12.63% of the market by users' tendency among all 617 restaurants. As a result, Assumption (A1) or (B1) appears to be satisfied for this example. Based on our testing procedure, $(S_{p,\hat{\mathscr{A}}} - \Lambda_T)/\sqrt{2c_{N,T}U_{N,\mathscr{A}}} = 9.21$, which is highly significant compared with the standard normal null distribution. Subsequently, we perform the multiple change-point detection. Lavielle (2005) suggested an intuitive method by first plotting the segmentation cost function versus the number of change-points and then finding an "elbow" in the plot, which would suggest the most suitable segmentation. The intuition is that as more true change-points are detected the cost function would continue to decrease, while at the same time it is likely to be detecting more false positives and thus the cost function may start to decrease slowly or level off. Figures 6 (a)–(b) present the plots of the penalized objective function based on segmentation, corresponding to equations (8)–(9), versus the number of change-points L. Figure 6 (a) clearly suggests that the model

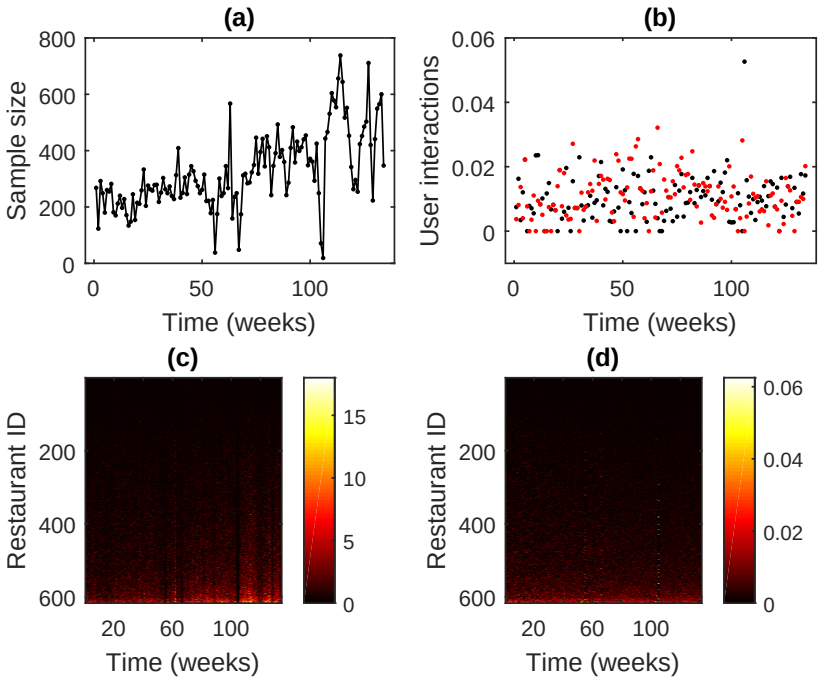

FIG 4. *Description of the raw data over time. (a): Sample size over time; (b): Scatter plot of the proportions of two randomly chosen restaurants over 134 weeks; (c): Heatmap of user interactions of 617 restaurants with brightness representing the frequency of users' final choices; (d): Heatmap of the proportions allotted in all the restaurants with time.*

with three change-points fit the data best on $\hat{\mathscr{A}}$, and the identified change-points are weeks 53, 54 and 90. Figure 6 (b) reveals that the model with two change-points fits the data best on $\hat{\mathscr{B}}$ in the sampling range $[55, 90)$, and the identified change-points are weeks 62 and 63. Figure 6 (c) presents the plot of $\mathcal{S}_{\hat{\mathscr{B}}}(\hat{\tau}^{(2)}_{L,1}, \ldots, \hat{\tau}^{(2)}_{L,L}) + L\hat{Q}_{\hat{\mathscr{B}}}(55, 90)$ versus the number of change-points L, which verifies the segmentation result in Figure 6 (b) as the rate of decline changes more sharply at the point $L = 2$. No change-points are found on $\hat{\mathscr{B}}$ in other sampling ranges, and thus in total five change-points are detected, i.e., weeks 53, 54, 62, 63 and 90. Our result delivers piecewise "stable" segmentations, i.e., within each segmentation users' tendency towards different types of restaurants can be regarded as unchanged. By identifying which restaurants become more preferable or less preferable at a change-point, we could explore potential factors, such as the food flavor, restaurant atmosphere and service quality, that may affect customers' choices, which would

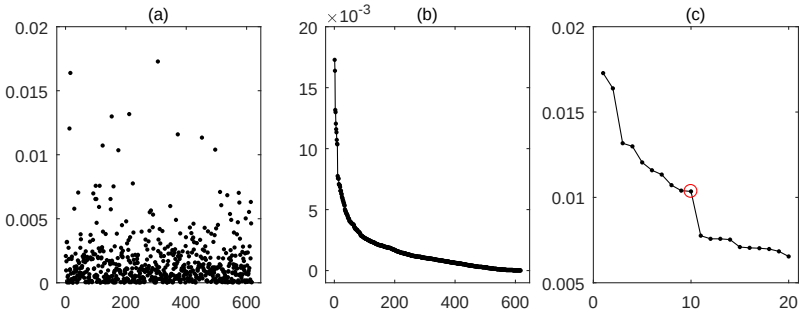

FIG 5. *Estimation of the sparsity pattern. (a): Scatter plot of the estimated proportions $\hat{q}_j$'s for $j = 1, \ldots, p$; (b)–(c): Plots of the estimated proportions in a decreasing order, i.e., $\hat{q}_{(j)}$'s, for $j = 1, \ldots, p$ and $j = 1, \ldots, 20$, respectively.*

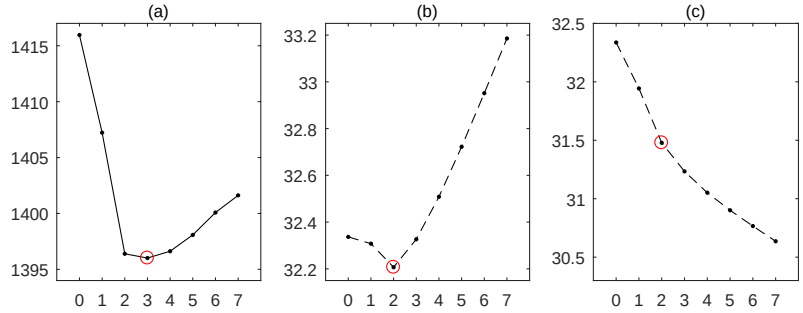

FIG 6. *(a)–(b): Plots of the penalized objective function based on segmentation, corresponding to equations (8)–(9), versus the number of change-points L, respectively. (c): Plot of $\mathcal{S}_{\hat{\mathscr{B}}}(\hat{\tau}^{(2)}_{L,1}, \ldots, \hat{\tau}^{(2)}_{L,L}) + L\hat{Q}_{\hat{\mathscr{B}}}(55, 90)$ versus the number of change-points L.*

in turn help to promote the development of catering industry.

5. Concluding remarks. A new approach to change-point detection is developed based on the estimated sparsity patterns, which gives a general yet tractable high-dimensional analogue to the classical Pearson's chi-squared statistic. The modified Pearson's chi-squared statistic in conjunction with the summation procedure is demonstrated to work well when the number of categories is large and the contingency table is sparse. A limitation of our method is the separation assumption that the estimators need to satisfy, which is a rather general issue in sparse estimation (Fan and Lv, 2008). In practice, it remains difficult to be assured that all significant proportions are distinguishable from the whole set. Nevertheless, our empirical

studies suggest that asymptotic p-values behave reasonably well even when the assumption may be possibly violated. In the analysis following change-point detection, it is important to incorporate knowledge on the discovered change-points to improve variable selection, inference, and prediction.

Acknowledgment. The authors are grateful to the referees, Associate Editor, and Editor for their insightful comments that have significantly improved the article. Wang and Zou were supported by NNSF of China grants 11690015, 11622104, 11431006 and 11371202. Yin's research was supported in part by a grant (17326316) from the Research Grants Council of Hong Kong.

Supplementary Material. The Supplementary Material contains all theoretical proofs of Theorems 1–5, Proposition 1 and Corollary 1, and additional simulation results.

REFERENCES

Agresti, A. (2013). *Categorical Data Analysis*. John Wiley & Sons.

Aue, A., Hörmann, S., Horváth, L. and Reimherr, M. (2009). Break Detection in the Covariance Structure of Multivariate Time Series Models. *The Annals of Statistics* **37** 4046–4087.

Bai, J. and Perron, P. (1998). Estimating and Testing Linear Models with Multiple Structural Changes. *Econometrica* **66** 47–78.

Bai, Z. and Saranadasa, H. (1996). Effect of High Dimension: By an Example of a Two Sample Problem. *Statistica Sinica* **6** 311–329.

Baranov, A. P. and Baranov, Y. A. (2005). A Power Divergence Test in the Problem of Sample Homogeneity for Large Numbers of Outcomes and Trials. *Discrete Mathematics and Applications* **15**.

Braun, J. V., Braun, R. K. and Müller, H. G. (2000). Multiple Changepoint Fitting via Quasilikelihood, with Application to DNA Sequence Segmentation. *Biometrika* **87** 301–314.

Bykov, S. I. and Ivanov, V. A. (1991). On the Conditions of Asymptotic Normality of Multidimensional Randomized Decomposable Statistics. *Discrete Mathematics and Applications* **1** 219–227.

Chen, J. and Gupta, A. K. (2000). *Parametric Statistical Change Point Analysis*. Birkhäuser Boston.

Chen, S. X. and Qin, Y.-L. (2010). A Two-Sample Test for High-Dimensional Data with Applications to Gene-Set Testing. *The Annals of Statistics* **38** 808–835.

Chen, H. and Zhang, N. R. (2013). Graph-Based Tests for Two-Sample Comparisons of Categorical Data. *Statistica Sinica* **23** 1479–1503.

Csörgö, M. and Horváth, L. (1997). *Limit Theorems in Change-Point Analysis*. Wiley.

Fan, J., Liao, Y. and Yao, J. (2015). Power Enhancement in High-Dimensional Cross-Sectional Tests. *Econometrica* **83** 1497–1541.

Fan, J. and Lv, J. (2008). Sure Independence Screening for Ultrahigh Dimensional Feature Space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70** 849–911.

Fryzlewicz, P. (2014). Wild Binary Segmentation for Multiple Change-Point Detection. *The Annals of Statistics* **42** 2243–2281.

Giné, E. and Nickl, R. (2016). *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge University Press.

Hall, P. and Heyde, C. C. (1980). *Martingale Limit Theory and Its Application*. Academic Press.

Hawkins, D. M. (2001). Fitting Multiple Change-Point Models to Data. *Computational Statistics & Data Analysis* **37** 323–341.

Holst, L. (1972). Asymptotic Normality and Efficiency for Certain Goodness-of-Fit Tests. *Biometrika* **59** 137–145.

Horváth, L. and Serbinowska, M. (1995). Testing for Changes in Multinomial Observations: The Lindisfarne Scribes Problem. *Scandinavian Journal of Statistics* **22** 371–384.

Ivchenko, G. and Levin, V. (1976). Asymptotic Normality of a Class of Statistics in the Multinomial Scheme. *Theory of Probability & Its Applications* **21** 188–192.

Kallenberg, W. C. M. (1985). On Moderate and Large Deviations in Multinomial Distributions. *The Annals of Statistics* **13** 1554–1580.

Killick, R., Fearnhead, P. and Eckley, I. A. (2012). Optimal Detection of Changepoints With a Linear Computational Cost. *Journal of the American Statistical Association* **107** 1590–1598.

Lavielle, M. (2005). Using Penalized Contrasts for the Change-Point Problem. *Signal Processing* **85** 1501–1510.

Morris, C. (1975). Central Limit Theorems for Multinomial Sums. *The Annals of Statistics* **3** 165–188.

Perron, P. and Vogelsang, T. J. (1992). Testing for a Unit Root in a Time Series With a Changing Mean: Corrections and Extensions. *Journal of Business & Economic Statistics* **10** 467–470.

Srivastava, M. S. and Worsley, K. J. (1986). Likelihood Ratio Tests for a Change in the Multivariate Normal Mean. *Journal of the American Statistical Association* **81** 199–204.

Srivastava, M. S. and Wu, Y. (1993). Comparison of EWMA, CUSUM and Shiryayev-Roberts Procedures for Detecting a Shift in the Mean. *The Annals of Statistics* **21** 645–670.

Yao, Y.-C. (1988). Estimating the Number of Change-Points via Schwarz' Criterion. *Statistics & Probability Letters* **6** 181–189.

Zou, C., Yin, G., Feng, L. and Wang, Z. (2014). Nonparametric Maximum Likelihood Approach to Multiple Change-Point Problems. *The Annals of Statistics* **42** 970–1002.

Guanghui Wang, Changliang Zou
Institute of Statistics and LPMC
Nankai University
China
E-mail: ghwang.nk@gmail.com
nk.chlzou@gmail.com

Guosheng Yin
Department of Statistics and Actuarial Science
The University of Hong Kong
Hong Kong
E-mail: gyin@hku.hk