Testing High-dimensional Multinomials with Applications to Text Analysis

T. Tony Cai

University of Pennsylvania, Philadelphia, Pennsylvania, United States.

Zheng Tracy Ke

Harvard University, Cambridge, Massachusetts, United States.

Paxton Turner†

Harvard University, Cambridge, Massachusetts, United States.

Summary. Motivated by applications in text mining and discrete distribution inference, we test for equality of probability mass functions of K groups of high-dimensional multinomial distributions. Special cases of this problem include global testing for topic models, two-sample testing in authorship attribution, and closeness testing for discrete distributions. A test statistic, which is shown to have an asymptotic standard normal distribution under the null hypothesis, is proposed. This parameter-free limiting null distribution holds true without requiring identical multinomial parameters within each group or equal group sizes. The optimal detection boundary for this testing problem is established, and the proposed test is shown to achieve this optimal detection boundary across the entire parameter space of interest. The proposed method is demonstrated in simulation studies and applied to analyze two real-world datasets to examine, respectively, variation among customer reviews of Amazon movies and the diversity of statistical paper abstracts.

Keywords: authorship attribution, closeness testing, customer reviews, martingale central limit theorem, minimax optimality, topic model

1. Introduction

Statistical inference for multinomial data has garnered considerable recent interest (Diakonikolas and Kane, 2016; Balakrishnan and Wasserman, 2018). One important application is in text mining. It is common to model the word counts in a text document by a multinomial distribution (Blei et al., 2003). As a motivating example, the study of online customer ratings and reviews is a trending topic in marketing research. Customer reviews are a good proxy to the overall "word of mouth" and can significantly influence customers' decisions. Research works aim to understand the patterns in online reviews and their impacts on sales. Classical studies only use numerical ratings but ignore the rich text reviews because of their unstructured nature. More recent works have revealed the importance of analyzing text reviews, especially for hedonic products such as books,

†*Address for correspondence:* Harvard University, Statistics. 1 Oxford St # 7 Cambridge, MA, USA 02138, Email: paxtonturner@g.harvard.edu

movies, and hotels (Chevalier and Mayzlin, 2006). A question of interest is to detect the heterogeneity in reviewers' response styles. For example, Leung and Yang (2020) discovered that younger travelers, women, and travelers with less review expertise tend to give more positive reviews and that guests staying in high-class hotels tend to have more extreme response styles than those staying in low-class hotels. Knowing such differences will offer valuable insights for hotel managers and online rating/review sites.

The aforementioned heterogeneity detection can be cast as a hypothesis test on multinomial data. Suppose reviews are written using a vocabulary of p distinct words. Let $X_i \in \mathbb{R}^p$ contain the word counts in review i. We assume X_i's are independent, and

$$X_i \sim \text{Multinomial}(N_i, \Omega_i), \qquad 1 \le i \le n, \tag{1}$$

where N_i is the total length of review i and $\Omega_i \in \mathbb{R}^p$ is a probability mass function (PMF) containing the population word frequencies. These reviews are divided into K groups by reviewer characteristics (e.g., age, gender, new/returning customer), product characteristics (e.g., high-class versus low-class hotels), and numeric ratings (e.g., from 1 star to 5 stars), where K can be presumably large. We view Ω_i as representing the 'true response' of review i. The "average response" of a group k is defined by a weighted average of the PMFs:

$$\mu_k = (n_k \bar{N}_k)^{-1} \sum_{i \in S_k} N_i \Omega_i, \qquad 1 \le k \le K. \tag{2}$$

Here $S_k \subset \{1, 2, \ldots, n\}$ is the index set of group k, $n_k = |S_k|$ is the total number of reviews in group k, and $\bar{N}_k = n_k^{-1} \sum_{i \in S_k} N_i$ is the average length of reviews in group k. We would like to test

$$H_0 : \quad \mu_1 = \mu_2 = \ldots = \mu_K. \tag{3}$$

When the null hypothesis is rejected, it means there exist statistically significant differences among the group-wise "average responses".

We call (1)-(3) the "K-sample testing for equality of average PMFs in multinomials" or "K-sample testing for multinomials" for short. As K varies, it includes several well-defined problems in text mining and discrete distribution inference as special cases.

(a) *Global testing for topic models.* Topic modeling (Blei et al., 2003) is a popular text mining tool. In a topic model, each Ω_i in (1) is a convex combination of M topic vectors. Before fitting a topic model to a corpus, it is often desirable to determine if the corpus indeed contains multiple topics. This boils down to the global testing problem, which tests $M = 1$ versus $M > 1$. In this case, we set $K = n$ and view each document as a separate group, so that Ω_i itself is the within-group average. Under the null hypothesis, all these Ω_i's are equal to a single topic vector. Under the alternative, the Ω_i's are not all equal. This is thus a special case of our problem with $K = n$ and $n_k = 1$.

(b) *Authorship attribution* (Mosteller and Wallace, 1963; Kipnis, 2022). In these applications, the goal is to determine the unknown authorship of an article from other articles with known authors. A famous example (Mosteller and Wallace, 2012) is to determine the actual authors of a few Federalist Papers written by three authors

but published under a single pseudonym. It can be formulated (Mosteller and Wallace, 1963; Kipnis, 2022) as testing the equality of population word frequencies between the article of interest and the corpus from a known author, a special case of our problem with $K = 2$.

(c) *Closeness between discrete distributions* (Chan et al., 2014; Bhattacharya and Valiant, 2015; Balakrishnan and Wasserman, 2019). There has been a surge of interest in discrete distribution inference. Closeness testing is one of most studied problems. The data from two discrete distributions are summarized in two multinomial vectors Multinomial(N_1, μ) and Multinomial(N_2, θ). The goal is to test $\mu = \theta$. It is a special case of our testing problem with $K = 2$ and $n_1 = n_2 = 1$.

In this paper, we provide a unified solution to all the aforementioned problems. The key to our methodology is a flexible statistic called DELVE (DE-biased and Length-assisted Variability Estimator). It provides a general similarity measure for comparing groups of discrete distributions such as count vectors associated with text corpora. Similarity measures (such as the classical cosine similarity, log-likelihood ratio statistic, and others) are fundamental in text mining and have been applied to problems in distribution testing (Kim et al., 2022), computational linguistics (Gomaa et al., 2013), econometrics (Hansen et al., 2018), and computational biology (Kolodziejczyk et al., 2015). Our method is a new and flexible similarity measure that is potentially useful in these areas.

We emphasize that our setting does not require that the X_i's in the same group are drawn from the same distribution. Under the null hypothesis (3), the group-wise means are equal, but the Ω_i's within each group can still be different from each other. As a result, the null hypothesis is composite and designing a proper test statistic is non-trivial.

1.1. Our results and contributions

The dimensionality of the testing problem is captured by (n, p, K) and $\bar{N} := n^{-1} \sum_{i=1}^{n} N_i$. We are interested in a high-dimensional setting where

$$n\bar{N} \to \infty, \quad p \to \infty, \quad \text{and} \quad n^2\bar{N}^2/(Kp) \to \infty. \tag{4}$$

In most places of this paper, we use a subscript n to indicate asymptotics, but our method and theory do apply to the case where n is finite and $\bar{N} \to \infty$. In text applications, $n\bar{N}$ is the total count of words in the corpus, and a large $n\bar{N}$ means either there are sufficiently many documents, or the documents are sufficiently long. Given that $n\bar{N} \to \infty$, we further allow (p, K) to grow with n at a speed such that $Kp \ll n^2\bar{N}^2$. In particular, our settings allow K to range from 2 to n, so as to cover all the application examples.

We propose a test that enjoys the following properties:

(a) *Parameter-free null distribution*: We shall define a test statistic ψ in (12) and show that $\psi \to N(0, 1)$ under the null H_0 in (3). Even under H_0, the model contains a large number of free parameters because the null hypothesis is only about the equality of "average" PMFs but still allows (N_i, Ω_i) to differ within each group. As an appealing property, the null distribution of ψ does not depend on these individual multinomial parameters; hence, we can always conveniently obtain the asymptotic p-value for our proposed test.

(b) *Minimax optimal detection boundary*: We define a quantity $\omega_n := \omega_n(\mu_1, \mu_2, \ldots, \mu_K)$ in (25) that measures the difference among the K group-wise mean PMFs. It satisfies that $\omega_n = 0$ if and only if the null hypothesis holds, and it has been properly normalized so that ω_n is bounded under the alternative hypothesis (provided some mild regularity conditions hold). We show that the proposed test has an asymptotic full power if $\omega_n^4 n^2 \bar{N}^2/(Kp) \to \infty$. We also provide a matching lower bound by showing that the null hypothesis and the alternative hypothesis are asymptotically indistinguishable if $\omega_n^4 n^2 \bar{N}^2/(Kp) \to 0$. Therefore, the proposed test is minimax optimal. Furthermore, in the boundary case where $\omega_n^4 n^2 \bar{N}^2/(Kp) \to c_0$ for a constant $c_0 > 0$, we show that $\psi \to N(0,1)$ under H_0, and $\psi \to N(c_1, 1)$, under a specific alternative hypothesis H_1 in (35), with c_1 being an explicit function of c_0.

To the best of our knowledge, this testing problem for a general K has not been studied before. The existing works primarily focused on closeness testing and authorship attribution (see Section 1.2), which are special cases with $K = 2$. In comparison, our test is applicable to any value of K, offering a unified solution to multiple applications. Even for $K = 2$, the existing works do not provide a test statistic that has a tractable null distribution. They determined the rejection region and calculated p-values using either a (conservative) large-deviation bound or a permutation procedure. Our test is the first one equipped with a tractable null distribution. Our results about the optimal detection boundary for a general K are also new to the literature. By varying K in our theory, we obtain the optimal detection boundary for different sub-problems. For some of them (e.g., global testing for topic models, authorship attribution with moderate sparsity), the optimal detection boundary was not known before; hence, our results help advance the understanding of the statistical limits of these problems.

1.2. Related literature

First, we make a connection to discrete distribution inference. Let $X \sim \text{Multinomial}(N, \Omega)$ represent a size-N sample from a discrete distribution with p categories. The one-sample closeness testing aims to test $H_0 : \Omega = \mu$, for a given PMF μ. Existing works focus on finding the minimum separation condition in terms of the ℓ^1-norm or ℓ^2-norm of $\Omega - \mu$. Balakrishnan and Wasserman (2019) derived the minimum ℓ^1-separation condition and proposed a truncated chi-square test to achieve it. Valiant and Valiant (2017) studied the "local critical radius", a local separation condition that depends on the "effective sparsity" of μ, and they proposed a "2/3rd + tail" test to achieve it. In the two-sample closeness testing problem, given $X_1 \sim \text{Multinomial}(N_1, \Omega_1)$ and $X_2 \sim \text{Multinomial}(N_2, \Omega_2)$, it aims to test $H_0 : \Omega_1 = \Omega_2$. Again, this literature focuses on finding the minimum separation condition in terms of the ℓ^1-norm or ℓ^2-norm of $\Omega_1 - \Omega_2$. When $N_1 = N_2$, Chan et al. (2014) derived the minimum ℓ^1-separation condition and proposed a weighted chi-square test to attain it. Bhattacharya and Valiant (2015) extended their results to the unbalanced case where $N_1 \neq N_2$, assuming $\|\Omega_1 - \Omega_2\|_1 \geq p^{-1/12}$. This assumption was later removed by Diakonikolas and Kane (2016), who established the minimum ℓ^1-separation condition in full generality. Kim et al. (2022) proposed a two-sample kernel U-statistic and showed that it attains the minimum ℓ^2-separation condition.

Since the two-sample closeness testing is a special case of our problem with $K = 2$

and $n_1 = n_2 = 1$, our test is directly applicable. An appealing property of our test is its tractable asymptotic null distribution of $N(0,1)$. In contrast, for the chi-square statistic in Chan et al. (2014) or the U-statistic in (Kim et al., 2022), the rejection region is determined by either an upper bound from concentration inequalities or a permutation procedure, which may lead to a conservative threshold or need additional computational costs. Regarding the testing power, we show in Section 4.3 that our test achieves the minimum ℓ^2-separation condition, i.e., our method is an optimal "ℓ^2 testor." Our test can also be turned into an optimal "ℓ^1 testor" (a test that achieves the minimum ℓ^1-separation condition) by re-weighting terms in the test statistic (see Section 4.3).

Another related problem is the independence testing (Diakonikolas and Kane, 2016; Berrett and Samworth, 2019). Given i.i.d. bivariate samples from the joint distribution of discrete variables I and J, it aims to test if I and J are independent. This is connected to our testing problem with $K = n$, as in this case our null hypothesis implies that the word distribution is independent of the document label. However, the data generating processes in two problems are not the same. In independence testing, it is assumed that the vectorization of X follows a multinomial distribution with $n\bar{N}$ trials and np possible outcomes. In our problem, each X_i follows a multinomial distribution with N_i trials and p possible outcomes. Hence, we cannot directly apply existing results from independence testing. In addition, we allow K to be any integer in $[2, n]$. When $K \neq n$, it is unknown how to relate independence testing to our problem.

Next, we make a connection to text mining. In this literature, a multinomial vector $X \sim \text{Multinomial}(N, \Omega)$ represents the word counts for a document of length N written with a dictionary containing p words. In a topic model, each Ω_i is a convex combination of M "topic vectors": $\Omega_i = \sum_{k=1}^{M} w_i(k) A_k$, where each $A_k \in \mathbb{R}^p$ is a PMF and the combination coefficient vector $w_i \in \mathbb{R}^K$ is called the "topic weight" vector for document i. Given a collection of documents $X_1, X_2, \ldots, X_n$, the global testing problem aims to test $M = 1$ versus $M > 1$. Interestingly, the optimal detection boundary for this problem has never been rigorously studied. As we have explained, this problem is a special case of our testing problem with $K = n$. Our results (a) provide a test statistic that has a tractable null distribution and (b) reveal that the optimal detection boundary is $\omega_n^2 \asymp (\sqrt{n}\bar{N})^{-1}\sqrt{p}$. Both (a) and (b) are new results. When comparing our results with those about estimation of A_k's (Ke and Wang, 2022), it suggests that global testing requires a strictly lower signal strength than topic estimation.

For authorship attribution, Kipnis (2022) treats the corpus from a known author as a single document and tests the null hypothesis that this combined document and a new document have the same population word frequencies. It is a two-sample closeness testing problem, except that sparsity is imposed on the difference of two PMFs. Kipnis (2022) proposed a test which applies an "exact binomial test" to obtain a p-value for each word and combines these p-values using Higher Criticism (Donoho and Jin, 2004). Donoho and Kipnis (2022) analyzed this test when the number of "useful words" is $o(\sqrt{p})$, and they derived a sharp phase diagram (a related one-sample setting was studied in Arias-Castro and Wang (2015)). In Section 4.2, we show that our test is applicable to this problem and has some nice properties: (a) tractable null distribution; (b) allows for $s \geq c\sqrt{p}$, where s is the number of useful words; and (c) does not require documents from the known author to have identical population word frequencies, making the setting

more realistic. On the other hand, when $s = o(\sqrt{p})$, our test is less powerful than the one in Kipnis (2022); Donoho and Kipnis (2022), as our test does not utilize sparsity explicitly. We can further improve our test in this regime by modifying the DELVE statistic to incorporate sparsity (see the remark in Section 4.2).

The rest of this paper is arranged as follows. In Section 2, we introduce the test statistic and explain the rationale behind it. We then present in Section 3 the main theoretical results, including the asymptotic null distribution, power analysis, a matching lower bound, the study of two special cases ($K = n$ and $K = 2$), and a discussion of the contiguity regime. Section 4 applies our results to text mining and discrete distribution testing. Simulations are in Section 5 and real data analysis is in Section 6. The paper is concluded with a discussion in Section 7. All proofs are in Cai et al. (2023).

2. The DELVE Test

Recall that $X_1, \ldots, X_n$ are independent, and $X_i \sim \text{Multinomial}(N_i, \Omega_i)$ for $1 \le i \le n$. There is a known partition $\{1, 2, \ldots, n\} = \cup_{k=1}^K S_k$. Write $n_k = |S_k|$, $\bar{N}_k = n_k^{-1} \sum_{i \in S_k} N_i$, and $\bar{N} = n^{-1} \sum_{i=1}^n N_i$. In (2), we have defined the group-wise mean PMF $\mu_k = (n_k \bar{N}_k)^{-1} \sum_{i \in S_k} N_i \Omega_i$. We further define the overall mean PMF $\mu \in \mathbb{R}^p$ by

$$\mu := \frac{1}{n\bar{N}} \sum_{k=1}^K n_k \bar{N}_k \mu_k = \frac{1}{n\bar{N}} \sum_{i=1}^n N_i \Omega_i. \tag{5}$$

We introduce a quantity $\rho^2 = \rho^2(\mu_1, \ldots, \mu_K)$ by

$$\rho^2 := \sum_{k=1}^K n_k \bar{N}_k \|\mu_k - \mu\|^2. \tag{6}$$

This quantity measures the variations across K group-wise mean PMFs. It is true that the null hypothesis (3) holds if and only if $\rho^2 = 0$. Inspired by this observation, we hope to construct an unbiased estimator of ρ^2 and develop it to a test statistic.

We can easily obtain the minimum variance unbiased estimators of μ_k and μ:

$$\hat{\mu}_k = \frac{1}{n_k \bar{N}_k} \sum_{i \in S_k} X_i, \qquad \text{and} \qquad \hat{\mu} = \frac{1}{n\bar{N}} \sum_{k=1}^K n_k \bar{N}_k \hat{\mu}_k = \frac{1}{n\bar{N}} \sum_{i=1}^n X_i. \tag{7}$$

For each $1 \le j \le p$, let μ_{kj}, μ_j, $\hat{\mu}_{kj}$ and $\hat{\mu}_j$ represent the jth entry of μ_k, μ, $\hat{\mu}_k$ and $\hat{\mu}$, respectively. A naive estimator of ρ^2 is

$$\widetilde{T} = \sum_{j=1}^p \widetilde{T}_j, \qquad \text{where} \quad \widetilde{T}_j = \sum_{k=1}^K n_k \bar{N}_k (\hat{\mu}_{kj} - \hat{\mu}_j)^2. \tag{8}$$

This estimator is biased. In Section F.1 of Cai et al. (2023), we show that $\mathbb{E}[\widetilde{T}_j] = \sum_{k=1}^K \big[n_k \bar{N}_k (\mu_{kj} - \mu_j)^2 + \big(\frac{1}{n_k \bar{N}_k} - \frac{1}{n\bar{N}} \big) \sum_{i \in S_k} N_i \Omega_{ij}(1 - \Omega_{ij}) \big]$. It motivates us to debias $\widetilde{T}_j$ by using an unbiased estimate of $\Omega_{ij}(1 - \Omega_{ij})$. By basic properties of multinomial

distributions, $\mathbb{E}[X_{ij}(N_i - X_{ij})] = N_i(N_i-1)\Omega_{ij}(1-\Omega_{ij})$. We thereby use $\frac{1}{N_i(N_i-1)}X_{ij}(N_i - X_{ij})$ to estimate $\Omega_{ij}(1-\Omega_{ij})$. It yields an unbiased estimator of ρ^2:

$$T = \sum_{j=1}^{p} T_j, \quad T_j = \sum_{k=1}^{K}\Big[n_k\bar{N}_k(\hat{\mu}_{kj} - \hat{\mu}_j)^2 - \Big(\frac{1}{n_k\bar{N}_k} - \frac{1}{n\bar{N}}\Big)\sum_{i\in S_k}\frac{X_{ij}(N_i - X_{ij})}{N_i - 1}\Big]. \quad (9)$$

LEMMA 1. *Under Models* (1)-(2)*, the estimator in* (9) *satisfies that* $\mathbb{E}[T] = \rho^2$.

To use T for hypothesis testing, we need a proper standardization of this statistic. In Sections D.1-D.2 of Cai et al. (2023), we study $\mathbb{V}(T)$, the variance of T. Under mild regularity conditions, it can be shown that $\mathbb{V}(T) = \Theta_n \cdot [1 + o(1)]$, where

$$\Theta_n := 4\sum_{k=1}^{K}\sum_{j=1}^{p} n_k\bar{N}_k(\mu_{kj} - \mu_j)^2\mu_{kj} + 2\sum_{k=1}^{K}\sum_{i\in S_k}\sum_{j=1}^{p}\Big(\frac{1}{n_k\bar{N}_k} - \frac{1}{n\bar{N}}\Big)^2\frac{N_i^3}{N_i - 1}\Omega_{ij}^2 \quad (10)$$

$$+ \frac{2}{n^2\bar{N}^2}\sum_{1\le k\ne\ell\le K}\sum_{i\in S_k}\sum_{m\in S_\ell}\sum_{j=1}^{p} N_iN_m\Omega_{ij}\Omega_{mj} + 2\sum_{k=1}^{K}\sum_{\substack{i\in S_k, m\in S_k,\\ i\ne m}}\sum_{j=1}^{p}\Big(\frac{1}{n_k\bar{N}_k} - \frac{1}{n\bar{N}}\Big)^2 N_iN_m\Omega_{ij}\Omega_{mj}.$$

In Θ_n, the first term vanishes under the null, so it suffices to estimate the other three terms in Θ_n. By properties of multinomial distributions, $\mathbb{E}[X_{ij}X_{mj}] = N_iN_m\Omega_{ij}\Omega_{mj}$, $\mathbb{E}[X_{ij}^2] = N_i^2\Omega_{ij}^2 + N_i\Omega_{ij}(1-\Omega_{ij})$, and $\mathbb{E}[X_{ij}(N_i - X_{ij})] = N_i(N_i-1)\Omega_{ij}(1-\Omega_{ij})$. It inspires us to estimate $\Omega_{ij}\Omega_{mj}$ by $\frac{X_{ij}X_{mj}}{N_iN_m}$ and estimate Ω_{ij}^2 by $\frac{X_{ij}^2}{N_i^2} - \frac{X_{ij}(N_i - X_{ij})}{N_i^2(N_i-1)} = \frac{X_{ij}^2 - X_{ij}}{N_i(N_i-1)}$. Define

$$V = 2\sum_{k=1}^{K}\sum_{i\in S_k}\sum_{j=1}^{p}\Big(\frac{1}{n_k\bar{N}_k} - \frac{1}{n\bar{N}}\Big)^2\frac{X_{ij}^2 - X_{ij}}{N_i(N_i-1)} + \frac{2}{n^2\bar{N}^2}\sum_{k\ne\ell}\sum_{i\in S_k}\sum_{m\in S_\ell}\sum_{j=1}^{p} X_{ij}X_{mj}$$

$$+ 2\sum_{k=1}^{K}\sum_{\substack{i\in S_k, m\in S_k,\\ i\ne m}}\sum_{j=1}^{p}\Big(\frac{1}{n_k\bar{N}_k} - \frac{1}{n\bar{N}}\Big)^2 X_{ij}X_{mj}. \quad (11)$$

The test statistic we propose is as follows (in the rate event $V < 0$, we simply set $\psi = 0$):

$$\psi = T/\sqrt{V}. \quad (12)$$

We call ψ the *DEbiased and Length-adjusted Variability Estimator (DELVE)*. In Section 3.1, we show that under mild regularity conditions, $\psi \to N(0,1)$ under the null hypothesis. For any fixed $\kappa \in (0,1)$, the asymptotic level-κ DELVE test rejects H_0 if

$$\psi > z_\kappa, \qquad \text{where } z_\kappa \text{ is the } (1-\kappa)\text{-quantile of } N(0,1). \quad (13)$$

REMARK 1 (OTHER TESTING IDEAS). The likelihood ratio (LR) test can only be applied when Ω_i's are equal within each group (in this case, the null/alternative hypotheses have much fewer free parameters). Moreover, the DELVE test attains the minimax optimal detection boundary in high-dimensional settings, but there is no such guarantee for

the LR test. From simulations in Section 5, when p is large, DELVE has better power than LR. Another idea is to use the ANOVA statistic $\widetilde{T}$ in (8) without de-biasing and apply a chi-square approximation or permutation procedure to compute the p-value. This test is unfortunately suboptimal. There are settings in which the bias term dominates the "signal" term in $\widetilde{T}$, causing the test to lose power (see Remark 4 for details).

REMARK 2. We have assumed $X_1, \ldots, X_n$ are independent. This is better interpreted as the conditional independence given Ω_i's. When Ω_i's are random and have some dependence structure, X_i's can be (marginally) dependent. We will see in Section 3 that the asymptotic null distribution of ψ does not depend on Ω_i's; then, the same asymptotic distribution also holds for random and dependent Ω_i. We have also assumed that the distribution of X_i is multinomial. However, our test only uses the first two moments of multinomials, not the likelihood. As a result, our method is relatively robust to model misspecification, and it is extendable to settings with under/over dispersion.

2.1. *The special cases of* $K = n$ *and* $K = 2$

As seen in Section 1, the application examples of $K = n$ and $K = 2$ are particularly intriguing. In these cases, we give more explicit expressions of our test statistic.

When $K = n$, we have $S_k = \{i\}$ and $\hat{\mu}_{kj} = N_i^{-1} X_{ij}$. The null hypothesis becomes $H_0 : \Omega_1 = \Omega_2 = \ldots = \Omega_n$. The statistic in (9) reduces to

$$T = \sum_{j=1}^{p} \sum_{i=1}^{n} \left[\frac{(X_{ij} - N_i \hat{\mu}_j)^2}{N_i} - \Big(1 - \frac{N_i}{n\bar{N}}\Big) \frac{X_{ij}(N_i - X_{ij})}{N_i(N_i - 1)} \right]. \tag{14}$$

Moreover, in the variance estimate (11), the last term is exactly zero, and it can be shown that the third term is negligible compared to the first term. We thereby consider a simpler variance estimator by only retaining the first term in (11):

$$V^* = 2 \sum_{i=1}^{n} \sum_{j=1}^{p} \Big(\frac{1}{N_i} - \frac{1}{n\bar{N}} \Big)^2 \frac{X_{ij}^2 - X_{ij}}{N_i(N_i - 1)}. \tag{15}$$

The simplified DELVE test statistic is $\psi^* = T/\sqrt{V^*}$.

When $K = 2$, we observe two collections of multinomial vectors, denoted by $\{X_i\}_{1 \le i \le n}$ and $\{G_i\}_{1 \le i \le m}$. We assume for $1 \le i \le n$ and $1 \le j \le m$,

$$X_i \sim \text{Multinomial}(N_i, \Omega_i), \qquad G_j \sim \text{Multinomial}(M_j, \Gamma_j). \tag{16}$$

Write $\bar{N} = n^{-1} \sum_{i=1}^{n} N_i$ and $\bar{M} = m^{-1} \sum_{i=1}^{m} M_i$. The null hypothesis becomes

$$H_0 : \quad \eta = \theta, \qquad \text{where } \eta = \frac{1}{n\bar{N}} \sum_{i=1}^{n} N_i \Omega_i, \text{ and } \theta = \frac{1}{m\bar{M}} \sum_{i=1}^{m} M_i \Gamma_i, \tag{17}$$

where θ and η are the two group-wise mean PMFs. We estimate them by $\hat{\eta} = (n\bar{N})^{-1} \sum_{i=1}^{n} X_i$ and $\hat{\theta} = (m\bar{M})^{-1} \sum_{i=1}^{m} G_i$. The statistic in (9) has an equivalent form as follows:

$$T = \frac{n\bar{N} m\bar{M}}{n\bar{N} + m\bar{M}} \left[\|\hat{\eta} - \hat{\theta}\|^2 - \sum_{i=1}^{n} \sum_{j=1}^{p} \frac{X_{ij}(N_i - X_{ij})}{n^2 \bar{N}^2 (N_i - 1)} - \sum_{i=1}^{m} \sum_{j=1}^{p} \frac{G_{ij}(M_i - G_{ij})}{m^2 \bar{M}^2 (M_i - 1)} \right]. \tag{18}$$

The variance estimate (11) has an equivalent form as follows:

$$V = \frac{4\sum_{i=1}^{n}\sum_{i'=1}^{m}\sum_{j=1}^{p} X_{ij}G_{i'j}}{(n\bar{N}+m\bar{M})^2} + \frac{2m^2\bar{M}^2\big[\sum_{i=1}^{n}\frac{X_{ij}^2-X_{ij}}{N_i(N_i-1)} + \sum_{1\le i\ne i'\le n} X_{ij}X_{i'j}\big]}{n^2\bar{N}^2(n\bar{N}+m\bar{M})^2}$$
$$+ \frac{2n^2\bar{N}^2\big[\sum_{i=1}^{m}\frac{G_{ij}^2-G_{ij}}{M_i(M_i-1)} + \sum_{1\le i\ne i'\le m} G_{ij}G_{i'j}\big]}{m^2\bar{M}^2(n\bar{N}+m\bar{M})^2}. \tag{19}$$

The DELVE test statistic is $\psi = T/\sqrt{V}$.

2.2. A variant: DELVE+

We introduce a variant of the DELVE test statistic to better suit real data. Let $\hat{\mu}$, T and V be as in (7), (9) and (11). Define

$$\psi^+ = T/\sqrt{V^+}, \qquad \text{where} \quad V^+ = V\cdot\big(1+\|\hat{\mu}\|_2 T/\sqrt{V}\big). \tag{20}$$

We call (20) the DELVE+ test statistic. In theory, this modification has little effect on the key properties of the test. To see this, we note that $\|\hat{\mu}\|_2 = o_{\mathbb{P}}(1)$ in high-dimensional settings. Suppose $T/\sqrt{V} \to N(0,1)$ under H_0. Since $\|\hat{\mu}\|_2 \to 0$, it is seen immediately that $V^+/V \to 1$; hence, the asymptotic normality also holds for ψ^+. Suppose $T/\sqrt{V} \to \infty$ under the alternative hypothesis. It follows that $V^+ \le 2\max\{V, \|\hat{\mu}\|_2 \cdot T\sqrt{V}\}$ and $\psi^+ \ge \frac{1}{\sqrt{2}}\min\{T/\sqrt{V},\, \|\hat{\mu}\|_2^{-1}(T/\sqrt{V})^{1/2}\} \to \infty$. We have proved the following lemma:

LEMMA 2. *As $n\bar{N}\to\infty$, suppose $\|\hat{\mu}\|_2 \to 0$ in probability. Under H_0, if $T/\sqrt{V} \to N(0,1)$, then $T/\sqrt{V^+} \to N(0,1)$. Under H_1, if $T/\sqrt{V}\to\infty$, then $T/\sqrt{V^+}\to\infty$.*

In practice, this modification avoids extremely small p-values. In some real datasets, V is very small and leads to an extremely small p-value in the original DELVE test. In DELVE+, as long as T is positive, ψ^+ is smaller than ψ, so that the p-value is adjusted.

In the numerical experiments, we consider both DELVE and DELVE+. For theoretical analysis, since these two versions have almost identical theoretical properties, we only focus on the original DELVE test statistic.

3. Theoretical Properties

We first present the regularity conditions. For a constant $c_0 \in (0,1)$, we assume

$$\min_{1\le i\le n} N_i \ge 2, \qquad \max_{1\le i\le n}\|\Omega_i\|_\infty \le 1-c_0, \qquad \max_{1\le k\le K}\frac{n_k\bar{N}_k}{n\bar{N}} \le 1-c_0. \tag{21}$$

In (21), the first condition is mild. Noting that $\|\Omega_i\|_1 = 1$, the second condition excludes those cases where one of the p categories has an extremely dominating probability in the PMF Ω_i, which is also mild. In the third condition, $n_k\bar{N}_k$ is the total number of counts in all multinomials of group k, and this condition excludes the extremely unbalanced case where one group occupies the majority of counts (in the special case of $K=2$, we further relax this condition to allow for severely unbalanced groups (see Section 3.4)).

Recall that $\mu_k = \frac{1}{n_k \bar{N}_k} \sum_{i \in S_k} N_i \Omega_i$ is the mean PMF within group k. We also define a 'covariance' matrix of PMFs for group k by $\Sigma_k = \frac{1}{n_k \bar{N}_k} \sum_{i \in S_k} N_i \Omega_i \Omega_i'$. Let

$$\alpha_n := \max \left\{ \sum_{k=1}^{K} \frac{\|\mu_k\|_3^3}{n_k \bar{N}_k}, \quad \sum_{k=1}^{K} \frac{\|\mu_k\|^2}{n_k^2 \bar{N}_k^2} \right\} \Big/ \left(\sum_{k=1}^{K} \|\mu_k\|^2 \right)^2, \tag{22}$$

and

$$\beta_n := \max \left\{ \sum_{k=1}^{K} \sum_{i \in S_k} \frac{N_i^2}{n_k^2 \bar{N}_k^2} \|\Omega_i\|_3^3, \quad \sum_{k=1}^{K} \|\Sigma_k\|_F^2 \right\} \Big/ (K \|\mu\|^2). \tag{23}$$

We assume that as $n\bar{N} \to \infty$,

$$\alpha_n = o(1), \qquad \beta_n = o(1), \qquad \text{and} \quad \frac{\|\mu\|_4^4}{K\|\mu\|^4} = o(1). \tag{24}$$

Here α_n and β_n only depend on group-wise quantities, such as μ_k, Σ_k and $\sum_{i \in S_k} N_i^2 \|\Omega_i\|_3^3$; hence, a small number of 'outliers' (i.e., extremely large entries) in Ω has little effect on α_n and β_n. Furthermore, in a simple case where $\max_k n_k \le C \min_k n_k$, $\max_k \bar{N}_k \le C \min_k \bar{N}_k$ and $\|\Omega\|_{\max} = O(1/p)$, it holds that $\alpha_n = O(\max\{\frac{1}{n\bar{N}}, \frac{Kp}{n^2\bar{N}^2}\})$, $\beta_n = O(\max\{\frac{K^2}{n^2 p}, \frac{1}{p}\})$ and $\frac{\|\mu\|_4^4}{K\|\mu\|^4} = O(\frac{1}{Kp})$. When $n\bar{N} \to \infty$ and $p \to \infty$, (24) reduces to $n^2\bar{N}^2/(Kp) \to \infty$. This condition is necessary for successful testing, because our lower bound in Section 3.3 implies that the two hypotheses are asymptotically indistinguishable if $n^2\bar{N}^2/(Kp) \to 0$.

3.1. The asymptotic null distribution

Under the null hypothesis, the K group-wise mean PMFs $\mu_1, \mu_2, \ldots, \mu_K$, are equal to each other, but this hypothesis is still highly composite, as (N_i, Ω_i) are not necessarily the same within each group. We show that the DELVE test statistic always enjoys a parameter-free asymptotic null distribution. Let T, Θ_n and V be as in (9)-(11). The next two theorems are proved in Cai et al. (2023).

THEOREM 1. *Consider Models* (1)-(2)*, where the null hypothesis* (3) *holds. Suppose* (21) *and* (24) *are satisfied. As* $n\bar{N} \to \infty$, $T/\sqrt{\Theta_n} \to N(0,1)$ *in distribution.*

THEOREM 2. *Under the conditions of Theorem 1, as* $n\bar{N} \to \infty$, $V/\Theta_n \to 1$ *in probability, and* $\psi := T/\sqrt{V} \to N(0,1)$ *in distribution.*

By Theorem 2, the asymptotic p-value is $1 - \Phi(\psi)$, where $\Phi(\cdot)$ is the CDF of $N(0,1)$. For any $\kappa \in (0,1)$, the rejection region of the asymptotic level-κ test is as given in (13).

The proofs of Theorems 1-2 contain two key steps. In the first step, we decompose T into mutually uncorrelated terms. Define a set of independent, mean-zero random vectors $\{Z_{ir}\}_{1 \le i \le n, 1 \le r \le N_i}$, where $Z_{ir} \sim \text{Multinomial}(1, \Omega_i) - \Omega_i$. Then, $X_i = N_i\Omega_i + \sum_{r=1}^{N_i} Z_{ir}$ (in distribution). We plug it into (9) to get $T = T_1 + T_2 + T_3 + T_4$, where T_1 is a linear form of $\{Z_{ir}\}$, T_2-T_4 are quadratic forms of $\{Z_{ir}\}$, and T_1-T_4 are uncorrelated (see Section D of Cai et al. (2023)). In the second step, we construct a martingale for each term T_j. This is accomplished by re-arranging the double-index sequence Z_{ir}

to a single-index sequence and successively adding terms in this sequence to T_j. We then apply the martingale central limit theorem (CLT) (Hall and Heyde, 2014) to prove asymptotic normality of each T_j. The asymptotic normality of T follows by identifying the dominating terms in T_1-T_4 (as model parameters change, the dominating terms also change) and studying their joint distribution. This step involves extensive calculations to bound conditional variances and verify the Lindeberg conditions of martingale CLT, as well as subtle uses of the Cauchy-Schwarz inequality to simplify moment bounds.

REMARK 3 (AN ADJUSTMENT WHEN $p = O(1)$). While we focus on high-dimensional settings, the case of $p = O(1)$ is still of interest. In this case, the variance estimator V may not be consistent. We propose a refined estimator $\widetilde{V}$ in Section H of Cai et al. (2023). When V is replaced by $\widetilde{V}$, $\psi \to N(0,1)$ continues to hold.

3.2. Power analysis

Under the alternative hypothesis, the PMFs $\mu_1, \mu_2, \ldots, \mu_K$ are not the same. In Section 2, we introduce a quantity ρ^2 (see (6)) to capture the total variation in μ_k's, but this quantity is not scale-free. We define a scaled version of ρ^2 as

$$\omega_n = \omega_n(\mu_1, \mu_2, \ldots, \mu_K) := \frac{1}{n\bar{N}\|\mu\|^2} \sum_{k=1}^{K} n_k \bar{N}_k \|\mu_k - \mu\|^2. \tag{25}$$

It is seen that $\omega_n \le \max_k\{\frac{\|\mu_k-\mu\|^2}{\|\mu\|^2}\}$, which is properly scaled.

THEOREM 3. *Consider Models* (1)-(2)*, where* (21) *and* (24) *are satisfied. Then,* $\mathbb{E}[T] = n\bar{N}\|\mu\|^2\omega_n^2$*, and* $\mathbb{V}(T) = O\big(\sum_{k=1}^K \|\mu_k\|^2\big) + \mathbb{E}[T] \cdot O\big(\max_{1\le k\le K} \|\mu_k\|_\infty\big)$.

For the DELVE test to have an asymptotically full power, we need $\mathbb{E}[T] \gg \sqrt{\mathbb{V}(T)}$. By Theorem 3, this is satisfied if $\mathbb{E}[T] \gg \sqrt{\sum_k \|\mu_k\|^2}$ and $\mathbb{E}[T] \gg \max_k \|\mu_k\|_\infty$. Between these two requirements, the latter one is weaker; hence, we only need $\mathbb{E}[T] \gg \sqrt{\sum_{k=1}^K \|\mu_k\|^2}$. It gives rise to the following theorem:

THEOREM 4. *Under the conditions of Theorem 3, we further assume that under the alternative hypothesis, as* $n\bar{N} \to \infty$*,*

$$\mathrm{SNR}_n := \frac{n\bar{N}\|\mu\|^2\omega_n^2}{\sqrt{\sum_{k=1}^K \|\mu_k\|^2}} \quad \to \quad \infty. \tag{26}$$

Under the alternative hypothesis, $\psi \to \infty$ *in probability. For any fixed* $\kappa \in (0,1)$*, the level-*κ *DELVE test has an asymptotic level of* κ *and an asymptotic power of* 1*. If we choose* $\kappa = \kappa_n$ *such that* $\kappa_n \to 0$ *and* $1 - \Phi(\mathrm{SNR}_n) = o(\kappa_n)$*, where* Φ *is the CDF of* $N(0,1)$*, then the sum of type I and type II errors of the DELVE test converges to* 0*.*

The detection boundary in (26) has simpler forms in some special cases. For example, if $\|\mu_k\| \asymp \|\mu\|$ for $1 \le k \le K$, then $\mathrm{SRN}_n \asymp n\bar{N}\omega_n^2\|\mu\|/\sqrt{K}$. If, furthermore, all entries of μ are at the same order, which implies $\|\mu\| \asymp p^{-1/2}$, then $\mathrm{SRN}_n \asymp n^2\bar{N}^2\omega_n^2/\sqrt{Kp}$. In this case, the detection boundary simplifies to $\omega_n^4 n^2 \bar{N}^2/(Kp) \to \infty$.

REMARK 4 (THE EFFECT OF DE-BIASING ON POWER). Let $\widetilde{T}$ be the statistic in (8) without bias correction. Under H_1, when $\mathrm{SNR}_n \to \infty$ but $n\bar{N} \ll Kp$, the bias in $\widetilde{T}$ can dominate the "signal" ρ^2. Consequentely, any test based on $\widetilde{T}$ has no power (details and examples are in Section C of Cai et al. (2023)). This shows that de-biasing is critical for achieving not only parameter-free limiting null but also good power.

3.3. A matching lower bound

We have seen that the DELVE test successfully separates two hypotheses if $\mathrm{SNR}_n \to \infty$, where SNR_n is as defined in (26). We now present a lower bound to show that the two hypotheses are asymptotically indistinguishable if $\mathrm{SNR}_n \to 0$.

Let $\ell_i \in \{1, 2, \ldots, K\}$ denote the group label of X_i. Write $\xi = \{(N_i, \Omega_i, \ell_i)\}_{1 \le i \le n}$. Let μ_k, α_n, β_n, and ω_n be the same as defined in (2), (22), (23), and (25), respectively. For each given $(n, p, K, \bar{N})$, we write $\mu_k = \mu_k(\xi)$ to emphasize its dependence on parameters, and similarly for $\alpha_n, \beta_n, \omega_n$. For any $c_0 \in (0, 1)$ and sequence ϵ_n, define

$$\mathcal{Q}_n(c_0, \epsilon_n) := \Big\{ \xi = \{(N_i, \Omega_i, \ell_i)\}_{i=1}^n : (21) \text{ holds for } c_0,\ \max(\alpha_n(\xi), \beta_n(\xi)) \le \epsilon_n \Big\} \quad (27)$$

Furthermore, for any sequence δ_n, we define a parameter class for the null hypothesis and a parameter class for the alternative hypothesis:

$$\begin{aligned} \mathcal{Q}_{0n}^*(c_0, \epsilon_n) &= \mathcal{Q}_n(c_0, \epsilon_n) \cap \{\xi : \omega_n(\xi) = 0\}, \\ \mathcal{Q}_{1n}^*(\delta_n; c_0, \epsilon_n) &= \mathcal{Q}_n(c_0, \epsilon_n) \cap \left\{ \xi : \frac{n\bar{N}\|\mu(\xi)\|^2 \omega_n^2(\xi)}{\sqrt{\sum_{k=1}^K \|\mu_k(\xi)\|^2}} \ge \delta_n \right\}. \end{aligned} \quad (28)$$

THEOREM 5. *Fix a constant $c_0 \in (0, 1)$ and positive sequences ϵ_n and δ_n such that $\epsilon_n \to 0$ as $n \to \infty$. For any sequence of $(n, p, K, \bar{N})$ indexed by n, consider Models (1)-(2) for $\Omega \in \mathcal{Q}_n(c_0, \epsilon_n)$. Let $\mathcal{Q}_{0n}^*(c_0, \epsilon_n)$ and $\mathcal{Q}_{1n}^*(\delta_n; c_0, \epsilon_n)$ be as in (28). If $\delta_n \to 0$, then* $\limsup_{n\to\infty} \inf_{\Psi \in \{0,1\}} \big\{ \sup_{\xi \in \mathcal{Q}_{0n}^*(c_0, \epsilon_n)} \mathbb{P}_\xi(\Psi = 1) + \sup_{\xi \in \mathcal{Q}_{1n}^*(\delta_n; c_0, \epsilon_n)} \mathbb{P}_\xi(\Psi = 0) \big\} = 1$.

3.4. The special case of $K = 2$

The special case of $K = 2$ is found in closeness testing and authorship attribution. We study this case more carefully. Given $\{X_i\}_{1 \le i \le n}$ and $\{G_i\}_{1 \le i \le m}$, we assume

$$X_i \sim \mathrm{Multinomial}(N_i, \Omega_i), \qquad G_j \sim \mathrm{Multinomial}(M_j, \Gamma_j). \quad (29)$$

Write $\bar{N} = n^{-1} \sum_{i=1}^n N_i$ and $\bar{M} = m^{-1} \sum_{i=1}^m M_i$. The null hypothesis becomes

$$H_0 : \quad \eta = \theta, \qquad \text{where } \eta = \frac{1}{n\bar{N}} \sum_{i=1}^n N_i \Omega_i, \text{ and } \theta = \frac{1}{m\bar{M}} \sum_{i=1}^m M_i \Gamma_i, \quad (30)$$

where θ and η are the two group-wise mean PMFs. In this case, the test statistic ψ has a more explicit form as in (18)-(19).

In our previous results for a general K, the regularity conditions (e.g., (21)) impose restrictions on the balance of sample sizes among groups. For $K = 2$, the severely unbalanced setting is interesting (e.g., in authorship attribution, $n = 1$ and m can be large). We relax the regularity conditions to the following ones:

CONDITION 1. *Let θ and η be as in* (30) *and define two matrices $\Sigma_1 = \frac{1}{n\bar{N}} \sum_{i=1}^{n} N_i \Omega_i \Omega_i'$ and $\Sigma_2 = \frac{1}{m\bar{M}} \sum_{i=1}^{m} M_i \Gamma_i \Gamma_i'$. We assume that the following statements are true (a) For $1 \le i \le n$ and $1 \le j \le m$, $N_i \ge 2$, $\|\Omega_i\|_\infty \le 1 - c_0$, $M_j \ge 2$, and $\|\Gamma_j\|_\infty \le 1 - c_0$, where $c_0 \in (0,1)$ is a contant, (b) $\max\big\{\big(\frac{\|\eta\|_3^3}{n\bar{N}} + \frac{\|\theta\|_3^3}{m\bar{M}}\big), \big(\frac{\|\eta\|_2^2}{n^2\bar{N}^2} + \frac{\|\theta\|_2^2}{m^2\bar{M}_2^2}\big)\big\} \big/ \big\|\frac{m\bar{M}}{n\bar{N}+m\bar{M}}\eta + \frac{n\bar{N}}{n\bar{N}+m\bar{M}}\theta\big\|^4 = o(1)$, (c) $\max\big\{\sum_i \frac{N_i^2}{n^2\bar{N}^2}\|\Omega_i\|_3^3, \sum_i \frac{M_i^2}{m^2\bar{M}^2}\|\Gamma_i\|_3^3, \|\Sigma_1\|_F^2 + \|\Sigma_2\|_F^2\big\} \big/ \|\mu\|^2 = o(1)$, and (d) $\|\mu\|_4^4/\|\mu\|^4 = o(1)$.*

Condition (a) is similar to (21), except that we drop the sample size balance requirement. Conditions (b)-(d) are equivalent to (24) but have more explicit expressions for $K = 2$.

THEOREM 6. *In Model* (29), *we test the null hypothesis H_0: $\theta = \mu$. As $\min\{n\bar{N}, m\bar{M}\} \to \infty$, suppose Condition 1 is satisfied. Under the alternative hypothesis, we further assume*

$$\frac{\|\eta - \theta\|^2}{\big(\frac{1}{n\bar{N}} + \frac{1}{m\bar{M}}\big)\max\{\|\eta\|, \|\theta\|\}} \to \infty. \tag{31}$$

Consider the DELVE test statistic $\psi = T/\sqrt{V}$. The following statements are true. Under the null hypothesis, $\psi \to N(0,1)$ in distribution. Under the alternative hypothesis, $\psi \to \infty$ in probability. Moreover for any fixed $\kappa \in (0,1)$, the level-κ DELVE test has an asymptotic level of κ and an asymptotic power of 1.

Compared with the theorems for a general K, first, Theorem 6 allows the two groups to be severely unbalanced and reveals that the detection boundary depends on the harmonic mean of $n\bar{N}$ and $m\bar{M}$. Second, the detection boundary is expressed using $\|\eta - \theta\|$, which is easier to interpret. We also note that, when $K = 2$, straightforward calculation yields $\mathbb{E}[T] = \rho^2 = (\frac{1}{n\bar{N}} + \frac{1}{m\bar{M}})^{-1}\|\eta - \theta\|^2$, which explains the appearance of the harmonic means in the detection boundary (31).

3.5. The special case of $K = n$

The special case of $K = n$ is interesting for two reasons. First, the application example of global testing in topic models corresponds to $K = n$. Second, for any K, when Ω_i's within each group are assumed to be the same (e.g., this is the case in closeness testing of discrete distributions), it suffices to aggregate the counts in each group, i.e., let $Y_k = \sum_{i \in S_k} X_i$ and operate on $Y_1, \ldots, Y_K$ instead of the original X_i's; this reduces to the case of $K = n$.

When $K = n$, the null hypothesis has a simpler form:

$$H_0: \quad \Omega_i = \mu, \qquad 1 \le i \le n. \tag{32}$$

Moreover, under the alternative hypothesis, the quantity ω_n^2 in (25) simplifies to

$$\omega_n = \omega_n(\Omega_1, \Omega_2, \ldots, \Omega_n) = \frac{1}{n\bar{N}\|\mu\|^2} \sum_{i=1}^{n} N_i \|\Omega_i - \mu\|^2. \tag{33}$$

The DELVE test statistic also has a simplified form as in (14)-(15). We can prove the same theoretical results under *weaker conditions*:

CONDITION 2. *We assume that the following statements are true: (a) For a constant* $c_0 \in (0,1)$, $2 \le N_i \le (1-c_0)n\bar{N}$ *and* $\|\Omega_i\|_\infty \le 1-c_0$, $1 \le i \le n$, *and (b)* $\max\big\{\sum_i \frac{\|\Omega_i\|_3^3}{N_i}, \sum_i \frac{\|\Omega_i\|^2}{N_i^2}\big\}/(\sum_i \|\Omega_i\|^2)^2 = o(1)$, *and* $(\sum_i \|\Omega_i\|_3^3)/(n\|\mu\|^2) = o(1)$

When $K = n$, Condition (a) is equivalent to (21); and Condition (b) is weaker than (24), as we have dropped the requirement $\frac{\|\mu\|_4^4}{K\|\mu\|^4} = o(1)$. We obtain weaker conditions for $K = n$ because the dominant terms in T differ from those for $K < n$.

THEOREM 7. *In Model* (1), *we test the null hypothesis* (32). *As* $n \to \infty$, *we assume that Condition 2 is satisfied. Under the alternative, we further assume that*

$$\frac{n\bar{N}\|\mu\|^2\omega_n^2}{\sqrt{\sum_{i=1}^n \|\Omega_i\|^2}} \to \infty. \tag{34}$$

Let T *and* V^* *be the same as in* (14)-(15). *Consider the simplified DELVE test statistic* $\psi^* = T/\sqrt{V^*}$. *Under the null hypothesis,* $\psi^* \to N(0,1)$ *in distribution. Under the alternative hypothesis,* $\psi^* \to \infty$ *in probability. Moreover, for any fixed* $\kappa \in (0,1)$, *the level-*κ *DELVE test has an asymptotic level of* κ *and an asymptotic power of* 1.

The detection boundary in (34) has a simpler form if $\sum_i \|\Omega_i\|^2 \asymp n\|\mu\|^2$. In this case, (34) is equivalent to $\sqrt{n}\bar{N}\|\mu\|\omega_n^2 \to \infty$. Additionally, if all entries of μ are at the same order, then $\|\mu\| \asymp 1/\sqrt{p}$, and (34) further reduces to $\sqrt{n\bar{N}^2/p}\cdot\omega_n^2 \to \infty$.

3.6. A discussion of the contiguity regime

Our power analysis in Section 3.2 concerns $\mathrm{SNR}_n \to \infty$, and our lower bound in Section 3.3 concerns $\mathrm{SNR}_n \to 0$. We now study the contiguity regime where SNR_n tends to a constant. For illustration, we consider a special choice of parameters, which allows us to obtain a simple expression of the testing risk.

Suppose $K = n$ and $N_i = N$ for all $1 \le i \le n$. Consider the pair of hypotheses:

$$H_0: \ \Omega_{ij} = p^{-1}, \qquad \text{v.s.} \qquad H_1: \ \Omega_{ij} = p^{-1}(1+\nu_n\delta_{ij}), \tag{35}$$

where $\{\delta_{ij}\}_{1\le i\le n, 1\le j\le p}$ satisfy that $|\delta_{ij}| = 1$, $\sum_{j=1}^p \delta_{ij} = 0$ and $\sum_{i=1}^n \delta_{ij} = 0$. Such δ_{ij} always exist.‡ The SNR_n in (26) satisfies that $\mathrm{SNR}_n \asymp (N\sqrt{n}/\sqrt{p})\nu_n^2$. We thereby set

$$\nu_n^2 = \frac{\sqrt{2p}}{N\sqrt{n}}\cdot a, \qquad \text{for a constant } a > 0. \tag{36}$$

Since $K = n$ here, we consider the simplified DELVE test statistic ψ^* as in Section 3.5.

‡For example, we can first partition the dictionary into two halves and then partition all the documents into two halves; this divides $\{1,2,\ldots,p\}\times\{1,2,\ldots,n\}$ into four subsets; we construct δ_{ij}'s freely on one subset and then specify the δ_{ij}'s on the other three subsets by symmetry.

THEOREM 8. *Consider Model* (1) *with* $N_i = N$. *For a constant* $a > 0$, *let the null and alternative hypotheses be specified as in* (35)-(36). *As* $n \to \infty$, *if* $p = o(N^2 n)$, *then* $\psi^* \to N(0,1)$ *under* H_0 *and* $\psi^* \to N(a,1)$ *under* H_1.

Let Φ be the cumulative distribution function of the standard normal. By Theorem 8, for any fixed constant $t \in (0, a)$, if we reject the null hypothesis when $\psi^* > t$, then the sum of type I and type II errors converges to $[1 - \Phi(t)] + [1 - \Phi(a - t)]$.

4. Applications to other statistical problems

As mentioned in Section 1, our testing problem includes global testing for topic models, authorship attribution, and closeness testing for discrete distributions as special examples. In this section, the DELVE test is applied separately to these three problems.

4.1. Global testing for topic models

Topic modeling (Blei et al., 2003) is a popular tool in text mining. It aims to learn a small number of "topics" from a large corpus. Given n documents written using a dictionary of p words, let $X_i \sim \text{Multinomial}(N_i, \Omega_i)$ denote the word counts of document i, where N_i is the length of this document and $\Omega_i \in \mathbb{R}^p$ contains the population word frequencies. In a topic model, there exist M topic vectors $A_1, A_2, \ldots, A_M \in \mathbb{R}^p$, where each A_k is a PMF. Let $w_i \in \mathbb{R}^M$ be a nonnegative vector whose entries sum up to 1, where $w_i(k)$ is the "weight" document i puts on topic k. It assumes

$$\Omega_i = \sum_{k=1}^{M} w_i(k) A_k, \qquad 1 \le i \le n. \tag{37}$$

Under (37), the matrix $\Omega = [\Omega_1, \Omega_2, \ldots, \Omega_n]$ admits a low-rank nonnegative factorization.

Before fitting a topic model, we would like to know whether the corpus indeed involves multiple topics. This is the global testing problem: $H_0 : M = 1$ v.s. $H_1 : M > 1$. When $M = 1$, by writing $A_1 = \mu$, the topic model reduces to the null hypothesis in (32). We can apply the DELVE test by treating each X_i as a separate group (i.e., $K = n$).

COROLLARY 1. *Consider Model* (1) *and define a vector* $\xi \in \mathbb{R}^n$ *by* $\xi_i = \bar{N}^{-1} N_i$. *Suppose that* $\Omega = \mu \mathbf{1}_n'$ *under the null hypothesis, with* $\mu = n^{-1}\Omega\xi$, *and that* Ω *satisfies* (37) *under the alternative hypothesis, with* $r := \text{rank}(\Omega) \ge 2$. *Suppose* $\bar{N}/(\min_i N_i) = O(1)$. *Denote by* $\lambda_1, \lambda_2, \ldots, \lambda_r > 0$ *the singular values of* $\Omega[\text{diag}(\xi)]^{1/2}$, *arranged in the descending order. We further assume that under the alternative hypothesis,*

$$\bar{N} \cdot \frac{\sum_{k=2}^{r} \lambda_k^2}{\sqrt{\sum_{k=1}^{r} \lambda_k^2}} \to \infty. \tag{38}$$

For any fixed $\kappa \in (0,1)$, *the level-*κ *DELVE test has an asymptotic level* κ *and an asymptotic power* 1.

The least-favorable configuration in the proof of Theorem 5 is in fact a topic model that follows (37) with $M = 2$. Transferring the argument yields the following lower bound that confirms the optimality of DELVE for the global testing of topic models.

COROLLARY 2. *Let $\mathcal{R}_{n,M}(\epsilon_n, \delta_n)$ be the collection of $\{(N_i, \Omega_i)\}_{i=1}^n$ satisfying the following conditions: 1) Ω follows the topic model (37) with M topics; 2) Condition 2 holds with $o(1)$ replaced by $\le \epsilon_n$; 3) $\bar{N}(\sum_{k=2}^r \lambda_k^2)/(\sum_{k=1}^r \lambda_k^2)^{1/2} \ge \delta_n$. If $\epsilon_n \to 0$ and $\delta_n \to 0$, then* $\limsup_{n\to\infty} \inf_{\Psi\in\{0,1\}} \Big\{ \sup_{\mathcal{R}_{n,1}(\epsilon_n,0)} \mathbb{P}(\Psi = 1) + \sup_{\cup_{M\ge 2}\mathcal{R}_{n,M}(\epsilon_n,\delta_n)} \mathbb{P}(\Psi = 0) \Big\} = 1.$

The detection boundary (38) can be simplified when $M = O(1)$. Following Ke and Wang (2022), we define $\Sigma_A = A'H^{-1}A$ and $\Sigma_W = n^{-1}WW'$, where $A = [A_1, A_2, \ldots, A_M]$, $W = [w_1, w_2, \ldots, w_n]$ and $H = \mathrm{diag}(A\mathbf{1}_M)$. Ke and Wang (2022) argued that it is reasonable to assume that eigenvalues of these two matrices are at the constant order. If this is true, with some mild additional regularity conditions, each λ_k is at the order of $\sqrt{n/p}$. Hence, (38) reduces to $\sqrt{n}\bar{N}/\sqrt{p} \to \infty$. In comparison, Ke and Wang (2022) showed that a necessary condition for any estimator $\hat{A} = [\hat{A}_1, \hat{A}_2, \ldots, \hat{A}_M]$ to achieve $\frac{1}{M}\sum_{k=1}^M \|\hat{A}_k - A_k\|_1 = o(1)$ is $\sqrt{n\bar{N}/p} \to \infty$. We conclude that consistent estimation of topic vectors requires strictly stronger conditions than successful testing.

4.2. Authorship attribution

In authorship attribution, given a corpus from a known author, we want to test whether a new document is from the same author. It is a special case of our testing problem with $K = 2$. We can directly apply the results in Section 3.4. However, the setting in Section 3.4 has no sparsity. Kipnis (2022); Donoho and Kipnis (2022) point out that the number of words with discriminating power is often much smaller than p. To see how our test performs under sparsity, we consider a sparse model. As in Section 3.4, let

$$X_i \sim \mathrm{Multinomial}(N_i, \Omega_i),\ 1 \le i \le n, \quad \text{and} \quad G_i \sim \mathrm{Multinomial}(M_i, \Gamma_i),\ 1 \le i \le m. \tag{39}$$

Let $\bar{N}$ and $\bar{M}$ be the average of N_i's and M_i's, respectively. Write $\eta = \frac{1}{n\bar{N}}\sum_{i=1}^n N_i\Omega_i$ and $\theta = \frac{1}{m\bar{M}}\sum_{i=1}^m M_i\Gamma_i$. We assume for some $\zeta_n > 0$,

$$\eta_j = \theta_j, \ \text{ for } j \notin S, \qquad \text{and} \qquad \big|\sqrt{\eta_j} - \sqrt{\theta_j}\big| \ge \zeta_n, \ \text{ for } j \in S. \tag{40}$$

COROLLARY 3. *Under the model (39)-(40), consider testing $H_0 : S = \emptyset$ v.s. $H_1 : S \ne \emptyset$, where Condition 1 is satisfied. Let η_S and θ_S be the sub-vectors of η and θ restricted to the coordinates in S. Suppose that under the alternative hypothesis,*

$$\frac{\zeta_n^2 \cdot (\|\eta_S\|_1 + \|\theta_S\|_1)}{\big(\frac{1}{n\bar{N}} + \frac{1}{m\bar{M}}\big)\max\{\|\eta\|, \|\theta\|\}} \to \infty. \tag{41}$$

As $\min\{n\bar{N}, m\bar{M}\} \to \infty$, the level-$\kappa$ DELVE test has an asymptotic level κ and an asymptotic power 1. Furthermore, if $n\bar{N} \asymp m\bar{M}$ and $\min_{j\in S}(\eta_j + \theta_j) \ge cp^{-1}$ for a constant $c > 0$, then (41) reduces to $n\bar{N}\zeta_n^2|S|/\sqrt{p} \to \infty$.

Donoho and Kipnis (2022) studied a case where $N = M$, $n = m = 1$, $p \to \infty$,

$$|S| = p^{1-\vartheta}, \qquad \text{and} \qquad \zeta_n = c \cdot N^{-1/2}\sqrt{\log(p)}. \tag{42}$$

When $\vartheta > 1/2$ (i.e., $|S| = o(\sqrt{p})$), they derived a phase diagram for the aforementioned testing problem (under a slightly different setting where the data distributions are Poisson instead of multinomial). They showed that when $\vartheta > 1/2$ and c is a properly large

constant, a Higher-Criticism-based test has an asymptotically full power. Donoho and Kipnis (2022) did not study the case of $\vartheta \leq 1/2$. By Corollary 3, when $\vartheta \leq 1/2$ (i.e., $|S| \geq C\sqrt{p}$), the DELVE test has asymptotically full power.

REMARK 5. When $\vartheta > 1/2$ in (42), the DELVE test loses power. However, we can borrow the idea of maximum test or Higher Criticism test (Donoho and Jin, 2004). For example, recalling T_j in (9), we may use $\max_{1\leq j\leq p}\{T_j/\sqrt{V_j}\}$ as the test statistic, where V_j is a proper estimator of the variance of T_j. We leave this to future work.

4.3. Closeness testing between discrete distributions

Two-sample closeness testing is a subject of intensive study in discrete distribution inference (Bhattacharya and Valiant, 2015; Chan et al., 2014; Diakonikolas and Kane, 2016; Kim et al., 2022). It is a special case of our problem with $K = 2$ and $n_1 = n_2 = 1$. We thereby apply both Theorem 6 and Theorem 7.

COROLLARY 4. *Let Y_1 and Y_2 be two discrete variables taking values on the same p outcomes. Let $\Omega_1 \in \mathbb{R}^p$ and $\Omega_2 \in \mathbb{R}^p$ be their corresponding PMFs. Suppose we have N_1 samples of Y_1 and N_2 samples of Y_2. The data are summarized in two multinomial vectors: $X_1 \sim \text{Multinomial}(N_1, \Omega_1), X_2 \sim \text{Multinomial}(N_2, \Omega_2)$. We test $H_0 : \Omega_1 = \Omega_2$. Write $\mu = \frac{1}{N_1+N_2}(N_1\Omega_1 + N_2\Omega_2)$. Suppose $\min\{N_1, N_2\} \geq 2$, $\max\{\|\Omega_1\|_\infty, \|\Omega_2\|_\infty\} \leq 1-c_0$, for a constant $c_0 \in (0, 1)$. Suppose $\frac{1}{(\sum_{k=1}^2 \|\Omega_k\|^2)^2} \max\big\{\sum_{k=1}^2 \frac{\|\Omega_k\|_3^3}{N_k}, \sum_{k=1}^2 \frac{\|\Omega_k\|^2}{N_k^2}\big\} = o(1)$, and $\frac{1}{n\|\mu\|^2}\sum_{k=1}^2 \|\Omega_k\|_3^3 = o(1)$. We assume that under the alternative hypothesis,*

$$\frac{\|\Omega_1 - \Omega_2\|^2}{\big(N_1^{-1} + N_2^{-1}\big)\max\{\|\Omega_1\|, \|\Omega_2\|\}} \to \infty. \tag{43}$$

As $\min\{N_1, N_2\} \to \infty$, the level-$\kappa$ DELVE test has level κ and power 1, asymptotically.

The requirement (43) matches with the minimum ℓ^2-separation condition for two-sample closeness testing (Kim et al., 2022, Proposition 4.4). Hence, our test is an optimal ℓ^2-testor. Other optimal ℓ^2-testors (Chan et al., 2014; Bhattacharya and Valiant, 2015; Diakonikolas and Kane, 2016) are not equipped with tractable null distributions.

REMARK 6. We can modify DELVE to incorporate frequency-dependent weights. Define $T(w) := \sum_{j=1}^p w_j T_j$, where T_j is the same as in (9) and let $w_j = \big(\max\{1/p,\ \hat{\mu}_j\}\big)^{-1}$. Such weights were used in discrete distribution inference (Balakrishnan and Wasserman, 2019; Chan et al., 2014) to turn an optimal ℓ^2 testor to an optimal ℓ^1 testor. We can similarly study the power of the test based on $T(w)$, except that we need an additional assumption $n\tilde{N} \gg p$ to guarantee that $\hat{\mu}_j$ is a sufficiently accurate estimator of μ_j.

5. Simulations

We investigate the numerical performance of DELVE in simulations. Recall that we introduced a variant of DELVE, DELVE+, in Section 2.2. DELVE+ has similar theoretical properties but is more suitable for real data. We include both versions in simulations.

Experiment 1 (Asymptotic normality). Given $(n, p, K, N_{\min}, N_{\max}, \phi)$, we generate data as follows: first, divide $\{1, \ldots, n\}$ into K equal-size groups. Next, we draw $\Omega_1^{alt}, \ldots, \Omega_n^{alt}$

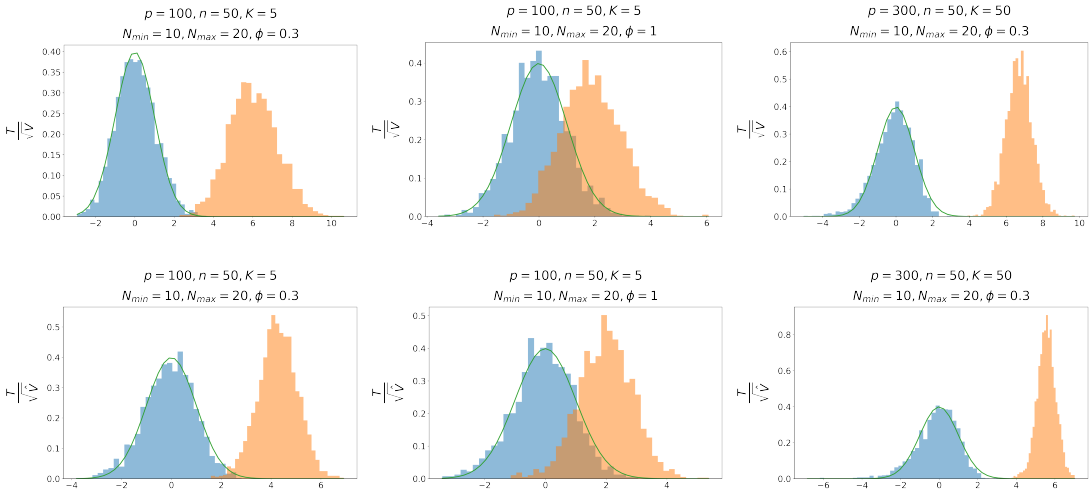


Fig. 1: Histograms of DELVE (top panels) and DELVE+ (bottom panels) statistics in Experiments 1.1-1.3. In each plot, the blue and orange histograms correspond to the null and alternative hypotheses, respectively; and the green curve is the density of $N(0,1)$.

i.i.d. from Dirichlet$(p, \phi \mathbf{1}_p)$. Third, we draw $N_i \overset{iid}{\sim} \text{Uniform}[N_{\min}, N_{\max}]$ and set $\Omega_i^{null} = \mu$, where $\mu := \frac{1}{nN}\sum_i N_i \Omega_i^{alt}$. Last, we generate $X_1, \ldots, X_n$ using Model (1). We consider three sub-experiments. In Experiment 1.1, $(n, p, K, N_{\min}, N_{\max}, \phi) = (50, 100, 5, 10, 20, 0.3)$. In Experiment 1.2, ϕ is changed to 1, and the other parameters are the same. When $\phi = 1$, Ω_i^{alt} are drawn from the uniform distribution of the standard probability simplex; in comparison, $\phi = 0.3$ puts more mass near the boundary of the standard probability simplex. In Experiment 1.3, we keep all parameters the same as in Experiment 1.1, except that (p, K) are changed to $(300, 50)$. For each sub-experiment, we generate 2000 data sets under the null hypothesis and plot the histogram of the DELVE test statistic ψ (in blue); similarly, we generate 2000 data sets under the alternative hypothesis and plot the histogram of ψ (in orange). The results are contained in Figure 1.

In all sub-experiments, when the null hypothesis holds, the histograms of DELVE and DELVE+ fit the standard normal density reasonably well. This supports our theory in Section 3.1. Second, when (p, K) increase, the finite sample effect becomes slightly more pronounced (c.f., Experiment 1.3 versus Experiment 1.1). Third, the tests have power in differentiating two hypotheses. As ϕ decreases or K increases, the power increases, and the two histograms become further apart. Last, in the alternative hypothesis, DELVE+ has smaller mean and variance than DELVE. By Lemma 2, they have similar asymptotic behaviors. The simulations suggest that they have noticeable finite-sample differences.

Experiment 2 (Power curve). Similarly as in Experiment 1, we divide $\{1, 2, \ldots, n\}$ into K equal-size groups and draw $N_i \sim \text{Uniform}[N_{\min}, N_{\max}]$. In this experiment, Ω_i's are generated in a different way. Under H_0, we draw $\mu \sim \text{Dirichlet}(p/2, \phi \mathbf{1}_{p/2})$ and set $\Omega_i^{null} = \tilde{\mu}$, where $\tilde{\mu}_j = \frac{1}{2}\mu_j$ for $j \le p/2$ and $\tilde{\mu}_j = \frac{1}{2}\mu_{j-p/2}$ for $j \ge p/2{+}1$. Under H_1, fixing some $\tau_n \in [0, 1]$, we draw $z_1, \ldots, z_K,\ b_1, \ldots, b_{p/2} \overset{iid}{\sim} \text{Rademacher}(1/2)$ and let $\Omega_{ij}^{alt} = \tilde{\mu}_j(1 + \tau_n z_k b_j)$, for i in group k and $1 \le j \le p/2$, and let $\Omega_{ij}^{alt} = \tilde{\mu}_j(1 - \tau_n z_k b_{j-p/2})$ for

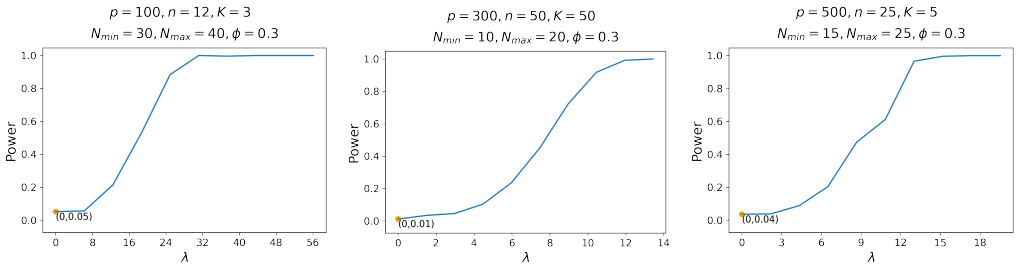


Fig. 2: Power of the level-5% DELVE test (x-axis represents the SNR $\lambda(\tau_n) = \frac{n\bar{N}\|\mu\|\tau_n^2}{\sqrt{K}}$).

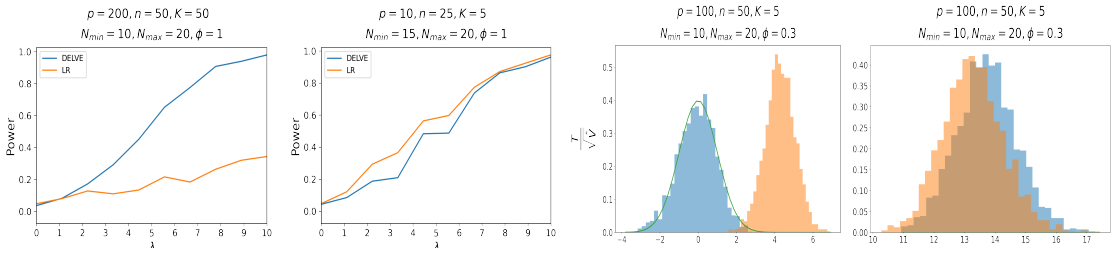


Fig. 3: Comparison of DELVE+, LR, and ANOVA (details are in Experiment 3).

$p/2+1 \le j \le p$. By applying our theory in Section 3.2 together with some calculations, the signal-to-noise ratio is captured by $\lambda(\tau_n) := K^{-1/2}n\bar{N}\|\mu\|\tau_n^2$. In particular, it holds that $\omega_n^2(\Omega^{alt}) = \tau_n^2$, for the ω_n^2 defined in (25). We consider three sub-experiments, Experiment 2.1-2.3, where the parameter values of $(n, p, K, N_{\min}, N_{\max}, \phi)$ are the same as in Experiments 1.1-1.3. For each sub-experiment, we consider a grid of 10 equally-spaced values of λ. When $\lambda = 0$, it corresponds to H_0; when $\lambda > 0$, it corresponds to H_1. For each λ, we generate 500 data sets and compute the fraction of rejections of the level-5% DELVE test. This gives a power curve for the level-5% DELVE test, in which the first point associated with $\lambda = 0$ is the actual level of the test. The results are in Figure 2. We repeat the same experiments for the DELVE+ test; owing to space limit, the plots are in Cai et al. (2023). In all three experiments, the actual level of our proposed tests is $\le 5\%$, suggesting that our tests perform well at controlling the type-I error. As λ increases, the power gradually increased to 1, suggesting that λ is a good metric of the signal-to-noise ratio. This supports our theory in Section 3.2.

Experiment 3 (Comparison with the LR and ANOVA tests). This experiment contains two sub-experiments. In Experiment 3.1, we compare DELVE+ with the likelihood ratio (LR) test. The LR test is only well-defined in the special case where Ω_i's are equal within each group. In this case, $T^{\mathrm{LR}} = \sum_k n_k \bar{N}_k \sum_j \hat{\mu}_{kj} \log\left(\frac{\hat{\mu}_{kj}}{\hat{\mu}_j}\right)$, where $\hat{\mu}_k$ and $\hat{\mu}$ are the same as in (5), and $\log(0/0) = 0$. Given $(n, p, K, N_{\min}, N_{\max}, \phi)$, we generate data in the same way as in Experiment 2 (these settings guarantee that Ω_i's are equal within-group, hence favoring the LR test). Since no asymptotic normality result is known for T^{LR}, we use an ideal threshold for the LR test - drawing 500 data sets from the null model ($\lambda = 0$) and computing the empirical 95%-quantile of T^{LR}. The power curves for two representative settings ($p = 200$ and $p = 10$) are shown in the left two panels of Figure 3. More

settings can be found in Section A.2 of Cai et al. (2023). We observe that DELVE+ significantly outperforms LR when p is large/moderate compared to n, and they perform similarly (with LR being slightly better) when p is small. In Experiment 3.2, we compare DELVE+ with the ANOVA test that uses $\widetilde{T}$ in (8) as the test statistic. The simulation settings are the same as in Experiment 1.1. The third panel of Figure 3 is a replication of the bottom left panel of Figure 1 and shows the histograms of DELVE+ test statistics under two hypotheses. The fourth panel of Figure 3 contains the histograms of $\widetilde{T}$. We see that $\widetilde{T}$ fails to distinguish two hypotheses while DELVE+ is able to do so. As explained in Remark 4, the naive ANOVA test can lose power due to the lack of de-biasing.

6. Real Data Analysis

We consider two real corpora consisting of statistical paper abstracts and Amazon movie reviews, respectively. We use them to showcase: Although testing the null hypothesis (3) is only a binary decision problem, it can be used to answer various questions of interest by simply varying the definition of "groups" in (3). For example, we may define "groups" of movie reviews by movie title, star rating, posting time, reviewer characteristics, etc.. Then, our test can detect many different kinds of heterogeneity in movie reviews (the same holds for other product reviews). In Section 2, we proposed DELVE and DELVE+ and explained that the latter is more suitable for real data; hence, we use DELVE+ here.

6.1. Abstracts of statisticians

The data set from Ji and Jin (2016) contains the bibtex information of published papers in four top-tier statistics journals, *Annals of Statistics*, *Biometrika*, *Journal of the American Statistical Association*, and *Journal of the Royal Statistical Society - Series B*, from 2003 to the first half of 2012. In the pre-processing step, we first remove common stop words such as "for", "also", "can", and "the", and common domain-specific words such as "statistician", "estimate", and "sample". We then perform *stemming*, which maps together words with a common prefix such as "play", "player", and "playing". Finally, we perform tokenization, which maps each abstract to its vector of word (stem) counts.

We conduct two experiments. In the first one, we fix an author and treat the collection of his/her co-authored abstracts as a corpus. We apply DELVE+ with $K = n$, where n is the number of abstracts written by this author. The Z-score measures the "diversity" or "variability" of this authors' abstracts. An author with a high Z-score possesses either diverse research interests or a variable writing style. A number of authors have only 1–2 papers, and the variance estimator V is often negative; we remove all those authors. In Figure 4 (left), we plot the histogram of Z-scores of retained authors. The mean is 4.52 and the standard deviation is 2.94. In Figure 4 (middle), we show the plot of Z-score versus logarithm of the number of abstracts written by this author. The most prolific author has 82 papers and a Z-score larger than 20, implying a huge diversity in his/her abstracts. There is also a positive association between Z-score and number of papers. It suggests that senior authors have more diversity in their abstracts, which is intuitive.

In the second experiment, we further divide an author's abstracts into smaller groups by publication year. Owing to space limit, we only show the results for the most prolific author who has 82 papers, but we keep in mind that the same analysis can be done for each author in the data set (see Cai et al. (2023)). We divide this author's abstracts into

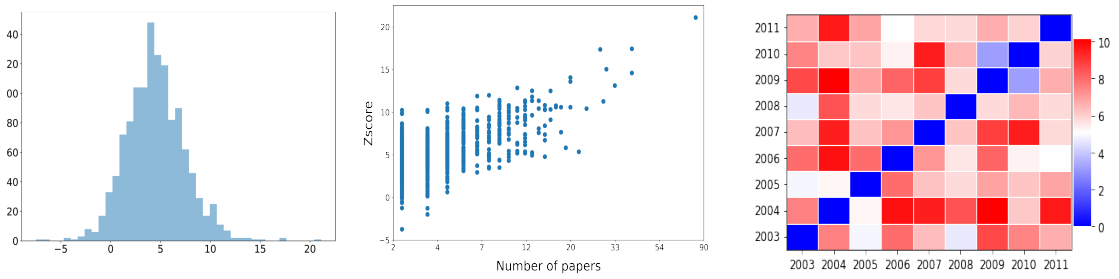


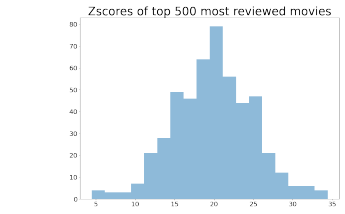
Fig. 4: Results about statistical abstracts. Left: Histogram of author Z-scores (mean is 4.52, and standard deviation is 2.94). Middle: Author Z-score versus number of papers. Right: Pairwise Z-score plot for a representative author.

9 groups, each group corresponding to one year. For each pair of groups, we implement DELVE+ with $K = 2$. This yields a pairwise plot of Z-scores, as shown in Figure 4 (right). It reveals the temporal patterns of this author in abstract writing. The group consisting of 2004-2005 abstracts has comparably large Z-scores in the pairwise comparison with other groups. To interpret the results, we read titles and abstracts of all of this author's papers and found that in 2004-2005 he/she extensively studied topics related to bandwidth selection in the context of nonparametric estimation.

Remark 7. The asymptotic normality in Section 3.1 is established under the condition $n^2\bar{N}^2 \gg Kp$. It is worth checking if this holds in real data. We compute $DR := n^2\bar{N}^2/(Kp)$ for all the corpora analyzed in the above two experiments (see Section B.2 of Cai et al. (2023)). These DR values are quite large. Therefore, it would be appropriate to apply the asymptotic normality result, and we think the Z-scores and p-values are trustworthy.

6.2. Amazon movie reviews

The dataset in Maurya (2018) contains 1,924,471 reviews of 143,007 visual media products (ie, DVDs, Bluray, or streams). We cleaned and stemmed these review text similarly as in Section 6.1. In the first experiment, given a movie, we consider the corpus consisting of all reviews of this movie and apply DELVE+ with $K = n$. The results are in the top panels of Figure 5. First, we plot the histogram of Z-scores for the top 500 most reviewed movies. The mean is 19.97 and the standard deviation is 5.07. Compared with the histogram of Z-scores for statistics paper abstracts, there is much larger diversity in movie reviews. Next, we list the 4 movies with the highest Z-scores and lowest Z-scores out of the 20 most reviewed movies. Each movie has more than 800 reviews, but some have surprisingly low Z-scores. The works by the comedian Jeff Dunham have the lowest Z-scores, suggesting strong homogeneity among the reviews. The 2012 horror film *Prometheus* has the highest degree of review diversity among the 20 most reviewed movies. In the second experiment, we further divide each movie's reviews into 5 groups by star rating. We compare each pair of groups using DELVE+ with $K = 2$, resulting in a pairwise Z-score plot. In the bottom panels of Figure 5, we plot this for 3 popular movies. We see a variety of polarization patterns among the scores. In *Harry Potter and the Deathly Hallows Part I*, DELVE+ signifies that the reviews with ratings in the



Rank	Title	Z-Score	Total reviews
1	Prometheus	34.44	813
2	Expelled: No Intelligence Allowed	34.17	830
3	V for Vendetta	32.24	815
4	Sin City	31.72	828
⋮	⋮	⋮	⋮
17	Cars	19.98	902
18	Food, Inc.	17.81	876
19	Jeff Dunham: Arguing with Myself	4.96	860
20	Jeff Dunham: Spark of Insanity	4.46	877

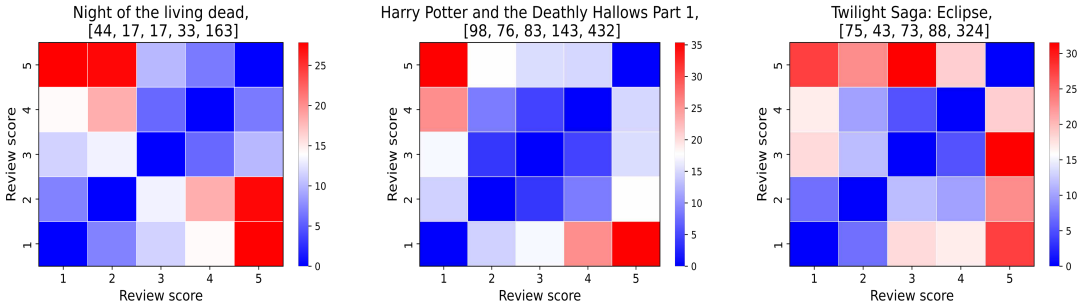


Fig. 5: Results about movie reviews. Top left: Histogram of Z-scores for the 500 most-reviewed movies (mean is 19.97, and standard deviation is 5.07). Top right: Information and Z-scores for the top 20 most reviewed movies. Bottom: Pairwise Z-score plots for 3 representative movies (the title lists the number of reviews of each rating from 1–5).

range 2–4 stars are all similar. We see a smooth gradation in how the 1-star reviews differ from those from 2–4 stars, and similarly for 5-star reviews versus those from 2–4 stars. *Twilight Saga: Eclipse* shows three clusters: 1–2 stars, 3–4 stars, and 5 star, while *Night of the living dead* shows two clusters: 1–2 stars and 3–5 stars.

As mentioned in Section 1, the marketing research aims to understand patterns of online customer reviews. Our DELVE testing framework is a flexible approach to detecting many kinds of heterogeneity in review text. If reviewer characteristics (e.g., gender) are available, we can group reviews by these characteristics and answer questions such as if female and male reviewers have different styles in writing review text. In the experiments here, we showcase how to use DELVE to find patters in movie ratings. Although many literature works have studied patterns of movie reviews (Baek et al., 2012), most are based on the distribution of numeric ratings. The three movies in Figure 5 have similar distributions of numerical ratings, but the patters in text reviews are considerably different. Such plots will be useful for improving rating systems, recommending movies to customers, and detecting fake reviews.

7. Discussions

We examine the testing for equality of PMFs of K groups of high-dimensional multinomial distributions. The proposed DELVE statistic has a parameter-free limiting null that allows for computation of Z-scores and p-values on real data. DELVE achieves the optimal detection boundary over the whole range of parameters $(n, p, K, \bar{N})$, including the high-dimensional case $p \to \infty$, which is very relevant to applications in text mining.

This work leads to interesting questions for future study. Recall that the ρ^2 defined

in (6) is a measure of heterogeneity among the group-wise means. So far, the focus is on testing $\rho^2 = 0$, but we may also consider estimation and inference of ρ^2. Assuming $\rho^2 = 0$, we have obtained a consistent variance estimator for the DELVE metric in (9) and established it asymptotic normality. To construct a confidence interval for ρ^2, we will need such results under the alternative hypothesis (where $\rho^2 \neq 0$). From Figure 1, the asymptotic normality still holds when $\rho^2 \neq 0$, except that stronger regularity conditions may be required. Inspired by the authorship attribution problem (Kipnis and Donoho, 2021; Kipnis, 2022), it is interesting to consider a sparse alternative hypothesis where the group mean vectors are equal except on a small set of "giveaway words". As discussed in Section 4.2, we may combine DELVE with the idea of higher criticism.

Another exciting future direction is to extend our methods from the 'bag-of-words' model to more realistic sequence-based models. One approach is to consider the counts of adjacent words (bi-grams) instead of raw word counts. More generally, one can consider the counts of short sequences of words, which are known as m-grams. It is possible that a suitably modified version of DELVE would perform well in a setting where the next word is generated according to a Markov transition kernel whose input is the previous $m-1$ observed words (Jurafsky and Martin, 2023). A final idea is to combine words that have similar meanings or are close in a word embedding into 'superwords' and to use these superword counts as the basis for DELVE. We leave them to future work.

Funding: The research of T. Tony Cai was supported in part by NSF Grant DMS-2015259 and NIH grant R01-GM129781. The research of Zheng Tracy Ke was supported in part by NSF CAREER Grant DMS-1943902.

Data availability: The data that support the findings of this study are available in GitHub at `https://github.com/ZhengTracyKe/DELVE`.

Conflict of interest: The authors claim that there is no conflict of interest.

References

Arias-Castro, E. and Wang, M. (2015) The sparse poisson means model. *Electronic Journal of Statistics*, **9**, 2170–2201.

Baek, H., Ahn, J. and Choi, Y. (2012) Helpfulness of online consumer reviews: Readers' objectives and review cues. *International Journal of Electronic Commerce*, **17**, 99–126.

Balakrishnan, S. and Wasserman, L. (2018) Hypothesis testing for high-dimensional multinomials: A selective review. *The Annals of Applied Statistics*, **12**, 727–749.

— (2019) Hypothesis testing for densities and high-dimensional multinomials: Sharp local minimax rates. *The Annals of Statistics*, **47**, 1893–1927.

Berrett, T. B. and Samworth, R. J. (2019) Nonparametric independence testing via mutual information. *Biometrika*, **106**, 547–566.

Bhattacharya, B. and Valiant, G. (2015) Testing closeness with unequal sized samples. *Advances in Neural Information Processing Systems*, **28**.

Blei, D., Ng, A. and Jordan, M. (2003) Latent dirichlet allocation. *Journal of Machine Learning Research*, **3**, 993–1022.

Cai, T., Ke, Z. T. and Turner, P. (2023) Supplementary material for "Testing high-dimensional multinomials with applications to text analysis". *Manuscript*.

Chan, S.-O., Diakonikolas, I., Valiant, P. and Valiant, G. (2014) Optimal algorithms for testing closeness of discrete distributions. In *Proc. 25th symposium on discrete algorithms*. SIAM.

Chevalier, J. A. and Mayzlin, D. (2006) The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research*, **43**, 345–354.

Diakonikolas, I. and Kane, D. M. (2016) A new approach for testing properties of discrete distributions. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, 685–694. IEEE.

Donoho, D. and Jin, J. (2004) Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*, 962–994.

Donoho, D. L. and Kipnis, A. (2022) Higher criticism to compare two large frequency tables, with sensitivity to possible rare and weak differences. *The Annals of Statistics*, **50**, 1447–1472.

Gomaa, W. H., Fahmy, A. A. et al. (2013) A survey of text similarity approaches. *International Journal of Computer Applications*, **68**, 13–18.

Hall, P. and Heyde, C. C. (2014) *Martingale limit theory and its application*. Academic press.

Hansen, S., McMahon, M. and Prat, A. (2018) Transparency and deliberation within the fomc: a computational linguistics approach. *The Quarterly Journal of Economics*, **133**, 801–870.

Ji, P. and Jin, J. (2016) Coauthorship and citation networks for statisticians. *The Annals of Applied Statistics*, **10**, 1779–1812.

Jurafsky, D. and Martin, J. H. (2023) *Speech and Language Processing*. 3rd edn. URL: `https://web.stanford.edu/~jurafsky/slp3/`. Online textbook.

Ke, Z. T. and Wang, M. (2022) Using SVD for topic modeling. *Journal of the American Statistical Association*, 1–16.

Kim, I., Balakrishnan, S. and Wasserman, L. (2022) Minimax optimality of permutation tests. *The Annals of Statistics*, **50**, 225–251.

Kipnis, A. (2022) Higher criticism for discriminating word-frequency tables and authorship attribution. *The Annals of Applied Statistics*, **16**, 1236–1252.

Kipnis, A. and Donoho, D. L. (2021) Two-sample testing of discrete distributions under rare/weak perturbations. In *2021 IEEE Int'l Symposium on Information Theory*. IEEE.

Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C. and Teichmann, S. A. (2015) The technology and biology of single-cell RNA sequencing. *Molecular cell*, **58**, 610–620.

Leung, X. Y. and Yang, Y. (2020) Are all five points equal? Scaling heterogeneity in hotel online ratings. *International Journal of Hospitality Management*, **88**, 102539.

Maurya, D. (2018) Web data: Amazon movie reviews. Electronic. Https://www.kaggle.com/datasets/dm4006/amazon-movie-reviews/metadata.

Mosteller, F. and Wallace, D. L. (1963) Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed federalist papers. *Journal of the American Statistical Association*, **58**, 275–309.

— (2012) *Applied Bayesian and classical inference: the case of the Federalist papers*. Springer Science & Business Media.

Valiant, G. and Valiant, P. (2017) An automatic inequality prover and instance optimal identity testing. *SIAM Journal on Computing*, **46**, 429–455.

A list of figure legends

- Figure 1: Histograms of DELVE (top panels) and DELVE+ (bottom panels) statistics in Experiments 1.1-1.3. In each plot, the blue and orange histograms correspond to the null and alternative hypotheses, respectively; and the green curve is the density of $N(0,1)$.
- Figure 2: Power of the level-5% DELVE test (x-axis represents the SNR $\lambda(\tau_n) = \frac{n\bar{N}\|\mu\|\tau_n^2}{\sqrt{K}}$).
- Figure 3: Comparison of DELVE+, LR, and ANOVA (details are in Experiment 3).
- Figure 4: Results about statistical abstracts. Left: Histogram of author Z-scores (mean is 4.52, and standard deviation is 2.94). Middle: Author Z-score versus number of papers. Right: Pairwise Z-score plot for a representative author.
- Figure 5: Results about movie reviews. Top left: Histogram of Z-scores for the 500 most-reviewed movies (mean is 19.97, and standard deviation is 5.07). Top right: Information and Z-scores for the top 20 most reviewed movies. Bottom: Pairwise Z-score plots for 3 representative movies (the title lists the number of reviews of each rating from 1–5).