Text analysis and testing of high-dimensional multinomials

Zheng Tracy Ke (Harvard University)

BIRS-IASM Workshop: Harnessing the power of latent structure models and modern Big Data learning, Dec. 12, 2023

Collaborators

- Tony Cai (U Penn Wharton)
- Paxton Turner (Harvard)

Cai, T., Ke, Z., & Turner, P. (2023). Testing High-dimensional Multinomials with Applications to Text Analysis. *JRSSB (to appear)*.

Problem: between-group variability in text

- Detecting variability in online customer reviews

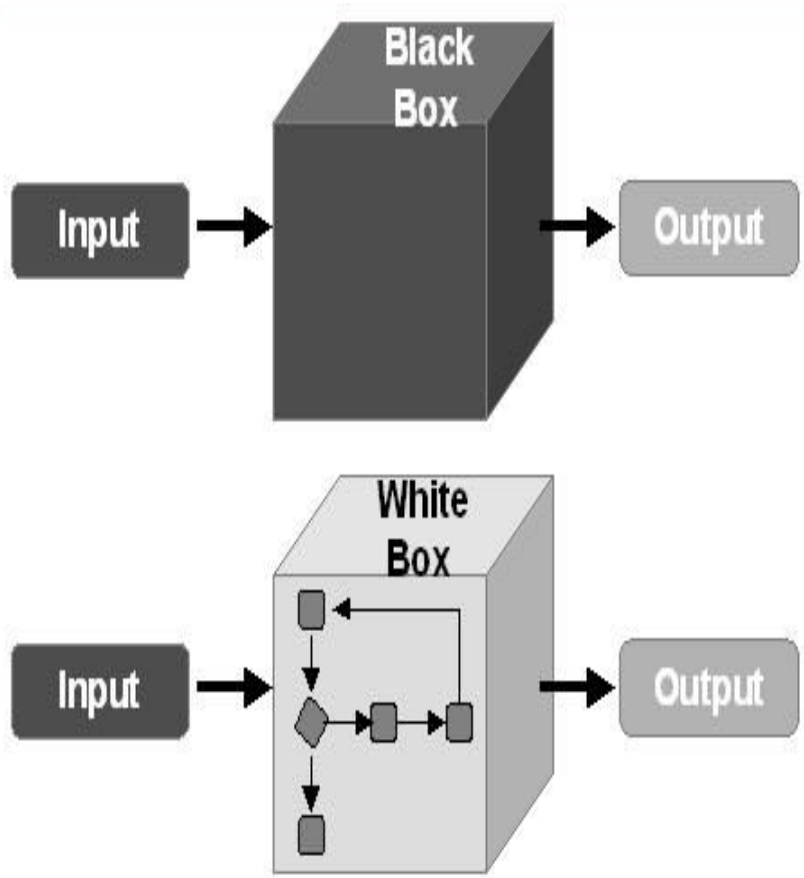
  "Younger travelers, women, and travelers with less review expertise tend to give more positive reviews to hotels" (Leung and Yang, 2020)

- Classical studies use numerical ratings, but text reviews are more informative, especially for hedonic products such as books, movies, and hotels (*Chevalier and Mayzlin, 2006*)

DNN models v.s. Statistical models

Successes of LLM were only reported under:

- Supervised learning
- Strong signals
- Extremely large (pre-)training data
- Interpretability is not required



Multinomial modeling of word counts

- The dictionary has p distinct words
- The N words in a document are sampled with replacement, using a probability mass function (PMF) $\Omega = (\Omega(1), \Omega(2), \ldots, \Omega(p))'$
- Let $X(j)$ denote the total count of word j

It follows that

$$X \sim \text{Multinomial}\Big(\underbrace{N}_{\#\text{words}}, \underbrace{\Omega}_{\text{PMF}}\Big)$$

The hypothesis testing problem

There are n documents:

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

- Documents are divided into K known groups
- The within-group mean PMF:

$$\mu_k = \frac{1}{\sum_{i \in S_k} N_i} \left(\sum_{i \in S_k} N_i \Omega_i \right) \quad \in \quad \mathbb{R}^p$$

- Test the null hypothesis

$$H_0 : \ \mu_1 = \mu_2 = \ldots = \mu_K$$

The challenges

$$X_i \sim \text{Multinomial}(N_i, \Omega_i),\ i \in S_1 \cup S_2 \cup \cdots \cup S_K.$$

$$H_0 : \quad \mu_1 = \mu_2 = \ldots = \mu_K, \qquad \mu_k : \text{within-group mean}$$

- H_0 is **highly composite**, since Ω_i's can be different from each other within each group
- High dimensionality, especially, allowing $p \gg \bar{N}$
- K can be any number between 2 and n
- Word frequencies are unbalanced (Zipf's law)

Re-formulation to testing against $\rho^2 = 0$

Let $n_k = |S_k|$ and $\bar{N}_k = \frac{1}{n_k} \sum_{i \in S_k} N_i$. Define

$$\rho^2 := \frac{1}{n\bar{N}} \sum_{k=1}^{K} n_k \bar{N}_k \|\mu_k - \mu\|^2, \quad \text{with } \mu = \frac{1}{n\bar{N}} \sum_{i=1}^{n} N_i \Omega_i$$

Our plan:

- An unbiased estimator of ρ^2
- A test statistic with asymptotic distribution of $N(0, 1)$ under H_0 (where $\rho^2 = 0$)

An unbiased estimator of ρ^2

Let $\hat{\mu}=\frac{1}{n\bar{N}}\sum_{i=1}^{n} X_i$ and $\hat{\mu}_k=\frac{1}{n_k\bar{N}_k}\sum_{i\in S_k} X_i$. A raw estimator is

$$T_0 = \frac{1}{n\bar{N}}\sum_{k=1}^{K} n_k\bar{N}_k\|\hat{\mu}_k - \hat{\mu}\|^2$$

Lemma: Suppose $X_i \sim \mathrm{Multinomial}(N_i, \Omega_i)$ and $X_1, X_2, \ldots, X_n$ are independent.

- $\mathbb{E}[T_0]=\rho^2+\sum_{k=1}^{K}\left(\frac{1}{n_k\bar{N}_k}-\frac{1}{n\bar{N}}\right)\sum_{i\in S_k}\sum_{j=1}^{p} N_i\Omega_{ij}(1-\Omega_{ij})$.
- $\Omega_{ij}(1-\Omega_{ij}) = \mathbb{E}\left[\frac{1}{N_i(N_i-1)}X_{ij}(N_i-X_{ij})\right]$.

An unbiased estimator of ρ^2

DEbiased and Length-adjusted Variability Estimator

$$T = T_0 - \sum_{k=1}^{K} \left(\frac{1}{n_k \bar{N}_k} - \frac{1}{n\bar{N}} \right) \sum_{i \in S_k} \sum_{j=1}^{p} \frac{X_{ij}(N_i - X_{ij})}{N_i - 1}$$

Theorem 1 (unbiasedness): $\mathbb{E}[T] = \rho^2$.

Regularity conditions

Condition 1:

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

Condition 2:

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

where $\alpha_n := \max\{\sum_{k=1}^K \frac{\|\mu_k\|_3^3}{n_k\bar{N}_k}, \sum_{k=1}^K \frac{\|\mu_k\|^2}{n_k^2\bar{N}_k^2}\} / \big(\sum_{k=1}^K \|\mu_k\|^2\big)^2$,

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

Using DELVE to test against H_0

Theorem 2 (parameter-free limiting null): Let

$$V = 2\sum_{k=1}^{K}\sum_{i\in S_k}\sum_{j=1}^{p}\Big(\frac{1}{n_k\bar{N}_k} - \frac{1}{n\bar{N}}\Big)^2\Big[\frac{N_i X_{ij}^2}{N_i - 1} - \frac{N_i X_{ij}(N_i - X_{ij})}{(N_i-1)^2}\Big]$$

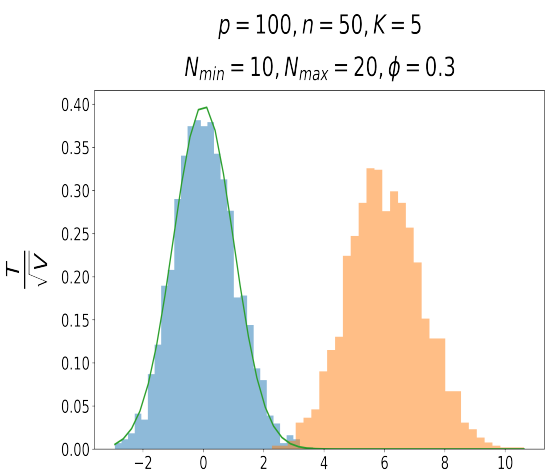
$$+ \frac{4}{n^2\bar{N}^2}\sum_{k\neq\ell}\sum_{i\in S_k}\sum_{m\in S_\ell}\sum_{j=1}^{p} X_{ij}X_{mj} + 2\sum_{k=1}^{K}\sum_{\substack{(i,m)\in S_k^2\\ i\neq m}}\sum_{j=1}^{p}\Big(\frac{1}{n_k\bar{N}_k} - \frac{1}{n\bar{N}}\Big)^2 X_{ij}X_{mj}.$$

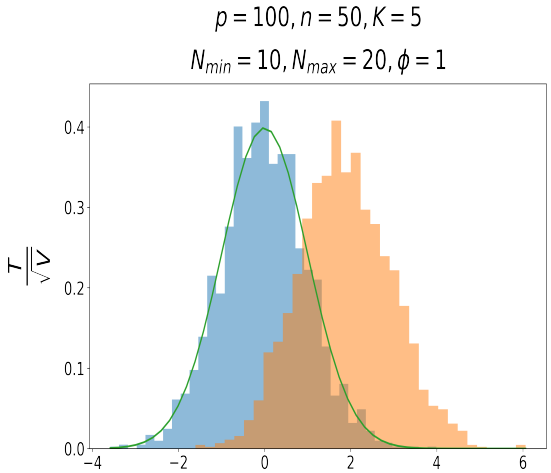
Suppose H_0 is satisfied. As $n\bar{N}\to\infty$ and $p\to\infty$,

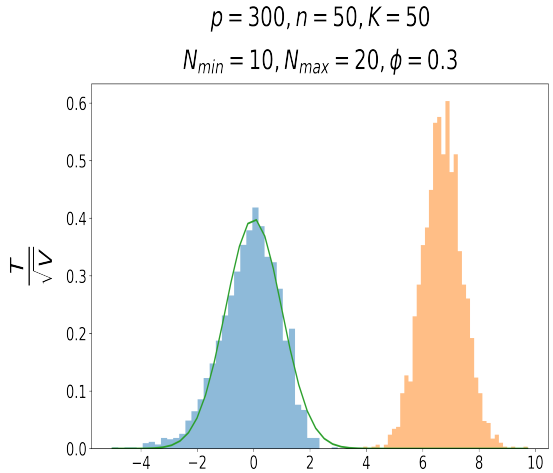
$$T/\sqrt{V} \quad \to_d \quad N(0,1)$$

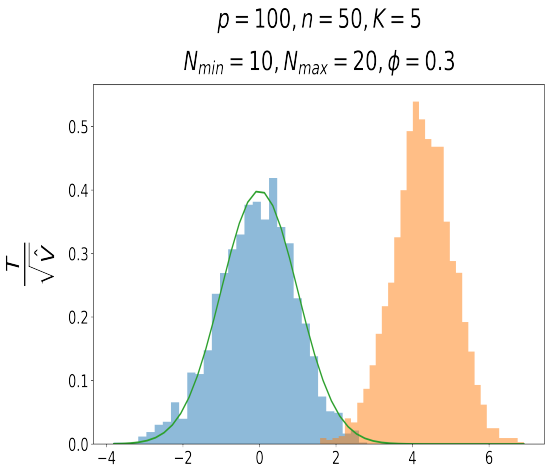
Simulations

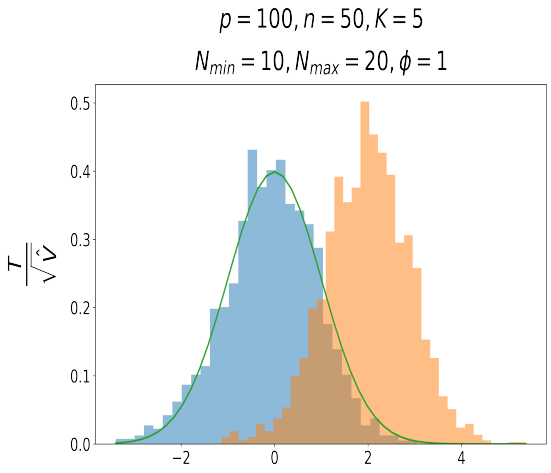
Ω_i are i.i.d. drawn from Dirichlet($\phi \mathbf{1}_p$) within each group.

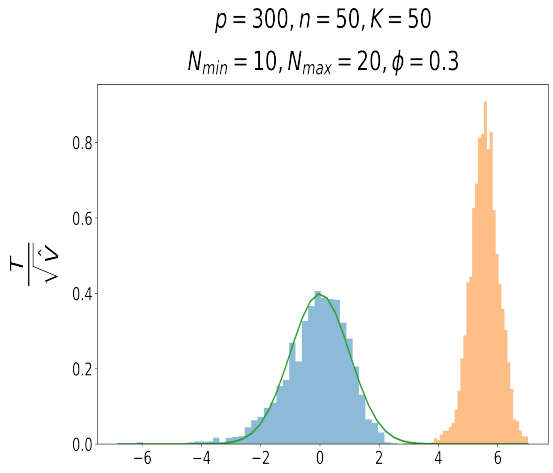












Minimax optimality of DELVE

Theorem 3: Define $\omega_n^2 = \rho^2 / \|\mu\|^2$ and

$$\mathrm{SNR}_n \equiv \frac{n\bar{N}\|\mu\|^2\omega_n^2}{\sqrt{\sum_{k=1}^K \|\mu_k\|^2}} \quad \left(\asymp \frac{n\bar{N}\omega_n^2}{\sqrt{Kp}}, \text{ if } \|\mu_k\| \asymp \frac{1}{\sqrt{p}} \right)$$

- *(Power)* If $\mathrm{SNR}_n \to \infty$, then $T/\sqrt{V} \to_{\mathbb{P}} \infty$; and the level-$\alpha$ DELVE test has an asymptotic level of α and an asymptotic power of 1.
- *(Lower bound)* If $\mathrm{SNR}_n \to 0$, for any test, the sum of type I and type II errors converges to 1.

Comparison with other testing ideas

- The likelihood ratio (LR) test is only applicable in a special case:

$$\Omega_i = \mu_k, \quad \text{for } i \in S_k \quad \Longrightarrow \quad H_0 : \quad \Omega_i \equiv \mu$$

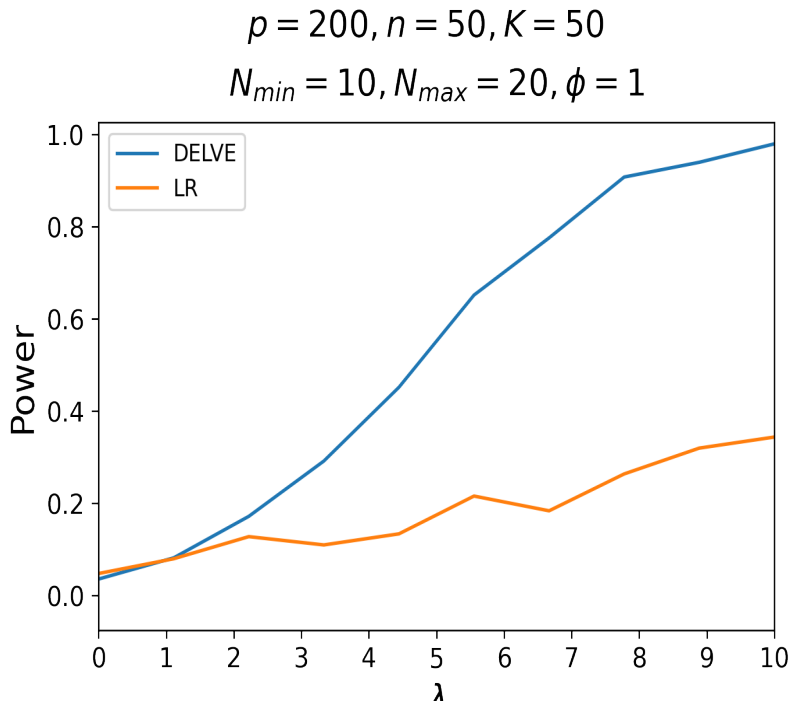
- The naive test based on T_0 is non-optimal:

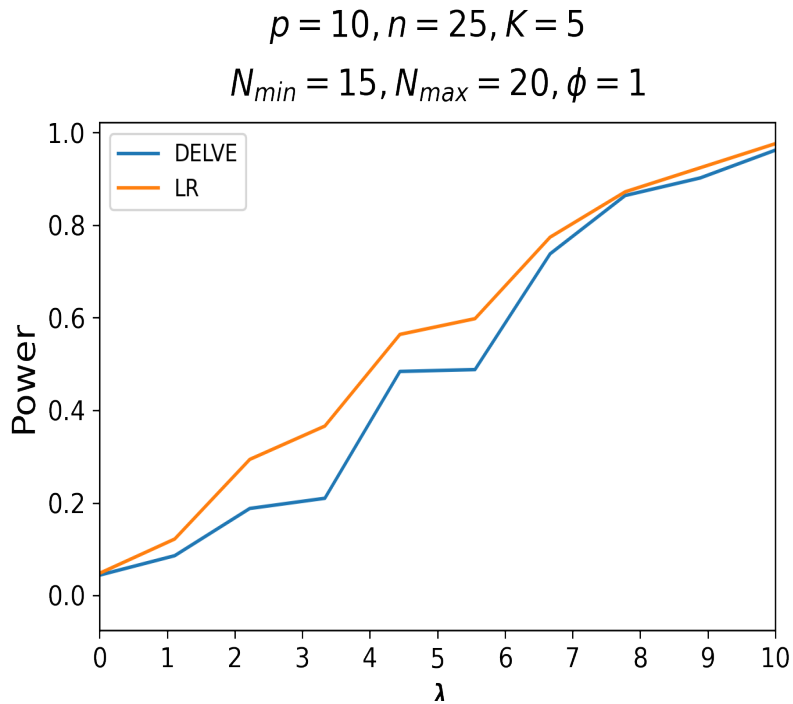
$$T_0 = T + \text{“bias”}$$

There exists a parameter regime in which the “bias” term dominates the “signal” in T

Power comparison with the LR test

The LR test is only defined in the special case where $\Omega_i = \mu_k$ within each group. Under H_0, $\mu \sim \text{Dirichlet}(p/2, \phi \mathbf{1}_{p/2})$. Under H_1, perturb half entries in μ by $\pm\tau_n$ to obtain $\mu_1, \ldots, \mu_K$. Write $\lambda = n\bar{N}\|\mu\|\tau_n^2/\sqrt{K}$. For DELVE, the cut-off is the 0.95 quantile of $N(0,1)$. For LR, we use the optimal cut-off value from simulating H_0 for 500 times.





Applications in statistical inference

We show that DELVE can be used for:

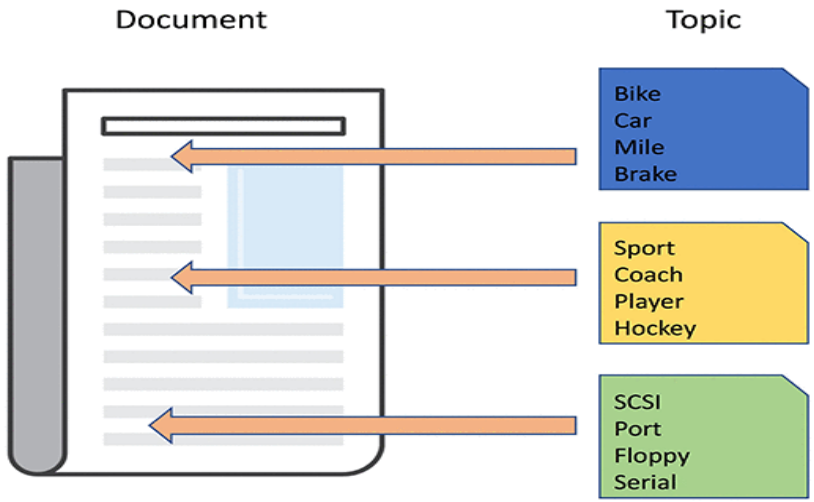
- Global testing for a topic model ($K = n$)

  Hofmann (1999), Blei, Ng, and Jordan (2003)
- Authorship attribution challenge ($K = 2$)

  Mosteller & Wallace (1963), Kipnis (2021), Donoho & Kipnis (2021)
- Closeness testing between two discrete distributions ($K = 2$)

  Chan, Diakonikolas, Valiant, and Valiant (2014), Bhattacharya and Valiant (2015), Balakrishnan and Wasserman (2019)

Example 1: global detection of topics



The topic model *(Hoffman, 1999; Blei et al., 2003)*:

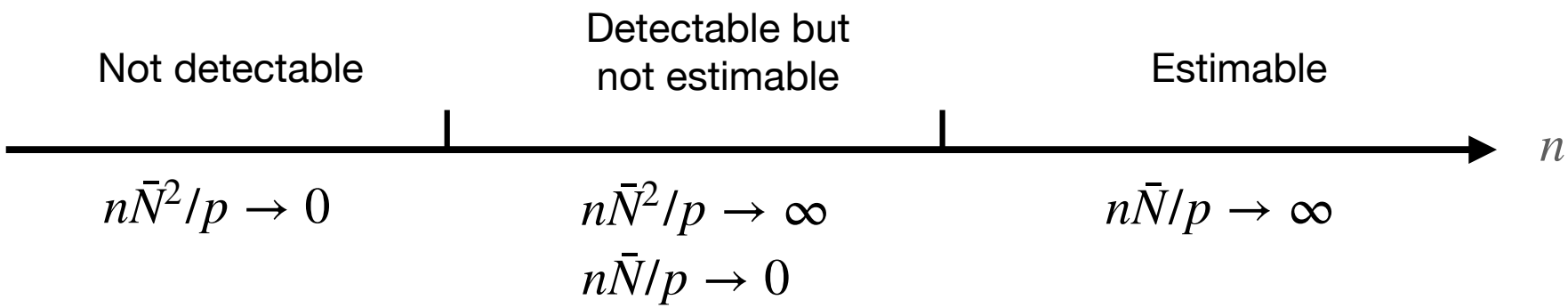
$$\Omega_i = \sum_{k=1}^{r} w_i(k) A_k \iff \Omega = \underbrace{[A_1, \ldots, A_r]}_{p \times r} \underbrace{[w_1, \ldots, w_n]}_{r \times n}$$

Global testing: $r = 1$ v.s. $r > 1$

A special case of our problem with $K = n$:

$$\Omega_i \begin{cases} \text{are all equal to } A_1, & \text{when } r = 1 \\ \text{range in the convex hull of } A_1, \ldots, A_r, & \text{when } r > 1 \end{cases}$$

- DELVE has a full power if $n\bar{N}^2/p \to \infty$. Furthermore, a matching lower bound is proved.
- Previous works (e.g., Ke and Wang; Bing et al.) showed that A is estimable if $n\bar{N}/p \to \infty$



Example 2: Authorship attribution

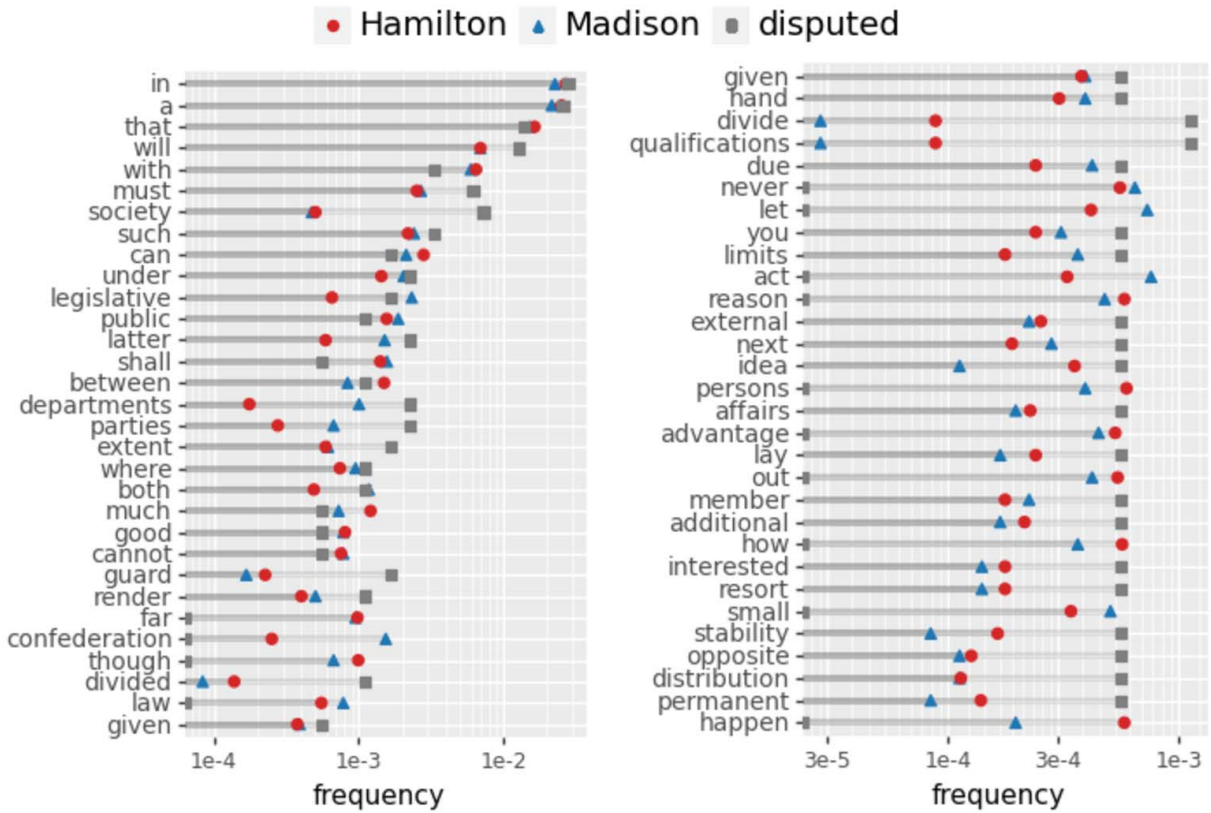


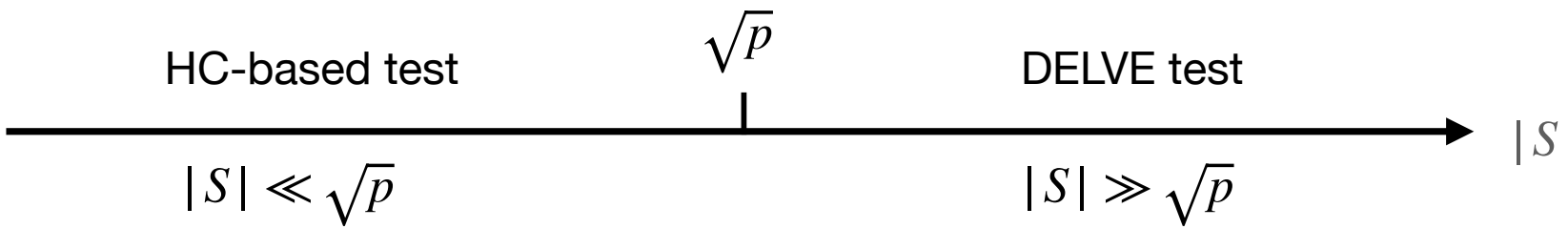
FIG. 1. *Word frequencies of three authorship sources. The two panels show a random sample of* 60 *words out of the* 500 *most common ones in one of the disputed articles* (*gray*), *the corpus of known Hamilton articles* (*blue*), *and the corpus of known Madison articles* (*red*) *out of the first* 77 *Federalist Papers. We attribute the disputed article by measuring the global discrepancy between its word frequencies to each corpus of known authorship.*

Mosteller & Wallace (1963), Kipnis (2022)

$$X_i \sim \text{Multinomial}(N_i, \Omega_i), \quad \eta = (n\bar{N})^{-1} \sum_{i=1}^{n} N_i \Omega_i$$
$$G_i \sim \text{Multinomial}(M_i, \Gamma_i), \quad \theta = (m\bar{M})^{-1} \sum_{i=1}^{m} M_i \Gamma_i$$

- H_0: $\eta = \theta$
- H_1: rare/weak signals *(Donoho and Kipnis, 2021)*:

$$\left|\sqrt{\eta_j} - \sqrt{\theta_j}\right| \begin{cases} = 0, & j \notin S, \\ \geq \beta_n, & j \in S, \end{cases} \qquad \text{with } |S| \ll p$$



Applications in real corpora

- Amazon movie reviews
- MADStat: abstracts of 83K statistical papers from 36 journals in 1975-2015

Ji and Jin (2016) The co-authorship and citation networks of statisticians (with discussions). Annals of Applied Statistics.

Ji, Jin, Ke & Li (2022) Co-citation and co-authorship networks of statisticians (with discussions). Journal of Business & Economic Statistics.

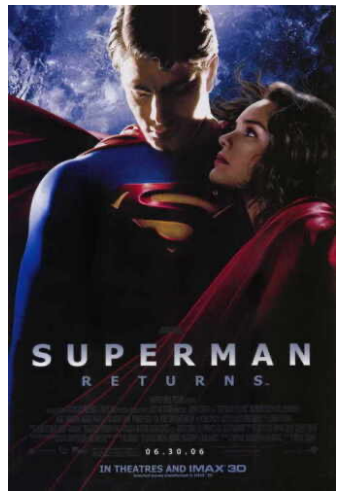
Ke, Ji, Jin & Li (2023+) Recent advances in text analysis. Annual Review of Statistics and its Applications (in press).

https://github.com/ZhengTracyKe/MADStat

Example 1: Amazon movie reviews

Maurya (2018)

(581 reviews)



(731 reviews)

(204 reviews)

For each movie, we apply DELVE in two ways:

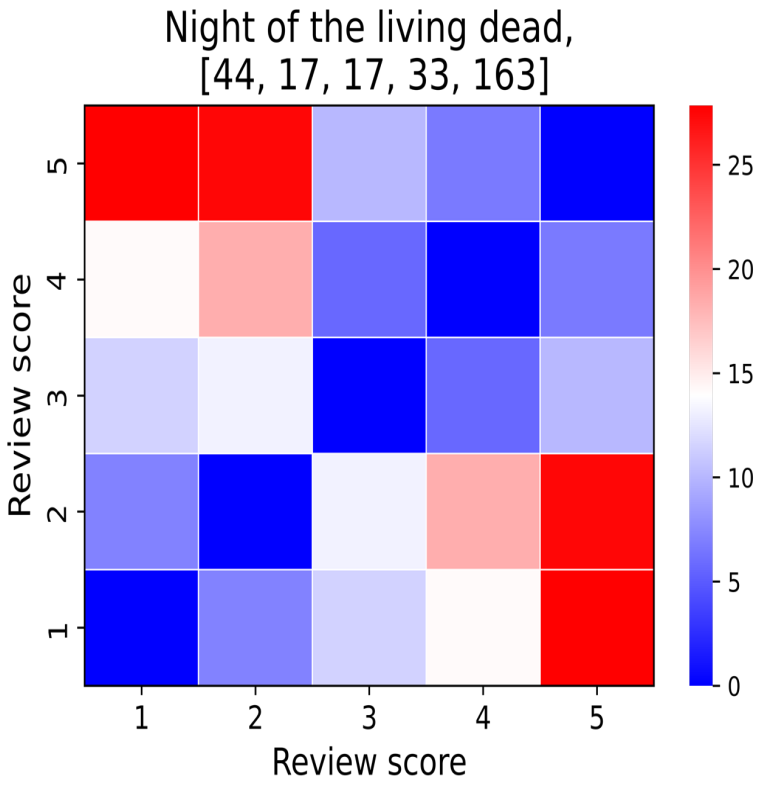
- Diversity of reviews: each review is a group ($K = n$).
- Difference between star ratings: all reviews with the same star rating are treated as a group ($K = 2$).

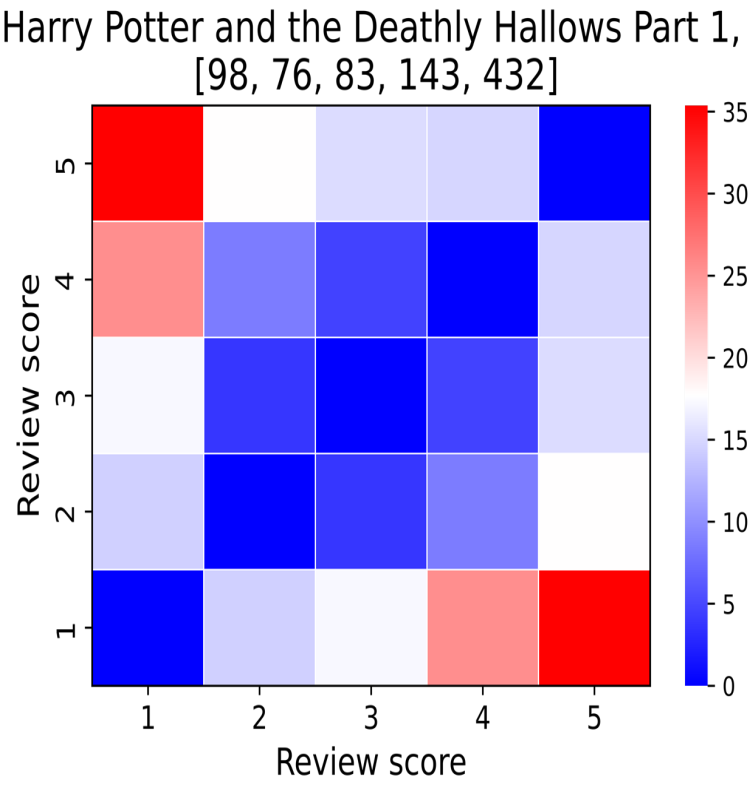
Diversity of reviews

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
5	No Country for Old Men	30.57	819
⋮	⋮	⋮	⋮
16	John Adams	20.78	857
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

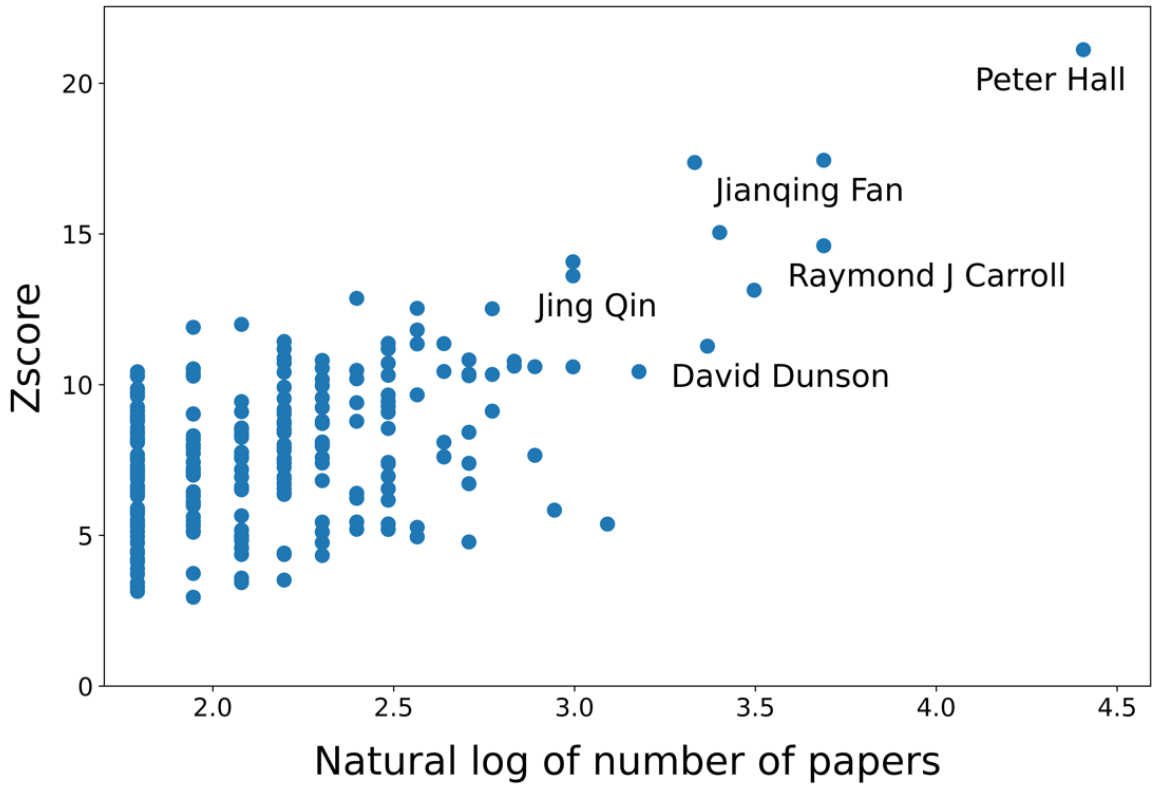
For the 500 most-reviewed movies, the mean of DELVE z-scores is 19.97, and the standard deviation of 5.07.

Difference between star ratings





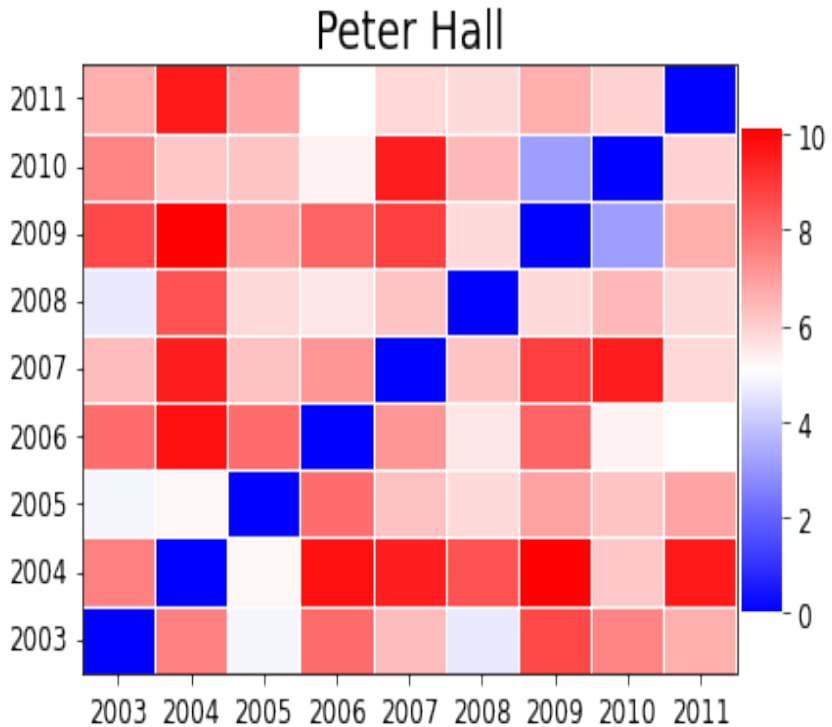
Example 2: MADStat (Phase I by Ji and Jin (2016))



For each author, we apply DELVE with $K = n$. The Z-score measures the semantic diversity of an author's abstracts. For all authors in the dataset, the mean of DELVE Z-scores is 4.52 and the standard deviation is 2.94.

Difference between years

In the cell (x, y), we compare Peter Hall's abstracts from time x with his abstracts from time y. The heatmap shows the value of the DELVE Z-score with $K = 2$ for each cell.



Year	Title	Journal
2004	Low order approximations in deconvolution and regression with errors in variables	JRSSB
2004	Nonparametric inference about service time distribution from indirect measurements	JRSSB
2004	Cross-validation and the estimation of conditional probability densities	JASA
2004	Nonparametric confidence intervals for receiver operating characteristic curves	Biometrka
2004	Bump hunting with non-Gaussian kernels	AOS
2004	Attributing a probability to the shape of a probability density	AOS

Summary

- Between-group variability in multinomials
  - H_0: $\mu_1 = \mu_2 = \ldots = \mu_K$ (K ranges from 2 to n)
- The DELVE test
  - parameter-free limiting null
  - optimal detection boundary
- This simple idea is widely useful
  - Global detection of topics
  - Authorship attribution
  - Closeness testing of two discrete distributions
  - Patterns in online customer reviews
  - Evolution of text abstracts of an author

Cai, T. T., Ke, Z. T., & Turner, P. (2023). Testing High-dimensional Multinomials with Applications to Text Analysis. *JRSS-B (to appear).*