A Parametric Bootstrap Test for Topic Models

Skyler Seto*, Sarah Tan*, Giles Hooker, Martin T. Wells

*Authors contributed equally

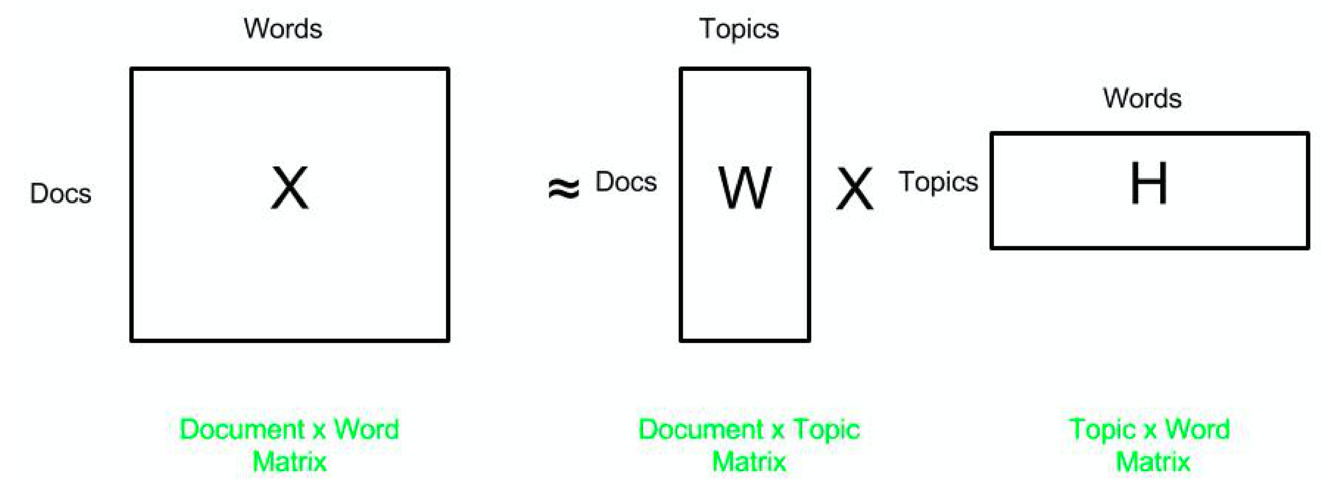
Cornell University

Introduction

Non-negative matrix factorization (NMF) is a technique for **decomposing a matrix X with non-negative entries into a low-rank approximation** $\hat{X} = WH$ where both W and H have a low rank $\leq k$ and contain no negative entries. The method has been applied to corpora to construct topic models. However, NMF has likelihood assumptions which are often violated by real document corpora. **We present a double parametric bootstrap test for evaluating the fit of an NMF-based topic model based on the duality of the KL divergence and Poisson maximum likelihood estimation.**

Topic Models via Nonnegative Matrix Factorization

In NMF, a matrix $X \in \mathbb{R}^{V \times M}$ of all non-negative entries is decomposed into two non-negative factors W and H that have latent dimensionality k, such that $X \approx WH = \hat{X}$



The optimal W and H matrices are found by minimizing $D(X \| \hat{X})$, i.e., the distance between X and its low-rank approximation.

$$D(X\|\hat{X}) = -\left(\sum_{i,j} x_{ij} \log\left(\frac{\hat{x}_{ij}}{x_{ij}}\right) - \hat{x}_{ij} + x_{ij}\right)$$

Duality Between Divergence and Maximum Likelihood

Generalized KL Divergence

$$D(X\|\hat{X}) = -\left(\sum_{i,j} x_{ij} \log\left(\frac{\hat{x}_{ij}}{x_{ij}}\right) - \hat{x}_{ij} + x_{ij}\right)$$

Maximum likelihood for $X_{ij} \sim$ Pois(λ_{ij})

$$\log P(X|\Lambda) = \sum_{i,j} x_{ij} \log(\lambda_{ij}) - \lambda_{ij} - \log(\Gamma(x_{ij}+1))$$

Double Parametric Bootstrap Hypothesis Test

Hypotheses:

$H_0 : X_{ij}$ is distributed as Pois(λ_{ij}) $H_A : X_{ij}$ is not distributed as Pois(λ_{ij})

Algorithm:

Data: X
Result: p-value ρ
Compute $\hat{X}$ for the observed X and let $\ell = \frac{D(X\|\hat{X})}{\sqrt{V \cdot D}}$;
Sample B_1 bootstrap samples $X_1^*, X_2^*, \ldots, X_{B_1}^* \sim$ Pois$(\hat{X})$;
for $i = 1 : B_1$ **do**
    Compute $\hat{X}_i^*$ and $\ell_i^* = \frac{D(X_i^*\|\hat{X}_i^*)}{\sqrt{V \cdot D}}$;
end
Compute $\rho^*(\ell) = 2\min\left\{\frac{1}{B_1}\sum_{i=1}^{B_1} \mathbb{1}[\ell_i^* \leq \ell], \frac{1}{B_1}\sum_{i=1}^{B_1} \mathbb{1}[\ell_i^* > \ell]\right\}$;
for $i = 1 : B_1$ **do**
    Sample B_2 bootstrap samples $X_{i1}^{**}, \ldots X_{iB_2}^{**} \sim$ Pois$(\hat{X}_i^*)$;
    for $j = 1 : B_2$ **do**
        Compute $\hat{X}_{ij}^{**}$ and $\ell_{ij}^{**} = \frac{D(X_{ij}^{**}\|\hat{X}_{ij}^{**})}{\sqrt{V \cdot D}}$;
    end
    Compute $\rho_i^{**}(\ell_i^*) = 2\min\left\{\frac{1}{B_2}\sum_{j=1}^{B_2} \mathbb{1}[\ell_{ij}^{**} \leq \ell_i^*], \frac{1}{B_2}\sum_{j=1}^{B_2} \mathbb{1}[\ell_{ij}^{**} > \ell_i^*]\right\}$;
end
return $\rho = 2\min\left\{\frac{1}{B_1}\sum_{i=1}^{B_1} \mathbb{1}[\rho^* \leq \rho_i^{**}], \frac{1}{B_1}\sum_{i=1}^{B_1} \mathbb{1}[\rho^* > \rho_i^{**}]\right\}$.

Algorithm 1: Double Parametric Bootstrap for Topic Models

Why Double Bootstrap?

Goal: construct sampling distribution for X_{ij} to test if it is Poisson

But... there is no natural pivotal test statistic to test this hypothesis, no exact test, need an approximate test - bootstrap

Aside: A pivot is a function of the data and unknown parameters whose distribution does not depend on the unknown parameters. E.g. z-score ($z = \frac{x-\mu}{\sigma}$) has distribution $N(0,1)$ which does not depend on the parameters

But... true p-value depends on unknown underlying sampling distribution; bootstrap p-value depends on bootstrap distribution. These 2 distributions differ when test statistic not pivotal (i.e. depends on the parameters) and parameters used in bootstrap data generating process different from true parameters

Solution: if test statistic is *asymptotically* pivotal, double bootstrap distribution will converge to true sampling distribution as sample size increases

More Advantages: double bootstrap p-value converges to p-value at rate faster than asymptotic p-value

Disadvantages: computationally very costly. For each of B_1 bootstrap samples, need to compute $B_2 + 1$ test statistics. Total # test statistics: $1 + B_1 + B_1 B_2$.

Simulation Studies

Simulation Procedure:

Poisson

- Simulate three differently sized corpora with $W = 16, 23, 23$ words, $M = 10, 23, 92$ documents, and $K = 5, 10, 10$ topics.
- Generate $W \sim$ Gamma$(10, 0.1)$ and $H \sim$ Gamma$(1, 100)$ and compute $\hat{X} = WH$.
- Fix $B_1, B_2 = 25$.
- Perform Kolmogorov-Smirnov (KS) test for uniformity.

Other Distributions

- Sample 25 $X \in \mathbb{R}^{10 \times 16}$ with $k = 5$ from a Zero Inflated Poisson with $p = 0.5$, Gamma, and Normal distribution with negative values replaced by zero.
- Report the average distance ℓ and KS test p-value

Probability-Prbability Plots for Poisson Data

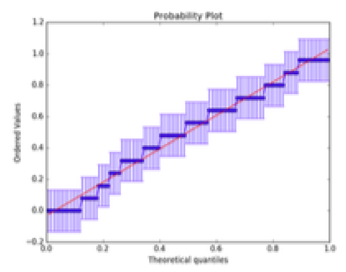


Figure 1: P-P plot for $W = 10, M = 16, K = 5, T = 0.133$

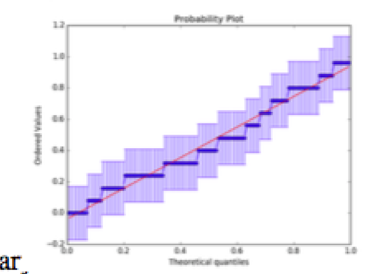


Figure 2: P-P plot for $W = 23, M = 23, K = 10, T = 0.17$

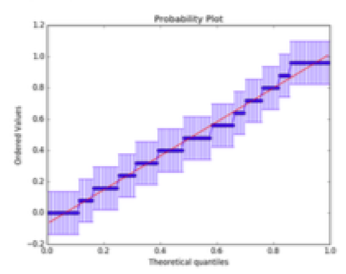


Figure 3: P-P plot for $W = 23, M = 92, K = 10, T = 0.137$

KS test p-values and distances

Size	p-value ρ
10×16	0.3685
23×23	0.1902
92×23	0.3687

Distribution	p-value ρ	ℓ
Poisson	0.4416	2.0277
Gamma	0.0	859.9227
Normal	0.0	442.5251
Zero Inflated Poisson	0.0	471.7884

Detecting Group Structure Across Documents

To detect variation in word usage within topics by structure (e.g. time, author), perform bootstrap test on entire X matrix vs. X matrix broken down by structure

Matrix	Scope	p-value ρ	Rejects
X		0.0	10
X_1	2004	0.32	2
X_2	2005	0.02	8
X_3	2006	0.07	8
X_4	2007	0.76	0

Matrix	Scope	p-value ρ	Rejects
X		0.0	10
X_1	Foreign	0.49	1
X_2	Business	0.06	8
X_3	Arts and Culture	0.0	10
X_4	National	0.73	0

Left: testing for temporal structure in NYT **Foreign Desk** articles. **Right:** testing for desk structure in **Jan 05** articles

Matrix	Scope	p-value ρ	Rejects
X		0.0	10
X_1	2004	0.15	6
X_2	2005	0.0	10
X_3	2006	0.09	7
X_4	2007	0.04	9

Matrix	Scope	p-value ρ	Rejects
X		0.04	9
X_1	Foreign	0.648	0
X_2	Business	0.728	1
X_3	Arts	0.624	0
X_4	National	0.0	10

Left: testing for temporal structure in articles from **all desks**. **Right:** testing for desk structure in articles from **all years**

Conclusion

- NMF likelihood assumptions often violated by real data
- Alternative divergence metrics could fit your data better
- Double parametric bootstrap test to check if the fit is good
- Checking assumptions helpful for interpretability