# Test Selection

**Bryan Kelly** **Asaf Manela** **Alan Moreira**

Presented by: Rui Qiu

September 2019

# Overview

- Purpose of text selection
- Model’s detail
- Model’s application
- Related work

# Prework of THIS PAPER

- **Multinomial inverse regression for text analysis – Matt Taddy**
- **Distributed multinomial regression – Matt Taddy**

# Background Knowledge

- **Regression**
  1. Linear Regression
  2. Logistic Regression (Multi-class Logistic Regression)
  3. Poisson Regression
  4. Inverse Regression

# Poisson Regression

- **Poisson distribution**

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

- **Poisson regression**
  - Used to model counts and rates like expected number of events over a given period of time
  - Why? - Distribution of data

- **Poisson Zero-Inflated Model**

$$p(y = 0 \mid x; \theta) = e^{-e^{\theta' x}}$$

# Inverse Regression

- **Word/Information embedding**
  - **First train y = k1x1 + k2x2 = z1 + z2**
- **Forward regression**
  - **Then use z1, z2 to model subpart of x**

# Multinomial distribution

- **A generalization of the binomial distribution.**
  - E.g. it models the probability of counts of each side for rolling a k-sided die n times.
- **MN PDF**

$$f(x_1, \ldots, x_k; p_1, \ldots, p_k) = \frac{\Gamma(\sum_i x_i + 1)}{\prod_i \Gamma(x_i + 1)} \prod_{i=1}^{k} p_i^{x_i}.$$

- **Key feature**

$$MN\left(\boldsymbol{c}_i; \boldsymbol{q}_i, m_i\right) = \frac{\prod_j Po\left(c_{ij}; e^{\eta_{ij}}\right)}{Po\left(m_i; \sum_{j=1}^{d} e^{\eta_{ij}}\right)} \approx \prod_j Po\left(c_{ij}; m_i e^{\eta_{ij}}\right)$$

# Purpose of Text Selection

- **Increasingly available with difficulties**
  - TB amount of data
  - Ultra-high dimension
  - Sparse in Matrix
- **Upgrade from DMR to HDRM**
  - First step model extensive
  - Second step model intensive

$\{X\} =$

| | c1 | c2 | c3 | c4 | c5 | m1 | m2 | m3 | m4 |
|---|---|---|---|---|---|---|---|---|---|
| human | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| interface | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| computer | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| user | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| system | 0 | 1 | 1 | 2 | 0 | 0 | 0 | 0 | 0 |
| response | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| time | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| EPS | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| survey | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| trees | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 |
| graph | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| minors | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |

r (human.user) = -.38

r (human.minors) = -.29

# Current Model - DMR

- **Multinomial logistic regression**

$$\mathrm{p}\left(\boldsymbol{c}_i|\boldsymbol{v}_i, m_i\right) = MN\left(\boldsymbol{c}_i; \boldsymbol{q}_i, m_i\right) \text{ for } i = 1 \ldots n,$$

$$q_{ij} = \frac{e^{\eta_{ij}}}{\sum_{k=1}^{d} e^{\eta_{ik}}} \text{ for } j = 1 \ldots d,$$

$$\eta_{ij} = \alpha_j + \boldsymbol{v}_i' \boldsymbol{\varphi}_j.$$

- **MN decomposition & the theory foundation of DMR**

$$\mathrm{p}\left(\boldsymbol{c}_i|\boldsymbol{v}_i, m_i\right) = MN\left(\boldsymbol{c}_i; \boldsymbol{q}_i, m_i\right) \approx \prod_j Po\left(c_{ij}; m_i e^{\eta_{ij}}\right).$$

# Current Model - DMR

- **Log-likelihood**

$$l\left(\alpha_j, \boldsymbol{\varphi}_j | \boldsymbol{c}_j, \boldsymbol{v}\right) = \sum_{i=1}^{n} \left[ m_i e^{\alpha_j + \boldsymbol{v}_i' \boldsymbol{\varphi}_j} - c_{ij} \left( \alpha_j + \boldsymbol{v}_i' \boldsymbol{\varphi}_j \right) \right].$$

- **Pre-process of Inverse/ forward regression**

$$\boldsymbol{z}_i = \hat{\boldsymbol{\varphi}}' \boldsymbol{c}_i$$

$$\mathbb{E}\left[v_{iy}\right] = \beta_0 + \left[z_{iy}, \boldsymbol{v}_{i,-y}, m_i\right]' \boldsymbol{\beta}$$

# HDMR Model – two part hurdle model $c_{ij}$

- **General format**
  - **First: model extensiveness(cover or not)**

$$h_{ij}^{*} = \kappa_j + \boldsymbol{w}_i'\boldsymbol{\delta}_j + \upsilon_{ij},$$

$$h_{ij} = \mathbf{1}\left(h_{ij}^{*} > 0\right),$$

  - **Second: model intensiveness (# of times cover)**

$$c_{ij}^{*} = \lambda\left(\alpha_j + \boldsymbol{v}_i'\boldsymbol{\varphi}_j\right) + \varepsilon_{ij} > 0,$$

$$c_{ij} = h_{ij}c_{ij}^{*}.$$

# HDMR Model – two part hurdle model $c_{ij}$

- **General format**

$$
\begin{aligned}
h_{ij}^{*} &= \kappa_j + \boldsymbol{w}_i'\boldsymbol{\delta}_j + \upsilon_{ij}, \\
h_{ij} &= \mathbf{1}\left(h_{ij}^{*} > 0\right), \\
c_{ij}^{*} &= \lambda\left(\alpha_j + \boldsymbol{v}_i'\boldsymbol{\varphi}_j\right) + \varepsilon_{ij} > 0, \\
c_{ij} &= h_{ij}c_{ij}^{*}.
\end{aligned}
$$

$$
\mathrm{p}\left(h_{ij} = 0 | \boldsymbol{w}_i\right) = \Pi_0\left(\kappa_j + \boldsymbol{w}_i'\boldsymbol{\delta}_j\right)
$$

$$
\mathrm{p}\left(c_{ij} | \boldsymbol{v}_i, h_{ij} = 1\right) = P^{+}\left(c_{ij}; \lambda\left(\alpha_j + \boldsymbol{v}_i'\boldsymbol{\varphi}_j\right)\right)
$$

$$
\mathrm{p}\left(c_{ij} | \boldsymbol{v}_i, \boldsymbol{w}_i\right) = \left[\Pi_0\left(\kappa_j + \boldsymbol{w}_i'\boldsymbol{\delta}_j\right)\right]^{1-h_{ij}} \left\{\left[1 - \Pi_0\left(\kappa_j + \boldsymbol{w}_i'\boldsymbol{\delta}_j\right)\right] P^{+}\left(c_{ij}; \lambda\left(\alpha_j + \boldsymbol{v}_i'\boldsymbol{\varphi}_j\right)\right)\right\}^{h_{ij}}
$$

# Application: Backcasting the intermediary capital ratio

- **Data: Front page of WSJ from 1926-2016**

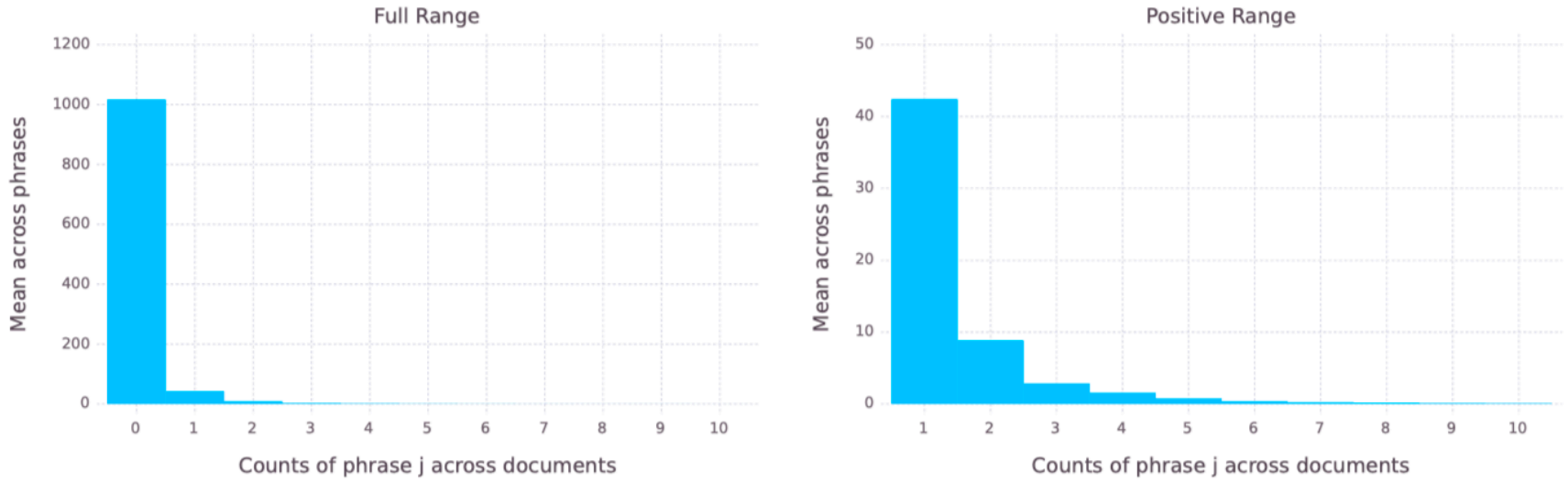
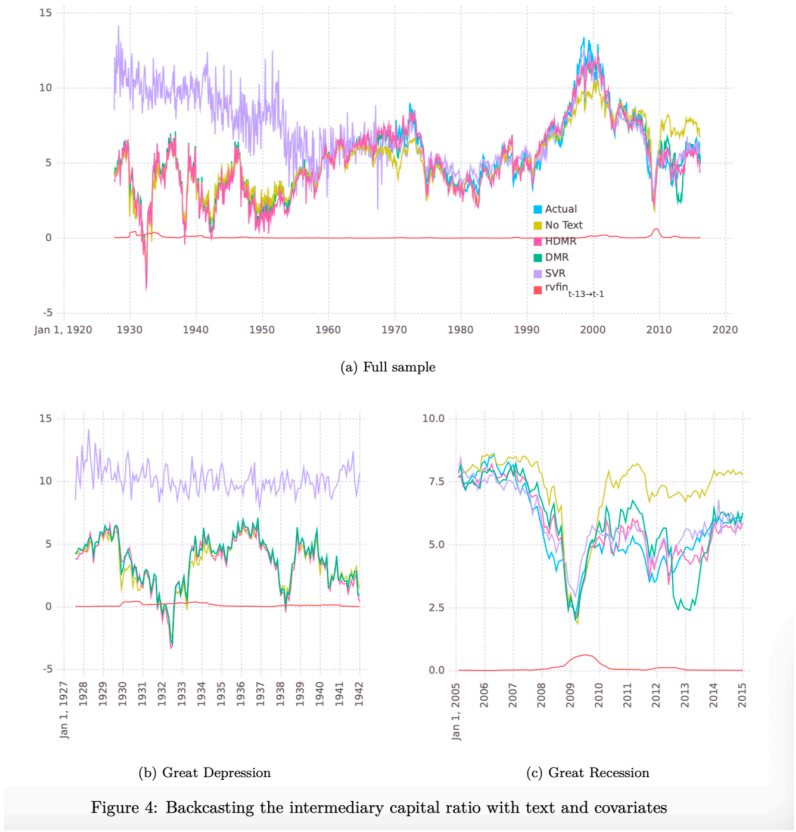



Figure 1: Mean distribution of WSJ front page articles monthly phrase counts

# HDMR Performance in Application 1

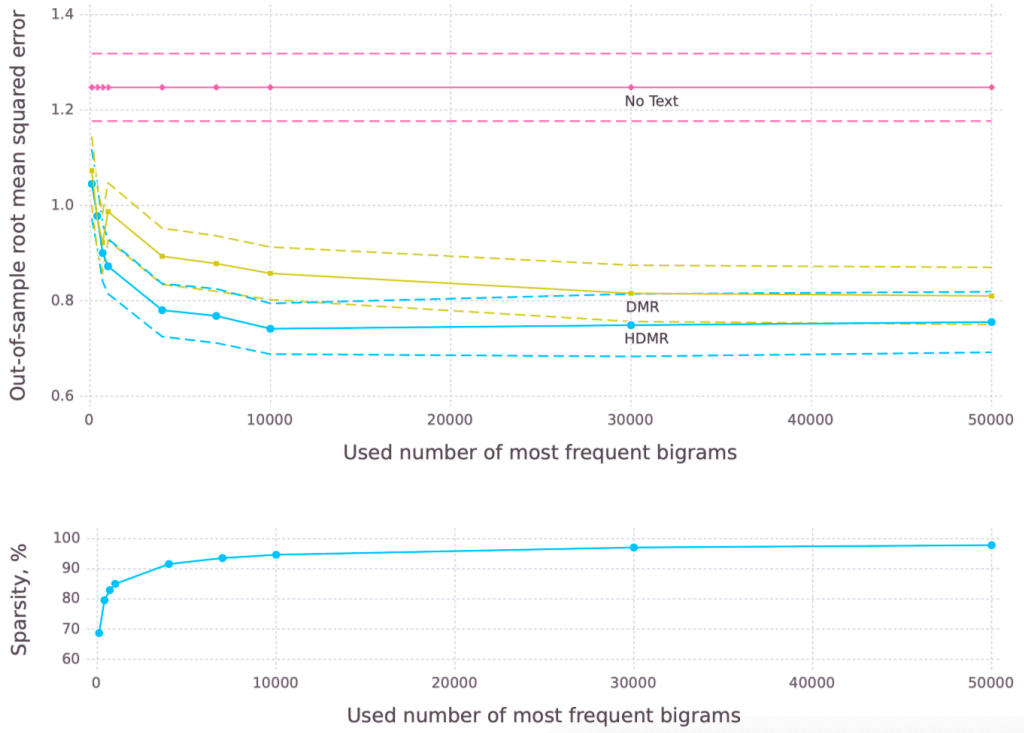



(a) Full sample

(b) Great Depression

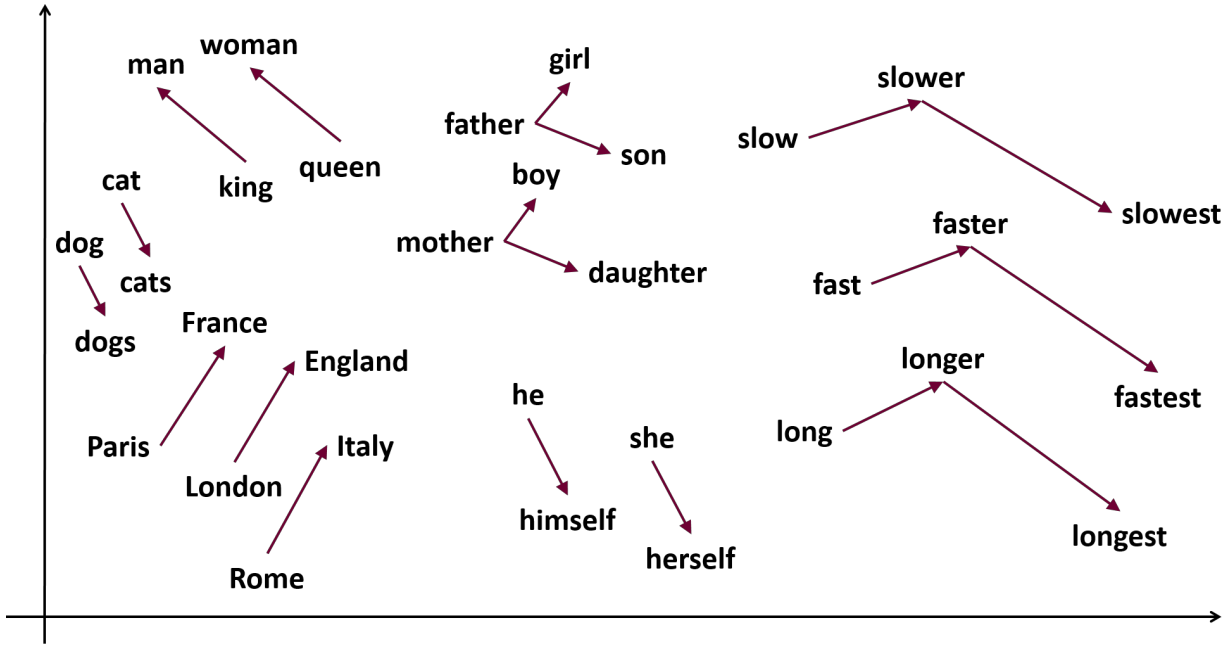
(c) Great Recession

Figure 4: Backcasting the intermediary capital ratio with text and covariates



# Related research in Text mining

- **Word Embedding:**
  - **Word2Vec**
  - **Elmo**



# Reference

- **Text selection Bryan Kelly** - Asaf Manela, Alan Moreira
- **Multinomial inverse regression for text analysis** – Matt Taddy
- **Distributed multinomial regression** – Matt Taddy
- **improving and evaluating topic models and other models of text** - Airoldi, Edoardo M., and Jonathan M. Bischof