

Text Selection

Bryan Kelly

Yale University

Asaf Manela

Washington University in St. Louis

Alan Moreira

University of Rochester

November 2018

Motivation

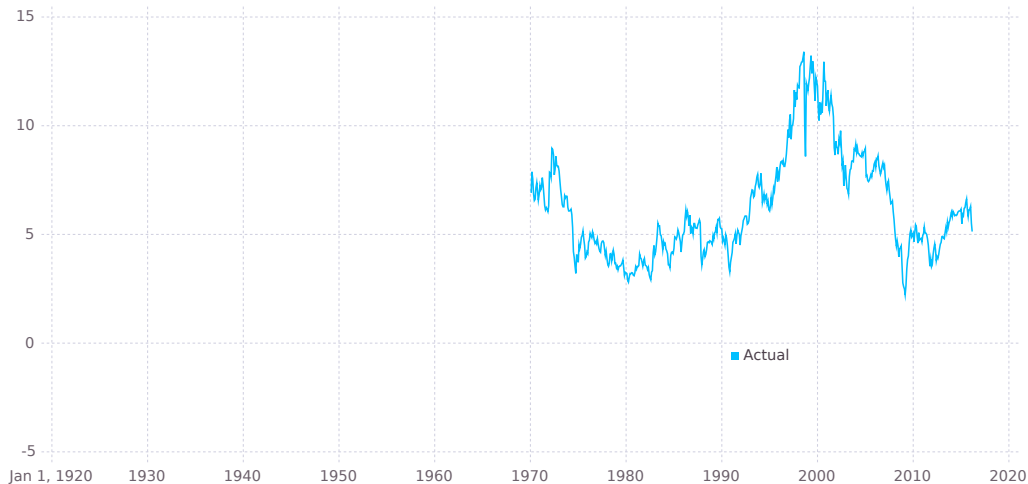
- ▶ Digital text is increasingly available to social scientists
 - ▶ Newspapers, blogs, regulatory filings, congressional records ...
- ▶ Unlike data often used by economists
 - ▶ Text is ultra high-dimensional
 - ▶ Phrase counts are sparse
- ▶ Statistical learning from text requires
 - ▶ Machine learning techniques
 - ▶ Scalable algorithms

This paper

- ▶ Text is often selected by journalists, speechwriters, and others who cater to an audience with limited attention
- ▶ Hurdle Distributed Multiple Regression (HDMR)
 - ▶ Highly scalable approach to inference from big counts data
 - ▶ Includes an economically-motivated [selection equation](#)
 - ▶ Especially useful when inclusion choice is separate or more interesting than repetition
- ▶ Applications using [newspaper coverage for prediction](#)

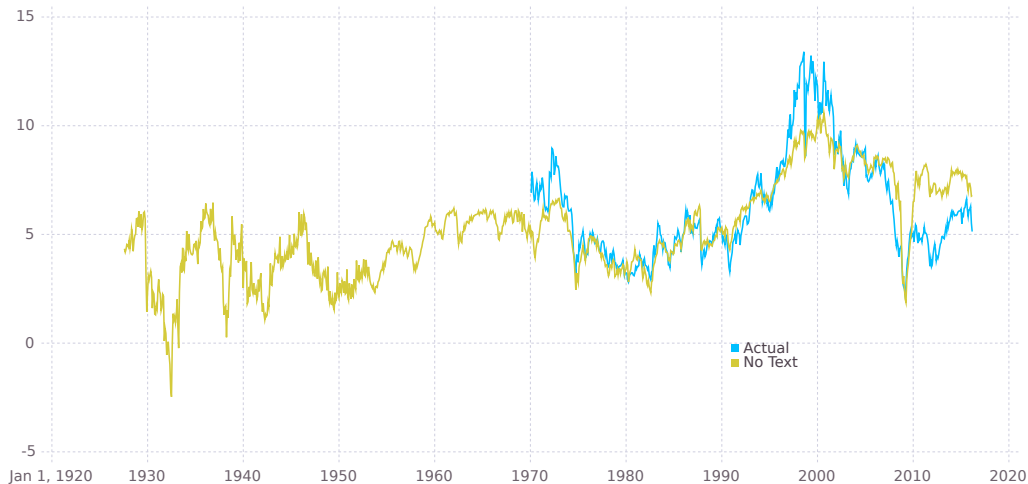
Intermediary capital ratio (He-Kelly-Manela, 2017 JFE)

Available only since 1970 because primary dealers used to be private



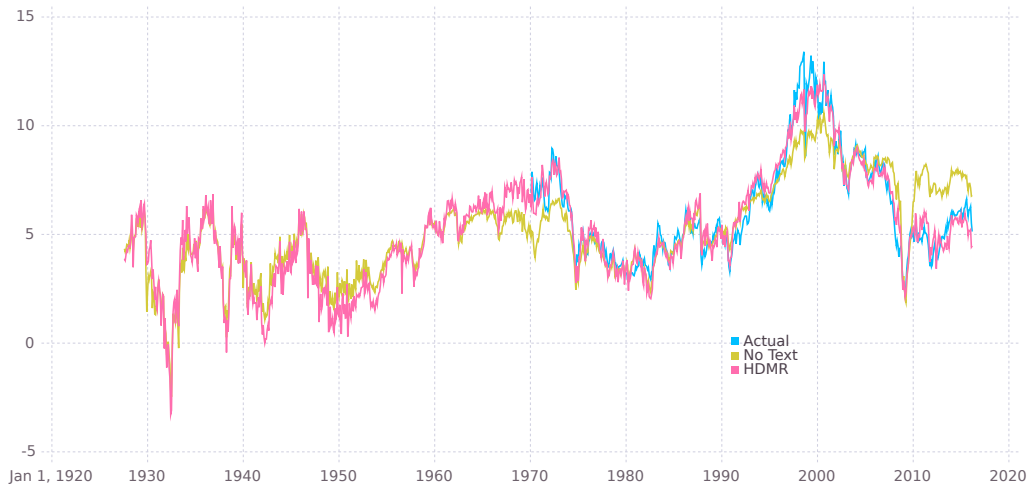
Backcasting the intermediary capital ratio

First stab may be to fit using realized variance and price-dividend ratio without text

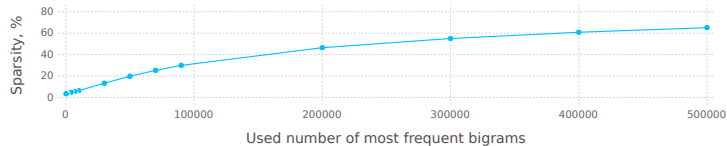
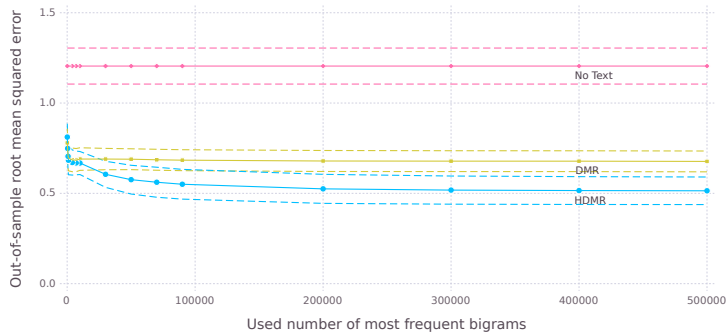


Backcasting the intermediary capital ratio with text

HDMR with historical newspaper text gives a different predicted series exploiting text inclusion and repetition



HDMR's advantage increases with sparsity



Newspaper coverage helps forecast macroeconomic series

- ▶ Stock-Watson (2012 JBES) show that macro forecasts of a simple dynamic factor model (DFM-5) are hard to beat
- ▶ We use their data + *Wall Street Journal* text to forecast 1–12 months ahead
- ▶ Large OOS fit improvement using text with HDMR relative to DFM-5 for macroeconomic fundamentals
 - ▶ Nonfarm payroll employment forecast is 23–44% better
 - ▶ Housing starts forecast is 45–52% better
- ▶ WSJ text helps predict asset prices directly (stocks, treasuries, currencies) in quarterly/annual horizon but not monthly

Conclusion

- ▶ Adding economics to machine learning can improve out-of-sample prediction
- ▶ Hurdle Distributed Multiple Regression (HDMR)
 - ▶ Highly scalable approach to inference from big counts data
 - ▶ Includes an economically-motivated selection equation
 - ▶ Useful where extensive margin is interesting or more important than intensive margin
- ▶ Applications using newspaper coverage for prediction
 1. Backcast intermediary capital ratio
 2. Forecast macroeconomic series