Unstructured Data Analysis for Macroeconomics and Monetary Policy

Stephen Hansen, stephen.hansen@imperial.ac.uk

1 Textbooks / Overview Material

The about of unstructured data in the world continues to grow rapidly, and is increasingly being incorporated into economic analysis. However the very nature of unstructured data makes it difficult to handle using traditional econometric tools since it is typically high dimensional. A related problem is that many unstructured data sources are naturally occurring, which generates potential samples biases. This short course presents various tools for handling these challenges, as well as numerous applications mainly in macroeconomics and monetary policy. An additional goal is to allow attendees to experiment with hands on analysis via a sequence of practical sessions with code demonstrations.

There is no one source that covers all of the material in the course. Grimmer and Stewart (2013), Bholat et al. (2015), and Gentzkow et al. (2019a) are survey articles that provide accessible introductions to text mining. Manning et al. (2008) is an information retrieval textbook that is referenced below as MRS, and Murphy (2012) is a machine learning textbook written from a probabilistic, and in particular Bayesian, perspective referenced below as KM.

In the references below, material in blue refers to core methodological background, material in black refers to applications, and material in green refers to readings outside the scope of the course related to extensions of the core ideas.

2 Theme I: Happenstance Data and Economic Statistics

- The DELVE Initiative (2020)
- Baker and Kueng (2021)
- Carvalho et al. (2021)

3 Theme II: Unstructured Data in Empirical Economics

3.1 Bag-of-Words Model

- MRS 1, 2.2, 6.1-6.3
- Tetlock (2007), Loughran and Mcdonald (2011), Shapiro et al. (2020), Nyman et al. (2021)
- Baker et al. (2016)
- Shapiro and Wilson (2021)
- Deming and Kahn (2018)
- Hassan et al. (2019)

3.2 Word Embeddings

- MRS 18
- Deerwester et al. (1990)
- Mikolov et al. (2013a,b)
- Ash et al. (2020)
- Hansen et al. (2021)
- Rudolph et al. (2016), Ruiz et al. (2020)
- Goldberg (2016)
- Devlin et al. (2019)

3.3 Probability Models for Discrete Data

- MRS 13
- KM 2.5.4, 3.3-3.4
- Taddy (2013, 2015)
- Gentzkow et al. (2019b)
- Davis et al. (2020)

3.4 Latent Variable Models

- KM 27.1-27.3.2, 27.3.1-27.3.6
- Blei et al. (2003)
- Hansen et al. (2018)
- Mueller and Rauh (2018)
- Larsen and Thorsrud (2019), Thorsrud (2020)
- Hansen and McMahon (2016), Hansen et al. (2019)
- Roberts et al. (2014, 2016)
- Neal (2012), Betancourt (2018)
- Srivastava and Sutton (2017)

4 Theme III: Survey Data

- Erosheva et al. (2007)
- Bandiera et al. (2020)
- Munro and Ng (2020)
- Draca and Schwarz (2021)
- Sacher et al. (2021)

References

Ash, E., Chen, D. L., and Ornaghi, A. (2020). Gender attitudes in the judiciary : Evidence from U.S. circuit courts. https://warwick.ac.uk/fac/soc/economics/research/workingpapers/2020/twerp_1256_-_ornaghi.pdf.

Baker, S. R., Bloom, N., and Davis, S. J. (2016). Measuring Economic Policy Uncertainty. *The Quarterly Journal of Economics*, 131(4):1593–1636.

Baker, S. R. and Kueng, L. (2021). Household Financial Transaction Data. Working Paper 29027, National Bureau of Economic Research.

Bandiera, O., Prat, A., Hansen, S., and Sadun, R. (2020). CEO Behavior and Firm Performance. *Journal of Political Economy*, 128(4):1325–1369.

Betancourt, M. (2018). A Conceptual Introduction to Hamiltonian Monte Carlo. *arXiv:1701.02434 [stat]*.

Bholat, D., Hans, S., Santos, P., and Schonhardt-Bailey, C. (2015). *Text Mining for Central Banks*. Centre for Central Banking Studies, Bank of England.

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

Carvalho, V., Garcia, J., Hansen, S., Ortiz, A., Rodrigo, T., Rodriguez Mora, S., and Ruiz, P. (2021). Tracking the COVID-19 crisis with high-resolution transaction data. *Royal Society Open Science*, 8.

Davis, S. J., Hansen, S., and Seminario-Amez, C. (2020). Firm-Level Risk Exposures and Stock Returns in the Wake of COVID-19. Working Paper 27867, National Bureau of Economic Research.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Deming, D. and Kahn, L. B. (2018). Skill Requirements across Firms and Labor Markets: Evidence from Job Postings for Professionals. *Journal of Labor Economics*, 36(S1):S337–S369.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the*

2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Draca, M. and Schwarz, C. (2021). How Polarized are Citizens? Measuring Ideology from the Ground-Up. SSRN Scholarly Paper ID 3154431, Social Science Research Network, Rochester, NY.

Erosheva, E. A., Fienberg, S. E., and Joutard, C. (2007). Describing disability through individual-level mixture models for multivariate binary data. *The Annals of Applied Statistics*, 1(2):502–537.

Gentzkow, M., Kelly, B., and Taddy, M. (2019a). Text as Data. *Journal of Economic Literature*, 57(3):535–574.

Gentzkow, M., Shapiro, J. M., and Taddy, M. (2019b). Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech. *Econometrica*, 87(4):1307–1340.

Goldberg, Y. (2016). A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, 57(1):345–420.

Grimmer, J. and Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(3):267–297.

Hansen, S. and McMahon, M. (2016). Shocking language: Understanding the macroeconomic effects of central bank communication. *Journal of International Economics*, 99:S114–S133.

Hansen, S., McMahon, M., and Prat, A. (2018). Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach. *The Quarterly Journal of Economics*, 133(2):801–870.

Hansen, S., McMahon, M., and Tong, M. (2019). The long-run information effect of central bank communication. *Journal of Monetary Economics*, 108:185–202.

Hansen, S., Ramdas, T., Sadun, R., and Fuller, J. (2021). The Demand for Executive Skills. Technical Report 28959, National Bureau of Economic Research, Inc.

Hassan, T. A., Hollander, S., van Lent, L., and Tahoun, A. (2019). Firm-Level Political Risk: Measurement and Effects. *The Quarterly Journal of Economics*, 134(4):2135–2202.

Larsen, V. H. and Thorsrud, L. A. (2019). The value of news for economic developments. *Journal of Econometrics*, 210(1):203–218.

Loughran, T. and Mcdonald, B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*, 66(1):35–65.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, illustrated edition edition.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013b). Distributed Representations of Words and Phrases and their Compositionality. *arXiv:1310.4546 [cs, stat]*.

Mueller, H. and Rauh, C. (2018). Reading Between the Lines: Prediction of Political Violence Using Newspaper Text. *American Political Science Review*, 112(2):358–375.

Munro, E. and Ng, S. (2020). Latent Dirichlet Analysis of Categorical Survey Responses. *Journal of Business & Economic Statistics*, pages 1–16.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press, Cambridge, MA, illustrated edition edition.

Neal, R. M. (2012). MCMC using Hamiltonian dynamics. *arXiv:1206.1901 [physics, stat]*.

Nyman, R., Kapadia, S., and Tuckett, D. (2021). News and narratives in financial systems: Exploiting big data for systemic risk assessment. *Journal of Economic Dynamics and Control*, 127:104119.

Roberts, M. E., Stewart, B. M., and Airoldi, E. M. (2016). A Model of Text for Experimentation in the Social Sciences. *Journal of the American Statistical Association*, 111(515):988–1003.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., and Rand, D. G. (2014). Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science*, 58(4):1064–1082.

Rudolph, M., Ruiz, F., Mandt, S., and Blei, D. (2016). Exponential Family Embeddings. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Ruiz, F. J. R., Athey, S., and Blei, D. M. (2020). SHOPPER: A probabilistic model of consumer choice with substitutes and complements. *The Annals of Applied Statistics*, 14(1):1–27.

Sacher, S., Battaglia, L., and Hansen, S. (2021). Hamiltonian Monte Carlo for Regression with High-Dimensional Categorical Data. *arXiv:2107.08112 [econ, stat]*.

Shapiro, A. H., Sudhof, M., and Wilson, D. J. (2020). Measuring news sentiment. *Journal of Econometrics*.

Shapiro, A. H. and Wilson, D. (2021). Taking the Fed at its Word: A New Approach to Estimating Central Bank Objectives using Text Analysis. Technical Report 2019-02, Federal Reserve Bank of San Francisco.

Srivastava, A. and Sutton, C. (2017). Autoencoding variational inference for topic models. In *ICLR*.

Taddy, M. (2013). Multinomial Inverse Regression for Text Analysis. *Journal of the American Statistical Association*, 108(503):755–770.

Taddy, M. (2015). Distributed Multinomial Regression. *The Annals of Applied Statistics*, 9(3):1394–1414.

Tetlock, P. C. (2007). Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *The Journal of Finance*, 62(3):1139–1168.

The DELVE Initiative (2020). Data readiness: Lessons from an emergency. Technical report, The Royal Society.

Thorsrud, L. A. (2020). Words are the New Numbers: A Newsy Coincident Index of the Business Cycle. *Journal of Business & Economic Statistics*, 38(2):393–409.