

# Extended Topic Models with Numerical Features

Gökhan Çapan, Ali Caner Türkmen

March 21, 2016

# Introduction

- ▶ Unsupervised learning, recover *latent* topics in documents
- ▶ Can be thought of as clustering. Loosely equivalent to link prediction.

# Latent Dirichlet Allocation

- ▶ (Blei et al., 2003)

# Comparison of Topic Models

- ▶ (Blei et al., 2003)

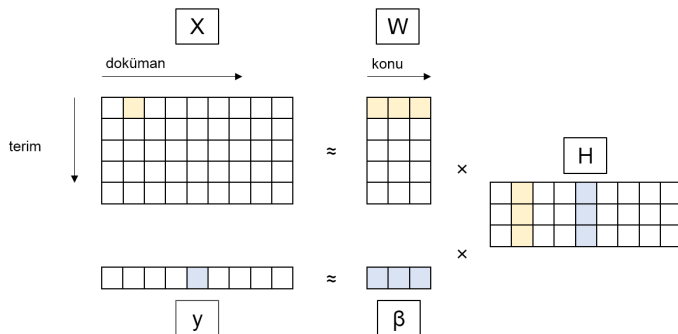
# Coupled Topic Model Applications

- ▶ Gokhan lit survey

# LDA $\equiv$ Bayesian NMF up to parameterization

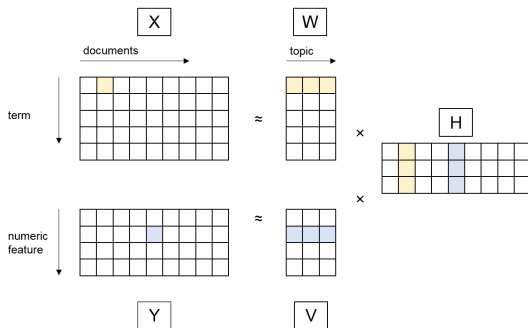
- ▶ (Jordan Blei) (Cemgil, 2009)

# Coupled Matrix Factorization for Recovering Topics



- ▶ Represent data with well-known algebraic structures
- ▶ Jointly guide topic assignments from multimodal datasets, in a probabilistically sound framework
- ▶ Easily extensible to semi-supervised learning, kernel methods
- ▶ (T. and Cemgil, 2016)

# Extended Coupled NMF for Topic Learning with Count Features



- ▶  $y_{ij} | W, H \sim GPO(\sum_t w_{it} h_{tj}, \phi)$
- ▶  $x_{kj} | V, H \sim GPO(\sum_t v_{kt} h_{tj}, \gamma)$
- ▶ Guide topic modeling with numeric count data
- ▶ Can assume priors  $p(W), p(H), p(V)$  for Bayesian learning



# Data Set and Features

- ▶ **Data Set:** News articles sampled from Anadolu Agency website.
- ▶ **Features:** Complexity features such as word count, sentence count, average sentence length, comma count.
- ▶ **Novel Features:** Etymological counts. Count the number of words from their etymological origins. Number of Arabic, Farsi, French words, etc.

# Learning

- ▶ EM-like updates with multiplicative NMF update rules
- ▶ (out of scope for this project) : Gibbs sampling

# Conclusion

We propose several contributions

- ▶ Put the topic modeling problem in a coupled NMF framework, extending with numerical features
- ▶ Use etymological counts for the Turkish language