Probabilistic Programming for Regression with High-Dimensional Categorical Data

Szymon Sacher
Columbia University

Laura Battaglia
Barcelona GSE

Stephen Hansen
Imperial College

Early draft: `https://arxiv.org/abs/2107.08112`

Monash-Warwick-Zurick Text-as-Data Workshop
February 17, 2022

Overview

Motivation

High dimensional data is increasingly common

- Text, surveys, shopping baskets.

Reasearchers typically follow a 2-step approach:

1. Use a latent variable model to extract low-dimensional representation of the data.
2. Use the extracted latent variables in OLS, discrete choice model, etc.

Motivation

High dimensional data is increasingly common

- Text, surveys, shopping baskets.

Reasearchers typically follow a 2-step approach:

1. Use a latent variable model to extract low-dimensional representation of the data.
2. Use the extracted latent variables in OLS, discrete choice model, etc.

This approach is black-box, statistically invalid, and inefficient.

Motivation

High dimensional data is increasingly common

- Text, surveys, shopping baskets.

Reasearchers typically follow a 2-step approach:

1. Use a latent variable model to extract low-dimensional representation of the data.
2. Use the extracted latent variables in OLS, discrete choice model, etc.

This approach is black-box, statistically invalid, and inefficient.

We show that Probabilistic Programming provides an alternative that can be easy to use, valid, efficient, and fast.

Probabilistic Programming

Probabilistic programming languages (PPLs) are domain-specific languages that describe probabilistic models and the mechanics to perform inference in those models.

PPLs implement variety of general-purpose inference algorithms (NUTS, BBVI, HMCES, etc.)

Often utlize automatic differentiation and accelerators (GPUs, TPUs).

Examples: Stan, Turing, Pyro, PyMC3, Numpyro, Greta, and more.

Literature

Latent variable models are common in economics:

- **Text** Macro/finance forecasting (Larsen and Thorsrud 2019, Bybee et al. 2020, Thorsrud 2020, Ellingsen et al. 2021); conflict forecasting (Mueller and Rauh 2018); asset pricing (Hanley and Hoberg 2019, Lopez Lira 2019); political deliberation (Hansen et al. 2018, Stiglitz and Caspi 2020); media economics (Nimark and Pitschner 2019, Bertsch et al. 2021, Widmer er al. 2022).
- **Other data** Surveys (Bandiera et al. 2020, Munro and Ng 2020, Draca and Schwarz 2021); Networks (Nimczik 2017, Olivella et al. 2021); IO (Sorensen et al, 2021, Han et al 2021)

Integrated models are less common.

- Notable exceptions: Vafa et al 2021 for sentiment, Olivella et al. 2021 for networks, Munro and Ng 2020 for survey data, Roberts et al (2013) for text.

PPLs are gaining popularity, but on very low-dimensional models:

- Examples: Meta-Analysis (Meager, 2019; Bandiera et al 2021); survey data (Angelucci & Prat, 2021); IO (Olenski and Sacher, 2022)

Overview

Topic models

LDA decomposes document-term matrix into a set of topics and a set of document-specific topic shares.

- The topics, β_k are probability distributions over terms.
- The topic shares, θ_d are probability distributions over topics.

Various extensions have been proposed, each with a custom inference algorithm.

- Examples: DTM, STM, Supervised TM.

In practice the plain LDA is often used and its' outputs are fed into downstream tasks.

We propose and estimate a model where topics θ_d depend on covariates q_d and together with regressors z_d influence the outcome Y_d.

Figure: Plate Diagram: Simple topic model

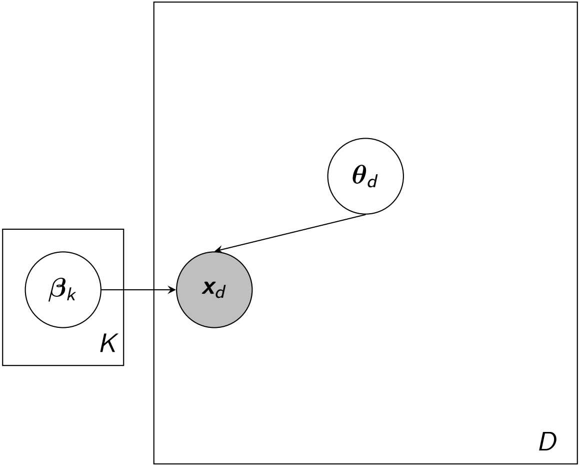


Figure: Plate Diagram: Regressions of topics on covariates

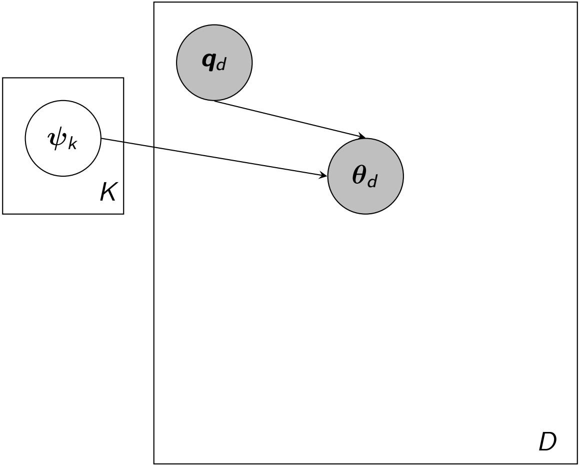


Figure: Plate Diagram: Regressions of outcomes on topics

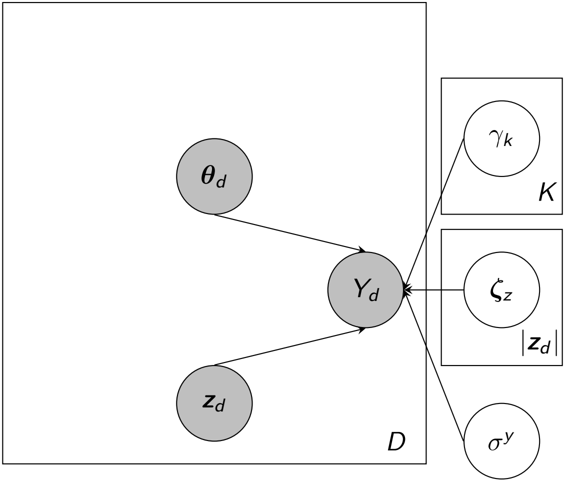
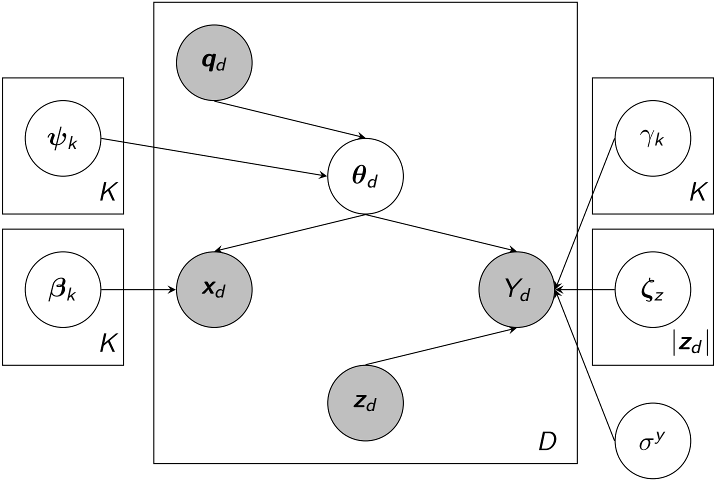


Figure: Plate diagram: Supervised topic model with covariates (Sup-TM-C)



Math | Code

Overview

Simulation study

We implement the Sup-TM-C model with Numpyro using Hamiltonian Monte Carlo (HMC) on a GPU.

- HMC only requires the value and gradient of log-joint.
- Guaranteed to converge, typically much faster than Metropolis-Hastings. Guaranteed to be correct.

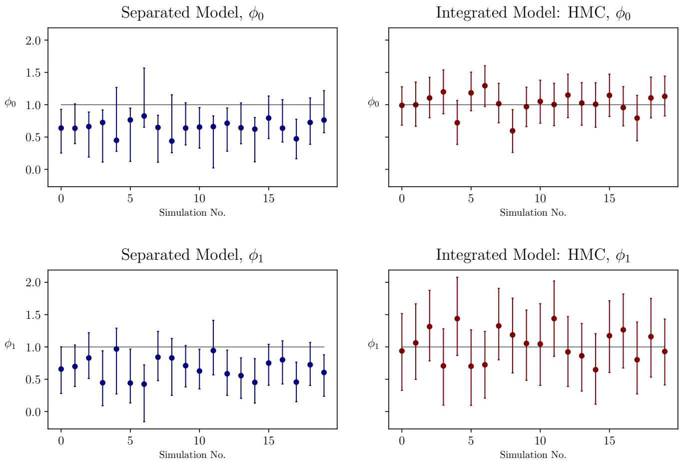
First we perform 20 simulations with $D = 1000$ documents of length $N = 300$ words each, $K = 2$ topics and $V = 500$ words in the dictionary.

We set the true value of regression coefficients ϕ_0, ϕ_1 and γ to 1.

We compare the results obtained with NUTS with ones using ''separated'', 2-step approach.

- Estimation time with NUTS: approx 3 min per simulation.

Simulation: effect of covariates on topic selection



Effect of topics on outcomes

Empirical application

Bandiera et al (2020) conducted a survey of 1114 CEOs and recorded their daily activities.

- Data represented as 654-dimensional vectors.

They use LDA to obtain posterior distribution of 2-dimensional CEO behavior index, θ_d.

Mean posterior value of $\theta_{d,1}$, $\hat{\theta}_{d,1}$, is used downstream in OLS/IV:

$$\hat{\theta}_{d,1} = g_d^T \gamma + \varepsilon_d \qquad \text{CEO characteristics and CEO index}$$

$$y_d = \hat{\theta}_{d,1} + q_d^T \zeta + \epsilon_d \qquad \text{CEO index on firm performance}$$

The CEO index is found to correlate with CEO and firm characteristic, including performance.

Correlations with observables

	Dependent Variable:		
	Log(sales)		Un-normalized CEO Index
	Sup-TM (1)	Sup-TM-C (2)	Sup-TM-C (3)
CEO Index, $\theta_{d,1}$	0.282 (0.119, 0.417)	0.317 (0.178, 0.488)	
Log Employment	0.945 (0.902, 1.008)	0.95 (0.911, 0.985)	0.438 (0.38, 0.499)
MBA			0.346 (0.21, 0.471)
Family CEO			-0.728 (-0.85, -0.595)
Public Firm			-0.986 (-1.172, -0.819)
MNE			1.081 (0.927, 1.265)
Controls	X	X	X

Posterior means and credible intervals of regression coefficients ψ (columns 1-2) and γ (column 3).

Correlations with observables

	Dependent Variable:		
	Log(sales)		Un-normalized CEO Index
	Sup-TM (1)	Sup-TM-C (2)	Sup-TM-C (3)
CEO Index, $\theta_{d,1}$	0.282	0.317	
	(0.119, 0.417)	(0.178, 0.488)	
Log Employment	0.945	0.95	0.438
	(0.902, 1.008)	(0.911, 0.985)	(0.38, 0.499)
MBA			0.346
			(0.21, 0.471)
Family CEO			-0.728
			(-0.85, -0.595)
Public Firm			-0.986
			(-1.172, -0.819)
MNE			1.081
			(0.927, 1.265)
Controls	X	X	X

Posterior means and credible intervals of regression coefficients ψ (columns 1-2) and γ (column 3).

Overview

BBVI for large corpora

In many applications with large corpora Hamiltonian Monte Carlo may be impractical.

In such cases we propose to use Black Box Variational Inference (BBVI) (Ranganath et al 2013, Kingma & Welling 2013):

- Researcher specifies an approximating parametrized *variational family*.
- The objective is to minimize distance between the approximating distribution and the true posterior.
- Optimization uses *reparametrization trick*, autodiff and efficient gradient-based optimizers (e.g. Adam).

BBVI converges fast and is scalable for large corpora.

Caveat Covariance is typically underestimated. The posterior is under-dispersed

Example: Topics in NBER abstracts

We collect abstracts from NBER Working Papers for 1980-2021 and the associated 1 digit JEL codes

The corpus contains $D = 27k$ documents, $V = 7113$ distinct terms, $Z = 50$ non-exclusive JEL codes. We estimate the model with $K = 30$ topics.

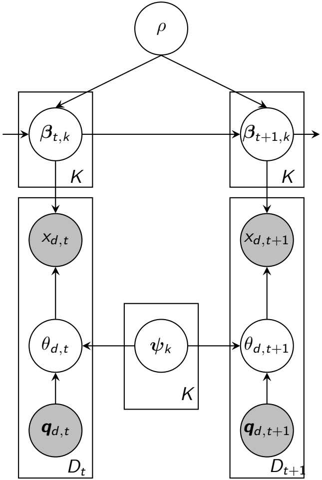
The model we propose (D-TM-C) has the following features:

- Topics evolve smoothly over time (see Blei&Lafferty, 2006).
- Documents' topic shares depend on JEL codes (see Roberts et al, 2013)

Goal

Recover interpretable, changing topics and their associations with JEL fields.

Figure: Plate diagram: Dynamic topic model with covariates.



Examples of topics

Examples of topics

Examples of topics

Examples of topics

Examples of topics

Examples of topics

Examples of trends within topics

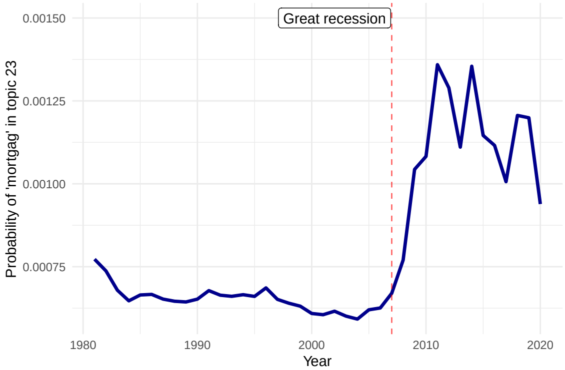


Figure: Probability of word *mortgage* given topic *23*

Examples of trends within topics

Figure: Probability of word *sovereign* given topic *23*

Examples of trends within topics

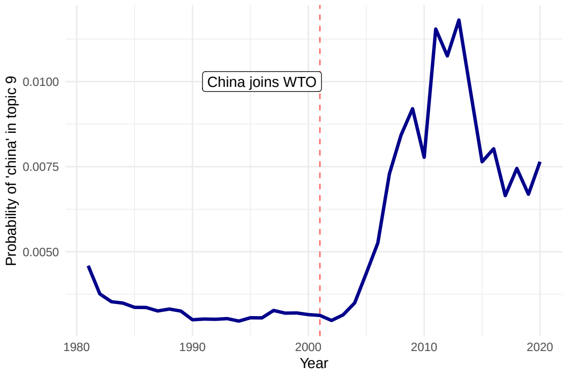


Figure: Probability of word *china* given topic *9*

Topics most associated with JEL codes

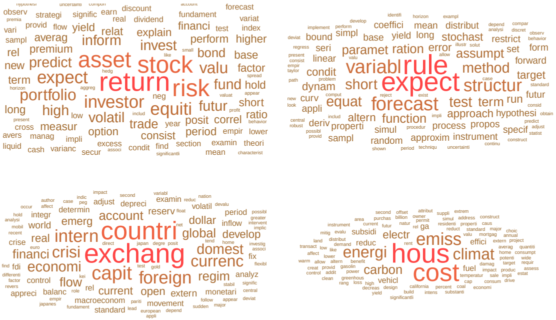


Figure: The topics with the largest γ coefficients for the JEL code *G1: Financial Economics – General Financial Markets*

Topics most associated with JEL codes

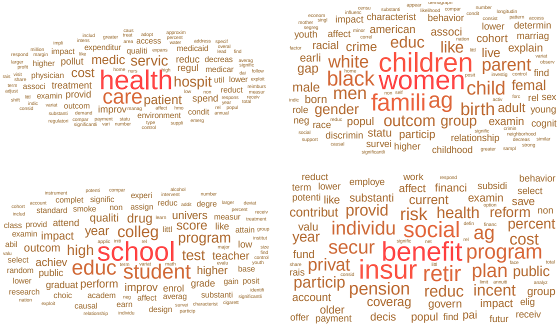


Figure: The topics with the largest γ coefficients for the JEL code *I1: Health, Education, and Welfare – Health*

Topics most associated with JEL codes

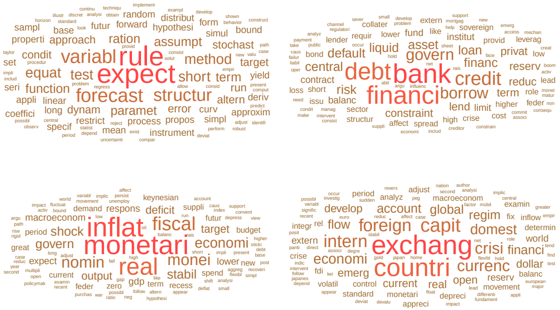


Figure: The topics with the largest γ coefficients for the JEL code *E2: Macroeconomics – Consumption, Saving, Production, Investment, Labor Markets, and Informal Economy*

Topics most associated with JEL codes

Figure: The topics with the largest γ coefficients for the JEL code *E3: Macroeconomics and Monetary Economics – Prices, Business Fluctuations, and Cycles*

Overview

Conclusion

We show that PPLs can be useful to model high-dimensional categorical data.

- We develop two new models, Sup-TM-C and D-TM-C, and apply them to survey and text data.

In small applications we recommend inference by sampling using HMC/NUTS

- Inference guaranteed to be correct. Few parameters are needed to be set.

When HMC/NUTS not feasible we recommend inference by black-box variational inference.

- Very fast and scalable but caution needed when interprating the posterior variance.

Probabilistic Programming for Regression with High-Dimensional Categorical Data

Szymon Sacher
Columbia University

Laura Battaglia
Barcelona GSE

Stephen Hansen
Imperial College

Early draft: `https://arxiv.org/abs/2107.08112`

Monash-Warwick-Zurick Text-as-Data Workshop
February 17, 2022

Overview

Extra Slides

Structural-Supervised LDA

$$\overbrace{\boldsymbol{\beta}_k \sim \text{Dirichlet}(\eta)}^{\text{CEO behavior types}}$$

$$\boldsymbol{\psi}_k \sim \text{Normal}(\mathbf{0}, \mathrm{I}\sigma^{\gamma})$$

$$\underbrace{\boldsymbol{\theta}_d \sim \text{LogisticNormal}\left[(\boldsymbol{q}_d^T \boldsymbol{\psi}_1, \ldots, \boldsymbol{q}_d^T \boldsymbol{\psi}_K)^T, \mathrm{I}\sigma^{\theta}\right]}_{\text{Latent variable -- CEO index}}$$

$$\underbrace{\boldsymbol{x}_d \sim \text{Multinomial}\left(\sum_k \boldsymbol{\beta}_k \theta_{d,k}, N_d\right)}_{\text{Categorical data -- CEO behaviour}}$$

$$\boldsymbol{\gamma} \sim \text{Normal}(\mathbf{0}, \mathrm{I}\sigma^{\gamma})$$

$$\boldsymbol{\zeta} \sim \text{Normal}(\mathbf{0}, \mathrm{I}\sigma^{\zeta})$$

$$\sigma_y \sim \text{Gamma}(s_0, s_1)$$

$$\underbrace{y_d \sim \text{Normal}(\boldsymbol{\theta}_d^T \boldsymbol{\gamma} + \boldsymbol{z}_d^T \boldsymbol{\zeta}, \sigma_y^2)}_{\text{Numerical data -- firm performance}}$$

A CEO is a mixture θ_d of various CEO types, β_k. Probability of different activities, x_d depend on CEO type θ_d and CEO/firm covariates $\boldsymbol{q}_d$
Firm performance, y_d depends on CEO index θ_d and other characteristics, $\boldsymbol{z}_d$.

Back

Maths vs Code

$$\beta_k \sim \text{Dirichlet}(\eta)$$

$$\psi_k \sim \text{Normal}(\mathbf{0}, \mathrm{I}\sigma^{\gamma})$$

$$\boldsymbol{\theta}_d \sim \text{LogisticNormal}[(\boldsymbol{q}_d^T \psi_1, \ldots, \boldsymbol{q}_d^T \psi_K)^T, \mathrm{I}\sigma^{\theta}]$$

$$\boldsymbol{x}_d \sim \text{Multinomial}\left(\sum_k \beta_k \theta_{d,k}, N_d\right)$$

$$\gamma \sim \text{Normal}(\mathbf{0}, \mathrm{I}\sigma^{\gamma})$$

$$\zeta \sim \text{Normal}(\mathbf{0}, \mathrm{I}\sigma^{\zeta})$$

$$\sigma_y \sim \text{Gamma}(s_0, s_1)$$

$$y_d \sim \text{Normal}(\boldsymbol{\theta}_d^T \gamma + \boldsymbol{z}_d^T \zeta, \sigma_y^2)$$

```python
import jax.numpy as jnp
import numpyro.distributions as dist
from numpyro import sample, plate
from jax.nn import softmax


def structural_slda(Y, X, N, Z, Q, K, eta = .1, alpha = 1):
    # Y : regression outcomes
    # X : document-word matrix of BoWs
    # N : total word counts per document
    # Z : matrix of non-text covariates
    # Q : matrix of covariates entering topic selection
    # K : number of topics
    # eta, alpha : Dirichlet hyperparamters

    D, V = X.shape
    z,q = Z.shape[1], Q.shape[1]

    #### LDA part of model

    with plate("topics", K):
        # Topic-word distributions
        beta = sample("beta", dist.Dirichlet(eta * jnp.ones(V)))

    phis = sample("phis", dist.Normal(0,2).expand([q, K-1]))

    with plate("docs", D, dim = -2):
        A = sample("A", dist.Normal(jnp.matmul(Q, phis) , alpha))

    # document-topic distributions
    theta = softmax(jnp.hstack([A, jnp.zeros([D, 1])]), axis = -1)

    distMultinomial = dist.Multinomial(total_count=N, probs = jnp.matmul(theta, beta))
    with plate("hist", D):
        sample("obs_x", distMultinomial, obs = X)

    #### Regression part of model
    gammas = sample("gammas", dist.Normal(0, 2).expand([K-1]))
    zetas = sample("zetas", dist.Normal(0,2).expand([z]))
    sigma = sample("sigma", dist.Exponential(1.))

    mean = jnp.matmul(theta[:,:(K-1)], gammas) + jnp.matmul(Z, zetas)

    with plate("y", D):
        sample("obs_y",  dist.Normal(mean, sigma), obs = Y)
```

Back

Simulation: effect of topics on outcomes