Exchange Rate Supervised Topic Modeling

Aaron Mora
Darla Moore School of Business
University of South Carolina

This version: August, 2025

Abstract

This paper shows how to use a hybrid of supervised and unsupervised learning models to go from text from news articles to an FX news index that can be used to enhance a traditional FX regression of the order flow. To do so, I adapt a Supervised Latent Dirichlet Allocation model that combines information about a supervising variable with topic extraction over a corpus of text in a single-stage estimation. The empirical application I present estimates a text-augmented order flow regression using transaction-level data for the Costa Rican colón and the United States dollar exchange rate along with news articles from the main Costa Rican newspapers. Results show that news articles can be used to construct covariates with relevant explanatory power for the dynamics of the exchange rate over short frequencies. This empirical relevance of text-based covariates from news articles reinforces the notion that order flow is driven by private information and facilitates building narratives around episodes of high volatility of the exchange rate.

Acknowledgments: I am deeply thankful with Nikolai Roussanov, Francis X. Diebold, Frank Schorfheide, Jesús Fernández Villaverde and Karun Adusumilli for great advice on this project. For valuable comments and suggestions I also thank seminar participants at the Penn Econometrics lunch seminar, the Central Bank of Costa Rica Research Seminars and conference participants of the 2022 Graduate Student Conference at the Washington University at St. Louis, the 2022 Rising Scholars Conference at University of Chicago Booth School of Business, the 2023 International Assoc. for Applied Econometrics Annual Conference at BI Norwegian Business School. All remaining errors are mine.
Key words: FX market, Natural Language Processing, Supervised Latent Dirichlet Allocation.
JEL codes: C82; G15; G17; O24.
Contact: aaron.mora@moore.sc.edu

1 Introduction

Standard macroeconomic theory supports the view that exchange rates are determined mainly by fundamental macroeconomic variables such as, interest rate differentials, productivity differences, terms of trade, and others. However, empirical evidence has repeatedly shown that such models lack explanatory power when dealing with the short run dynamics of the exchange rate.

Under short periods of intense activity on FX markets, fundamentals can be silent for a practical reason: data on them is published at a much longer frequency, meaning that when confronted with data at short frequencies, fundamentals may not have enough variation to act as covariates in explaining the exchange rate. Despite the lack of new information on fundamentals, it could be that agents are acting on the expectations they have over fundamentals which are essentially latent variables changing at a much faster frequency than published data.

News articles could be an important source of information capable of causing agents to revise their expectations. If the expectation formation process is a signal extraction problem (e.g. Beaudry and Portier (2014)), some news articles contain new information related to fundamentals, so when agents manage to filter a positive signal correctly and act accordingly, fundamentals should eventually reflect their position. On the other hand, when agents respond positively to a signal that turns out to be noise, fundamentals will prove them wrong, and agents should revise their expectations. True news is therefore information that should have predictive power for future developments in the economy.

This idea suggests *filtering* news articles to extract the signal that is relevant for predicting developments in the economy. This paper argues that this can be done by extracting the topics a news article writes about and determining which topics correlate with a macroeconomic variable of interest tracking an aspect of the economy. For example, when considering an exchange rate, to the extent that certain topics in news articles provide a relevant description of the economy, the more intensely such topics are represented in the newspaper at a given point in time the more likely it is that such topics represent something important for the economy's development and may cause revised expectations for some agents. This suggests that the information in news articles might be relevant in explaining the short term dynamics in the exchange rate.

This paper aims to gather evidence on whether news articles have explanatory power for exchange rates in short term frequencies. The empirical evidence will be centered

around the Monex Market, the main Costa Rican platform for FX trade. Accordingly, news articles will be gathered from the main Costa Rican newspapers. The exchange rate of interest is that between the Costa Rican Colón (CRC), the local currency, and the United States dollar (USD). As the United States is Costa Rica's main trade partner, the CRC/USD exchange rate receives almost exclusive attention in FX matters from the general Costa Rican public and the monetary authorities in Costa Rica.

Methodologically, the strategy of this paper uses a hybrid of unsupervised and supervised learning models to go from text data in the news articles to an FX news index that can be used to enhance traditional models from the FX literature. To do so, I use an adaptation of the Supervised Latent Dirichlet Allocation (sLDA) model (Blei and McAuliffe 2008) which combines information about a supervising variable with topic extraction over a corpus of text in a single-stage estimation.

It is possible from a methodological perspective to divide the problem into two stages: a first stage that extracts a low-dimensional feature from the unstructured text data and a second stage that assesses how this low-dimensional feature correlates with the response variable. For example, in the context of exchange rates, we could divide the problem into a first stage that uses standard Latent Dirichlet Allocation (LDA) (Blei, Ng and Jordan 2003) to extract topics from the news articles and a second stage that uses the estimated topic frequencies as regressors on a metric y of the exchange rate (for example the daily log return). However, via a Monte Carlo simulation I provide evidence that opting for a single-stage estimation, like sLDA, has efficiency gains over a two-stage approach.

Another strategy to process text from newspapers into indices is sentiment extraction. In the standard approach sentiment metrics from news articles are computed based on positive, negative, or particular word counts according to a pre-specified dictionary. Results can be strongly driven by the subjective choice of this dictionary that dictates which words should count. A topic extraction approach over the news articles, like in a sLDA model, is a less restrictive choice for producing news measures, since topics are extracted in an agnostic fashion. News measures based on positive or negative counts implicitly assume a two topic structure (positive and negative) and are more prone to a *market's animal spirits* interpretation where changes in expectations can be self-fulling and not necessarily rooted in economic fundamentals.

Using the estimates from sLDA it is possible to summarize the information from the supervised extracted topics into an FX news index. This FX news index in turn can be studied as a covariate controlling for public information from news articles in

traditional models looking at FX dynamics. Construction of the index relies on the parameter estimates from a single-stage approach, which improves parameter precision. With a similar objective, Mora (2020)[1] uses a two-stage approach combining LDA and a supervised model with dynamic variable relevance to study the role of news articles in the CRC/USD exchange rate from the Monex market. Evidence is provided that the inclusion of an FX news index improves traditional order flow models. In this framework, order flow remains an important driver in short-term dynamics for the exchange rate after controlling for public information using a two-stage methodology for topic extraction. This reinforces the notion that order flow is driven by private information.

Related literature This project is related to at least four strands of literature. The first one is about the use of text data in economics and finance. Gentzkow et al. (2019) survey different approaches to incorporate textual analysis into answering economic questions as well present applications of such approaches in recent papers. One example of a paper which relies on topic models is Larsen and Thorsrud (2019) which pursues a similar, though methodologically different, goal to this project. They use news articles from Norwegian news outlets and assess their information content to predict a set of macroeconomic variables in Norway. Thorsrud (2020) builds on this dataset of Norwegian news articles to construct a daily business cycle index combining the text data with quarterly GDP growth. Using news articles from the Wall Street Journal, Bybee, Kelly, Manela and Xiu (2020) combine a LDA topic model with a vector autoregressive model to track the state of the U.S. economy via text analysis of business news.

Examples of text models in finance include Ke et al. (2021) who use a supervised sentiment extraction model on labeled Dow Jones Newswires to predict asset returns. Lopez-Lira (2023) uses a LDA model on annual reports to elicit the risk factors that firms themselves identify in their annual reports. It is worth remarking that all of the papers mentioned above follow a two-stage approach to estimation where the first stage focuses on extracting numerical features of text data that will be subsequently used in a second stage that relates such features to economics variables of interest.

The second related literature as mentioned in the introduction deals with studying short term dynamics of the exchange rate. This paper is related to the literature of market microstructure in exchange rates and the role of the order flow, the difference between buyer-initiated and seller-initiated trades, as a high frequency covariate for the

[1]This is a previous version of this project. It can be consulted at https://www.dropbox.com/s/4s4g7owbwdu266p/Mora_TextIntoFX.pdf?dl=0.

exchange rate. Many of the key insights of this literature can be found in Lyons (1995) and Evans and Lyons (2002), surveys of the literature include Vitale (2007) and King et al. (2013). Many papers have suggested explanations for the empirical relevance of the order flow in exchange rate dynamics (e.g. Bacchetta and Van Wincoop (2006), Vitale (2011) Evans and Lyons (2012)). However, despite the fact that these papers propose different hypotheses for why the order flow has explanatory power for short-run exchange rate dynamics, a definite answer backed by empirical evidence remains an open question.

A third related literature has to do with topic modeling. The LDA model from Blei, Ng and Jordan (2003) has become an industry standard in recent years and has sprouted a series of variations extending different aspects of the original model. Blei, Griffiths, Jordan and Tenenbaum (2003) propose hierarchical LDA (hLDA) to include topic hierarchies by using the Chinese restaurant process as a prior for such topic hierarchies[2]. Andrzejewski et al. (2009) also extends LDA at the level of topic priors to included domain knowledge, that is external information about the composition of words that should have high or low probability in various topics. Jagarlamudi, Daumé III and Udupa (2012) propose seedeed LDA that also incorporates domain knowledge on the topic priors to guide topic models and learn about topics of specific interest to a user by providing sets of seed words that a user believes are representative of the underlying topics in a corpus. Ramage, Hall, Nallapati and Manning (2009) also extend LDA to incorporate external information and propose Labelled LDA to study the problem of credit attribution when documents have user tags in addition to a topic structure. A dynamic extension of LDA is presented in Blei and Lafferty (2006), where topic and topic membership for each document evolves over time as a function of previous periods' values. All these papers reflect the convenient modularity of LDA, demonstrating how its basic topic structure can be part of richer models that work with text corpora with different features. In this paper I introduce slight modifications to sLDA so it fits our goal of introducing text data into a regression of the exchange rate; details are shown in Section 2.

Finally a fourth related literature has to do with approximate inference in models with latent variables, in particular with the variational inference methods. Topic models, and in particular sLDA, posit latent features (variables) in the text structure. Given the high-dimensional nature of text data, it is common to resort to approximate inference.

[2]In this extension, topics follow a L-dimensional tree structure that represent their hierarchy. Then a document is drawn by first choosing an L-level path through the restaurants and then drawing the words from the L topics which are associated with the restaurants along that path.

sLDA, as presented below, performs estimation using variational inference with a mean-field approximation. A good introduction to the techniques of variational inference can be found in Blei, Kucukelbir and McAuliffe (2017); more through treatments can be found in Bishop (2006) and Wainwright and Jordan (2008).

The rest of this paper is organized as follows: section 2 shows how to extract supervised topics from a database of news articles using a sLDA model. Section 3 shows the simulation study that documents the efficiency gains of using sLDA as opposed a two-stage strategy of extracting topics using the traditional LDA and then using its output in a regression. Section 4 presents the data for our empirical application, the CR News database and the Monex Market, the source of the exchange rate data. Section 5 presents the strategy for constructing the FX news index. Finally, Section 6 concludes.

2 Supervised Topic Extraction

Given a corpus of news articles, our premise is that each word in each document has a latent assignment into one of K possible topics and moreover this latent assignment determines a response variable y. In this context each document is seen as a bag of words, that is a sequence of words $\{w_n\}_{n=1}^{N}$ where each word w_n is indexed in a dictionary of size V and N is the total number of words in a document. Once indexed, a document becomes a sequence of positive integers. To make this sequence human-readable, we would need to replace index $w_n \in \{1, ..., V\}$ with the word corresponding to the w_n-th index in the dictionary.

Topics are considered as clusters of semantic content and modeled as unknown distributions over a dictionary V. To estimate these latent assignments and their relationship with the response variable y, I will use a modified version of Supervised Latent Dirichlet allocation (Blei and McAuliffe (2008)) (sLDA) that builds on the industry standard methodology of Latent Dirichlet Allocation (LDA) (Blei, Ng and Jordan (2003)) to include in a single-stage estimation information of the response variable to improve topic extraction. In a nutshell, sLDA is based on three premises:

i. A topic is a distribution over words.

ii. A document is a mixture of topics. That is, there is mixed membership of documents into topics.

iii. The empirical topic membership of a document along with covariates W_t, determine a response y_t, a label for the document.

In this context data is composed of document-response triplets $\{(w_{t,1:N_t}, y_t, W_t)\}_{t=1}^{T}$. In our exchange rate application, t would index daily dates: y_t will represent the daily log return using the exchange rate between day t and day $t-1$. The variable $w_{t,1:N_t}$ stands for the news articles available at the beginning of day t, hence produced in day $t-1$, pooled into a single document. In this aggregation we are exploiting the mixed membership feature of LDA models[3] Finally, W_t are other covariates known to be relevant to determine short-term dynamics of the exchange rate, like the order flow.

Two modifications are done with respect to sLDA as presented in Blei and McAuliffe (2008). The first modification is to include covariates other than text in the supervising equation; which is important for the exchange rate application. The second, is to include a Dirichlet prior for the vector of term frequencies of each topic. This facilitates the estimation process and it is line with the *smooth*-LDA (see Blei, Ng and Jordan (2003) section 5.4) implemented in most programming packages. The next subsection presents the generative for the modified sLDA.

2.1 Generative Model

Given a corpus of T documents indexed by $t = 1, ..., T$ compose of words from a dictionary of size V. The supervised model used assumes that there are K underlying topics that generate the text in the following fashion:

i. Draw K vectors $\beta_k \sim Dir(\eta)$ so $\beta_k \in \Delta^{V-1}$.

ii. Draw T vectors $\theta_t \sim Dir(\alpha)$ so $\theta_t \in \Delta^{K-1}$

iii. For each word $w_{t,n}$ with $n = 1, ..., N_t$

- Draw topic assignment $z_{t,n}|\theta_t \sim Mult(\theta_t)$. So $z_{t,n} \in \{1, ..., K\}$.
- Draw word $w_{t,n}|z_{t,n}, \beta_{z_{t,n}} \sim Mult(\beta_{z_{t,n}})$. So $w_{t,n} \in \{1, ..., V\}$.

iv. Compute the empirical topic frequency $\bar{Z}_t = \frac{1}{N_d}\sum_{n=1}^{N_t} z_{t,n}$; with $z_{t,n}$ written as an one-hot vector of size K.

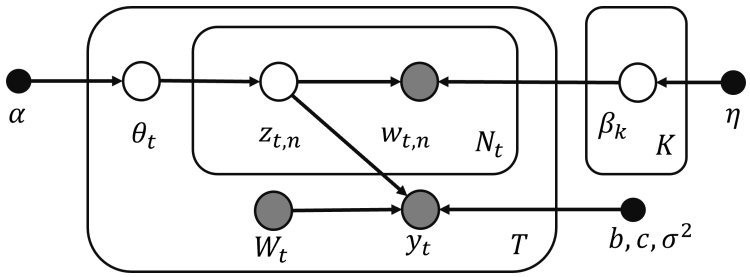
[3] Alternatively we can keep each news article as a document and then pool the empirical topic frequencies of each document with timestamp t into a pooled topic frequency for day t.

v. Given exogenous regressors W_d and parameters generate a response y_d as

$$y_t = b'\bar{Z}_t + c'W_t + \varepsilon_t, \qquad \varepsilon_t \sim N(0, \sigma^2) \tag{1}$$

The variables dependencies can be summarized in the graphical model depicted in Figure 1.

Figure 1: sLDA Graphical Model



Notes: Graphical model for the generative model of Supervised Latent Dirichlet Allocation.

In specifying the response equation (1) there are several modeling choices. Here I use a basic Gaussian response; however, Blei and McAuliffe (2008) show how this response equation can be generalized to other members of the exponential family. A response with time varying parameters and with possible stochastic volatility is left for future research.

In the generative model I have assumed that the topic structure determines the response. It is possible to write a version of equation (1) that uses θ instead of $\bar{Z}$. This would imply that documents and responses are exchangeable in the sense that once θ is drawn we can first generate the document or the response. It makes more sense in the context of an application that seeks to extract information from news articles capable of changing market behavior and hence market outcomes to assume that documents are determined first followed by responses. Notice that so far (and in the details below) that the strategy would be the same if y is a response containing information for other economic variables, suggesting that we can think of supervised topic extraction with GDP, inflation, trade, unemployment, and other macroeconomic variables.

2.2 Variational Inference

Given data,T the next natural step is to wonder how we can estimate the parameters involved in the framework above. The main challenge is the presence of latent variables, namely θ that controls document topic membership, $\{z_n\}$ the words' topic assignments and $\beta_{1:K}$ the term frequencies for each topic. The posterior distribution of the latent variables is given by:

$$p(z_{1:N}, \theta, \beta_{1:K} | w_{1:N}, y, W; b, c, \sigma^2; \alpha, \eta) = \frac{p(y|z_{1:N}, W, b, c, \sigma^2) \left[\prod_{n=1}^{N} p(w_n|z_n, \beta_{1:K}) p(z_n|\theta)\right] p(\theta|\alpha) p(\beta_{1:K}|\eta)}{\sum_{z_{1:n}} \int_{\theta} \int_{\beta_{1:K}} \left\{p(y|z_{1:N}, W, b, c, \sigma^2) \left[\prod_{n=1}^{N} p(w_n|z_n, \beta_{1:K}) p(z_n|\theta)\right] p(\theta|\alpha) p(\beta_{1:K}|\eta)\right\} d\theta d\beta}$$

The normalizing term in the posterior distribution of the latent variables is the likelihood of the data $p(w_{1:N}, y|\pi, W)$ where $\pi = \{b, c, \sigma^2, \alpha, \eta\}$. Unfortunately this likelihood is not efficiently computable, as documented in Blei, Ng and Jordan (2003), making the posterior of the latent variables also not efficiently computable [4]. If this posterior were computationally tractable we could use a standard EM algortihm (Dempster et al. (1977)) to perform maximum likelihood and estimate the parameters.

Given this limitation, I resort to approximate maximum-likelihood estimation using a variational expectation maximization (VEM) strategy where I treat parameters π as constants to be estimated rather than random variables. The VEM strategy follows the spirit of the traditional EM algorithm but I use a variational approximation to the complete likelihood to find a maximizing set of parameters as our estimates. Notice the hybrid approach to parameters in this strategy: latent variables that can be thought of as observation-specific parameters that have a Bayesian structure with priors on them, whereas common latent variables or global parameters are treated as unknown constants. For a discussion on when to use variational methods versus exact inference strategies like MCMC see Blei et al. (2017).

2.3 Evidence Lower Bound and Mean Field Approximation

To obtain an approximation of the complete likelihood in the presence of latent variables, I resort to the Evidence Lower Bound (ELBO), a key object in variational inference. Let $q(\theta, z_{1:N}, \beta_{1:K})$ denote the *variational distribution* for our latent variables (topic member-

[4] Notice that in the the normalizing term, just the sum is over K^N terms.

ships, word assignments and topic term frequencies). By appropriately selecting the family of the variational distribution, we are able to approximate the likelihood. The log likelihood of the observed data is given by:

$$\begin{aligned}
\log p(w_{1:n}, y|\pi, W) &= \log \left[\sum_{z_{1:n}} \int_{\theta} \int_{\beta_{1:K}} p(w_{1:n}, y, z_{1:n}, \theta, \beta_{1:K}|\pi, W) d\theta d\beta_{1:K} \right] \\
&= \log \left[\sum_{z_{1:n}} \int_{\theta} \int_{\beta_{1:K}} p(w_{1:n}, y, z_{1:n}, \theta, \beta_{1:K}|\pi, W) \frac{q(z_{1:n}, \theta, \beta_{1:K})}{q(z_{1:n}, \theta, \beta_{1:K})} d\theta d\beta_{1:K} \right] \\
&\geq E_q \left[\log p(w_{1:n}, y, z_{1:n}, \theta, \beta_{1:K}|\pi, W) \right] - E_q \left[\log q(\theta, z_{1:N}, \beta_{1:K}) \right]
\end{aligned}$$

Expectations in the last line are with respect to $q(\theta, z_{1:N}, \beta_{1:K})$. This last line is the Evidence Lower Bound (ELBO). Denoting the entropy of the variational distribution as $H(q) = -E_q[\log q(\theta, z_{1:N}, \beta_{1:K})]$ the ELBO expands into:

$$\begin{aligned}
\text{ELBO} &= E_q[\log p(w_{1:N}, y, \theta, z_{1:N}, \beta_{1:K}|\pi, W)] + H(q) \\
&= E_q[\log p(\theta, z_{1:N}, \beta_{1:K}|w_{1:N}, y, \pi)] + E_q[\log p(w_{1:N}, y|\pi, W)] + H(q) \\
&= E_q[\log p(\theta|\alpha)] + \sum_{k=1}^{K} E_q[\log p(\beta_k|\eta)] + \sum_{n=1}^{N} E_q[\log p(z_n|\theta)] \\
&+ \sum_{n=1}^{N} E_q[\log p(w_n|z_n, \beta_{1:K})] + E_q[\log p(y|z_{1:N}, W, b, c, \sigma^2)] + H(q) \\
&+ \log p(w_{1:N}, y|\pi)
\end{aligned}$$

Notice that the ELBO is just a function of data $(w_{1:N}, y)$ and the variational distribution q. This bound is tight when q is the posterior of the latent variables, since maximizing the ELBO is equivalent to minimizing the Kullback-Leibler divergence between the posterior and q, (see Blei et al. (2017)). However, this posterior is computationally intractable so we approximate it.

Next, I take a stance on the family of the variational distribution and use mean-field variational inference to specify a fully factorized distribution:

$$q(\theta, z_{1:N}, \beta_{1:K}|\gamma, \phi_{1:N}, \lambda_{1:K}) = q(\theta|\gamma) \prod_{n=1}^{N} q(z_n|\phi_n) \prod_{k=1}^{K} q(\beta_k|\lambda_k). \tag{2}$$

Within this family, our goal is to find $(\gamma, \phi_{1:N}, \lambda_{1:K})$ that maximize the ELBO. Since the last term of the ELBO does not depend on the variational distribution our objective

function is:

$$\mathcal{L}(\gamma, \phi_{1:N}, \lambda_{1:K}) = E_q[\log p(\theta|\alpha)] + \sum_{k=1}^{K} E_q[\log p(\beta_k|\eta)] + \sum_{n=1}^{N} E_q[\log p(z_n|\theta)] + \sum_{n=1}^{N} E_q[\log p(w_n|z_n, \beta_{1:K})] + E_q[\log p(y|z_{1:N}, W, b, c, \sigma^2)] + H(q) \quad (3)$$

Detailed derivations of how to maximize the ELBO can be found in Appendix A.1 The optimization is done by block coordinate-ascent variational inference (CAVI) (see Bishop (2006) for an exposition), which implies iteratively maximizing $\mathcal{L}$ with respect to each variational parameter. In this maximization we have the additional constraint that each multinomial parameter vector ϕ_n should sum up to one. Optimization boils down to iteratively updating $\{\gamma, \phi_{1:N}, \lambda_{1:K}\}$ as:

$$\gamma^{new} \leftarrow \alpha + \sum_{n=1}^{N} \phi_n \quad (4)$$

$$\phi_n^{new} \propto \exp\left[E_q[\log \theta] + E_q[\log \beta_{w_n,\cdot}] + \frac{\tilde{y}b}{N\sigma^2} - \frac{1}{2\sigma^2 N^2}[2(b'\phi_{-n})b + b \cdot b]\right] \quad (5)$$

with $\tilde{y} := y - c'W$; $\phi_{-n} := \sum_{m \neq n}^{N} \phi_n$ and proportionality means to compute each entry and then normalize to one, and finally $\cdot$ stands for element-wise multiplication; details can be found in Appendix A.1.2. The variational parameters for each topic, λ_k, are updated as

$$\lambda_{k,v} \leftarrow \eta_v + \sum_{t=1}^{T} \sum_{n=1}^{N_t} \mathbb{I}\{w_{t,n} = v\} \phi^*_{t,n,k}$$

$$\text{or} \quad \lambda_k \leftarrow \eta + \sum_{t=1}^{T} \sum_{n=1}^{N_t} \phi^*_{t,n,k} \omega_{t,n} \quad (6)$$

with $\omega_{t,n}$ a one-hot representation of $w_{t,n}$ as a vector of size V. Using the update above for $\{\gamma, \phi_{1:N}, \lambda_{1:K}\}$ I can maximize the ELBO for each document in the corpus to obtain a corpus-level ELBO:

$$L(b, c, \sigma^2) = \sum_{t=1}^{T} E_{q_t^*}[\log p(\theta_t, z_{t,1:N_t}, \beta_{1:K}, w_{t,1:N_t}, y_t | W_t, \pi)] + H(q_t)$$
$$= \sum_{t=1}^{T} \mathcal{L}(\gamma_t^*, \phi_{t,1:N}^*, \lambda_{1:K}^* | b, c, \sigma^2), \tag{7}$$

where t indexes the sample data, that is the document-response pair for day t, and q_t^* stands for the document-specific variational distribution with corresponding maximizing $\{\gamma_t^*, \phi_{t,1:N}^*\}$ and $\{\lambda_{1:K}^*\}$. Notice that the corpus-level ELBO is a just a function of parameters from the supervising equation, this is because (α, η) are treated as hyperparameters selected by the econometrician.

2.4 Parameter Estimation

Next I look for parameters $\pi = \{b, c, \sigma^2; \alpha, \eta\}$ that maximize the corpus-level ELBO in (7); this is done via the VEM algorithm, that iteratively performs the E-Step and M-Step described below.

E-step: In the E-step given a set of parameters from the previous iteration, we approximate the likelihood of the data by maximizing each document-specific ELBO, and then summing them into $L(b, c, \sigma^2)$ as described in the previous section. In this step the goal is to find a collection $\{\gamma_t^*, \phi_{t,1:N}^*, \lambda_{1:K}^*\}$ that maximizes each document-specific ELBO using CAVI. This amounts to iteratively perform the updates in (4)-(6).

M-step: In the the M-step we maximize the corpus ELBO with respect to the sLDA parameters $\pi = \{c, b, \sigma^2; \alpha, \eta\}$. Detailed derivations for this step are shown in Appendix A.1.3 The updates in this step boil down to:

$$b^{new} \leftarrow \left(\sum_{t=1}^{T} E_{q^*}[\bar{Z}_t \bar{Z}_t'] \right)^{-1} \left(\sum_{t=1}^{T} E_{q^*}[\bar{Z}_t] \tilde{y}_t \right) \tag{8}$$

$$c^{new} \leftarrow \left(\sum_{t=1}^{T} W_t W_t' \right)^{-1} \left(\sum_{t=1}^{T} W_t (y_t - E_{q^*}[\bar{Z}_t]) b^{new} \right) \tag{9}$$

$$\sigma^{2,new} \leftarrow \frac{1}{T} \sum_{t=1}^{T} \left(y_t^2 - y_t E_{q^*}[\bar{Z}_t'] b^{new} - y_t W_t' c^{new} \right), \tag{10}$$

where q^* stands for the variational distribution with the optimal variational parameters from the previous E-step. As mentioned above, the topic membership hyper parameter, α, and the hyperparameter for topic term frequencies, η, are treated as a hyper parameter chosen by the econometrician. For example it is common to set $\alpha = (1/K) \cdot \mathbf{1}_K$ and $\eta = 0.1 \cdot \mathbf{1}_V$ guided by the suggestion in Griffiths and Steyvers (2004).

The VEM algorithm iteratively alternates these steps until convergence in the corpus-level ELBO in (7) is achieved. Algorithm 1 in Appendix B sketches the steps for estimating the modified sLDA model and obtaining parameter estimates $\hat{\pi} = \{\hat{b}, \hat{c}, \hat{\sigma}^2; \alpha, \eta\}$.[5]

2.5 Prediction

In sLDA given estimates for topics, $\hat{\beta}_{1:K}$, parameters $\hat{\pi}$ and a new document $w_{1:n}$ with covariates W, we can predict the response variable y associated with the new document. Using a quadratic loss function, the prediction simplifies to computing the conditional expected value:

$$\begin{aligned} E[y|w_{1:N}, W, \hat{\pi}, \hat{\beta}_{1:K}] &= E[E[y|w_{1:N}, z_{1:n}, \theta, W, \hat{\pi}, \hat{\beta}_{1:K}]] \\ &= \hat{b}' E[\bar{Z}|w_{1:N}, \theta, \hat{\pi}, \hat{\beta}_{1:K}] + \hat{c}' W \end{aligned}$$

To compute the last conditional expectation we resort again to variational inference and compute the posterior mean of $\bar{Z}$. The goal is to find a variational distribution $\tilde{q}(\theta, z_{1:N})$ over the latent variables that minimizes the KL divergence between this distribution and the posterior distribution $p(\theta, z_{1:n}|w_{1:n}, W, \hat{b}, \hat{c}, \hat{\sigma}^2, \hat{\beta}_{1:K})$; this is, again, equivalent to maximizing the corresponding ELBO. This is similar to the ELBO maximization in sLDA but the difference is that terms involving the response variable y will not appear since this variable is averaged out. The variational updates are the same as in an unsupervised LDA model for $(\theta, z_{1:N})$.

Using a fully factorized variational distribution, the updates are given by:

$$\begin{aligned} \tilde{\gamma}^{new} &\leftarrow \alpha + \sum_{n=1}^{N} \tilde{\phi}_n \\ \tilde{\phi}_n^{new} &\propto \exp\left[E_{\tilde{q}}[\log \theta] + \log \hat{\beta}_{w_n,\cdot}\right] \end{aligned}$$

[5]Code written in Julia language is available on request.

Once the variational parameters have been optimized, then we obtain a predicted response $\hat{y}$ as:

$$\hat{y} = \hat{b}' E[\bar{Z}|w_{1:N}, \hat{\pi}, \hat{\beta}_{1:K}] + \hat{c}' W \approx \hat{b}' E_{\tilde{q}}[\bar{Z}] + \hat{c}' W = \hat{b}' \left(\frac{1}{N} \sum_{n=1}^{N} \tilde{\phi}_n \right) + \hat{c}' W \tag{11}$$

Algorithm 2 in Appendix B summarizes the steps for predicting a response given a document and sLDA estimates.

3 A Simulation Example

When modeling how text data can be used as a covariate for a supervising variable of interest, the problem can divided into two stages: a first stage that extracts a low-dimensional feature from the unstructured text data and a second stage that assesses how this low-dimensional feature correlates with the response variable. For example, in our context of exchange rates, we could divide the problem into a first stage that uses LDA to extract topics from the news articles and a second stage that uses the estimated topic frequencies as regressors on a metric y of the exchange rate (for example the daily log return). In a more complex two stage example, Larsen and Thorsrud (2019) use LDA in the first stage followed by a Latent Dynamic Threshold model (LTM) (Nakajima and West (2013)) in the second stage to assess how news article text can be filtered into different indices over macroeconomic variables. In this section I argue, via a Monte Carlo simulation, that opting for single-stage estimation has efficiency gains over a two-stage approach; namely the Mean Square Errors (MSE) on the estimated parameters and the Mean Square Prediction Error (MSPE) over the test sets are smaller when using the sLDA versus the two stage approach of LDA followed by a regression.

3.1 Simulation Design

We start by describing how to generate a labeled corpus. We fix the dictionary size a $V = 200$, which means that words are just indices $w \in 1, 2, 3, ..., 200$. We also match these indices with the 200 most common words in English to facilitate comparison between the results. We consider $K = 5$ underlying topics and the topic membership hyperparameter α_0 is set to $\alpha_0 = (1/K) \cdot 1_K$ where 1_K is a vector of ones of size K. Topics are drawn according to $\beta_k^0 \sim Dir(\eta_0)$ with $\eta_0 = 0.1 \cdot 1_V$. Each is a vector of size $V = 200$ with entries

summing to one. We put these topics into a (V, K) matrix $\beta^0_{1:K}$. The supervised weights b were set to $b_0 = [2, 1, 0, -1, -2]'$. Next we generate $T = 1000$ document-response pairs following the blueprint described in the generative model in section 2:

(i) Draw how many words are in the document: $N_t \sim Unif(100, 300)$.

(ii) Draw topic membership $\theta_t \sim Dir(\alpha_0)$.

(iii) For each word in the sample $n = 1, ..., N_t$:

- Draw topic assignment: $z_n \sim Mult(\theta_t)$. Hence $z_n \in \{1, ..., K\}$.
- Draw a word index $w_n \sim Mult(\beta_{z_n})$. Hence $w_n \in \{1, ..., V\}$.

(iv) Rewrite z_n in one-out-K form and compute the empirical topic frequency $\bar{Z}_t \in \mathbb{R}^K$ for this document:

$$\bar{Z}_t = \frac{1}{N_t} \sum_{n=1}^{N_t} z_n$$

(v) Generate $\varepsilon_t \sim N(0, \sigma_0^2 = 1)$ and using $b_0 \in \mathbb{R}^K$ generate a response:

$$y_t = \bar{Z}_t' b_0 + \varepsilon_t$$

The previous steps generate a dataset of document-response pairs $(w_{t,1:N_t}, y_t)$ with $t = 1, ..., 1000$. These documents-response pairs will be split into a training set formed by 900 pairs and a test set composed of the 100 remaining pairs. In this simulation we have assumed no complementary covariates W_t determining y_t along with $w_{t,1:N_t}$ as we are mainly interested in contrasting the estimates related to text in single versus two-stage estimation procedures.

3.2 Simulation Results

We generated a hundred datasets, each containing $T = 1000$ document-response pairs. Using each of these datasets, we perform single-stage and two-stage estimation on the training set and prediction over the test set. Single-stage estimation refers to sLDA using the VEM algorithm whereas two-stage estimation refers to running LDA with the VEM algorithm followed by regressing the dependent variable y on the empirical topic frequencies per document extracted from LDA. Details about the execution of the VEM algorithms are shown in Appendix C.

Table 1 shows the bias, variance and MSE for this repeated exercise for the supervised parameters b and σ^2. These metrics for single-stage estimation present smaller numbers and larger relative efficiency, defined as the MSE from two-stage estimates divided the MSE from single stage estimates, for all the parameters. In particular for σ^2 the MSE from two-stage is 5.8 times larger than in single-stage estimation. These results suggest that even though topics are generated before responses and that both LDA and sLDA model topics the same way, having supervised information improves the topic extraction process.

During prediction we also observe efficiency gains. Figure 2 shows distributions of Mean Square Prediction Errors (MSPE) for both strategies over a test sample. Specifically, each MSPE is computed over a test set of 100 documents, for which the response variable y was predicted using estimates from a training set of 900 document-response pairs. This figure shows that MSPE from single-stage forecasts generally take smaller values than the two-stage predictions; this is consistent with efficiency gains in favor of the single-stage strategy during estimation.

4 Data Sources

4.1 News Articles

For this paper more than 56,000 news articles were collected from La Nación (LN) (Grupo-Nación-LN (2025)), the main Costa Rican newspaper, and from El Financiero (EF) (Grupo-Nación-EF (2025)), the main Costa Rican business newspaper. Article dates range from January 2015 to March 2021, with almost all articles written in Spanish. In this period of 2,271 days there is a median of 21 news articles, 15,193 words and 13 newspaper sections per day. For every article, the section, the author (if available), publishing date up to milliseconds, subject tags, header, subheader, and the main text of the article were recorded.

It is important to mention that despite the fact that this database contains only news articles from a Costa Rican newspaper, news about the United States economy is closely followed. As Costa Rica is a small open economy, the US is its main trade partner, and the USD is the vehicle currency for international trade, the CRC/USD exchange rate is a macroeconomic variable closely followed by the media. Hence news about developments in the United States will typically be covered in the newspapers.

Table 1: Bias, MSE and Efficiency for Supervised Parameters

Parameter	Estimation	Bias	Variance	MSE	RE
σ^2	Two-Stage	0.111	0.023	0.035	5.835
	Single-Stage	0.038	0.005	0.006	
b_1	Two-Stage	-0.376	0.074	0.214	3.557
	Single-Stage	-0.185	0.026	0.060	
b_2	Two-Stage	-0.250	0.097	0.158	2.276
	Single-Stage	-0.134	0.052	0.069	
b_3	Two-Stage	0.021	0.058	0.058	1.815
	Single-Stage	0.033	0.031	0.032	
b_4	Two-Stage	0.163	0.041	0.067	1.652
	Single-Stage	0.090	0.033	0.041	
b_5	Two-Stage	0.337	0.066	0.179	2.585
	Single-Stage	0.189	0.034	0.069	

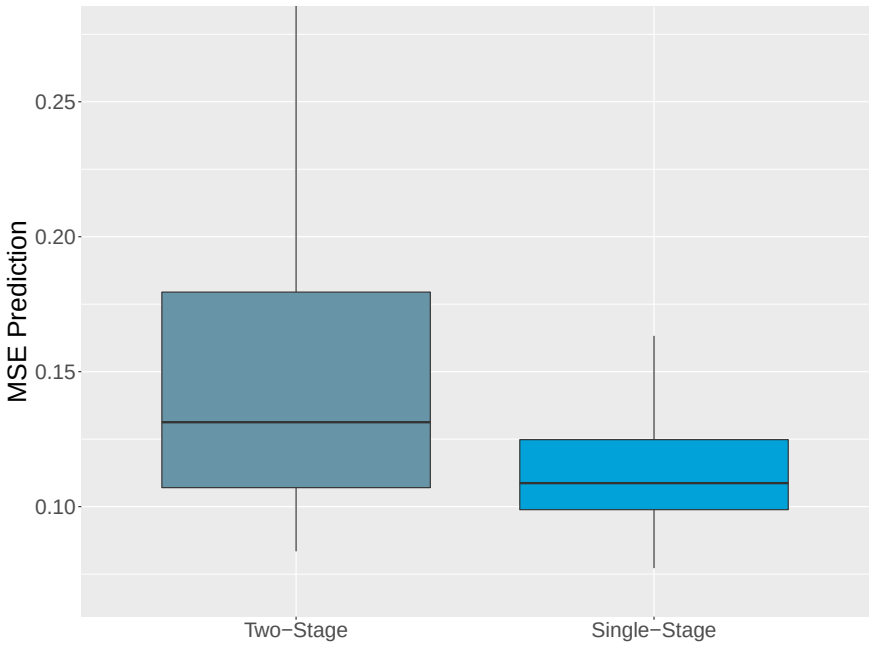
Notes: Results for 100 Monte Carlo replications. Each generated dataset has $K = 5$ underlying topics for a dictionary composed of the top $V = 200$ most common nouns in English. Topics were drawn as $\beta_k \sim Dir(0.1 \cdot \mathbf{1}_V)$ and each simulated training set is composed of $T = 900$ documents, each with a number of words drawn as $N_t \sim Unif(200, 300)$. RE is the efficiency of Single-Stage estimation relative to Two-Stage estimation: RE = MSE(Two-Stage)/MSE(Single-Stage).

4.2 Text Preprocessing

Before the text in each news article can be used as an input to estimate a topic model it needs to be preprocessed. This is done to clean the text and reduce its dimension to facilitate the extraction of semantic content. This is done in several steps: tokenization, filtering out special terms, stemming, removing stop words, and a final frequency filter.

Tokenization requires transforming a long string of text into a vector with just one word in each entry; for example a phrase comprised of 10 words would be tokenized into a vector of size 10 with each word of the phrase occupying one entry. In a second step we filter out special terms; this removes numbers, punctuation, words not in Spanish, and special characters. The next step is to stem each word, this is done with Porter (2001)'s `snowball` stemmer. This step reduces the dimension of the dictionary by *collaps-*

Figure 2: MSE Prediction



Notes: Boxplots for the Mean Square Prediction Error (MSPE) over 100 Monte Carlo replications. Each MSPE is computed over a test set of 100 documents, for which the response variable y was predicted using estimates from a training set of 900 document-response pairs.

ing different words with the same underlying concept[6] into a single stem, for example the words: *bank, banks, banking* will be stemmed into *bank*.

Once words have been stemmed, we further reduce the dimension of the dictionary by removing stop words, which are common words such as connectors, that are so generic that removing them does not change the semantic content of a phrase but reduces its dimension. This step is done by filtering out words from a list of terms considered to be stop words. Spanish stop words come from `Python`'s package `stopwords-ISO` (Diaz (2020)). A final frequency filter is applied based on the *tf-idf* measure. This metric ranks each term in the dictionary according to the multiplication of the term frequency (tf) and the inverse document frequency (idf)[7], and will assign a low ranking to terms that are *too*

[6]For instance, in a language like Spanish where verbs are conjugated differently according to each grammatical tense and subject, stemming simplifies the dictionary of the corpus while preserving the main concepts.

[7]For each term v, tf is the number of times term v appears in document d divided by the number of terms in document d. idf is defined as the logarithm of the number of documents in the corpus minus the logarithm of the number of documents that contain term v. Finally *tf-idf* is given by $tf \times idf$.

common or *too rare* in the corpus, and hence they do not provide representative semantic content to distinguish topics. Using the *tf-idf* metric we only keep the top 50,000 stems in the dictionary.

Figure 3a shows a wordcloud with the size of the word representing how frequently that word appears in the CR News database; note that many economic terms can be identified. Figure 3b shows an example of a news paragraph before and after preprocessing it using the steps described above. After preprocessing the CR News database of articles, we have over 56,000 news articles comprising more than 15 million stems and 50,000 unique stems.

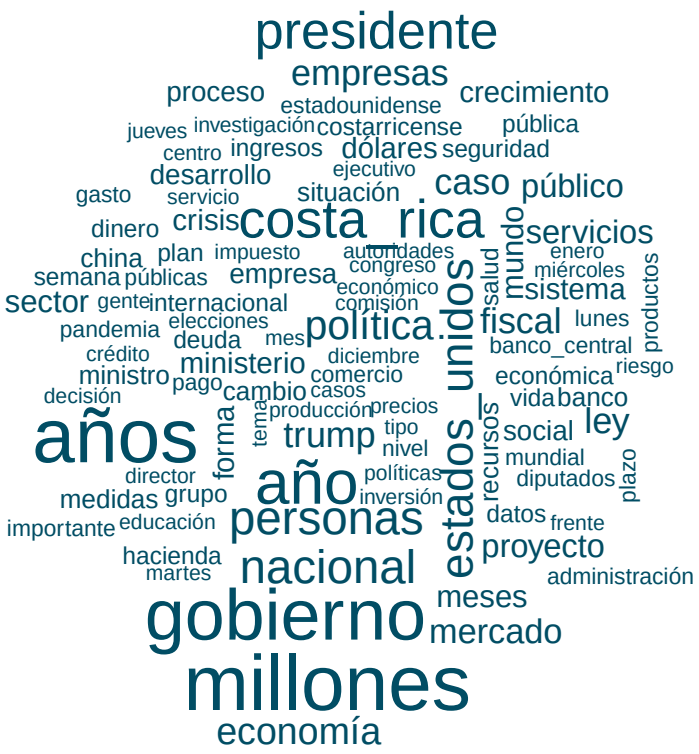
4.3 Costa Rica's Monex Market

The Monex market is a digital platform created and operated by the Central Bank of Costa Rica (BCCR) and it constitutes the official wholesale market where FX dealers exchange Costa Rican colónes (CRC), the local currency, against United States dollars (USD). As the United States is Costa Rica's main trade partner, the CRC/USD exchange rate receives almost exclusive attention in FX matters from the general public and the monetary authorities in Costa Rica. Moreover, the role of the USD as a vehicle currency for international trade is exacerbated in Costa Rica: almost all exports and imports in Costa Rica are invoiced in USD dollars despite that approximate 40% of the exports are destined to the United States and around 42% of the imports come from the United States (Boz et al. 2020).

The nature and design of the Monex market in Costa Rica is inherently related to the FX regime operating in the country. In 2006 the Monex market was created to play a fundamental role in the crawling band regime which ran 2006 to 2015. This paper concentrates on the managed floated regime that began in February 2015, where in principle dealers' incentives are not explicitly tied to the level of the exchange rate, as may happen in a crawling band regime when the exchange rate hits the bands.

The Monex market can be considered a second tier of the Costa Rican FX market; the first tier is constituted by over-the-counter windows (OTC Windows) between the general public and financial institutions. The second tier, reserved for FX interdealer trades, accounts for around 10% of all USD trades registered in the Costa Rican economy while the rest is dispersed through the OTC windows of financial institutions. As Cubero et al. (2019) remarks, approximately 90% of the transactions occurring through the OTC windows with the general public are for amounts less than 1000 USD, which is the

Figure 3: Costa Rica News Database



(a) Word cloud for most frequent words in the news articles

Original string

```
"De los nuevos recursos que prestó el sector financiero costarricense en el p
rimer trimestre de este año, el Gobierno se llevó el 60%, 14 puntos porcentua
les más que en el primer trimestre del 2015. Lo que se llevó de más el Gobier
no se había prestado el año pasado a personas o empresas privadas. La partici
pación privada pasó de 52,6% a 35,3%, según los datos que publica el Banco Ce
ntral en su página web. La mayor cantidad de recursos que obtuvo el Gobierno
son para financiar el gasto que no cubre con los ingresos de los impuestos. P
or otro lado, el crédito al sector privado se dirige a consumo y producción."
```

After Preprocessing

```
[“de”, “los”, “nuevos”, “recurs”, “que”, “prest”, “el”, “sector”, “financi
er”, “costarricens”, “en”, “el”, “primer”, “trimestr”, “de”, “este”, “año”
, “el”, “gobierno”, “se”, “llevó”, “el”, “60%”, “14”, “punt”, “porcentual
es”, “mas”, “que”, “en”, “el”, “primer”, “trimestr”, “del”, “2015”, “lo”, “que
”, “se”, “llevó”, “de”, “mas”, “el”, “gobiern”, “se”, “había”, “prest”, “ado”,
“el”, “año”, “pasado”, “a”, “person”, “o”, “empres”, “privadas”, “la”, “pa
rticipacion”, “privada”, “paso”, “de”, “52,6%”, “a”, “35,3%”, “segun”, “los”,
“dat”, “que”, “public”, “el”, “banc”, “central”, “en”, “su”, “pagin”, “w
eb”, “la”, “mayor”, “cantid”, “de”, “recurs”, “que”, “obtuv”, “el”, “gob
iern”, “son”, “para”, “financi”, “el”, “gast”, “que”, “no”, “cubr”, “con
”, “los”, “ingres”, “de”, “los”, “impuest”, “por”, “otro”, “lado”, “el”,
“credit”, “al”, “sector”, “privado”, “se”, “dirige”, “a”, “consum”, “y”, “p
roduccion”]
```

(b) Example of preprocessing outcome

Notes: Panel (a) a word cloud with the most frequent terms in the CR News Database; the size of the word represents its relative frequency in the corpus. Panel (b) shows the before and after text preprocessing for a news article extract.

minimum trade amount of the Monex market. This suggests that the Monex market is reserved for relatively specialized trade and therefore its turnover share.

Despite the fact that the Monex market is reserved for specialized trade, it has a crucial role in the price discovery process of the CRC/USD exchange rate. As Osler et al. (2011) hypothesizes, this could be though a 3-stage process: in the first stage the FX dealers extract information from their OTC transactions with informed agents, in the second stage the FX dealers will trade on this information in the interdealer market, in our case the Monex market, where this new information would be incorporated into the bid/ask exchange rates. Finally, after such private information is progressively shared and compounded into the equilibrium exchange rate of the interdealer market, in the third stage all FX dealers would react to this new information by updating their quotes in OTC windows. This suggests that the Monex market is a critical place where information is incorporated into the equilibrium exchange rate. Furthermore the Monex does not only signal private information coming to a specific customer segment served by participating FX dealers, but as it is also the market where BCCR's interventions take place, it can potentially signal current or future information on fundamental variables.

Participants in the Monex market are mainly financial institutions authorized by the BCCR to act as FX dealers. In order to be authorized to participate in the Monex market, dealers should comply with the ROOC (*Reglamento para las Operaciones Cambiarias de Contado*) regulation promulgated by the BCCR, that specifies requirements on behalf of FX dealers such as appropriate accounting systems, reports on FX activity to the BCCR and compliance with risk measures related to FX dealership (BCCR (2017a)).

Furthermore, due to regulation on FX dealers that puts prudential limits on their open net FX positions, both at absolute levels and in daily variation rates, any excess demand or supply of USD that FX dealers receive through their OTC windows outside these limits is transferred to the Monex market. Thus, in addition to being a place for price discovery, there is an incentive to participate in the Monex market, since it also represents a valuable tool for financial institutions to comply with FX regulation.

The number of participants in the Monex market varies from year to year depending on how FX dealers weight the costs and benefits of keeping their authorization from the BCCR. In the managed float period, 38 FX dealers can be identified as regular participants in the Monex market: the BCCR, the CD platform[8], 4 state-own banks, 12 private

[8]A "public" dealer for the general public introduced in 2009 where any individual or firm with registered accounts on the CD platform can put their offers in the Monex Market as long as the currencies obtained are not used for FX dealership (BCCR (2010)).

banks, 2 credit unions, 5 cooperatives, 2 currency exchange offices, 8 brokerage firms, and 3 financial firms. If the general public that participates through the CD platform is disaggregated, then number of participants during the managed float period, is on average 77, with a minimum of 47 and a maximum of 137 participants.

The Monex market trades an average of 15.3 million USD daily with a median of 161 transactions[9]. This paper makes use of access to proprietary data from the Monex Market, including every transaction made in the Monex market from February 2nd 2015 until February 26th 2021, with more than 255,000 observations on 1520 trading days. Information about every transaction records: name of both parties involved, the amount of USD traded, the corresponding exchange rate, whether the trade is buyer or seller initiated, and a timestamp up to milliseconds.

5 Exchange Rate Supervised Topics and FX News index

In this section we use text data from news articles to augment a standard model from the FX literature. We work with an specification that relates the daily exchange rate return and the order flow from currency transactions. This market-microstructure approach has the advantage that it can be estimated at relative higher frequencies than traditional fundamentals-based models, and hence shed more light about short term dynamics of the exchange rate.

Viewing FX activity using market microstructure tools, in particular the order flow, allows us to articulate a story based on the informational content of exchange rates and market frictions in order to map changes in fundamentals into new equilibria. Evidence in favor of an order flow with explanatory power even in short frequencies in the Monex market suggests an informational imbalance on the part of the FX dealers. In qualifying this informational imbalance, a news index with explanatory power is relevant.

We consider a classic order flow regression inspired by Evans and Lyons (2002), but augmented with the text data, given by the following supervising equation:

$$\begin{aligned} y_t &= b'\bar{Z}_t + c'W_t + \varepsilon_t \\ &= b'\bar{Z}_t + c' \underbrace{\begin{bmatrix} \mathcal{O}_t & \Delta rd_t & N_t & m_t \end{bmatrix}'}_{W_t} + \varepsilon_t \end{aligned} \tag{12}$$

[9]These statistics are computed using transactions data from the start of the managed float period, February 2nd 2015, until February 26th 2021, for a total of 1520 trading days.

where y_t is the daily log-return $y_t := \log(ER_t) - \log(ER_{t-1})$; and ER_t is the amount-weighted average exchange rate (expressed as CRC per USD) among all the trades in the Monex market for day t. $\bar{Z}_t$ are the empirical topic frequencies that follow the sLDA generative model in section 2.1 using news articles available at the beginning of day t; that is, news articles produced in $t-1$ and up to day t (excluding day-t news)[10]. The covariates in W_t include: $\mathcal{O}_t$ as the order flow, the difference between the number of buyer-initiated and seller-initiated trades. $\Delta rd_t = \Delta(i_t^* - i_t)$ is the first difference in the interest rate differential using benchmark interest rates for Costa Rica and the United States. In particular i_t stands for the Costa Rican TBP (*Tasa básica pasiva*, an index of passive short-term interest rates paid in Costa Rica) and i_t^* for the U.S. prime rate, both at daily level. N_t number of daily transactions. Finally, m_t is net amount traded (total purchases minus total sales in USD).

In our application, the estimation of sLDA will yield an output composed of (i) exchange rate supervised topics, (ii) topic memberships that can be interpreted as time series on the daily attention each topic receives in the newspapers and (iii) parameters estimates from the supervising equation. Looking at these three outputs together will be key to interpret the supervised extracted topics.

5.1 Base Results

The base results are based on 31,000 news articles from the CR News database ranging from January 2018 to December 2020. The daily number of terms ranges from $3,000$ to $30,000$ with an average $12,000$ terms and the corpus is composed of 8.6 million terms after preprocessing. Pairing these text data with FX data produces $T = 750$ document-response triplets that will be used in this base estimation. The algorithm is run with $K = 25$ topics[11], a dictionary size of $V = 10,000$ and hyperparameters set to $\alpha = (1/K) \cdot \mathbf{1}_K$ and $\eta = 0.1 \cdot \mathbf{1}_V$.

Figure 4 shows word clouds, attention series and words associate with the most frequent terms for two of the extracted supervised topics. In the sLDA model each topic is represented as a vector of dimension $V = 10,000$, where each entry represents the

[10]Since the Monex trades on weekdays, $\{w_{t,1:N_t}\}$ will be one or three days worth of news articles. For example, if date t happens to be a Monday, w_t will include news articles from the previous Friday, Saturday and Sunday, that is news available at the beginning of Monday. Notice that in concatenating articles into a single article for day t we are exploiting the mixed membership feature in the LDA part of sLDA.

[11]Other model specifications with 30, 35 and 40 topics were estimated but $K = 25$ yielded the best balance regression fit and interpretable topics.

probability of sampling the term in that entry according to the topic. A common way to represent topics is to use word clouds with the most probable terms in each topic; the size of the word represents its relative term frequency according to the topic.

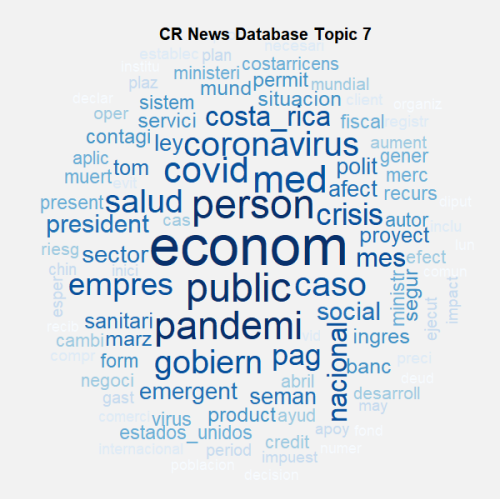
Panels (a) and (c) from Figure 4 show word clouds with the most frequent terms for the Exchange Rate Supervised Topics 7 and 20. Similarly, Panels (b) and (c) show daily series of topic memberships for these topics; such time series can be interpreted as the attention level for these topics in the newspapers. Additionally a set of the most probable words corresponding to the most probable terms (word stems) in each topic are listed. In conjunction, the word cloud of most probable terms, the attention series and the list of most probable words suggest that topic 20 refers to the 2018 Tax Reform Discussion in Costa Rica, whereas Topic 7 refers to the Covid-19 Pandemic. Looking at the attention series validates this interpretation: we see that in the case of Topic 20, its attention peaks around the start of strikes in Costa Rica triggered by a tax reform debated and approved in 2018. For topic 7, the most probable words and the time when its attention peaks supports labeling it as the Covid-19 Pandemic topic.

Notice that the labeling is innocuous for estimation; what identifies topics are the corresponding $(\beta; \theta)$ mixtures; labels are just a convenient way to refer to topics when presenting results. A complete list of labeled topics extracted from the sLDA estimation with their top most probable stems in Spanish and their English translations can be found in Appendix D.

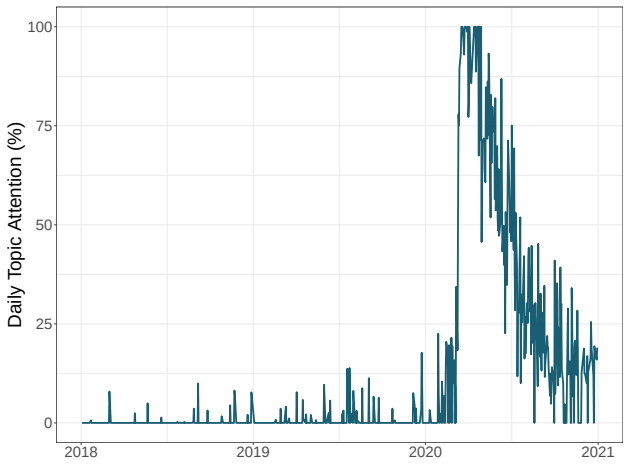
Table 2 shows sLDA estimates for the supervising equation in (12). Results for a traditional order flow regression without text data are shown in column (1). Results for the sLDA model with the order flow regression supervising equation are shown in column (2). Supervising parameters for selected extracted topics in column (2) show the average effects in basis points for the daily log return in the exchange rate if the topics receives a 100% of the attention in the newspaper. Positive (negative) values of these coefficients represent a partial effect consistent with short-term depreciations (appreciation) of the CRC/USD exchange rate.

Figure 4: Selected Exchange Rate Supervised Topics

Topic 7-Covid-19 Pandemic



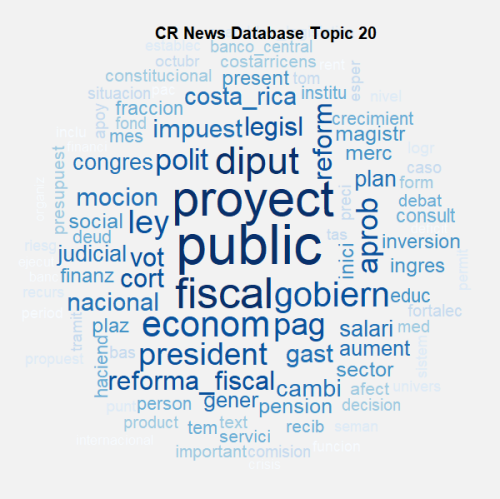
(a) Word cloud Topic 7



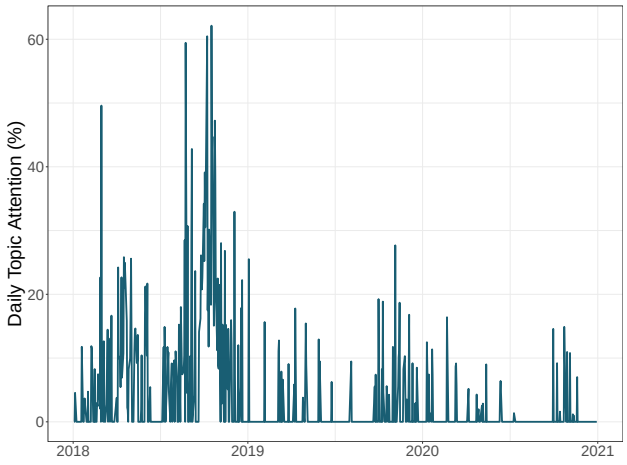
(b) Attention Series Topic 7

Topic 7 Top words: economy, medicine, pandemic, person, covid, coronavirus, public, firm, government, health, crisis, national, emergency, month.

Topic 20 - 2018 Tax Reform



(c) Word cloud Topic 20



(d) Attention Series Topic 20

Topic 20 Top words: public, project, fiscal_reform, government, president, national, law, services, social, judicial, case, legislature.

Notes: Panels (a) and (c) show word clouds with the most frequent terms for the Exchange Rate Supervised Topics 2 and 20; the size of the term represents its relative frequency according to each topic. Panels (b) and (c) show daily topic memberships for Topic 2 and 20; these time series can be interpreted as the attention level for these topics in the newspapers. Additionally a set of the most probable words corresponding to the most probable terms (word stems) in each topic is listed. In conjuction, the word cloud, the attention series and the list of most probable words suggest how to interpret each topic.

In the case of the selected topics we observe the expected coefficient sign. For example, topic 20 and 25 refer to tax reform that was debated and approved in Costa Rica in 2018, respectively. The discussion of tax reform was put forward to correct the rapid growth of the government debt and restore investors' confidence that the Costa Rican government would be able to meet its debt payments. Topic 25 refers to the general discussion and strikes that were triggered during the debate of the tax reform. This topic has a positive coefficient which supports the interpretation that the strike jeopardized the likelihood of approval of the tax reform and clouded investors' trust on the long term sustainability of the CRC currency. Investors in face of such uncertainty will have a relative greater taste for holding USD and the exchange rate will depreciate. Topic 25 refers to the coverage in the newspaper of the approval of tax reform towards the end of 2018. This topic has a negative coefficient which is consistent with the interpretation that the approval of the tax reform eased concerns about the fiscal sustainability of the Costa Rican Government which in turn will return confidence in the CRC currency and an appreciation followed.

Table 2 also shows that the inclusion of text data improves the fit of the order flow regression. R^2 almost double compared with an specification that does not include text data from news articles. This pattern is also present in the adjusted-R^2 of the model which improves the fit in more than 10% with respect to an order flow that does not controls for the role of public information in determining the short-term dynamics of the exchange rate.

Once the sLDA have been estimated it is possible to summarize the supervised informational signal from news articles into an FX news index that can be used in further anaylis and even as a covariate in other model specifications of the FX literature. With sLDA estimates $\hat{\pi} = \{\hat{b}, \hat{c}, \hat{\sigma}^2; \alpha, \eta\}$ and $\{\hat{\gamma}_t, \hat{\phi}_{t,1:N_t}, \hat{\beta}_{1:K}\}$; we can construct a FX news index as:

$$NI_t := \hat{b}'\hat{Z}_t = \hat{b}' \left(\frac{1}{N_t} \sum_{i=1}^{N_t} \hat{\phi}_{t,i} \right), \tag{13}$$

notice that we estimate $\bar{Z}_t = \frac{1}{N_t} \sum_{i=1}^{N_t} z_{t,i}$ with the expected value under the estimated variational distribution $E_{\hat{q}_t}[\bar{Z}] = \frac{1}{N_t} \sum_{i=1}^{N_t} \hat{\phi}_{t,i}$. Positive (negative) values of the index can be interpreted as a positive correlation for a depreciation (appreciation) for the CRC/USD exchange rate (larger daily return for holding USD) for period t. The larger the magnitude of the FX news index, the larger the correlation for a depreciation or an

Table 2: Supervising Equation Parameter Estimates

	Dependent Variable: Daily Log Return ER_t (b.p.)	
	Model without Text	Model with Text
Δrd_t (b.p)	0.133	0.140
	(0.091)	(0.094)
Order Flow	0.277	0.248
	(0.039)	(0.028)
N. Transactions	0.019	-0.070
	(0.007)	(0.021)
Selected Topics (b.p. for 100% attention)		
Topic 7: Covid 19 Pandemic		15.165
		(5.077)
Topic 20: 2018 Tax Reform Discussion		25.904
		(10.325)
Topic 25: 2018 Tax Reform Approval		-15.558
		(12.739)
Mean Daily Log Return	1.056	1.056
σ	23.051	21.213
Number of Topics	-	25
Observations	750	750
R^2	0.159	0.288
R^2-Adjusted	0.156	0.258
Aikaike Information Criterion (x10^6)	291.989	247.020
Bayesian Information Criterion (x10^6)	291.989	247.020

Notes: Results for a traditional order flow regression are shown in the column labelled as "Model without text". Results for an order flow augmented with text data from newspapers and estimated with sLDA are shown in the column "Model with Text". The sLDA model is estimated with $K = 25$ underlying topics and a dictionary size of $V = 10,000$. Supervising parameters for selected extracted topics are shown in column (2); these coefficients show the average effects in basis point for the daily log return in the exchange rate if the topics receives a 100% of the attention in the newspaper. Standard errors are compute as $\widehat{var}(\hat{\beta}) = T^{-1}(T^{-1}\sum_{t=1}^{T} x_t x_t')^{-1}(T^{-1}\sum_{t=1}^{T} \hat{u}_t^2 x_t x_t')(T^{-1}\sum_{t=1}^{T} x_t x_t')^{-1}$ where x_t stands for the vector of regressors included in the regression model and $\hat{u}_t$ the corresponding residuals.

appreciation on the CRC/USD exchange rate.

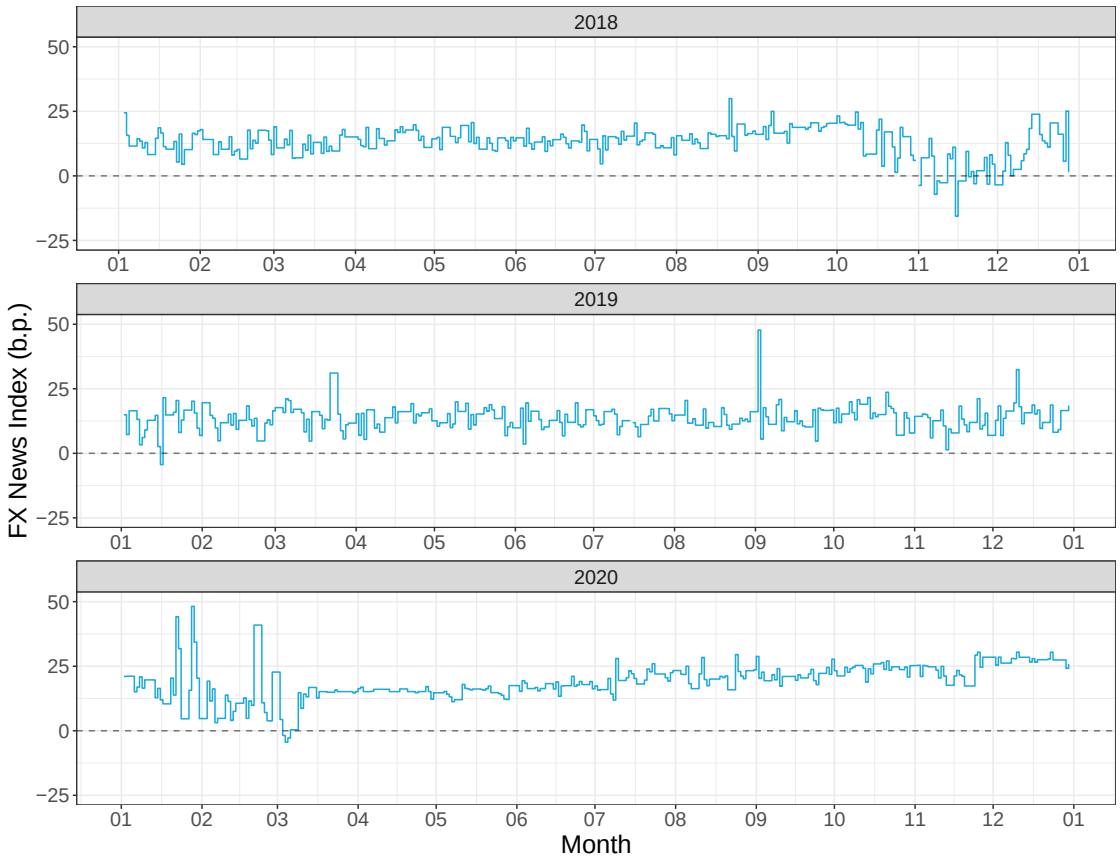
Figure 5 shows the FX news index constructed from the sLDA estimation. Two dynamics are worth remarking from the evolution of the index. The first one is the appreciation of the CRC towards the end of 2018, as mentioned above it is likely related to the approval of the tax reform and a restore confidence in the currency. The second is the sustained depreciation trend for most of 2020 and that started around March 2020. This depreciation trend is consistent with the interpretation that investors, in times of generalized global economic uncertainty, substitute away currencies from emerging market economies, in this case the CRC to hold hard currencies like the USD. In this case such generalized global economic uncertainty was triggered by the onset of the Covid-19 pandemic.

5.2 Prediction

We can also assess the predictive power that adding text data from news articles has on an order flow regression. In this out-of-sample prediction exercise we use base the estimates from models presented in Table 2 to forecast the daily log return of the CRC/USD exchange rate over period of two months. Specifically we used observed data from January and February 2021, in this period we compare the out-of-sample prediction and order flow regression will make with the forecast that the augmented model with text data from news articles will make. Given that we observe the realized daily return for the exchange rate we can compare prediction errors under both specifications.

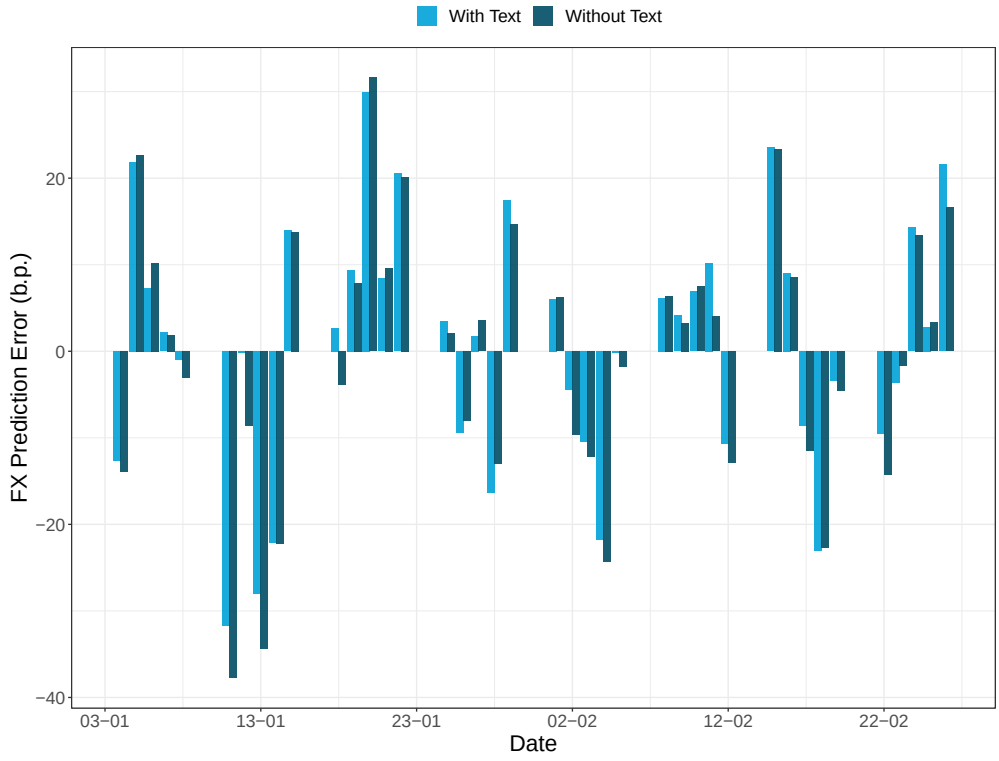
Results for the 40 trading sessions contained in this two-month period are shown in Figure 6. The MSPE for the model with no text is 233.99 whereas MSPE the sLDA model is 208.47. These numbers yield a relative efficiency of prediction of 1.12, that is prediction errors of a model without text are 12% higher than the sLDA model. This suggest that incorporating a control for public information in the form of supervised extracted topics from news articles improves the prediction power of an order flow regression in short-term frequencies.

Figure 5: FX News Index for the CRC/USD exchange rate



Notes: FX News Index for the CRC/USD exchange rate constructed from the sLDA estimation. It corresponds to the partial effect of news data in the supervising equation of sLDA: $NI_t := \hat{b}'\hat{Z}_t = \hat{b}'\left(\frac{1}{N_t}\sum_{i=1}^{N_t}\hat{\phi}_{t,i}\right)$. Positive values of the index can be interpreted as a positive correlation for a depreciation on the CRC/USD exchange rate (larger daily return for holding USD) for period t. Analogously negative values can be interpreted as a correlation for an appreciation of the CRC/USD exchange rate.

Figure 6: Prediction Errors for the Exchange Rate



Notes: Out-of-sample prediction errors from the models presented in Table 2. The label 'with text' refers for the sLDA model that uses text data from news articles and the label 'without text' refers to a standard order flow regression that does not uses text data. The out-of-sample period comprises 40 trading sessions in the months of January and February 2022. The MSPE for the model with no text is 233.99 whereas MSPE the sLDA model is 208.47. These numbers yield a relative efficiency of prediction of 1.12, that is prediction errors of a model without text are 12% higher than the sLDA model.

6 Conclusions

This paper shows how to use a hybrid of unsupervised and supervised learning models to go from text from news articles to an FX news index that can be used to enhance traditional models from the FX literature. We relied on Supervised Latent Dirichlet Allocation (sLDA) which combines information about a supervising variable with topic extraction over a corpus of text in a single-stage estimation.

Via a Monte Carlo simulation we document the efficiency gains of single-stage strategies like sLDA over two-stage estimation approaches that perform LDA follow by a regression. The message from this simulation study is threefold: (i) efficiency measured by the ratio of MSE from single-stage to MSE two-stage estimates of supervised parameters are the largest; this is consistent with an explicit acknowledge in sLDA of the connection between text and response. (ii) Even though topics are generated before responses and both LDA and sLDA model topics the same way, having supervised information improves the topic extraction process. This is reflected in the estimation of word probabilities. (iii) MSPE are also generally smaller when using a single-stage approach; the relative more efficient estimation of sLDA also improves prediction.

The empirical application is centered around the Monex Market, the main Costa Rican platform for FX trade and accordingly news articles are gathered from the main Costa Rican newspapers. Using the CRC/USD exchange rate we estimated an sLDA model with a supervising equation form the market-microstructure approach to exchange rates. This equation relates the the daily exchange rate return, the order flow from currency transactions, other control covariates and it is augmented to include text data in the form of topic membership series; all estimated in the single-stage procedure of the modified sLDA used in this paper. Estimates from sLDA show that the inclusion of text data improves the fit of the order flow regression. R^2 almost double compared with an specification that does not include text data from news articles. Adjusted-R^2 of the sLDA model also shows an improved the fit, with a more than 10% increase with respect to an order flow that does not controls for the role of public information in determining the short-term dynamics of the exchange rate.

The extracted supervised topics show high degree of interpretability. Most of them can be related to events that can be interpreted in connection with exchange rate dynamics. The inclusion of external information during the topic extraction process facilitates a parsimonious text structure. This is reflected in a relative lower number of topics, compared with a unsupervised topic extraction, as our focus is to extract topics that

matter for the supervising variable, in this case the exchange rate. The supervised informational signal from news articles can be summarized into an FX news index. We observed that this index correlates with depreciation and appreciation dynamics that can be interpreted in the light of macroeconomic events.

When performing analysis centered around prediction we presented evidence that in an out-of-sample prediction exercise, the forecast errors of a traditional order flow regression without text are 12% higher than the sLDA model. This suggest that incorporating a control for public information in the form of supervised extracted topics from news articles improves the prediction power of an order flow regression in short-term frequencies.

Future research. There are extensions to both data and methodology which could benefit from future research. In terms of expanding the data, next step is to include news articles from US newspapers; they would contain information about the US economy that could influence the demand and supply for USD in the Monex Market. The main challenge with including US newspapers is to ensure topic cohesiveness when the corpus is written in more than one language.

Another data expansion is to consider a panel of countries and currencies, for example when considering the USD vs the pound sterling (GBP), the USD vs the Mexican peso (MXN) and the USD vs the Canadian dollar (CAD), all fit the methodology presented in this paper: exchange rates in flotation for which news articles can be found in English or Spanish. This collection of exchange rates also receive more attention given that these currencies represent a larger share of international trade.

In term of methodology, there a research opportunity is to consider time-varying parameters in sLDA, specifically the supervised parameters (a time varying version of LDA already exists in Blei and Lafferty (2006)). A first step would be to gather evidence that time-varying sLDA improves empirical applications, for doing so can perform static sLDA with estimation over a rolling window. If results provide evidence of dynamic relevance of topics to explain the exchange rate at a daily frequency, it would be interesting to pursue a time-varying sLDA.

Another methodology aspect that requires further research is to compare the estimation of sLDA using variational inference that approximates the likelihood versus using exact inference in the form of a MCMC strategy. This can be done by assuming priors for the supervise parameters and writing out a corresponding Gibbs sampling strategy; then a Monte Carlo simulation contrasting both approaches can be designed.

References

Andrzejewski, D., Zhu, X. and Craven, M. (2009), Incorporating domain knowledge into topic modeling via Dirichlet forest priors, *in* 'Proceedings of the 26th Annual International Conference on Machine Learning', pp. 25–32.

Bacchetta, P. and Van Wincoop, E. (2006), 'Can information heterogeneity explain the exchange rate determination puzzle?', *American Economic Review* **96**, 552–576.

BCCR (2010), 'Mercado cambiario en costa rica y pasivos con costo en colones del banco central de costa rica, octubre 2006-marzo 2010', *San José: Banco Central de Costa Rica* . https://bit.ly/2Dx6jbl, Consulted on 9-25-2018.

BCCR (2017a), 'Reglamento de operaciones cambiarias de contado', *San José: Banco Central de Costa Rica* .

Beaudry, P. and Portier, F. (2014), 'News-driven business cycles: Insights and challenges', *Journal of Economic Literature* **52**, 993–1074.

Bishop, C. (2006), *Pattern Recognition and Machine Learning*, Springer.

Blei, D. M., Griffiths, T., Jordan, M. I. and Tenenbaum, J. (2003), 'Hierarchical topic models and the nested chinese restaurant process', *Advances in Neural Information Processing Systems* **16**.

Blei, D. M., Kucukelbir, A. and McAuliffe, J. D. (2017), 'Variational inference: A review for statisticians', *Journal of the American Statistical Association* **112**, 859–877.

Blei, D. M. and Lafferty, J. D. (2006), Dynamic topic models, *in* 'Proceedings of the 23rd International Conference on Machine Learning', pp. 113–120.

Blei, D. M. and McAuliffe, J. D. (2008), Supervised topic models, *in* 'Advances in Neural Information Processing Systems', Vol. 20, pp. 121–128.

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

Boz, E., Casas, C., Georgiadis, G., Gopinath, G., Le Mezo, H., Mehl, A. and Nguyen, T. (2020), 'Patterns in invoicing currency in global trade', *IMF Working Papers* .

Bybee, L., Kelly, B. T., Manela, A. and Xiu, D. (2020), 'The structure of economic news'. National Bureau of Economic Research Working paper. Available at https://www.nber.org/papers/w26648.

Cubero, R., Lankester, V. and Munoz, E. (2019), 'Chapter 10 Costa Rica: Learning to float', pp. 155–194. in *Foreign Exchange Interventions in Inflation Targeters in Latin America* edited by M. Chamon, D. Hofman, N. Magud and A. Werner. International Monetary Fund.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977), 'Maximum likelihood from incomplete data via the em algorithm', *Journal of the Royal Statistical Society: Series B (Methodological)* **39**, 1–22.

Diaz, G. (2020), 'Stopwords ISO: The most comprehensive collection of stopwords for multiple languages.', https://github.com/stopwords-iso/stopwords-iso.

Evans, M. D. and Lyons, R. K. (2002), 'Order flow and exchange rate dynamics', *Journal of Political Economy* **110**, 170–180.

Evans, M. D. and Lyons, R. K. (2012), 'Exchange rate fundamentals and order flow', *The Quarterly Journal of Finance* **2**, 1250018.

Gentzkow, M., Kelly, B. and Taddy, M. (2019), 'Text as data', *Journal of Economic Literature* **57**, 535–74.

Griffiths, T. L. and Steyvers, M. (2004), 'Finding scientific topics', *Proceedings of the National Academy of Sciences* **101**, 5228–5235.

Grupo-Nación-EF (2025), 'Periódico el financiero'. https://www.elfinancierocr.com/.

Grupo-Nación-LN (2025), 'Periódico la nación'. https://www.nacion.com/.

Jagarlamudi, J., Daumé III, H. and Udupa, R. (2012), Incorporating lexical priors into topic models, *in* 'Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics', pp. 204–213.

Ke, Z. T., Kelly, B. T. and Xiu, D. (2021), 'Predicting returns with text data'. Working paper available at: https://ssrn.com/abstract=3389884.

King, M. R., Osler, C. L. and Rime, D. (2013), 'The market microstructure approach to foreign exchange: Looking back and looking forward', *Journal of International Money and Finance* **38**, 95–119.

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

Lopez-Lira, A. (2023), 'Risk factors that matter: Textual analysis of risk disclosures for the cross-section of returns'. Working paper available at: https://ssrn.com/abstract=3313663.

Lyons, R. K. (1995), 'Tests of microstructural hypotheses in the foreign exchange market', *Journal of Financial Economics* **39**, 321–351.

Mora, A. (2020), Text into FX: Exchange rate supervised topic extraction., Working paper, University of Pennsylvania, Philadelphia.

Nakajima, J. and West, M. (2013), 'Bayesian analysis of latent threshold dynamic models', *Journal of Business and Economic Statistics* **31**, 151–164.

Osler, C. L., Mende, A. and Menkhoff, L. (2011), 'Price discovery in currency markets', *Journal of International Money and Finance* **30**, 1696–1718.

Porter, M. F. (2001), 'Snowball: A language for stemming algorithms'. Available at: http://snowball.tartarus.org/.

Ramage, D., Hall, D., Nallapati, R. and Manning, C. D. (2009), Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora, *in* 'Proceedings of the 2009 conference on empirical methods in natural language processing', pp. 248–256.

Thorsrud, L. A. (2020), 'Words are the new numbers: A newsy coincident index of the business cycle', *Journal of Business and Economic Statistics* **38**, 393–409.

Vitale, P. (2007), 'A guided tour of the market microstructure approach to exchange rate determination', *Journal of Economic Surveys* **21**, 903–934.

Vitale, P. (2011), 'The impact of fx intervention on fx markets: a market microstructure analysis', *International Journal of Finance and Economics* **16**, 41–62.

Wainwright, M. J. and Jordan, M. I. (2008), *Graphical models, exponential families, and variational inference*, Now Publishers Inc.

Appendices

A Mathematical derivations

A.1 Computation details for the ELBO

The objective in the E-step of the VEM algorithm is to find the values $\{\gamma_t^*, \phi_{t,1:N_t}^*\}$ for each document $t = 1, ..., T$ and $\{\lambda_k^*\}$ for each topic $k = 1, ..., K$ that maximize the corpus-level ELBO. To do such maximization we start by providing detailed expressions of the ELBO at each document:

$$\mathcal{L}(\gamma, \phi_{1:N}, \lambda_{1:K}) = E_q[\log p(\theta|\alpha)] + \sum_{k=1}^{K} E_q[\log p(\beta_k|\eta)] + \sum_{n=1}^{N} E_q[\log p(z_n|\theta)] + \sum_{n=1}^{N} E_q[\log p(w_n|z_n, \beta_{1:K})] + E_q[\log p(y|z_{1:N}, W, b, c, \sigma^2)] + H(q)$$

A.1.1 Components of $\mathcal{L}(\gamma, \phi_{1:N}, \lambda_{1:K})$

1. $E_q[\log p(\theta|\alpha)]$: Since $\theta|\alpha \sim Dir(\alpha)$ we have that $p(\theta|\alpha) = \left[\frac{\prod_{k=1}^{K}\Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K}\alpha_k)}\right]\prod_{k=1}^{K}\theta_k^{\alpha_k-1}$ then:

$$\log p(\theta|\alpha) = \log\Gamma\left(\sum_{k=1}^{K}\alpha_k\right) - \sum_{k=1}^{K}\log\Gamma(\alpha_k) + \sum_{k=1}^{K}(\alpha_k - 1)\log\theta_k$$

$$\text{and } E_q[\log p(\theta|\alpha)] = \log\Gamma\left(\sum_{k=1}^{K}\alpha_k\right) - \sum_{k=1}^{K}\log\Gamma(\alpha_k) + \sum_{k=1}^{K}(\alpha_k - 1)E_q[\log\theta_k]$$

   where $E_q[\log\theta_k] = \psi(\gamma_k) - \psi\left(\sum_{k=1}^{K}\gamma_k\right)$ and $\psi(\cdot)$ stands for the digamma function.

2. $\sum_{k=1}^{K} E_q[\log p(\beta_k|\eta)]$: Since $\beta_k|\eta \sim Dir(\eta)$ we have that $p(\beta_k|\eta) = \left[\frac{\prod_{v=1}^{V}\Gamma(\eta_v)}{\Gamma(\sum_{v=1}^{V}\eta_v)}\right]\prod_{v=1}^{V}\beta_{k,v}^{\eta-1}$ then (similar to computation for $\theta|\alpha$):

$$E_q[\log p(\beta_k|\eta)] = \log\Gamma\left(\sum_{v=1}^{V}\eta_v\right) - \sum_{v=1}^{V}\log\Gamma(\eta_v) + \sum_{v=1}^{V}(\eta_v - 1)E_q[\log\beta_{k,v}]$$

$$\text{and } \sum_{k=1}^{K} E_q[\log p(\beta_k|\eta)] = K\log\Gamma\left(\sum_{v=1}^{V}\eta_v\right) - K\sum_{v=1}^{V}\log\Gamma(\eta_v) + \sum_{k=1}^{K}\sum_{v=1}^{V}(\eta_v - 1)E_q[\log\beta_{k,v}]$$

with $E_q[\log \beta_{k,v}] = \psi(\lambda_{k,v}) - \psi\left(\sum_{v=1}^{V} \lambda_{k,v}\right)$.

3. $\sum_{n=1}^{N} E_q[\log p(z_n|\theta)]$: We assumed that $z_n|\theta \sim Mult(\theta)$. Coding z_n in one-in-K form we can write $p(z_n|\theta) = \prod_{k=1}^{K} \theta_k^{z_{n,k}}$, then $\log p(z_n|\theta) = \sum_{k=1}^{K} z_{n,k} \log \theta_k$ and

$$E_q[\log p(z_n|\theta)] = \sum_{k=1}^{K} E_q[z_{n,k}] E_q[\log \theta_k] = \sum_{k=1}^{K} \phi_{n,k} E_q[\log \theta_k]$$

$$\text{and hence } \sum_{n=1}^{N} E_q[\log p(z_n|\theta)] = \sum_{n=1}^{N} \sum_{k=1}^{K} \phi_{n,k} E_q[\log \theta_k]$$

4. $\sum_{n=1}^{N} E_q[\log p(w_n|z_n, \beta_{1:K})]$: Recall that $w_n \in \{1, ..., V\}$ and that $\beta_{1:K}$ is $V \times K$ matrix so we can write $p(w_n|z_n, \beta_{1:K}) = \prod_{k=1}^{K} \beta_{w_n,k}^{z_{n,k}}$ so we have:

$$E_q[\log p(w_n|z_n, \beta_{1:K})] = \sum_{k=1}^{K} \phi_{n,k} E_q[\log \beta_{w_n,k}]$$

$$\text{and } \sum_{n=1}^{N} E_q[\log p(w_n|z_n, \beta_{1:K})] = \sum_{n=1}^{N} \sum_{k=1}^{K} \phi_{n,k} E_q[\log \beta_{k,w_n}]$$

5. $E_q[\log p(y|z_{1:N}, W, b, c, \sigma^2)]$: Given the gaussian response for $y = b'\bar{Z} + c'W + \varepsilon$ with $\varepsilon \sim N(0, \sigma^2)$ we have:

$$p(y|z_{1:N}, W, b, c, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y - \bar{Z}'b - c'W)^2}{2\sigma^2}\right)$$

$$\Rightarrow \log p(y|z_{1:N}, W, b, c, \sigma^2) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y - c'W - \bar{Z}'b)^2$$

Denoting $\tilde{y} = y - c'W$, we have:

$$E_q[\log p(y|z_{1:N}, W, b, c, \sigma^2)] = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left[\tilde{y}^2 - 2\tilde{y}E_q[\bar{Z}]'b + b'E_q[\bar{Z}\bar{Z}']b\right]$$

with

$$E_q[\bar{Z}] = \frac{1}{N}\sum_{n=1}^{N}\phi_n$$
$$E_q[\bar{Z}\bar{Z}'] = \frac{1}{N^2}\left(\sum_{n=1}^{N}\phi_n\phi'_{-n} + \sum_{n=1}^{N}\text{diag}(\phi_n)\right)$$
$$\phi_{-n} = \sum_{m\neq n}\phi_m$$

6. $H(q)$: the entropy of the variational distribution is given by

$$\begin{aligned}H(q) = &-\log\Gamma\left(\sum_{k=1}^{K}\gamma_k\right) + \sum_{k=1}^{K}\log\Gamma(\gamma_k) - \sum_{k=1}^{K}(\gamma_k - 1)E_q[\log\theta_k] \\ &- \sum_{n=1}^{N}\sum_{k=1}^{K}\phi_{n,k}\log(\phi_{n,k}) \\ &- \sum_{k=1}^{K}\log\Gamma\left(\sum_{v=1}^{V}\lambda_{k,v}\right) + \sum_{k=1}^{K}\sum_{v=1}^{V}\log\Gamma(\lambda_{k,v}) - \sum_{k=1}^{K}\sum_{v=1}^{V}(\lambda_{k,v} - 1)E_q[\log\beta_{k,v}]\end{aligned}$$

A.1.2 Maximization of $\mathcal{L}(\gamma, \phi_{1:N}, \lambda_{1:K})$

The optimization is done by block coordinate-ascent variational inference, that means iteratively maximizing $\mathcal{L}$ with respect to each variational parameter. In this maximization we need to include the constraint that each of the multinomial parameters ϕ_n should sum up to one. First we obtain the partial derivative of $\mathcal{L}$ with respect to γ_k for $k = 1, ..., K$

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\gamma_k} &= \frac{\partial}{\partial\gamma_i}E_q[\log p(\theta|\alpha)] + \frac{\partial}{\partial\gamma_k}\sum_{n=1}^{N}E_q[\log p(z_n|\theta)] + \frac{\partial}{\partial\gamma_k}H(q) \\ &= (\alpha_k - 1)\frac{\partial E[\log\theta_k]}{\partial\gamma_k} + \sum_{n=1}^{N}\phi_{n,k}\frac{\partial E[\log\theta_k]}{\partial\gamma_k} - \psi\left(\sum_{k=1}^{K}\gamma_k\right) \\ &+ \psi(\gamma_k) - E_q[\log\theta_k] - (\gamma_k - 1)\frac{\partial E[\log\theta_k]}{\partial\gamma_k} \\ &= \left[\alpha_k - \gamma_k + \sum_{n=1}^{N}\phi_{n,k}\right]\frac{\partial E[\log\theta_k]}{\partial\gamma_k}\end{aligned}$$

therefore

$$\Rightarrow \frac{\partial \mathcal{L}}{\partial \gamma_k} = 0 \Leftrightarrow \gamma_k = \alpha_k + \sum_{n=1}^{N} \phi_{n,k},$$

So we can update the whole vector:

$$\gamma^{new} \leftarrow \alpha + \sum_{n=1}^{N} \phi_n \tag{14}$$

Notice that this step is exactly the same as in the LDA VEM update. Next we obtain the partial derivative with respect to $[\phi_{n,k}]$ for $k = 1, ..., K$ and $n = 1, ..., N$

$$\frac{\partial \mathcal{L}}{\partial \phi_{n,k}} = E_q[\log \theta_k] + E_q[\log \beta_{w_n,k}] - \log \phi_{n,k} - 1 + \lambda + \frac{\partial}{\partial \phi_{n,k}} E_q[\log p(y|z_{1:N}, b, c, \sigma^2)]$$

where λ is the Lagrange multiplier for the sum up to one constraint. The last partial derivative requires more work:

$$\frac{\partial}{\partial \phi_{n,k}} E_q[\log p(y|z_{1:N}, b, c, \sigma^2)] = \frac{(y - c'W)b_k}{N\sigma^2} - \frac{1}{2\sigma^2} \frac{\partial}{\partial \phi_{n,k}} \left(b' E_q[\bar{Z}\bar{Z}']b \right)$$

Let $\phi_{-n} := \sum_{m \neq n} \phi_m$; the trick here is to express $b' E_q[\bar{Z}\bar{Z}']b$ as a function of just ϕ_n:

$$\begin{aligned} f(\phi_n) &= \frac{1}{N^2} b'[\phi_n \phi'_{-n} + \phi_{-n} \phi'_n + diag(\phi_n)]b + \text{const.} \\ &= \frac{1}{N^2}[2(b'\phi_{-n})b'\phi_n + (b \cdot b)'\phi_n] + \text{const.} \\ \Rightarrow \frac{\partial}{\partial \phi_{n,k}} \left(b' E_q[\bar{Z}\bar{Z}']b \right) &= \frac{1}{N^2}[2(b'\phi_{-n})b + b \cdot b] \end{aligned}$$

Substituting this and noticing that λ acts as a normalization constant then:

$$\frac{\partial \mathcal{L}}{\partial \phi_{n,k}} = 0 \Leftrightarrow \phi_{n,k} \propto \exp\left[E_q[\log \theta_k] + E_q[\log \beta_{w_n,k}] + \frac{(y - c'W)b_k}{N\sigma^2} - \frac{1}{2\sigma^2 N^2}[2(b'_k \phi_{-n})b_k + b_k \cdot b_k] \right]$$

and updating the whole vector is done by:

$$\phi_n^{new} \propto \exp\left[E_q[\log\theta] + E_q[\log\beta_{w_n,\cdot}] + \frac{(y - c'W)b}{N\sigma^2} - \frac{1}{2\sigma^2N^2}[2(b'\phi_{-n})b + b \cdot b]\right] \tag{15}$$

here the proportional stands for compute each entry and then normalize to one. Finally the derivative with respect to $[\lambda_{k,v}]$ is given by:

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\lambda_{k,v}} &= \frac{\partial}{\partial\lambda_{k,v}}\sum_{k=1}^{K} E_q[\log(\beta_k|\eta)] + \frac{\partial}{\partial\lambda_{k,v}}\sum_{n=1}^{N} E_q[\log p(w_n|z_n,\beta_{1:K})] + \frac{\partial H(q)}{\partial\lambda_{k,v}} \\ &= \left[\eta_v - \lambda_{k,v} + \sum_{n=1}^{N}\mathbb{I}\{w_n = v\}\phi_{n,k}\right]\frac{\partial}{\partial\lambda_{k,v}}E_q[\log\beta_{k,v}]\end{aligned}$$

so accounting for commonality of $\lambda_{k,v}$ among all documents and equating the derivative to zero yields:

$$\lambda_{k,v} = \eta_v + \sum_{t=1}^{T}\sum_{n=1}^{N_t}\mathbb{I}\{w_{t,n} = v\}\phi_{t,n,k}$$

And updating the vector is done by:

$$\lambda_k = \eta + \sum_{t=1}^{T}\sum_{n=1}^{N_t}\phi_{t,n,k}\omega_{t,n} \tag{16}$$

with $\omega_{t,n}$ a one-hot representation of $w_{t,n}$ as a vector of size V.

A.1.3 Maximization of $L(b, c, \sigma^2)$

Once we have maximizers $\{\gamma_t^*, \phi_{t,1:N_t}^*\}$ of the ELBO for each given document-response and $\{\lambda_{1:K}^*\}$ for the topics, we are ready to update the global parameters $\pi = \{b, c, \sigma^2\}$ in the M-step by maximizing $L(b, c, \sigma^2)$, the corpus-level ELBO. In this step we want to maximize over π the objective function:

$$L(b, c, \sigma^2) = \sum_{t=1}^{T}\mathcal{L}(\gamma_t^*, \phi_{t,1:N}^*, \lambda_{1:K}^*|b, c, \sigma^2)$$

Here each term in the sum is the $(w_{t,1:N_t}, y_t, W_t)$ specific ELBO evaluated at the corresponding maximizing variational parameters from the E-step. Optimizing with respect to b we have:

$$\begin{aligned}
\frac{\partial L}{\partial b} &= \sum_{t=1}^{T} \frac{\partial}{\partial b} E_{q^*} \left[\log p(y_t | z_{t,1:N_t}, b, c, \sigma^2) \right] \\
&= \sum_{t=1}^{T} \frac{1}{\sigma^2} \left[(y_t - c' W_t) E_{q^*}[\bar{Z}_t]' - b' E_{q^*} \left[\bar{Z}_t \bar{Z}_t' \right] \right] = 0 \\
\Rightarrow b^{new} &= \left(\sum_{t=1}^{T} E_{q^*} \left[\bar{Z}_t \bar{Z}_t' \right] \right)^{-1} \left(\sum_{t=1}^{T} E_{q^*} \left[\bar{Z}_t \right] (y_t - c' W_t) \right)
\end{aligned}$$

Next, optimizing with respect to c:

$$\begin{aligned}
\frac{\partial L}{\partial c} &= \sum_{t=1}^{T} \frac{\partial}{\partial c} E_{q^*} \left[\log p(y_t | z_{t,1:N_t}, b, c, \sigma^2) \right] \\
&= \sum_{t=1}^{T} \frac{1}{\sigma^2} \left[(y_t - E_{q^*}[\bar{Z}_d]' b) W_t' - c' W_t W_t' \right] = 0 \\
\Rightarrow c^{new} &= \left(\sum_{t=1}^{T} W_t W_t' \right)^{-1} \left(\sum_{t=1}^{T} W_t (y_t - E_{q^*}[\bar{Z}_d]' b^{new}) \right)
\end{aligned}$$

Finally the response dispersion coefficient σ^2 is optimized by:

$$\begin{aligned}
\frac{\partial L}{\partial \sigma^2} &= \sum_{t=1}^{T} \frac{\partial}{\partial \sigma^2} E_{q^*}[\log p(y_t | z_{t,1:N_t}, b, \sigma^2)] \qquad (17) \\
&= \sum_{t=1}^{T} -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4} E_{q^*}[(y_t - \bar{Z}_t' b^{new} - W_t' c^{new})^2] = 0 \\
\Rightarrow \sigma^{2,new} &= \frac{1}{T} \sum_{t=1}^{T} \left(y_t^2 - y_t E_{q^*}[\bar{Z}_t'] b^{new} - y_t W_t' c^{new} \right). \qquad (18)
\end{aligned}$$

B Pseudo Code for Algorithms

Algorithm 1 sketches the steps necessary for estimation of the sLDA model using the Variational Expectation Maximization (VEM) algorithm. Algorithm 2 shows pseudo code for the steps necessary to use an estimated sLDA model for prediction.

Algorithm 1: VEM algorithm for sLDA estimation

Input: Document and Response variables $\{w_{t,1:N_t}, y_t, W_t\}$ for $t = 1, ..., T$; and hyperparameters (α, η).
Output: A set of estimated parameters $\hat{\pi} = \{\hat{b}, \hat{c}, \hat{\sigma}^2\}$.
Initialize: π^0, $\{\gamma_t^0, \phi_{t,1:N_t}^0\}$ for $t = 1, ..., T$ and $\{\lambda_k^0\}$ for $k = 1, ..., K$.
while *Corpus ELBO has not converged* **do**
  `/* E-Step */`
  for $t = 1$ **to** T **do**
    while *Document-Response ELBO has not converged* **do**
      `/* Update Variational Parameters */`
      $\gamma^{new} \leftarrow \alpha + \sum_{n=1}^{N} \phi_n$;
      $\phi^{new} \propto \exp\left[E_q[\log\theta] + E_q[\log\beta_{w_n,\cdot}] + \frac{\tilde{y}b}{N\sigma^2} - \frac{1}{2\sigma^2N^2}[2(b'\phi_{-n})b + b \cdot b]\right]$;
    end
  end
  for $k = 1$ **to** K **do**
    `/* Update Topic Variational Parameters */`
    $\lambda_{k,v} = \eta_v + \sum_{t=1}^{T}\sum_{n=1}^{N_t} \mathbb{I}\{w_{t,n} = v\}\phi_{t,n,k}^*$
  end
  `/* M-Step: Update Parameters */`
  $b^{new} \leftarrow \left(\sum_{t=1}^{T} E_{q^*}[\bar{Z}_t\bar{Z}_t']\right)^{-1}\left(\sum_{t=1}^{T} E_{q^*}[\bar{Z}_t]\tilde{y}_t\right)$;
  $c^{new} \leftarrow \left(\sum_{t=1}^{T} W_tW_t'\right)^{-1}\left(\sum_{t=1}^{T} W_t(y_t - E_{q^*}[\bar{Z}_t])b^{new}\right)$;
  $\sigma^{2,new} \leftarrow \frac{1}{T}\sum_{t=1}^{T}\left(y_t^2 - y_tE_{q^*}[\bar{Z}_t']b^{new} - y_tW_t'c^{new}\right)$
  `/* Compute Corpus ELBO */`
  for $t = 1$ **to** T **do**
    $L(\pi^{new}) += \mathcal{L}(\gamma_t^*, \phi_{t,1:N_t}^*, \lambda_{1:K}^*; w_{t,1:N_t}, y_t, W_t; \pi^{new})$
  end
  if *Corpus ELBO meets convergence criteria* **then**
    Stop Execution
  end
end

Algorithm 2: VEM algorithm for sLDA prediction

Input: Document $w_{1:N}$, covariates W, hyperparameters α, η and sLDA estimates $\{\hat{b}, \hat{c}, \hat{\sigma}^2, \hat{\beta}_{1:K}\}$

Output: A predicted response $\hat{y}$.

Initialize: Variational parameters $(\tilde{\gamma}_0, \tilde{\phi}_{0,1:N})$.

while *Document ELBO has not converged* **do**

$\tilde{\gamma}^{new} \leftarrow \alpha + \sum_{n=1}^{N} \tilde{\phi}_n$;

$\tilde{\phi}_n^{new} \propto \exp\left[E_{\tilde{q}}[\log\theta] + \log\beta_{w_n,\cdot}\right]$;

if *Document ELBO meets convergence criteria* **then**

$\hat{y} = \hat{b}'\left(N^{-1}\sum_{n=1}^{N}\tilde{\phi}_n\right) + \hat{c}'W$;

Return: $\hat{y}$

end

end

Table C1: VEM Algorithm Execution Parameters

	Single-Stage	Two-Stage
Max VEM iterations	100	100
Max ELBO iterations by document	25	25
VEM-Corpus Tolerance	10^{-5}	10^{-5}
ELBO-Document Tolerance	10^{-5}	10^{-5}
Documents in the training sample	900	900
Documents in the test sample for prediction	100	100

Notes: Execution parameters used for both versions of the VEM Algorithm on the simulation study. See text for details.

C Additional Details of the Monte Carlo Simulation

In the VEM algorithm for supervised topic extraction there are four key parameters during implementation: (i) the max number of *VEM iterations*, that is the maximum number of alternations between the E-step and the M-step, (ii) the max number of *ELBO iterations*, which is the maximum number of times we use coordinate ascent to maximize the ELBO document by document, (iii) the *VEM-Corpus tolerance*, base on which we would stop the VEM algorithm when the change in the corpus ELBO between subsequent iterations is less than such tolerance and finally in a similar fashion (iv) the *ELBO-Document tolerance*, that controls the maximization of the ELBO document by document. In the single-stage estimation the value for these execution parameters are shown in Table C1.

The two-stage estimation is done by first using LDA on $\{w_{t,1:N_t}\}$ to estimate $\{z_{t,1:N_T}\}$, $\beta_{1:K}$ and $\{\theta_t\}$. For this step we use the VEM algorithm simplified to handle just the unsupervised LDA, but we need still to set during execution the max VEM iterations, the max ELBO iterations, the VEM-Corpus tolerance and the ELBO-Document tolerance; these are shown in Table C1. The second stage uses the estimates $\{\hat{z}_{t,1:N_T}\}$ to construct estimates of the empirical topic frequency for each document $\{\hat{Z}_t = N_t^{-1}\sum_{i=1}^{N_t}\hat{z}_{t,i}\}$ that would be used as regressors for $\{y_t\}$. Based on this second-stage regression we would obtain estimates for b and σ^2.

D Extracted Topics Details

Table D1: CR Exchange Rate Supervised Extracted Topics

Topic	Label	Spanish Top Stems	English Top Words
1	Health System Strike	public; polit; ley; huelg; president; ccss; econom; chin; servici; declar;	public; politics; law; strike; president; CCSS; economy; China; service; declare
2	Public Institutions	rabin; suges; lineamient; naranj; marcham; wong; mideplan; soa; mattis; garr;	Rabin; SUGESE; guideline; Naranjo; Auto Insurance; Wong; MIDEPLAN; SOA; Mattis; Garro
3	International News	econom; public; person; nacional; cambi; polit; banc; iran; pag; costa_rica;	Economy; public; person; national; change; policy; bank; Iran; payment; Costa Rica
4	Europe Economic Activity	public; gobiern; econom; costa_rica; gast; polit; nacional; proyect; pag; president;	public; government; economy; Costa Rica; expense; policy; national; project; payment; president
5	IMF Deal Strike	gobiern; public; president; nacional; huelg; servici; fiscal; person; proyect; pag;	government; public; president; national; strike; service; fiscal; person; project; payment
6	State-owned Companies	public; costa_rica; banc; person; empres; nacional; pag; econom; credit; gobiern;	public; Costa Rica; bank; person; firm; national; payment; economy; credit; government
7	Covid-19 Pandemic	econom; public; person; pandemi; med; covid; caso; coronavirus; gobiern; salud;	economy; public; person; pandemic; physician; covid; case; coronavirus; government; health
8	US Presidential Election	public; trump; biden; president; nacional; dialog; econom; sector; octubr; fiscal;	public; Trump; Biden; president; national; dialog; economy; sector; october; fiscal

Table D1: CR Exchange Rate Supervised Extracted Topics (continued)

Topic	Label	Spanish Top Stems	English Top Words
9	Consumer Finance Policy	credit; tas; banc; person; entidad; comision; financ; tarjet; sugef; merc;	credit; rate; bank; person; entity; comission; finance; card; SUGEF; market
10	Covid-19 Second Wave	president; public; polit; gobiern; moral; econom; nacional; octubr; social; proyect;	president; public; policy; government; Morales; economy; national; october; social; project
11	Tax Discussion	polit; econom; public; person; gobiern; president; social; nacional; estados_unidos; costa_rica;	policy; economy; public; person; government; president; social; national; United States; Costa Rica
12	US Economy News	gobiern; polit; president; trump; econom; estados_unidos; public; empres; chin; nacional;	government; policy; president; Trump; economy; United States; public; firm; China; national
13	Fiscal Policy	subast; contribuyent; lun; compr; deud; abort; ajust; chin; kong; hong;	auction; taxpayer; Monday; buy; debt; abort; adjust; China; Hong Kong
14	Covid-19 Vaccines	president; vacun; gobiern; econom; dolar; estados_unidos; public; pandemi; polit; person;	president; vaccine; government; economy; dollar; United States; public; pandemic; policy; person
15	Middle East News	econom; israel; netanyahu; polit; ley; gobiern; president; estados_unidos; cambi; proyect;	economy; Israel; Netanyahu; politics; law; government; president; United States; change; project
16	Congress Debates	polit; president; gobiern; public; econom; costa_rica; nacional; fiscal; diput; person;	Policy; president; government; public; economic; Costa Rica; fiscal; lawmaker; person

Table D1: CR Exchange Rate Supervised Extracted Topics (continued)

Topic	Label	Spanish Top Stems	English Top Words
17	Government Economic Policy	public; pag; empres; impuest; econom; servici; fiscal; caso; person; polit;	public; payment; firm; tax; economy; service; fiscal; case; person; policy
18	Brexit	presupuest; gast; ger; corrient; nicol; brexit; vuel; interes; johnson; europe;	budget; expenditure; current; Brexit; flight; interest; Johnson; Europe
19	Industrial Policy	proyect; empres; ley; electr; pension; energ; gobiern; cambi; person; pag;	project; firm; law; electric; pension; energy; government; change; person; payment
20	2018 Tax Reform Discussion	public; proyect; fiscal; diput; econom; gobiern; ley; pag; president; aprob;	public; project; fiscal; lawmaker; economy; government; law; payment; president; approval
21	Introduction of VAT	iva; impact; micas; servici; dueñ; public; empres; pag; caso; impuest;	VAT; impact; service; owner; public; firm; payment; case; tax
22	UPAD Goverment Scandal	dat; polit; president; public; caso; person; democrat; investig; febrer; ministr;	data; politics; president; public; case; person; democrat; investigation; february; minister
23	Exports	econom; gobiern; costa_rica; president; estados_unidos; trump; public; polit; chin; merc;	economy; government; Costa Rica; president; United States; Trump; public; policy; China; market
24	Rent Prices	candidat; polit; public; empres; gobiern; econom; gast; banc; tem; fiscal;	candidate; politics; public; firm; government; economy; economy; expense; bank; topic; fiscal
25	2018 Tax Reform Approval	public; fiscal; polit; econom; person; president; gobiern; merc; caso; pag;	public; fiscal; politics; economy; person; president; government; market; case; payment