Topic Factor Modelling: uncovering thematic structure in financial data

Anonymous Author(s)
Affiliation
Address
email

Abstract

We examine the task of finding thematic structure in a data corpus comprising text and time series, motivated by applications to financial data such as improving correlation estimation. We introduce a topic factor model (TFM): a joint generative model for both text and time series data which resembles supervised latent Dirichlet allocation. Our TFM allows the decomposition of time series into factors which also reflect the thematic content of the text. The structure is found using mean field variational inference, though this is complicated by the lack of a closed form update for some variables. The key modelling challenge is balancing the combination of continuous and discrete distributions. We use a corpus from the foreign exchange market to demonstrate improved likelihood of held out time series data.

In finance, decomposition of time series into the key driving forces is commonplace. PCA and related methods are used to identify features of comovement of prices. Alternatively, returns can be attributed to some set of economic variables by regression. The resulting, ubiquitous, probabilistic models of returns are called factor models (see, for example Fama and French [2]). We are not aware of any previous attempt at a middle ground of automatically detecting a time series decomposition where the components have economic meaning. Joint topic modelling of text and time series should be able to achieve just that. By extension to the above mentioned methods we call our model a topic factor model (or TFM). It differs from existing joint models in that the non-text variables are not sampled by first drawing a topic from the topic distribution, but rather depend directly on the topic proportions. This creates some issues around balancing discrete and continuous probability but gives rise to a model which both corresponds better to our intuition and is able to outperform existing models.

The canonical example of mixing text with continuous data is supervised latent Dirichlet allocation (sLDA) [1], which models text data and adds a response variable with distribution given by a generalized linear model. Under sLDA, however, topics can only affect the response variable when also used to explain associated text. For some applications one might require the model have the flexibility to allocate mass to topics and explain the response variable *without* using that mass to explain text. This is particularly the case for financial data, where thematic structure omitted from text indicates information not considered by the authors of the text and may indicate a valuable competitive advantage. Another option would be to use Dirichlet multinomial regression [3] which has features upstream of the document topic distribution. Sadly this doesn't generalize to new features, which is important in the context of financial time series because the distribution of tomorrow's returns is critical. For these reasons we use a TFM similar in form but slightly different to sLDA.

1 Topic Factor Models

The generative model for text in our topic factor model is taken directly from LDA. For each of a set of K topics, a categorical distribution β_k over all M words in a dictionary is sampled from a symmetric Dirichlet distribution with parameter η. Independently, a vector θ_d of dimension K is sampled from a symmetric Dirichlet distribution (with parameter α) for each document in a corpus. This makes up the probability density function for a categorical distribution from which variables $z_{d,n}$ are drawn. These in turn indicate the topics from which the words $w_{d,n}$ are to be drawn.

$$p(z_{d,n} = k|\theta_d) = \theta_{d,k} \tag{1}$$
$$p(w_{d,n} = m|z_{d,n}, \beta) = \beta_{z_{d,n},m} \tag{2}$$

To allow simultaneous emission of time series data, θ_d also acts as a latent variable in a time series model. This dual interpretation of θ, as both a distribution over topics and a weight vector in a factor model, allows inference on a joint corpus. Each topic has associated with it, not only a distribution over words but also a series of returns $R_{k,t}$ of length T. We combine these linearly at each time interval ($\theta_d^{\mathsf{T}} R_t$) to construct time series returns for each document. To this we add an idiosyncratic term $\epsilon_{d,t} \sim \mathcal{N}(0, 1)$. Figure 1 shows the combination of the K topic time series to give each document time series.

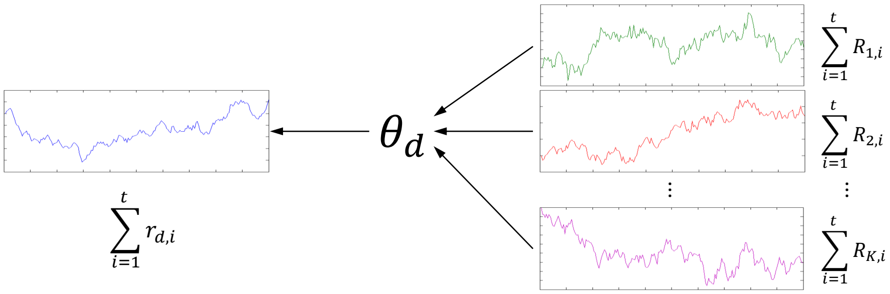

Figure 1: The time series model reinterprets the categorical distribution θ_d as a weight vector for the linear combination of topic factors to construct a document time series.

We want to have flexibility within the model to change the contributions from the topics and from the idiosyncratic component, which we achieve using a parameter ρ. The returns are thus given by

$$r_{d,t} = \frac{\rho}{\sqrt{v(\alpha)}} {\theta_d}^{\mathsf{T}} R_t + \sqrt{1 - \rho^2}\, \epsilon_{d,t} \tag{3}$$

where the scale factor v is added so that the expectation of the variance of the returns is one. The parameter ρ plays the role of tuning the significance of the time series in the model. At $\rho = 0$ the time series has no dependence on θ and the model corresponds to LDA. In contrast, for ρ near 1, the probability of the time series will be vanishingly small, and fitting θ to the time series will overwhelm the influence of the text.

The prior distribution of the returns of the latent time series are Gaussian with zero mean and unit covariance, $R_{k,t} \sim \mathcal{N}(0, 1)$. The mean of the observed returns will thus be zero and the variance given by

$$E[r_{d,t}^2] = \frac{\rho^2}{v(\alpha)} E\left[\sum_k \theta_{d,k}^2 \middle| \alpha \right] + 1 - \rho^2. \tag{4}$$

For the returns to have unit variance we therefore use the scale factor

$$v(\alpha) = \frac{\alpha + 1}{K\alpha + 1}. \tag{5}$$

We want a generative model whose returns have unit variance so that standardized return data can be used. We otherwise must simultaneously learn the variance of each document so that inference isn't biased towards higher variance documents.

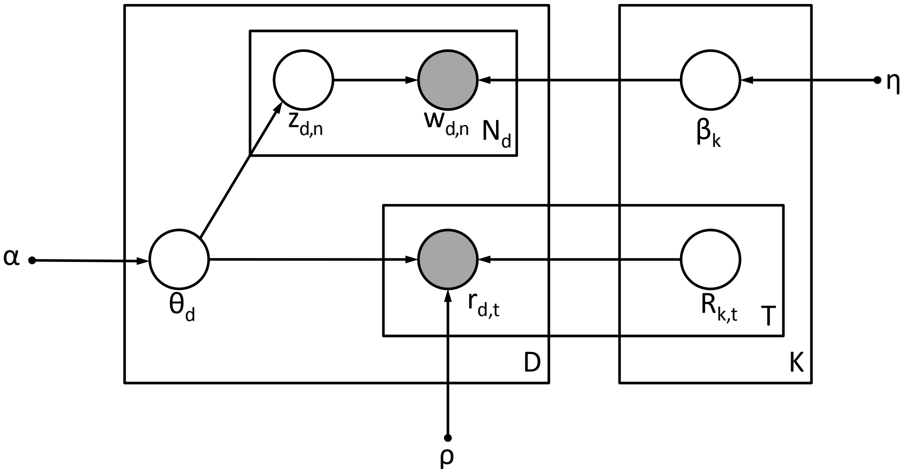

Figure 2: The Bayesian network for a topic factor model. The shaded nodes are the observed data: words $w_{d,n}$ and time series intervals $r_{d,t}$.

Putting together the model described above, the graphical model representing a TFM is shown in Figure 2. This corresponds to a factorization

$$p(w, r, z, \theta, \beta, R) = \prod_d p(\theta_d) \prod_n p(z_{d,n}|\theta_d) p(w_{d,n}|z_{d,n}) \prod_k p(\beta_k) \prod_t p(R_{k,t}) p(r_{d,t}|R_t, \theta_d). \quad (6)$$

With no text this looks like factor analysis, but with the weights constrained to be non-negative and with Dirichlet priors. With no time series it reduces to LDA.

2 Inference

The quantity of interest under this model is the posterior of the latent variables given a corpus:

$$p(\theta, \beta, z, R|w, r) = \frac{p(\theta, \beta, w, r, z, R)}{p(w, r)}. \quad (7)$$

is intractable because of the high dimensionality of z and R. We therefore opt to apply mean field variational inference. We approximate the posterior $p(\theta, \beta, z, R|w, r)$ using a simpler, variational distribution $q(\theta, \beta, z, R)$. The expectations of θ, β z and R under q can then be interpreted as an approximation to the MAP settings. The optimal variational distribution $q^*(\theta, \beta, z, R)$ should be as close to the true posterior as possible, by KL divergence:

$$\underset{q}{\operatorname{argmin}} \operatorname{KL}(q||p) = \underset{q}{\operatorname{argmin}} \; \operatorname{E}_q\big[\log q(\theta, \beta, z, R)\big] - \operatorname{E}_q\big[\log p(\theta, \beta, z, R|w, r)\big]. \quad (8)$$

Then using Bayes' rule and the fact that $p(w, r)$ is a constant

$$q^*(\theta, \beta, z, R) = \underset{q}{\operatorname{argmin}} \; \operatorname{E}_q\big[\log q(\theta, \beta, z, R)\big] - \operatorname{E}_q\big[\log p(\theta, \beta, z, R, w, r)\big]. \quad (9)$$

As for LDA we would like to take the variational factors for each component to be of the same (exponential family) form as the corresponding complete conditional in the true posterior and derive update rules. Unfortunately, although the form of the complete conditional in θ_d belongs to the exponential family, using a variational distribution of the same form doesn't give rise to tractable updates. It is possible, however, to choose a simpler variational distribution such that the objective and its gradients are tractable. Gradient descent can then be applied in the variational parameter space. We take the variational distributions of θ and β to be Dirichlet with parameter vectors ϕ and γ respectively, z to be categorical $q(z|\lambda) = \lambda_z$, and $R_{k,t}$ to be independently Gaussian distributed with mean $\mu_{k,t}$ and standard deviation $\sigma_{k,t}$. For λ, γ and $\{\mu, \sigma\}$, update rules allow us to find the minima with respect to each parameter explicitly. In the case of ϕ however, no such simple update exists, and a gradient following method is needed. We repeat these updates until the distribution converges and then take the expected parameter settings under the converged distribution as a point estimate of the latent variables.

3 Uncovering thematic structure in foreign exchange data

In the foreign exchange market there are some economic factors which are shared between currencies. For instance, exchange rates are likely to be impacted by trade deficit changes which are in turn affected by the fates of various industries shared across national borders. We construct a corpus by finding the daily log return of a set of 11 currencies (against a base of the US dollar) through 2012. The text data comprises Citi's global economic outlook summaries for each country in the corpus. To demonstrate the impact of the text on the validity of the time series model we compare the average held out likelihood of a day's log return for a model trained without text, with text and a benchmark of probabilistic PCA.

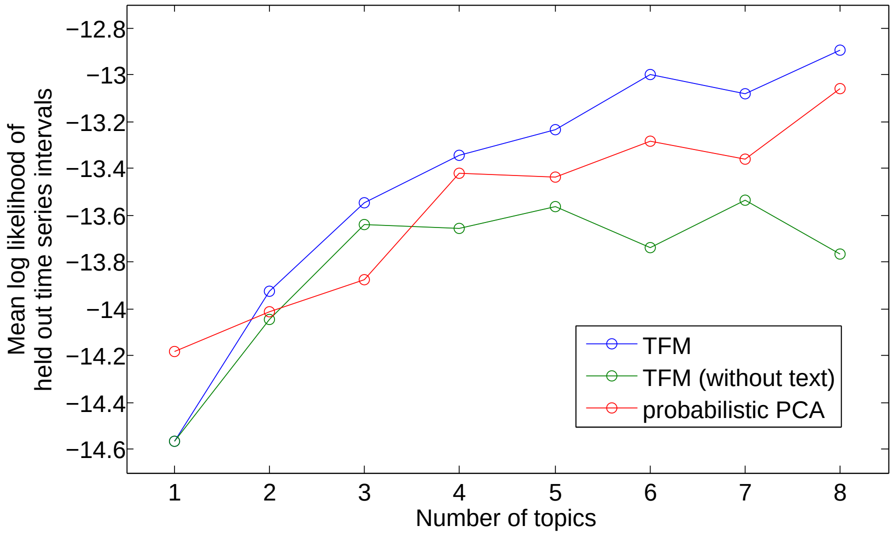

Figure 3: The likelihood of held out data for TFM, TFM without text and pPCA. Note that the highest log likelihood is achieved by the topic factor model including full text data, suggesting that the thematic structure of the text can help better model the time series.

Figure 3 shows that the text holds thematic structure that can help us to better model the comovement of exchange rates (which should be some comfort to the economists responsible for it!). We didn't find similar improvements in the held out likelihood of words, though performance was no worse than LDA. One interesting feature of the results is the poor performance of the text-free topic factor model relative to pPCA. This implies that the Dirichlet structure borrowed from LDA is perhaps less than ideal; having non-negative weights is clearly not the most expressive time series model.

If this ability to robustly estimate correlation proves to aid forward prediction it could be extremely valuable. One could use forward looking economic forecasts for the text. We are keen to apply this method to an equity market dataset, wherein the link between the price time series and analysis text would be more immediate and thus bigger gains in modelling might be possible. Appropriate text is, however, difficult to obtain. Other future directions include adding temporal variation to thematic structure or more structured relationships between topics.

References

[1] D. Blei and J. McAuliffe. Supervised topic models. In *Advances in Neural Information Processing Systems*, volume 20, pages 121–128. 2008.

[2] E.F. Fama and K.R. French. Multifactor explanations of asset pricing anomalies. *The Journal of Finance*, 51(1):55–84, 1996.

[3] D. Mimno and A. McCallum. Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. In *Proceedings of the 24th conference on Uncertainty in artificial intelligence*, pages 411–418, 2008.