TIME-EVOLVING MODELING OF SOCIAL NETWORKS

Eric Wang, Jorge Silva, Rebecca Willett and Lawrence Carin

Department of Electrical and Computer Engineering
Duke University
Durham, NC 27708-0291

ABSTRACT

A statistical framework for modeling and prediction of binary matrices is presented. The method is applied to social network analysis, specifically the database of US Supreme Court rulings. It is shown that the ruling behavior of Supreme Court judges can be accurately modeled by using a small number of latent features whose values evolve with time. The learned model facilitates the discovery of inter–relationships between judges and of the gradual evolution of their stances over time. In addition, the analysis in this paper extends previous results by considering automatic estimation of the number of latent features and other model parameters, based on a nonparametric–Bayesian approach. Inference is efficiently performed using Gibbs sampling.

Index Terms— Machine learning, Predictive models, Bayesian methods, Social networks

1. INTRODUCTION

The problem of modeling time-evolving social network behavior has been the subject of increasing research in recent years [1, 2]. For instance, we might observe every time a subset of people in the network act in concert (e.g. have a meeting or vote together), and from this data we want to estimate the underlying temporally evolving probability distribution.

One way to represent such data is with a large matrix in which each row corresponds to a different person in the network and each column corresponds to a different observation in time. Understanding the structure of this matrix and its connection to probabilistic models of social network behavior is thus closely related to the burgeoning field of matrix completion [3, 4, 5, 6, 7, 8, 9]. However, in general, direct application of these methods to the problem at hand presents several challenges.

1. Our observations are inherently binary – a person votes either "yes" or "no" – while most matrix completion methods focus on real- or complex-valued entries.
2. In many real world settings the behavior of network participants can be modeled using a small number of latent (unobserved) variables ($e.g.$, their political leanings or leisure activities), which are not accounted for in most matrix completion methods.
3. Many matrix completion methods are insensitive to the ordering of the columns (observations). However, in the context of social networks it is more reasonable to assume that behavior is consistent over short time spans but can exhibit drift over longer periods.

This research is funded by the US Office of Naval Research, NSF CAREER Award No. CCF-06-43947 and ARO Grant No. W911NF-09-1-0262.

While the low-rank structure underlying matrix completion methods seems highly appropriate for analyzing social networks, the above limitations necessitate the development of new methods which incorporate binary data, latent variables, and time dependence. In this paper, we describe a Bayesian framework that addresses these challenges.

Our analysis is focused on the database of US Supreme Court rulings, from 1953 to the present. This dataset is important in its own right and has been analyzed in the political science literature [2]. While existing work assumes the existence of only one dominant latent dimension ($e.g.$, political leanings), our model automatically estimates the number of latent dimensions, using a nonparametric–Bayesian methodology. Additionally, unlike most matrix completion methods, our approach learns a posterior density on all model parameters, rather than point estimates. Thus, in a setting where there are missing votes ($e.g.$, in the related problem of voting in a legislative body) which one would like to predict, our model yields estimates of the posterior distributions for the missing values. The statistical model is an extension of factor analysis where we impose sparsity of the factor scores via a beta–Bernoulli construction.

In the remainder of the paper we present a detailed explanation of the model in Section 2 and describe our experiments in Section 3. Concluding remarks are given in Section 4.

2. MODELING FRAMEWORK

Assume we are given a set of binary matrices, $\{\mathbf{B}_t\}_{t=1,T}$, with $\mathbf{B}_t \in \{0,1\}^{N_y^{(t)} \times N_x^{(t)}}$, in which the number of rows and columns, respectively $N_y^{(t)}$ and $N_x^{(t)}$, may vary with time. For example, for the Supreme Court data considered below, time index t corresponds to year; the number N_x of cases and possibly the number N_y of judges change with time (there are nine judges at any time point, but over a year the total number of judges may be greater than nine, if there is a change in Court membership).

The binary matrices are generated according to a probit model, with $\mathbf{B}_t(i,j) = 1$ if $\mathbf{X}_t(i,j) > 0$, and $\mathbf{B}_t(i,j) = 0$ otherwise. The values "0" and "1" have meanings "affirm" and "reverse", respectively, in our Supreme Court application. The latent real–valued matrices are defined as

$$\mathbf{X}_t(i,j) = \left\langle \boldsymbol{y}_i^{(t)}, \boldsymbol{x}_j^{(t)} \right\rangle + \beta_i^{(t)} + \alpha_j^{(t)} + \epsilon_{i,j}^{(t)} \tag{1}$$

where $\langle \cdot, \cdot \rangle$ denotes a vector inner product, and $\epsilon_{i,j}^{(t)} \sim \mathcal{N}(0,1)$. Here $\boldsymbol{y}_i^{(t)}$ corresponds to latent information about judge i at time t, while $\boldsymbol{x}_j^{(t)}$ corresponds to latent information about case j at time t. The dimensionality D of this information is referred to as the "latent dimensionality"; while existing work assumes the existence of only one dominant latent dimension, our model automatically estimates D. The inner product in (1) models the similarity between judge i's values and case issue j. For instance, for $D = 1$ ($y_i^{(t)}$ and $x_j^{(t)}$ scalars), $y_i^{(t)}$ might measure how much judge i wants to protect free speech and $x_j^{(t)}$ might measure how much case j may influence free speech rights. Thus when this inner product is high, judge i has relatively high likelihood of affirming case j.

However, this inner product alone may not accurately predict a judge's vote on a case. To address this shortcoming, we use random effects β_i and α_j which obey $\beta_i^{(t)} \sim \mathcal{N}(0, \lambda_\beta^{-1})$ and $\alpha_j^{(t)} \sim \mathcal{N}(0, \lambda_\alpha^{-1})$, with $\lambda_\alpha \sim \mu_\alpha \delta_\infty + (1 - \mu_\alpha)\text{Gamma}(a, b)$ and $\lambda_\beta \sim \mu_\beta \delta_\infty + (1 - \mu_\beta)\text{Gamma}(a, b)$; δ_∞ is a point measure located at ∞, corresponding to the absence of an associated random effect. The probability of a random effect being active is controlled by μ_β and μ_α, each of which is drawn from a beta distribution.

Random effect α_j is motivated by our example application, for which the index j denotes a specific case that is under examination by the Court; this parameter reflects the "difficulty" of the ruling, and if $|\alpha_j|$ is large, then all judges are likely to vote one way or the other (an "easy" ruling), while if $\alpha_j^{(t)}$ is small the details of the judge (defined by $\boldsymbol{y}_i^{(t)}$) and case (defined by $\boldsymbol{x}_j^{(t)}$) strongly impact the vote. Similarly, index i denotes a specific judge, and random effect β_i reflects the strength of a judge's case–independent preferences. In previous political science Bayesian analysis [1], researchers have simply set $\mu_\beta = 1$ and $\mu_\alpha = 0$, but here we consider a more general setting, and infer these relationships.

Additionally, in previous Bayesian analysis [1] the dimensionality of $\boldsymbol{y}_i^{(t)}$ and $\boldsymbol{x}_j^{(t)}$ has been set (usually to one or two). In related probabilistic matrix factorization (PMF) applied to real matrices [4, 6], priors/regularizers are employed to constrain the dimensionality of the latent features. Here we model each $\boldsymbol{x}_j^{(t)}$ as the product of a Gaussian $\hat{\boldsymbol{x}}_j^t \in \mathbb{R}^K$ and a sparse binary vector $\boldsymbol{b} \in \{0, 1\}^K$; thus $\hat{\boldsymbol{x}}_j^{(t)}$ contains a collection of possible weights, and $\boldsymbol{b}$ selects a sparse subset of those weights. The components b_k of $\boldsymbol{b}$ are drawn $b_k \sim \text{Bernoulli}(\pi_k)$, and $\pi_k \sim \text{Beta}(c/K, d(K-1)/K)$, for K set to a large integer. By setting c and d appropriately, this imposes sparsity by favoring solutions where most of the components of $\boldsymbol{b}$ are zero. Specifically, by integrating out the $\{\pi_k\}_{k=1,K}$, one may readily show that the number of non-zero components in $\boldsymbol{b}$ is a random variable drawn from $\text{Binomial}(K, c/(c + d(K-1)))$, and the expected number of ones in $\boldsymbol{b}$ is $cK/[c + d(K-1)]$. This is related to a draw from a truncated beta-Bernoulli process [10]. Hence, our estimate of the latent dimensionality D is equal to the number of nonzero components of $\boldsymbol{b}$.

We consider two types of matrix indices (row vs. column indices). Specifically, we assume that each row corresponds to a person/entity that may be present for matrix $t+1$ and matrix t; we also assume that each column corresponds to a question (in the examples, a case), and each question is unique. Since the columns are each unique, we assume

$$\boldsymbol{x}_j^{(t)} = \boldsymbol{b} \circ \hat{\boldsymbol{x}}_j^{(t)}, \quad \hat{\boldsymbol{x}}_j^{(t)} \sim \mathcal{N}(\boldsymbol{0}, \gamma_x^{-1}\mathbf{I}_K), \qquad \gamma_x \sim \text{Gamma}(e, f), \tag{2}$$

where $\circ$ denotes the elementwise/Hadamard vector product. If the person/entity associated with the ith row at time t is introduced for the first time, as in, *e.g.*, a newly appointed judge, its associated feature vector is similarly drawn $\boldsymbol{y}_i^{(t)} = \boldsymbol{b} \circ \hat{\boldsymbol{y}}_i^{(t)}$, $\hat{\boldsymbol{y}}_i^{(t)} \sim \mathcal{N}(\boldsymbol{0}, \gamma_y^{-1}\mathbf{I}_K)$, with $\gamma_y \sim \text{Gamma}(e, f)$. However, assuming $\boldsymbol{y}_i^{(t)}$ has already been drawn (person/entity i is active prior to time $t+1$), then a simple auto-regressive model is used to draw $\boldsymbol{y}_i^{(t+1)}$:

$$\boldsymbol{y}_i^{(t+1)} = \boldsymbol{b} \circ \hat{\boldsymbol{y}}_i^{(t+1)}, \quad \hat{\boldsymbol{y}}_i^{(t+1)} \sim \mathcal{N}(\boldsymbol{y}_i^{(t)}, \xi^{-1}\mathbf{I}_K), \qquad \xi \sim \text{Gamma}(g, h). \tag{3}$$

The prior on ξ is set to favor small/smooth changes in the features of an individual on consecutive years.

This model constitutes a relatively direct extension of existing techniques for real matrices [4, 6]. Specifically, we have introduced a probit link function and a simple auto-regression construction to impose statistical correlation in the traits of a person/entity at consecutive times. The introduction of random effects has also not been considered within much of the machine-learning matrix-analysis literature, but it is standard in political science Bayesian models [1]. A graphical representation of the model is shown in Figure 1.

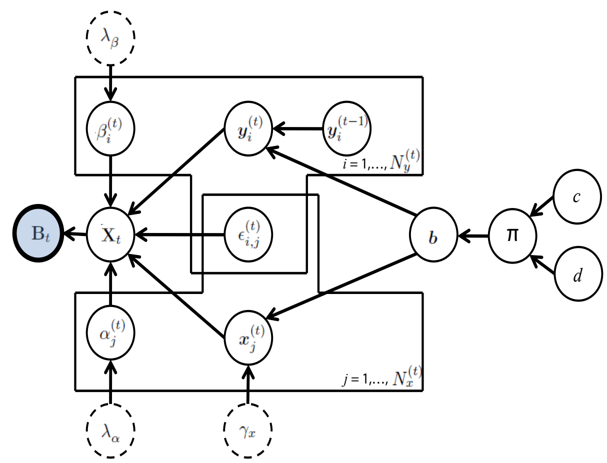


Fig. 1. Graphical representation of the proposed Bayesian model. The plates indicate repetition, and the filled circle around $\mathbf{B}_t$ means that its value is observed. The dashed circles indicate random variables that are gamma distributed, with the corresponding hyperparameters omitted from the plot for simplicity.

3. EXPERIMENTS

We have used the Supreme Court database of rulings from 1953 to 2008. The data is available from the Supreme Court Judicial Database [11] at `http://scdb.wustl.edu/data.php` and consists of 7217 cases and 31 judges. Following [2], we preprocessed the original data so that, instead of depending on a so–called *disposition* variable, which expresses agreement/disagreement with the majority, our binary values are "0" for "reverse" and "1" for "affirm". However, unlike [2], we did not remove unanimous cases, due the additional flexibility afforded by our random effect variables. The average number of cases per year was 128; 2007 had the fewest number of cases at 70, and 1967 had the most at 197 cases. We present results obtained using the procedure described in Section 2.

All gamma distribution hyperparameters (a, b, e, f, g and h) were set to 10^{-6}, which is a standard "broad" prior choice. The beta hyperparameters for π_k were set to $c = d = 1$, and the truncation level for the number of latent dimensions was set to nine. The model inferred two significant latent dimensions, one of which is dominant (has the largest variance). This first dimension seems to correspond

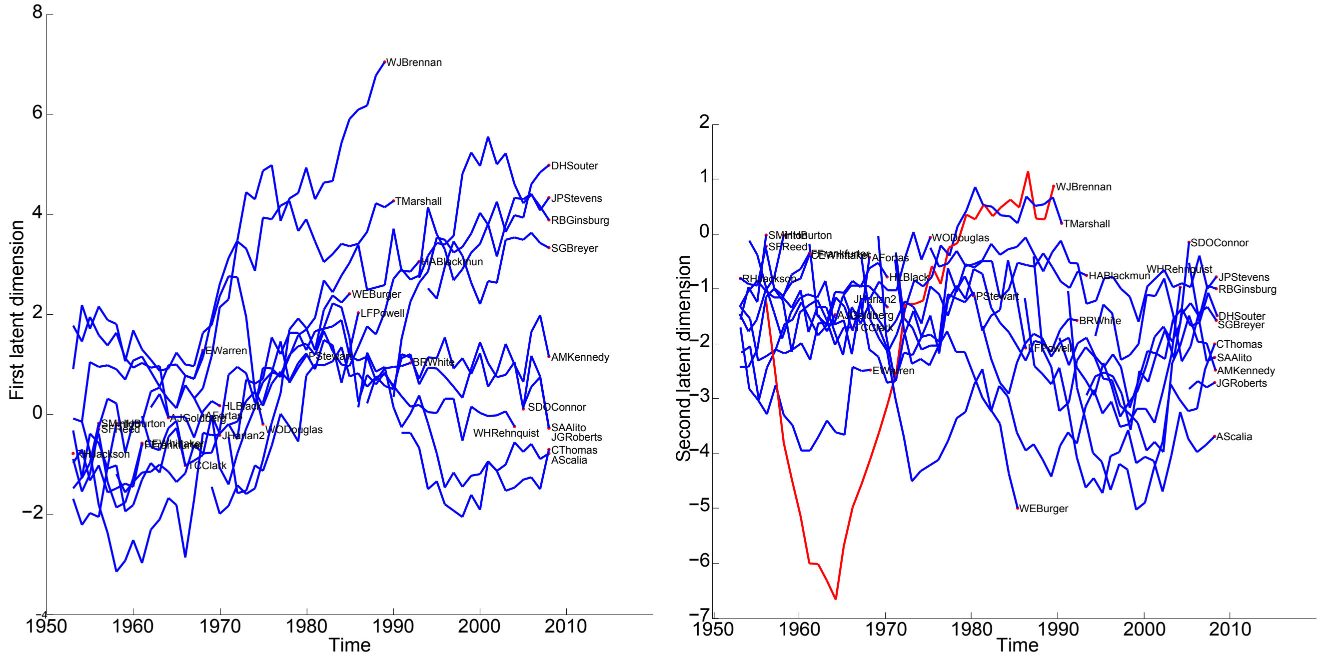


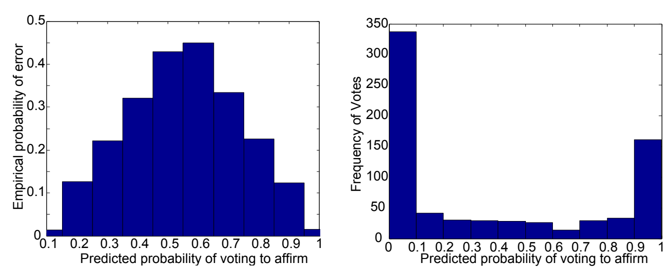
Fig. 2. (a) Trajectories of the judges in the first latent dimension, plotted vs. time. This latent dimension corresponds to a liberal–conservative axis. (b) Trajectories of the judges in the second latent dimension. The trajectory of Justice Brennan, which makes a large excursion during the decade of 1960–1970 (not apparent in the first latent dimension), is drawn in red.

to a liberal–conservative axis, in agreement with previous findings [2]. The second dimension also has significant variance, and its possible existence has been hinted at in [2] but, to our knowledge, our model is the first to automatically infer the second dimension for this dataset.

We executed the Gibbs inference procedure for 20,000 iterations in all cases, of which 15,000 are burn–in and 5,000 are collection iterations. The total run time was approximately eight hours for the entire dataset (the code is in Matlab, with no optimization, and the computations were run on a cluster of blade-based servers with 2.5 GHz clock frequency, eight CPU-cores and 16 Gb shared RAM), and Singular Value Decomposition (SVD) was employed to initialize each year.

Figure 2 shows the trajectories of each judge in the two inferred latent dimensions, plotted versus time. By examining Figure 2a it is possible to rank the judges according to their position in the liberal–conservative axis, corresponding to the first latent dimension. It is also clear that the preferences of the judges along this axis drift with time. Figure 2b shows the trajectories in the second latent dimension. We emphasize the trajectory of Justice W. J. Brennan, which shows considerable departure from the trajectories of most fellow Justices during the decade of 1960–1970.

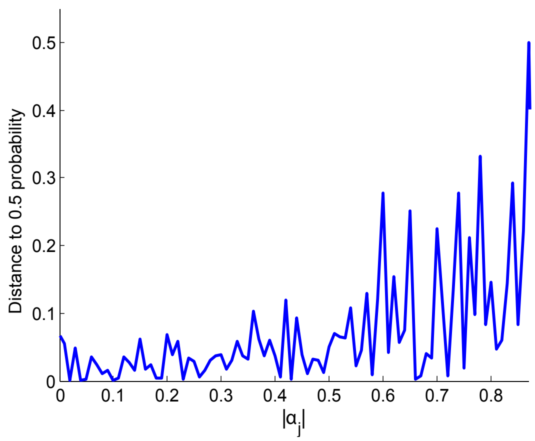
Figure 3 depicts the accuracy of the probit model in predicting the ruling behavior of the judges. In Figure 3a we plot the average empirical probabilities of error between the known binary matrix values $\mathbf{B}_t(i,j)$ and the predictions obtained by thresholding Gibbs samples of $\mathbf{X}_t(i,j)$ at zero. The predicted probit probabilities are computed as the cumulative density function (cdf), evaluated at $\mathbf{X}_t(i,j)$, of a standard Gaussian distribution. The plot in Figure 3a indicates that the model is correctly predicting the probabilities of the rulings, in the sense that when the predicted probability of voting "affirm" approaches 0.5, so does the empirical error probability; when the predicted probabilities of "affirm" approach either 0.1 or 0.9, the empirical error probabilities quickly decrease below 0.1, reflecting the fact that these rulings are less ambiguous (there is more certainty regarding the preferences of the judges). Moreover, Figure 3b indicates that most of the rulings exhibit high certainty, , and in these cases our method has particularly low error as displayed in Figure 3a.



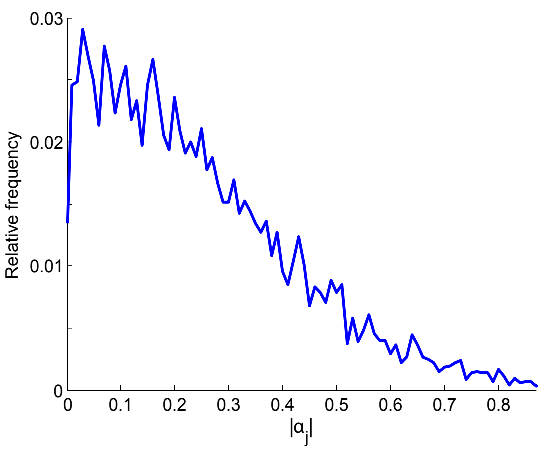
(a) (b)

Fig. 3. (a) Empirical probability of error vs. predicted probability of voting to affirm according to the probit model. The plot indicates that the empirical error probabilities are consistent with the probit model. (b) Frequency of cases vs. predicted probability. Confronting with (a), it is evident that the majority of votes fall in the low–error probability bins.

The random effects have a significant role in the model. We achieved best results by setting $\mu_\beta = 1$, which leads to the inactivity of the judge–related random efect β_i, while still inferring μ_α and the associated random effect α_j. Strong case–independent judge preferences do not seem evident in our analysis of this dataset (with this of course expected). Figure 4 illustrates the relationship between α_j and the probit probabilities, as well as a histogram of α_j. It is shown that the magnitude of α_j is larger for cases where the probabilities are farther from 0.5, $i.e.$, there is more certainty in the deliberation. Additionally, Figure 4b indicates that the random effect on the cases is often small, which is expected as well, as the Supreme Court often deals with cases that are "hard" and yield a high degree of disagreement (5-4 decisions).



(a)



(b)

Fig. 4. (a) Distance to 0.5 probit probability vs. the magnitude of the case–related random effect α_j. Larger $|\alpha_j|$ is associated with probabilities farther from 0.5. (b) Histogram of $|\alpha_j|$.

4. CONCLUSION

A new model for analysis of time–evolving social networks is presented and utilized in the analysis of the US Supreme Court dataset. The model is more flexible than previous work, allowing automatic inference of multiple latent dimensions as well as random effects. In a setting where some of the votes are missing, the proposed method offers advantages over most matrix completion methods by yielding estimates of the posterior distributions for the missing values. Moreover, the model considers time dependence and enforces smooth evolution of the network participants over time. Inference is efficiently performed using Gibbs sampling. Experiments show that the method automatically finds two latent dimensions, which appears consistent with political science literature.

5. REFERENCES

[1] J. Cinton, S. Jackman, and D. Rivers, "The statistical analysis of roll call data," *Am. Political Sc. Review*, 2004.

[2] A.D. Martin and K.M. Quinn, "Dynamic ideal point estimation via Markov chain Monte Carlo for the US Supreme Court, 1953–1999," *Political Analysis*, vol. 10, no. 2, pp. 134, 2002.

[3] E. Meeds, Z. Ghahramani, R. Neal, and S. Roweis, "Modeling dyadic data with binary latent factors," in *Advances in NIPS*, 2007, pp. 977–984.

[4] N. Srebro, J.D.M. Rennie, and T.S. Jaakkola, "Maximum-margin matrix factorization," in *Advances in NIPS*, 2005.

[5] J. Abernethy, F. Bach, T. Evgeniou, and J.-P. Vert, "A new approach to collaborative filtering: operator estimation with spectral regularization," *J. Machine Learning Research*, 2009.

[6] R. Salakhutdinov and A. Mnih, "Bayesian probabilistic matrix factorization with MCMC," in *Advances in NIPS*, 2008.

[7] R. Salakhutdinov and A. Mnih, "Probabilistic matrix factorization," in *Advances in NIPS*, 2008.

[8] K. Yu, J. Lafferty, S. Zhu, and Y. Gong, "Large-scale collaborative prediction using a nonparametric random effects model," in *Proc. Int. Conf. Machine Learning*, 2009.

[9] E.J. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2053–2080, 2010.

[10] R. Thibaux and M.I. Jordan, "Hierarchical beta processes and the Indian buffet process," in *International Conference on Artificial Intelligence and Statistics*, 2007.

[11] H. Spaeth, "United State Supreme Court Judicial Database, 1953-2000 Terms [Computer File], 16th ed.," *Ann Arbor, MI: Inter-University Consortium for Political and Social Research*, 2001.