Learning a Graphical Model of Bloomberg Financial and News Data

Kui Tang, Henrique Gubert, Rashmi Tonge, Anyi Wang, Liang Wu, Dwayne Campbell
Chris Kedzie, Liao Wang, Andelyn Russell, Anthony Kimball, Anju Kambadur, Gideon Mann
Stefano Pacifico, James Hodson, David Da-Wei Yao, Kathleen McKeown, Tony Jebara,
Columbia Data Science Institute and Bloomberg LLC.

Abstract

We build a Bayesian network that models interactions between heterogeneous data sources including news feeds, social media and financial indices. We also propose a method for temporal adjustment using conjugate priors. We use this network to do inference about the different variables of the model under stress conditions.

1 Problem Formulation and Proposed Method

The goal of this project is to apply Markov Random Fields to better understand and model interactions between news, social media and economy.

We extracted binary variables from news and social media using natural language processing methods, and financial and economic indicators. The data is structured as an hourly multi-variate time-series from 2009 to 2014.

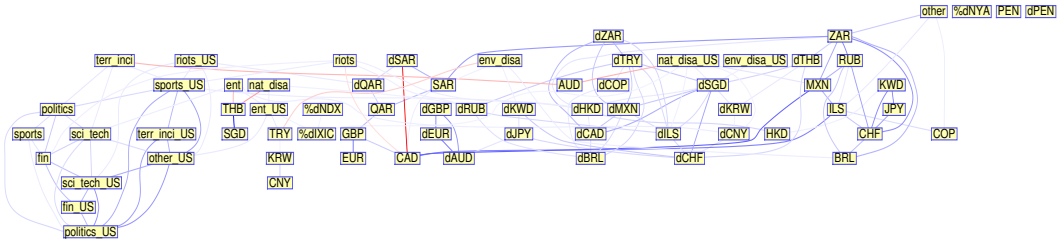


Figure 1: Estimated Graphical Model from partial hourly dataset; dark blue edges represent highly positive correlation, while dark red edges highly negative correlation. The true and approx. test log likelihood of this model w.r.t. dataset are -44.3310 and -40.1648 respectively.

1.1 Structure Learning

The first step to build a graphical model from the available samples is to learn the graph structure. We use the algorithm described in (1) consisting of using l1-regularized logistic regression trying to predict each of the variables as a function of the others. We learn the structure of the graph by thresholding the edge potentials.

1.2 Parameter Estimation and Model Selection

The parameter estimation problem is to learn the model parameters given the graphical network structure. We formulate the problem as maximum likelihood estimation (MLE) (2). The central challenge in MLE is computing the log-partition function $\log Z(X^{(m)}, \theta)$ for each sample m at each iteration. We use the Bethe Approximation, a standard approximation to the Gibbs free energy. We focus on the restricted set of counting numbers that yield *convex* reweighted free energies. The *ρ-reweighted free energy* is specified by a polytope approximation $\mathcal{T}$, a hypergraph $G = (V, \mathcal{A})$, an entropy approximation H^{η}_{ρ}, a parameter $\eta \in [0, 1/2)$, and a vector of counting numbers (a.k.a reweighting parameters) ρ:

$$\log F^{\eta}_{\rho}(\tau, X; \theta) \triangleq E(\tau, X; \theta) - H^{\eta}_{\rho}(\tau), \tag{1}$$

where the energy is given by

$$E(\tau, X; \theta) \triangleq -\sum_{i \in V} \langle \sum_{Y_i} \tau_i(Y_i) \phi_i(X, Y_i), \theta_V \rangle - \sum_{\alpha \in \mathcal{A}} \langle \sum_{Y_\alpha} \tau_\alpha(Y_\alpha) \phi_\alpha(X, Y_\alpha), \theta_{\mathcal{A}} \rangle \tag{2}$$

and the entropy approximation is given by

$$H^{\eta}_{\rho}(\tau) \triangleq -\sum_{i \in V} \sum_{y_i} \left(1 - \sum_{\alpha \supset i} \rho_\alpha\right) g_\eta(\tau_i(y_i)) - \sum_{\alpha \in \mathcal{A}} \sum_{y_\alpha} \rho_\alpha g_\eta(\tau_\alpha(y_\alpha)), \tag{3}$$

where $g_\eta(x) = x \log(x)$ and τ is restricted to lie in the local polytope,

$$\mathcal{T} \triangleq \{\tau \geq 0 : \text{for all } i \in V, \sum_{Y_i} \tau_i(Y_i) = 1, \text{for all } \alpha \in \mathcal{A}, i \in \alpha, Y_i, \sum_{Y_{\alpha \setminus \{i\}}} \tau_\alpha(Y_\alpha) = \tau_i(Y_i)\}. \tag{4}$$

The reweighted log-partition function is then computed by minimizing (1) over $\mathcal{T}$

$$\log Z^{\eta}_{\rho}(X; \theta) \triangleq -\min_{\tau \in \mathcal{T}} F^{\eta}_{\rho}(\tau, X; \theta). \tag{5}$$

Setting $\rho_\alpha = 1$ for each $\alpha \in \mathcal{A}$ recovers the typical Bethe free energy approximation. The reweighting parameters can always be chosen so that the approximate free energy is convex.

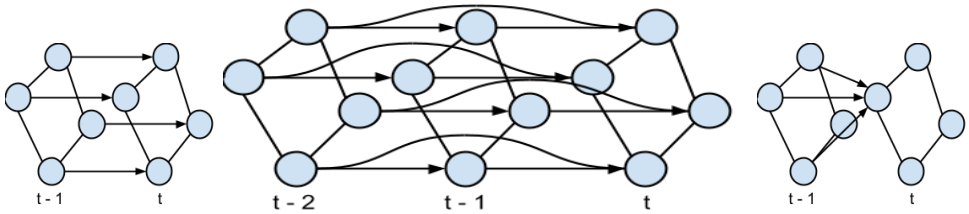
In structure learning and parameter estimation steps, multiple models are trained by over a grid of the regularization and threshold parameters. We select the model with the highest log test likelihood computed with the approximate partition function as follows:

$$\log \mathcal{L} = \frac{1}{N_{\text{test}}} \sum_{i \in V} \theta_i x_i + \sum_{ij \in E} \theta_{ij} x_i x_j - \log Z_B \tag{6}$$

Experimental analysis shows that both true and approximate likelihood vary similarly over hyper-parameters and using right heuristic we can get the model consistent with the highest true log test likelihood.

1.3 Adding Temporal Dependency

The method above assumes all the samples are independent and identically distributed (IID), which is practically unrealistic assumption for multivariate time series. We use temporal conjugate prior, where based on several different topologies shown in Figure 2, we are able to put a prior on the θ parameters, and model part of the temporal dependencies between the variables.



(a) (b) (c)

Figure 2: Figure shows different toplogies used for temporal modelling. Figures (a), (b) and (c) show First Order Markov Chain, Second Order Markov Chain and Neighborhood Chain respectively.

Addition of temporal dependencies this way does not change the algorithm complexity and has been verified to improve performance in higher level than more naive approaches like ensembles of temporal and non-temporal models.

References

[1] Pradeep Ravikumar, Martin J. Wainwright, and John D. Lafferty, *High-dimensional Ising model selection using l1-regularized logistic regression*, Annals of Statistics, Volume 38, Number 3, 2010, 1287-1319.

[2] Kui Tang, Nicholas Ruozzi, David Belanger, Tony Jebara, *Bethe Learning of Conditional Random Fields via MAP decoding*, arXiv:1503.01228v1, 2015.

[3] Martin Jaggi, *Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization*, ICML, Volume 28, 2013, 427-435.