

Estimation of Sparse Binary Markov Networks

Holger Höfling PhD thesis

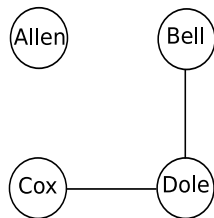
November 30, 2008

Graphical models

- Model of joint distribution of a set of random variables
- Graph represents dependencies among random variables
- Two main types of graphical models
 - Directed acyclic graph (DAG); known as Bayesian network
 - **Here:** Undirected graph; known as Markov network or Markov random field
- Very useful in many applications
 - Speech recognition
 - Modeling of gene regulatory networks
 - Modeling of genetic variation (e.g. HapMap data)

Example: Voting

- A board consists of 4 people that can vote
- Several rounds of voting data available
- Possible questions:
 - Do blocks of voters exists?
 - How strongly do the blocks vote together?



Allen	Bell	Cox	Dole
Yes	No	No	No
Yes	Yes	No	Yes
No	No	No	No
No	Yes	Yes	Yes

Pairwise Binary Markov network

- Underlying graph $\mathcal{G} = (V, E)$
- Data are binary random vectors $x = (x_1, \dots, x_p)^T \in \{0, 1\}^p$
- Parameter matrix $\Theta \in \mathbb{R}^{p \times p}$, symmetric
- $(u, v) \in E$ iff $\theta_{uv} \neq 0$
- Distribution given by

$$\log p(x, \Theta) = \sum_{s \geq t=1}^p \theta_{st} x_s x_t - \Psi(\Theta)$$

- $\Psi(\Theta)$ is the log-normalization constant; also known as partition function

Partition function

- Partition function defined as

$$\psi(\Theta) = \log \left(\sum_{x \in \{0,1\}^p} \exp \left(\sum_{s \geq t} \theta_{st} x_s x_t \right) \right)$$

- Partition function in general requires to sum over 2^p elements
- Inference in general model prohibitively expensive

Pairwise Binary Markov network

- Faster algorithms that exploit sparse graph structure exist
 - Exact: e.g. Junction Tree algorithm
 - Approximate: e.g. Loopy Belief Propagation; MCMC
- Use L_1 penalized log-likelihood
- (Lee, Ganapathi & Koller 2007) proposes using L_1 penalized log-likelihood to get sparse graphs
- Also derives an exact procedure to maximize penalized log-likelihood
- (Wainwright, Ravikumar & Lafferty 2007) suggest approximate procedure

Goals

- Develop a fast algorithm
- Find approximate procedure that can be extended to give exact results
- Compare accuracy of approximate procedures to exact results

(Wainwright et al. 2007) and (Lee et al. 2007)

(Wainwright et al. 2007): Estimate row i of Θ by a penalized logistic regression of X_i onto $X_{\setminus i}$, i.e.

$$X_i \sim \text{Bernoulli}(p) \quad \text{with} \quad \text{logit}(p) = \theta_{ii} + \sum_{j \neq i} x_j \theta_{ij}.$$

then symmetrize Θ . We use 2 methods to make Θ symmetric, referred to as Wainwright-min and Wainwright-max.

(Lee et al. 2007): Optimizes the L_1 penalized log-likelihood by optimizing over reduced variable set F , that is being extended by grafting.

Graphical lasso idea works, but is too slow. Have to make too many evaluations of the log-likelihood.

Pseudo-likelihood

- Use pseudo-likelihood function (see (Besag 1975)) instead of likelihood:

$$\tilde{l}(\Theta|x) = \sum_{s=1}^p \log p(x_s, \Theta|x_{\setminus s})$$

where

$$\log p(x_s, \Theta|x_{\setminus s}) = x_s(\theta_{ss} + \sum_{s \neq t} x_t \theta_{st}) - \Psi_s(x, \Theta)$$

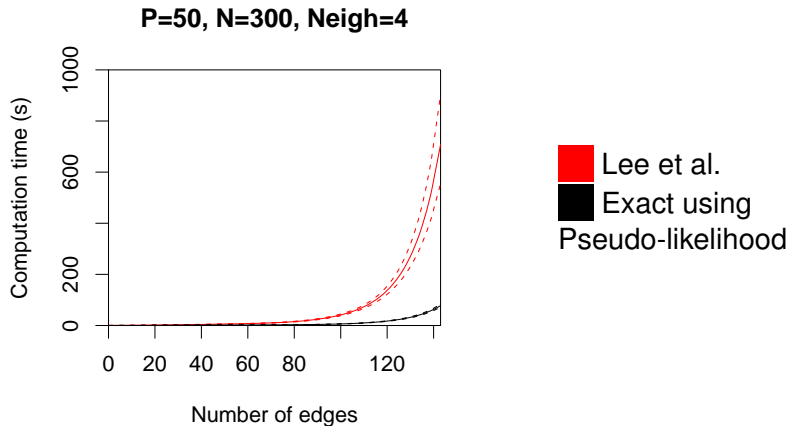
with $\Psi_s(x, \Theta) = \log(1 + \exp(\theta_{ss} + \sum_{t \neq s} x_t \theta_{st}))$, the normalization constant from logistic regression

- Different than (Wainwright et al. 2007) as optimization is jointly over all of Θ instead of only one vector at a time
- No need to use min or max rule

Simulation setup

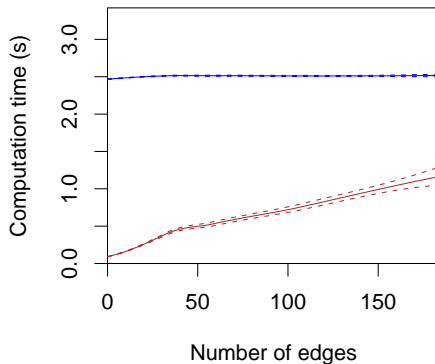
- Use $p = 50$ random variables
- Draw sparse random $\Theta \in \mathbb{R}^{50 \times 50}$ with
 - Diagonal elements uniformly from $\{-0.5, 0, 0.5\}$
 - Edges at random s.t. on average every node has 4 neighbours
 - Weights -0.5 or 0.5 uniformly on edges
- Using Θ generate $n = 300$ observations using Gibbs sampling

Speed comparison



Speed comparison

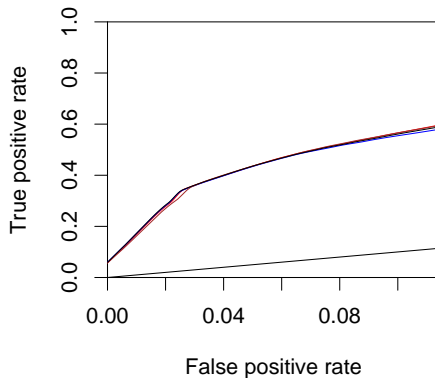
P=50, N=300, Neigh=4



Wainwright et al.
Pseudo-likelihood

ROC curve

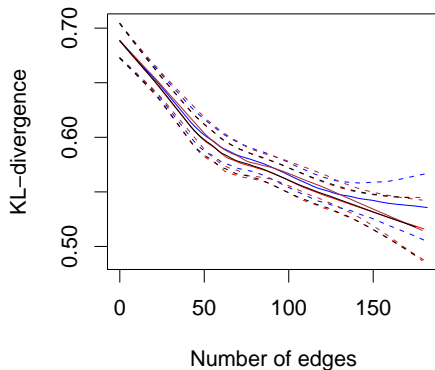
P=50, N=300, Neigh=4



- Exact
- Wainwright-min
- Wainwright-max
- Pseudo-likelihood

Kullback-Leibler divergence

P=50, N=300, Neigh=4







- Exact
- Wainwright-min
- Wainwright-max
- Pseudo-likelihood

Conclusion

- Our algorithm is faster than the competing exact method of (Lee et al. 2007)
- (Wainwright et al. 2007) and pseudo-likelihood methods are **much** faster and only slightly less accurate for sparse graphs
- In small models, use exact method
- If application time sensitive or model larger, use pseudo-likelihood method

Possible Extensions

- Belief Nets (Geoff Hinton)- multi-layer networks with layers of hidden (unobserved) binary units.

-  Besag, J. (1975), 'Statistical analysis of non-lattice data', *The Statistician* **24**(3).
-  Friedman, J., Hastie, T. & Tibshirani, R. (2008), 'Regularization paths for generalized linear models via coordinate descent', *Submitted*.
-  Lee, S.-I., Ganapathi, V. & Koller, D. (2007), Efficient structure learning of Markov networks using L1-regularization, *in* 'Advances in Neural Information Processing Systems (NIPS 2006)'.
-  Wainwright, M., Ravikumar, P. & Lafferty, J. (2007), High-dimensional graphical model selection using ℓ_1 -regularized logistic regression, *in* 'Presented at Advances in Neural Information Processing Systems 2006, Vancouver'.