Latent Variable Graphical Modeling in Generalized Linear Models

Armeen Taeb (Caltech)

Joint with

Venkat Chandrasekaran (Caltech), Parikshit Shah (Facebook)

CA Water Reservoir Network

Reservoirs are central source of water

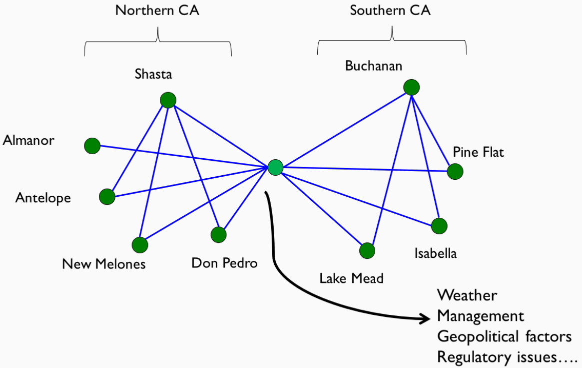
- buffer against severe drought

Question: *what is the likelihood of system-wide catastrophe (i.e. multiple reservoirs exhausting)*?

To answer this question, must characterize

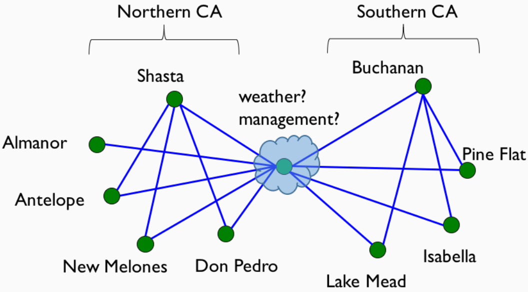
1. effect of external factors
2. reservoir interdependencies

Graphical Modeling



Nodes: random variables ; Edges: conditional dependencies

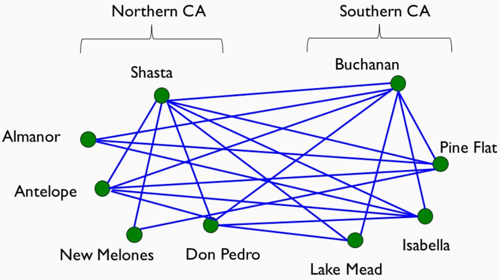
Graphical Modeling



Nodes: random variables ; Edges: conditional dependencies

Challenge: many external factors are not observed or latent

Graphical Modeling



Nodes: random variables ; Edges: conditional dependencies

Challenge: many external factors are not observed or latent

- Not accounting for them leads to false negatives and false positives

Gaussian Graphical Modeling with Latent Vars.

Gaussian graphical models with latent vars [**C** et al '12]

- Precision matrix = sparse + low-rank
- Convex estimator to find conditional graph structure + latent effect
- Approach to interpret latent vars. [**T** & Chandrasekaran '18]

Gaussian Graphical Modeling with Latent Vars.

Gaussian graphical models with latent vars [**C** et al '12]

- Precision matrix = sparse + low-rank
- Convex estimator to find conditional graph structure + latent effect
- Approach to interpret latent vars. [**T** & Chandrasekaran '18]

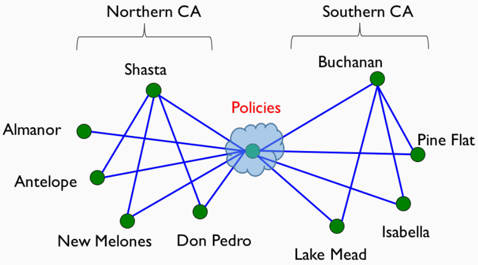
Model findings [**T** et al '18]

- reservoir dependency structure
- external factors: drought index, hydropower, snowpack
- system-wide response to these factors

Validity of the Gaussian Approximation

Some of the variables might deviate strongly from Gaussianity

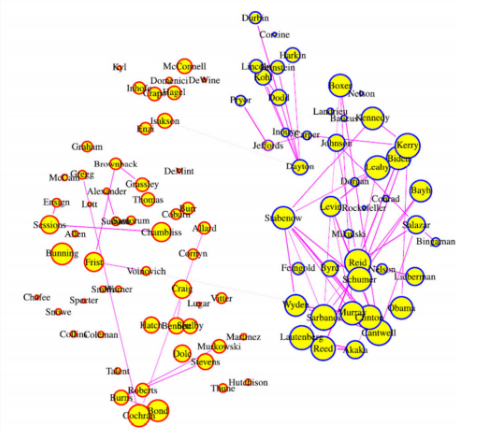
- Reservoirs



Validity of the Gaussian Approximation

Some of the variables might deviate strongly from Gaussianity

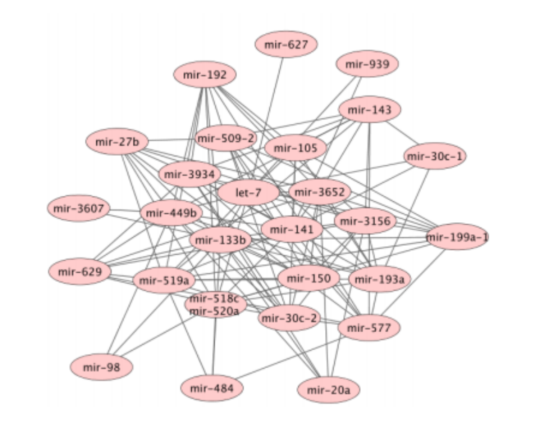
- Voter records data



Validity of the Gaussian Approximation

Some of the variables might deviate strongly from Gaussianity

- RNA sequence count data



Beyond Gaussian Graphical Models

State of the art suffers from at least one of these deficiencies:

- unable to handle non-Gaussianity
- non convexity (e.g. EM) or computationally intractable
- cannot account for latent variables

We address all three:

- based on Generalized linear models

Modeling Framework

Observed $x \in \mathbb{R}^p$; latent variables $z \in \mathbb{R}^k$ from the class

$$\mathscr{P} = \left\{ \Pr(x|z) = \exp\left(\left[\alpha^T x + \sum_{s=1}^{p} \Lambda_{s,s} f(x_s) \right] + \frac{1}{2} x^T K x + x^T B z + A \right) \right\}$$

- $\alpha \in \mathbb{R}^p$: parameters encoding linear effect of observed vars.
- $\Lambda \in \mathbb{R}^{p \times p}$: node potential
- K: graph structure $K_{s,s} = 0$ and $K_{s,t} = 0$ when s, t disconnected
- $B \in \mathbb{R}^{p \times k}$ encoding latent effect
- $A(\alpha, \Gamma, K, B)$ is a normalization constant

Modeling Framework

Observed $x \in \mathbb{R}^p$; latent variables $z \in \mathbb{R}^k$ from the class

$$\mathscr{P} = \left\{ \Pr(x|z) = \exp\left(\left[\alpha^T x + \sum_{s=1}^{p} \Lambda_{s,s} f(x_s) \right] + \frac{1}{2} x^T K x + x^T B z + A \right) \right\}$$

- Gaussian: $f(x_s) = x_s^2$, Λ diagonal encoding conditional variance
- Bernoulli: $f(x_s) = 0$
- Poisson: $f(x_s) = -\log(x_s)$, $K \leq 0$, $B \leq 0$, Λ identity
- Exponential: $f(x_s) = 0$, $K \leq 0$, $B \leq 0$, $\alpha < 0$, Λ identity

Modeling Framework

Observed $x \in \mathbb{R}^p$; latent variables $z \in \mathbb{R}^k$ from the class

$$\mathscr{P} = \left\{ \mathrm{Pr}(x|z) = \exp\left(\left[\alpha^T x + \sum_{s=1}^{p} \Lambda_{s,s} f(x_s) \right] + \frac{1}{2} x^T K x + x^T B z + A \right) \right\}$$

Conditional distribution is a *Generalized Linear Model*

$$\mathscr{P}_{\mathsf{cond}} = \left\{ p(x_s | x_{\sim s}, z) = \exp\Big(\Lambda_{s,s} f(x_s) + x_s \eta_s - D(\eta_s) \Big) \right\}$$

- predictor: $\eta_s = \alpha_s + e_s^T K x + e_s^T B z$
- convex link function: $\mathcal{D}(\eta_s)$

Naive Inference

Given observations x, z, minimize the negative log-likelihood

$$\hat{\theta} = \arg\min_{\theta} -\log\left[\Pr(x|z)\right]$$

Where

$$-\log\left[\Pr(x|z)\right] \quad = \quad \left[\alpha^T x + \sum_{s=1}^{p} \Lambda_{s,s} f(x_s)\right] + \frac{1}{2} x^T K x + x^T B z + A$$

Naive Inference

Given observations x, z, minimize the negative log-likelihood

$$\hat{\theta} = \arg\min_{\theta} -\log\left[\Pr(x|z)\right]$$

Where

$$-\log\left[\Pr(x|z)\right] = \left[\alpha^T x + \sum_{s=1}^{p} \Lambda_{s,s} f(x_s)\right] + \frac{1}{2} x^T K x + x^T B z + A$$

Challenge: norm. constant $A(K, B, z, \alpha)$ intractable to compute

- e.g. 2^p computations with Bernoulli variables

Psuedo-likelihood

Exact inference for full likelihood may be computationally costly

Psuedo-likelihood MLE [Besag '75]

$$\underbrace{\min_{\theta} - \log(\Pr(x|\theta))}_{\text{MLE}} \approx \underbrace{\min_{\theta} \sum_{s=1}^{p} - \log\left(\Pr(x_s|x_{-s};\theta)\right)}_{\text{psuedo-MLE}}$$

Psuedo-likelihood

Exact inference for full likelihood may be computationally costly

Psuedo-likelihood MLE [Besag '75]

$$\underbrace{\min_\theta - \log(\Pr(x|\theta))}_{\text{MLE}} \approx \underbrace{\min_\theta \sum_{s=1}^{p} - \log\left(\Pr(x_s|x_{-s};\theta)\right)}_{\text{psuedo-MLE}}$$

Computational and statistical tradeoff:

- Pseudo-MLE statistically consistent but may be less efficient than MLE [Liang & Jordan '08]
- Pseudo-MLE computationally more efficient

Commonly employed for graphical modeling without latent variables

- e.g. neighborhood selection [Meinshausen & Buhlmann '08]

Psuedo-likelihood Formulation

Psuedo-likelihood approximation for $\theta = (B, z, K, \alpha)$:

$$\min_\theta - \log(\Pr(x|z;\theta)) \approx \min_\theta \sum_{s=1}^{p} -x_s \eta_s(\theta) + D(\eta_s(\theta))$$

- linear predictor $\eta_s(\theta) = \alpha_s + e_s^T Kx + e_s^T Bz$
- convex link $D(\eta_s(\theta))$

Psuedo-likelihood Formulation

Psuedo-likelihood approximation for $\theta = (B, z, K, \alpha)$:

$$\min_\theta - \log(\Pr(x|z;\theta)) \approx \min_\theta \sum_{s=1}^{p} -x_s \eta_s(\theta) + D(\eta_s(\theta))$$

- linear predictor $\eta_s(\theta) = \alpha_s + e_s^T K x + e_s^T B z$
- convex link $D(\eta_s(\theta))$

Psuedo-likelihood estimator for observations $X \in \mathbb{R}^{p\times n}$ and $Z \in \mathbb{R}^{k\times n}$

$$\begin{array}{ll} \underset{\theta, M}{\arg\min} & \frac{1}{n}\sum_{i=1}^{n}\sum_{s=1}^{p} -X_{s,i}M_{s,i} + D(M_{s,i}) \\ \text{subject-to} & M = KX + BZ + \alpha\mathbf{1}' \;\; ; \;\; K \text{ symmetric} \;\; ; \;\; K_{s,s} = 0 \end{array}$$

Psuedo-likelihood Formulation

Psuedo-likelihood approximation for $\theta = (B, z, K, \alpha)$:

$$\min_{\theta} -\log(\Pr(x|z;\theta)) \approx \min_{\theta} \sum_{s=1}^{p} -x_s \nabla_{x_s} \mathcal{L}(\theta; x) + D(\mathcal{L}(\theta; x))$$

- linear predictor $\eta_s(\theta) = \alpha_s + e_s^T K x + e_s^T B z$
- convex link $D(\eta_s(\theta))$

Psuedo-likelihood estimator for observations $X \in \mathbb{R}^{p \times n}$ and $Z \in \mathbb{R}^{k \times n}$

$$\begin{array}{ll} \underset{\theta, M}{\arg\min} & \frac{1}{n} \sum_{i=1}^{n} \sum_{s=1}^{p} -X_{s,i} M_{s,i} + D(M_{s,i}) \\ \text{subject-to} & M = KX + BZ + \alpha \mathbf{1}' \;\; ; \;\; K \text{ symmetric} \;\; ; \;\; K_{s,s} = 0 \end{array}$$

convex objective but not a convex program

Convex Estimator

Observation: $L = BZ$ has rank less than or equal to k

Convex estimator:

$$\begin{array}{rc} \underset{K,\alpha,L,M}{\arg\min} & \frac{1}{n}\sum_{i=1}^{n}\sum_{s=1}^{p} -X_{s,i}M_{s,i} + D(M_{s,i}) + \lambda(\|K\|_1 + \gamma\|L\|_\star) \\ \text{subject-to} & M = KX + L + \alpha\mathbf{1}' \;\; ; \;\; K \text{ symmetric} \;\; ; \;\; K_{s,s} = 0 \end{array}$$

Convex Estimator

Observation: $L = BZ$ has rank less than or equal to k

Convex estimator:

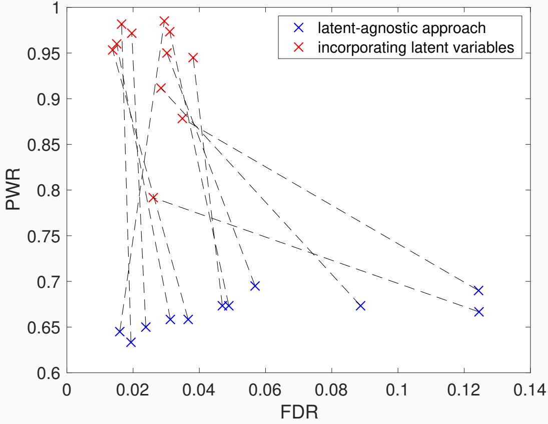
$$\begin{array}{rc} \underset{K,\alpha,L,M}{\arg\min} & \frac{1}{n}\sum_{i=1}^{n}\sum_{s=1}^{p} -X_{s,i}M_{s,i} + D(M_{s,i}) + \lambda(\|K\|_1 + \gamma\|L\|_\star) \\ \text{subject-to} & M = KX + L + \alpha\mathbf{1}' \;\; ; \;\; K \text{ symmetric} \;\; ; \;\; K_{s,s} = 0 \end{array}$$

Loss function: Bregman divergence d w.r.t. function ψ and map g

$$X_{s,i}M_{s,i} + D(M_{s,i}) = d_\psi(X_{s,i}, g(M_{s,i}))$$

- e.g. for Poisson: ψ is relative entropy, g is exponential map

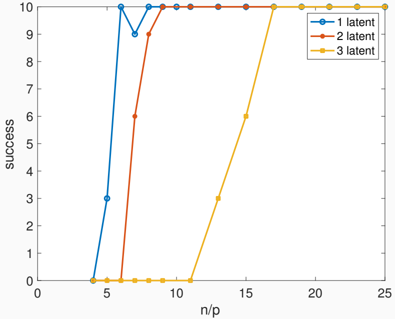
Incorporating Latent Variables → Better Structure Recovery



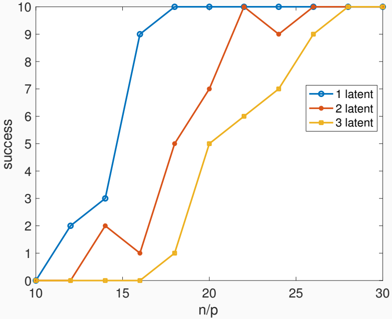
$$\mathrm{FDR} = \mathbb{E}\left[\frac{\#\text{ false edges}}{\#\text{ estimated edges}}\right] \quad ; \quad \mathrm{PWR} = \mathbb{E}\left[\frac{\#\text{ correct edges}}{\#\text{ estimated edges}}\right]$$

Consistency Experiment

Synthetic data: cycle graph structure and varying # latent variables



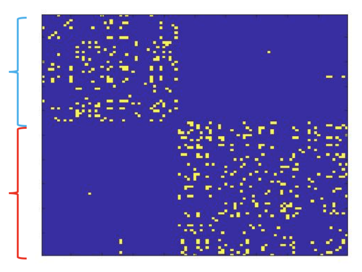
(a) Poisson-Bernoulli



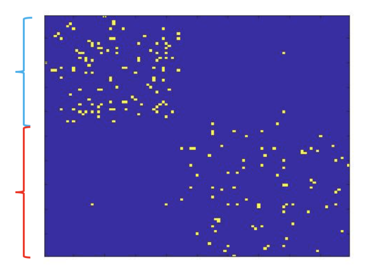
(b) Bernoulli-Gaussian

U.S. Senate Voter Records Dataset

108th Senate Voting Records: 44 democrats, 55 republicans



(c) 5% sparsity

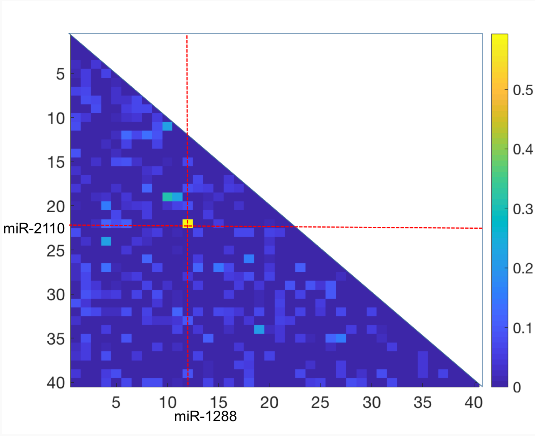


(d) 2% sparsity, 4 latent vars

RNA Sequence

Dataset: Dataset: Level III breast cancer miRNA expression

- count data $n = 544$, $p = 262$
- processed: $n = 544$, $p = 40$ [Allen & Liu, '13]



Latent Variable Model:

- 5 latent variables and 32.5% sparsity
- approximate likelihood ratio test against null: 34 p-values ≤ 0.05
- approximate likelihood ratio test: 32 p-values $\leq \frac{0.05}{40}$

Tailored Regularizers

$L = BZ$ may have additional structure beyond low-rank

For example: Z has positive entries for Poisson, ± 1 for Bernoulli

Tailored regularizers for latent structure:

- Gaussian: nuclear norm ; Bernoulli: max-2 norm
  Poisson: complete positive norm

Tailored Regularizers

$L = BZ$ may have additional structure beyond low-rank

For example: Z has positive entries for Poisson, ± 1 for Bernoulli

Tailored regularizers for latent structure:

- Gaussian: nuclear norm ; Bernoulli: max-2 norm
  Poisson: complete positive norm
- Natural semidefinite relaxation in each case

Example with Bernoulli:

$$\|L\|_{\text{relax}} = \min_{W_1 \in \mathbb{S}^p, W_2 \in \mathbb{S}^n} \quad \tfrac{1}{2}\text{trace}(W_1) + \tfrac{n}{2}\max(\text{diag}(W_2))$$

$$\text{subject-to} \qquad \begin{pmatrix} W_1 & L \\ L' & W_2 \end{pmatrix} \succeq 0$$

Tailored Regularizers

$L = BZ$ may have additional structure beyond low-rank

For example: Z has positive entries for Poisson, ± 1 for Bernoulli

Tailored regularizers for latent structure:

- Gaussian: nuclear norm ; Bernoulli: max-2 norm
  Poisson: complete positive norm

- Natural semidefinite relaxation in each case

Example with Poisson:

$$\|L\|_{\text{relax}} = \min_{W_1 \in \mathbb{S}^p, W_2 \in \mathbb{S}^n} \quad \tfrac{1}{2}\text{trace}(W_1) + \tfrac{1}{2}\text{trace}(W_2)$$

$$\text{subject-to} \quad \begin{pmatrix} W_1 & L \\ L' & W_2 \end{pmatrix} \succeq 0; W_2 \geq 0$$

Tailored Regularizers

$L = BZ$ may have additional structure beyond low-rank

For example: Z has positive entries for Poisson, ± 1 for Bernoulli

Tailored regularizers for latent structure:

- Gaussian: nuclear norm ; Bernoulli: max-2 norm
  Poisson: complete positive norm
- Natural semidefinite relaxation in each case

Example with Poisson with positive latent effects:

$$\|L\|_{\text{relax}} = \min_{W_1 \in \mathbb{S}^p, W_2 \in \mathbb{S}^n} \quad \tfrac{1}{2}\text{trace}(W_1) + \tfrac{1}{2}\text{trace}(W_2)$$

$$\text{subject-to} \quad \begin{pmatrix} W_1 & L \\ L' & W_2 \end{pmatrix} \succeq 0;\ W_2 \geq 0;\ W_1 \geq 0$$

Tailored Regularizers

Model $x = B^\star z + \epsilon$ for $p = 30$ where

- z Poisson random vector; $B^\star \geq 0$
- ϵ Gaussian random vector with independent entries

Let $\mathcal{C}^\star = \text{col-space}(B^\star)$, $\hat{\mathcal{C}}$ estimated column space

Regularizer	n = 30	n = 50
Nuclear norm FDR;PWR	0.61 ; 0.40	0.51; 0.48
Tailored FDR;PWR	0.38; 0.61	0.31; 0.68

$$\text{FDR} = \mathbb{E}\left[\frac{\text{trace}\left(\mathcal{P}_{\hat{\mathcal{C}}}\mathcal{P}_{\mathcal{C}^{\star\perp}}\right)}{\dim(\hat{\mathcal{C}})}\right] \quad ; \quad \text{PWR} = \mathbb{E}\left[\frac{\text{trace}\left(\mathcal{P}_{\hat{\mathcal{C}}}\mathcal{P}_{\mathcal{C}^{\star}}\right)}{\dim(\hat{\mathcal{C}})}\right]$$

How to Chose Regularization Parameters?

Cross-validation techniques are not appropriate

- will select models that are full-rank

Idea: choose low complexity and stable models

Approach to quantify stability:

1. obtain many subsampled bags of data
2. obtain model structure : tangent spaces to determinantal/sparse varieties at estimates for each bag
3. compute variability of the tangent spaces across bags

Come to my talk at MS195, part 3: 3pm-5pm!

Summary

Approach to identify a latent variable graphical model for GLM's

- a psuedo-likelihood approach based on convex optimization
- tailored regularizers based on the type of latent variable

Future: exact goodness of fit tests

Future: testing for presence of latent variables

Future: mixed latent variable graphical model

users.cms.caltech.edu/∼ataeb