Mixed and covariate-dependent graphical models

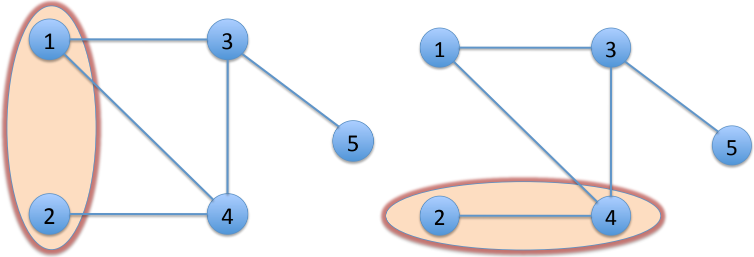
Liza Levina

Department of Statistics, University of Michigan

Joint work with Jie Cheng, Ji Zhu (University of Michigan) and Pei Wang (Fred Hutchinson Cancer Center)

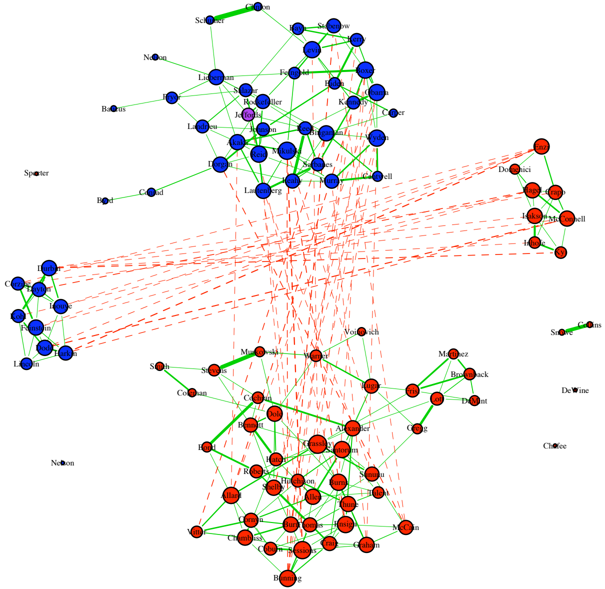
Graphical Models

- Represent conditional independence relationships between a set of random variables
- No edge between X_j and $X_{j'}$ $\iff$ X_j is independent of $X_{j'}$ conditional on all other variables



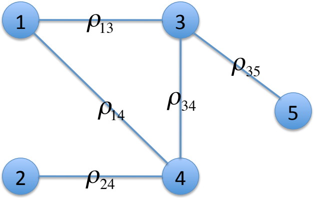
- Typically, estimated from n iid observations on p variables

Example: Senate votes



Gaussian Graphical Models

- $X_1, \ldots, X_p$ jointly follow $\mathcal{N}_p(\mu, \Omega^{-1})$
- Partial correlations ρ_{ij} are proportional to the entries of Ω
- Estimating the graph $\Longleftrightarrow$ estimating the zeros of Ω



$$\Longleftrightarrow \Omega = \begin{bmatrix} \omega_{1,1} & 0 & \omega_{1,3} & \omega_{1,4} & 0 \\ 0 & \omega_{2,2} & 0 & \omega_{2,4} & 0 \\ \omega_{3,1} & 0 & \omega_{3,3} & \omega_{3,4} & \omega_{3,5} \\ \omega_{4,1} & \omega_{4,2} & \omega_{4,3} & \omega_{4,4} & 0 \\ 0 & 0 & \omega_{5,3} & 0 & \omega_{5,5} \end{bmatrix}$$

Fitting Gaussian Graphical Models

Equivalent to estimating a sparse inverse covariance matrix

- Element-wise selection (Dempster, 1972; Drton & Perlman, 2004)
- Neighborhood selection: lasso regression of each node on its neighbors (Meinshausen & Bulhmann (2006))
- ℓ_1-penalized maximum likelihood and extensions: Yuan & Lin (2007), Banerjee et al. (2008), Rothman et al. (2008), Friedman et al (2008), Lam & Fan (2009), Ravikumar et al (2009), Zhou et al (2009), Rocha et al. (2008); Peng et al. (2009); Yuan (2010); Cai et al. (2011); for example

$$\max_{\Omega \succ 0} \quad \log(\det(\Omega)) - \text{trace}(\widehat{\Sigma}\Omega) - \lambda \sum_{j \neq j'} |\omega_{j,j'}|$$

where $\widehat{\Sigma}$ is the sample covariance matrix

Binary Markov networks (aka Ising models)

- The graphical model for binary and discrete data

$$f(X_1, \ldots, X_p) = \frac{1}{Z(\Theta)} \exp\Big(\sum_{j=1}^{p} \theta_{j,j} X_j + \sum_{1 \le j < j' \le p} \theta_{j,j'} X_j X_{j'} \Big).$$

- The dependence structure is determined by the interaction effects $\theta_{j,j'}$
- Higher-order interaction terms are typically omitted (in principle, they can be turned into order-2 interactions by adding more variables)

Fitting Ising models

- Likelihood is computationally intractable because of the normalizing constant
- Various approximations have been proposed – surrogate likelihood, pseudo-likelihood, etc (Banerjee et al 2008, Hoefling & Tibshirani 2009, Ravikumar et al 2009, Guo et al 2010)
- One approach is to run penalized logistic regression of each node on all others (analog of neighborhood selection)
- Alternatively can maximize penalized pseudo-likelihood

Covariate dependent graphical models

Motivation

- Standard assumption: the data $\{\mathbf{y}^i\}_{i=1}^n$ are i.i.d, from the same underlying graphical model.
- Data are often available in form of $\{(\mathbf{y}^i, \mathbf{x}^i)\}_{i=1}^n$, where $\mathbf{x}^i$ are additional covariates; the relationships between $\mathbf{y}$'s may depend on $\mathbf{x}$.
- A breast cancer study: $\mathbf{y}^i$ is the indicator of deletion event for various genes of a cancer patient and $\mathbf{x}^i$ is the patient's clinical phenotypes (tumor category, mutation status of TP53, estrogen receptors status).

Goals

- A graphical model for $\mathbf{y}^i$ which depends on $\mathbf{x}^i$
- Focus on Ising models for $P(\mathbf{y}|\mathbf{x})$ due to the motivating application; other cases can be developed similarly
- Subject-specific graphical models with interpretability and "continuity"
- Computational feasibility

Recent related work

- Yin & Li (2011), Cai, Li, Liu, Xie (2011): model the means in the Gaussian graphical model as covariate-dependent, but not the precision matrices
- Liu, Chen, Lafferty, Wasserman (2010): graph-valued regression partitions the covariate space non-parametrically and fits different graphical models to each part
- Guo, Levina, Michailidis, Zhu (2010): jointly fit graphical models in several categories (conditional on a single categorical covariate)

Covariate Dependent Ising Model

- Given covariate vector $\mathbf{x}$, assume

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\theta(\mathbf{x}))} \exp\left(\sum_{j=1}^{q} \theta_{jj}(\mathbf{x}) y_j + \sum_{1\le k<j\le q} \theta_{jk}(\mathbf{x}) y_j y_k \right)$$

- Parametrize $\theta_{jk}(\mathbf{x})$ as linear functions of $\mathbf{x}$

$$\begin{aligned} \theta_{jk}(\mathbf{x}) &= \theta_{jk0} + \theta_{jk}^T \mathbf{x}, \quad \text{where } \theta_{jk}^T = (\theta_{jk1}, \ldots, \theta_{jkp}) \\ \theta_{jk}(\mathbf{x}) &= \theta_{kj}(\mathbf{x}), \qquad \forall j \neq k \end{aligned}$$

- Benefits of linear parametrization:
  interpretability, continuity, convexity.

Optimization Criterion

- Loss function:
  - Directly maximizing the likelihood is computationally intractable due to the normalizing constant.
  - Focus on optimizing conditional likelihood

$$\ell_j(\theta;\mathbf{x},\mathbf{y}) = -\frac{1}{n}\sum_{i=1}^{n}\log P(y_j^i|\mathbf{x}^i,\mathbf{y}_{-j}^i)$$

- Regularization: use ℓ_1 penalty to select only the important covariates and edges.

Fitting the model

- Separate approach: estimate each $\theta_j, j = 1, \ldots, q$ separately by

$$\min_{\boldsymbol{\theta}_j \in \mathbb{R}^{(p+1)q}} \ell_j(\boldsymbol{\theta}_j, \mathscr{D}_n) + \boldsymbol{\lambda} \|\boldsymbol{\theta}_j\|_1$$

  Followed by ad hoc symmetrization (min or max of $\hat{\theta}_{jk}$ and $\hat{\theta}_{kj}$

- Joint approach: estimate the entire vector θ simultaneously by

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^{(p+1)q(q+1)/2}} \sum_{j=1}^{q} \ell_j(\boldsymbol{\theta}, \mathscr{D}_n) + \boldsymbol{\lambda} \|\boldsymbol{\theta}\|_1$$

- Optimization is done by a coordinate descent type algorithm (Fu, 1998).

Tumor suppressor genes and breast cancer study

- Deletion of tumor suppressor genes plays an important role in tumor initiation and development
- Goals of study:
  1. Characterize the conditional associations among deletion events of various genes
  2. Investigate how these association patterns vary across different types of patients

Data Description

- Data consists of $n = 143$ tumor samples, all from breast cancer patients at various stages before start of therapy.
- 39,632 DNA copy number profiles $\rightarrow$ 620 cytobands
- $\mathbf{y}^i$ is a 620-dimensional binary vector; $y^i_j = 1$ if the j^{th} cytoband has been deleted in the i^{th} tumor sample.
- $\mathbf{x}^i$ contains 3 clinical phenotypes:
  - TP53 mutation status (0/1)
  - Estrogen Receptor status (0/1): 1 means the sample is ER positive.
  - Tumor category (1, 2, 3, 4): ordinal variable, larger values indicate more advanced tumors.

Covariate dependent inter-chromosome interactions ranked by selection frequency

Gene1	Gene2	Freq	Gene1	Gene2	Freq
Main Effect (θ_{jk0})			TP53 Mutation (θ_{jk1})		
4q31.3	18q23	0.95	3p22.2	22q13.1	0.79
2p25.2	15q26.2	0.87	3p12.3	12p13.1	0.72
2q36.3	3p26.1	0.84	12q22	15q14	0.7
7q21.13	8q21.13	0.84	2p12	Xp22.33	0.69
6p21.32	16q12.2	0.83	6p21.32	8p11.22	0.68
3p21.1	17p13.2	0.81	1p34.2	3p24.1	0.67
4q24	12q21.1	0.81	2p21	Xp11.22	0.67
2q23.3	6p12.1	0.79	2p12	7p21.1	0.66
8p21.3	21q21.1	0.79	12q15	13q12.12	0.63
2q34	3q13.31	0.78	4q25	8p11.22	0.62
6p21.32	9q31.3	0.78	8p11.22	Xq23	0.62
6p21.32	13q21.1	0.78	9p21.2	16q22.1	0.61
ER Status (θ_{jk2})			Tumor stage (θ_{jk3})		
3q26.1	11p14.3	0.69	16q23.3	17p13.1	0.61
4q34.3	5q32	0.64	12p11.23	16q12.2	0.59
8p11.22	11p14.2	0.63	3q13.13	Xq23	0.57
3q24	22q11.23	0.57	7p21.3	12p11.23	0.56
4p14	11p15.3	0.55	9q34.13	15q21.1	0.55
1q31.1	Xq27.3	0.54	11q24.2	13q32.3	0.55
13q33.2	22q11.23	0.54	8q21.13	13q33.1	0.54
21q21.1	22q11.21	0.54	2p21	12p13.31	0.53
5q33.1	17q21.31	0.53	10q26.3	17p11.2	0.53
12q21.32	18q22.3	0.51	7p21.3	12p12.1	0.51
8p11.22	22q11.21	0.5	3q13.13	7p21.3	0.5
8q21.13	Xp22.11	0.5	9q34.13	15q22.1	0.5

Asymptotic behavior

- Focus on the separate approach
- Need standard assumptions on the design matrix, which now includes both $\mathbf{x}$ and $\mathbf{y}$ terms
- An exponential decay assumption on the tails of $\mathbf{x}$
- Get standard results on parameter estimation and model selection consistency
- Roughly, the rate is governed by $\sqrt{d\log(pq)/n}$, where d is the max # non-zero parameters per edge

Assumptions

- $\mathbf{x_j} \otimes \mathbf{y}_{-j}$: all terms in the j's logistic regression
- θ_j^*: true coefficients of the j-th logistic regression
- S_j: the set of non-zero elements of θ_j^*
- $\mathbf{I}_j^* = \mathbb{E}_{\theta^*}(\nabla^2 \log P_\theta(y_j|\mathbf{x}, \mathbf{y}_{-j}))$: information matrix
- $\mathbf{U}_j^* = \mathbb{E}_{\theta^*}\left((\mathbf{x} \otimes \mathbf{y}_{-j})(\mathbf{x} \otimes \mathbf{y}_{-j})^T\right)$

A1 There exists $\alpha \in (0,1]$, s. t.

$$\|\mathbf{I}^*_{S_j^c S_j}\left(\mathbf{I}^*_{S_j S_j}\right)^{-1}\|_\infty \leq (1-\alpha)$$

A2 There exist $\Delta_{\min}, \Delta_{\max} > 0$, s. t.

$$\begin{aligned} \Lambda_{\min}\left(\mathbf{I}^*_{S_j S_j}\right) &\geq \Delta_{\min} \\ \Lambda_{\max}(\mathbf{U}_j^*) &\leq \Delta_{\max} \end{aligned}$$

A3 $\forall \delta > 0$, $\forall M \geq M_0$,

$$P(\|\mathbf{x}\|_\infty \geq M) \leq \exp(-M^\delta)$$

Theorem

Let $d = \max_j |S_j|$, $C > 0$, $\gamma \in (0,1)$ constants. If A1, A2, A3 hold and $M_n \geq (C\lambda_n^2 n)^{\frac{1}{1+\delta}}$, $\lambda_n \geq CM_n\sqrt{\frac{\log(pq)}{n}}$, $n \geq CM_n^2 d^3 \log(pq)$, then with probability at least $1 - \exp^{-C(\lambda_n^2 n)^\gamma}$ for any $j \in \{1, \dots, q\}$ the following holds:

1. Uniqueness: $\hat{\theta}_j$ is the unique optimal solution.
2. Norm consistency: $\|\hat{\theta}_j - \theta_j^*\|_2 \leq 5\lambda_n\sqrt{d}/\Delta_{\min}$.
3. Sign consistency: $\hat{\theta}_j$ correctly identifies all zeros in θ_j^*, and the sign of non-zeros in θ_j^* whose absolute value is at least $10\lambda_n\sqrt{d}/\Delta_{\min}$.

Simulation: Effect of Sparsity

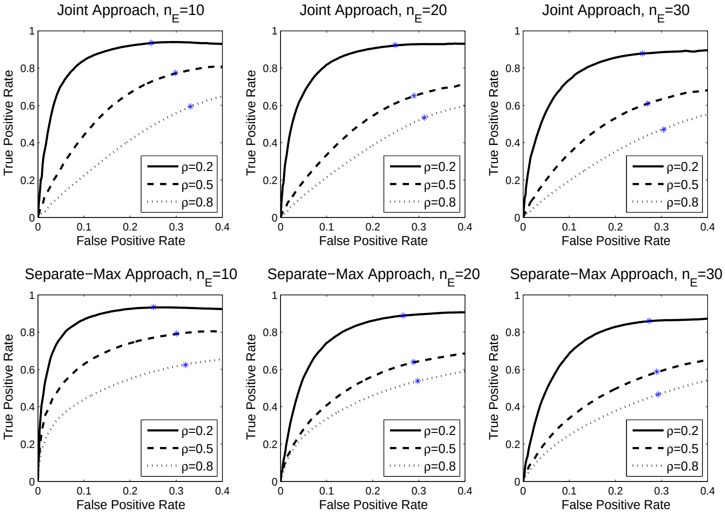
Sparsity can mean:

- Small number of edges in the graph
- Small number of non-zero parameters per edge

Simulation settings:

- $p = 20$ covariates, $q = 10$ binary variables, $n = 200$
- Proportion of non-zeros per edge $\rho = \{0.2, 0.5, 0.8\}$
- Total number of edges n_E = $\{$10, 20, 30$\}$.
- Results summarized in the form of ROC curves

Simulation results: Effect of Sparsity



Mixed graphical models

Motivation: In practice, many datasets contain both continuous and discrete variables!

- Let $X = (Z, Y)$, where $Z \in \{0,1\}^q$ and $Y \in \mathbb{R}^p$
- Suppose X has the conditional Gaussian distribution (CGD) (Lauritzen and Wermuth, 1989):

$$f(x) = f(z,y) = \exp\left(g_z + h_z^T y - \frac{1}{2} y^T K_z y\right),$$

where $\{(g_z, h_z, K_z) : g_z \in \mathbb{R}, h_z \in \mathbb{R}^p, K_z \in \mathbb{R}^+_{p\times p}, z \in \{0,1\}^q\}$ are the canonical parameters of the distribution.

Markovian conditional Gaussian distributions

Let Δ index Z, Γ index Y. The canonical parameters can be written as

$$g_z = \sum_{d:d\subseteq\Delta} \lambda_d(z), \quad h_z = \sum_{d:d\subseteq\Delta} \eta_d(z), \quad K_z = \sum_{d:d\subseteq\Delta} \Phi_d(z) \ ,$$

where functions indexed by d only depend on z through z_d.

Theorem (Lauritzen 1996): a CGD is Markovian with respect to a graph $\mathscr{G}$ iff the density has an expansion that satisfies

$$\begin{array}{rcll} \lambda_d(z) & \equiv & 0 & \text{unless } d \text{ is complete in } \mathscr{G}, \\ \eta_d^\gamma(z) & \equiv & 0 & \text{unless } d\cup\{\gamma\} \text{ is complete in } \mathscr{G}, \\ \Phi_d^{\gamma\mu}(z) & \equiv & 0 & \text{unless } d\cup\{\gamma,\mu\} \text{ is complete in } \mathscr{G}. \end{array}$$

A simplified CGD

- The full model has $O(2^q p^2)$ parameters – impossible to fit to high-dimensional data
- Consider instead a simplified model, with $\log f(z,y) =$

$$\sum_{d:d\subseteq\Delta,|d|\leq 2} \lambda_d(z) + \sum_{d:d\subseteq\Delta,|d|\leq 1} \eta_d(z)^T y - \frac{1}{2}\sum_{d:d\subseteq\Delta,|d|\leq 1} y^T \Phi_d(z) y =$$

$$(\lambda_0 + \sum_j \lambda_j z_j + \sum_{j>k} \lambda_{jk} z_j z_k) + y^T(\eta_0 + \sum_j \eta_j z_j) - \frac{1}{2} y^T (\Phi_0 + \sum_{j=1}^{q} \Phi_j z_j) y$$

- $O(\max(q^2, p^2 q))$ parameters
- Still includes all possible graphs

Model parameters and conditional independence

With the loglikelihood given by

$$\log f(y,z) = (\lambda_0 + \sum_j \lambda_j z_j + \sum_{j>k} \lambda_{jk} z_j z_k) + y^T(\eta_0 + \sum_j \eta_j z_j) - \frac{1}{2} y^T (\Phi_0 + \sum_{j=1}^{q} \Phi_j z_j) y$$

the conditional independencies are determined as follows:

$$\begin{aligned} Z_j \perp Z_k \mid X \backslash \{Z_j, Z_k\} &\iff \lambda_{jk} = 0, \\ Z_j \perp Y_\gamma \mid X \backslash \{Z_j, Y_\gamma\} &\iff \theta_{j\gamma} = \left(\eta_j^\gamma, \{\Phi_j^{\gamma\mu} : \mu \neq \gamma\}\right) = 0, \\ Y_\gamma \perp Y_\mu \mid X \backslash \{Y_\gamma, Y_\mu\} &\iff \theta_{\gamma\mu} = \left(\Phi_0^{\gamma\mu}, \{\Phi_j^{\gamma\mu} : j \in \Delta\}\right) = 0. \end{aligned}$$

Related recent work

- Lee and Hastie (2012): a special case of our model with covariance of Y independent of Z (all $\Phi^j = 0$).
- Fellinghauer et al (2011): neighborhood selection using random forests (no generative model)

Model fitting

- Likelihood involves intractable normalizing constant
- Instead look at conditional log-likelihood (neighborhood selection)
- Continuous variables ⇒ linear regression:

$$E(Y_\gamma | Y_{-\gamma}, Z) = \eta_0^\gamma + \sum_j \eta_j^\gamma Z_j - \sum_{\mu \neq \gamma} \left(\Phi_0^{\gamma\mu} + \sum_j \Phi_j^{\gamma\mu} Z_j \right) Y_\mu$$

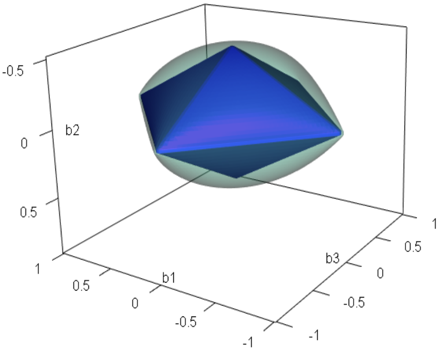
- Binary variables ⇒ logistic regression:

$$\log \frac{P(Z_j = 1 | Z_{-j}, Y)}{P(Z_j = 0 | Z_{-j}, Y)} = \lambda_j + \sum_{k \neq j} \lambda_{jk} Z_k + \sum_{\gamma=1}^{p} \eta_j^\gamma Y_\gamma - \frac{1}{2} \sum_{\gamma,\mu=1}^{p} \Phi_j^{\gamma\mu} Y_\gamma Y_\mu$$

Penalty

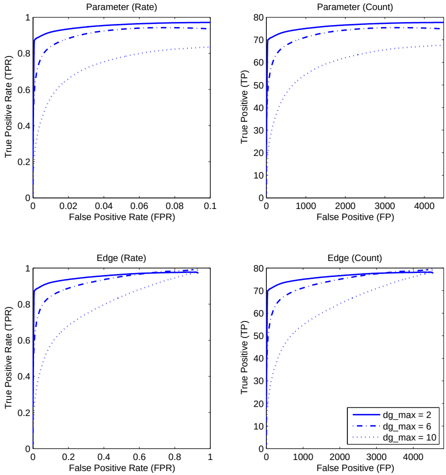
- Need a sparse estimate $\Rightarrow$ regularize
- Complication: parameters are in overlapping groups
- Regular lasso penalty: $\|\theta\|_1 = \sum_i |\theta_i|$
- Group lasso penalty: $\|\theta\|_2 = \sqrt{\sum_i \theta_i^2}$ - computationally difficult, especially with overlaps

"Approximate" the group penalty by an upper bound:
$\|\theta\|_2 \leq \|\theta\|_1$



Green (outside): $\{\theta : \sqrt{\theta_1^2 + \theta_2^2} + \sqrt{\theta_2^2 + \theta_3^2} = 1\}$
Blue (inside): $\{\theta : |\theta_1| + 2|\theta_2| + |\theta_3| = 1\}$

Simulated example: varying max node degree



- Max node degree varies in $\{2, 6, 10\}$
- 80 edges total (fixed)
- $p = 90$ (continuous), $q = 10$ (categorical)
- Sample size $n = 100$
- ROC curves averaged over 20 replications.

Asymptotic behavior

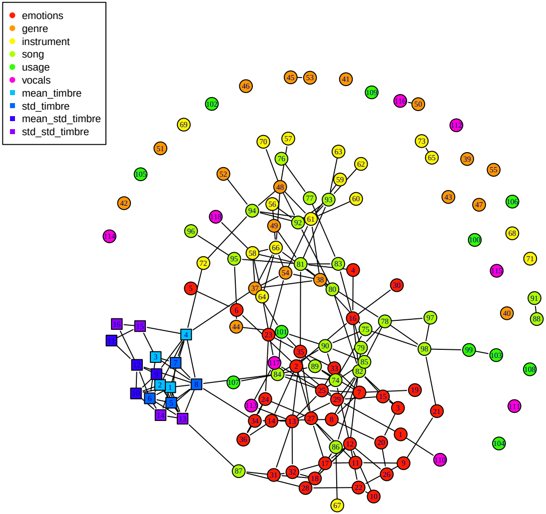
- We fit regular or logistic regressions with weighted L_1 penalties
- The weights are either 1 or 2, do not depend on data
- Standard results establish consistency of parameter estimation and model selection
- Only need to assume that the standard assumptions (such as irrepresentable condition) hold on a rescaled version of the design matrix

Example: music annotation dataset

- CAL500 data set: $n = 502$ observations, $q = 128$ discrete variables, and $p = 16$ continuous variables.
- The 128 discrete variables come from six categories: emotions, genres, instruments, song characteristics, usages, and vocal types; manually labelled by human experts.
- The continuous features are extracted from the time series of the audio signal and represent “brightness” of the music, noisiness, amplitude, etc.

Fitted edges for music data

Showing edges with stability selection frequency of at least 0.9



Some interesting findings

- Amplitude $\leftrightarrow$ "alternative rock"
- Noisiness $\leftrightarrow$ "negative feelings"
- Short period amplitude variation $\leftrightarrow$ popular likable songs
- Songs with positive feelings $\leftrightarrow$ piano
- Songs with high energy $\leftrightarrow$ optimistic emotions, dancable songs
- Fast tempo music $\leftrightarrow$ classic rock
- Likable or popular songs $\leftrightarrow$ driving, reading

Summary

- Graphical models are a popular exploratory tool but they need more flexibility
- Conditioning on covariates allows subject-specific models; linear models provide interpretation
- Mixed graphical models allow exploring relationships between continuous and categorical variables
- Other questions of interest: mixtures of graphical models (unsupervised learning), more complex covariate relationships, combining graphical models with network models

Cheng, J., Levina, E., and Zhu, J. (2013). Joint graphical models for discrete and continuous variables.
Cheng, J., Levina, E., Wang, P. and Zhu, J. Sparse Ising models with covariates. (2012).

Thank you