Sparse topology selection in graphical models of autoregressive time series

Jitkomut Songsiri, Chulalongkorn University, Thailand

Lieven Vandenberghe, UCLA

BIRS, January 21, 2011

Outline

- **Gaussian graphical models**
- Graphical models of autoregressive time series
- Algorithms

Gaussian graphical model

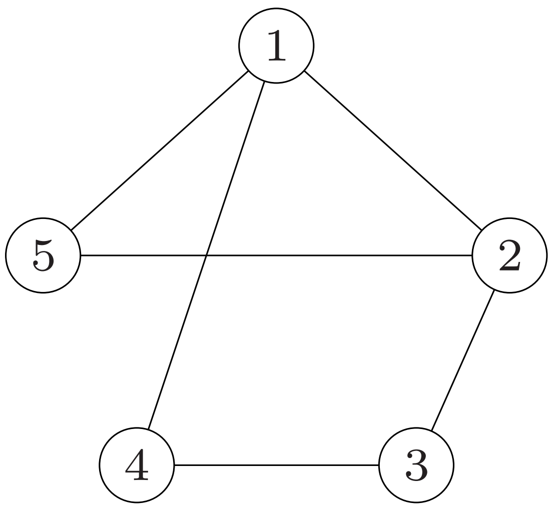
n-dimensional Gaussian vector

$$x = (x_1, \ldots, x_n) \sim \mathcal{N}(0, \Sigma)$$

x_i, x_j are **conditionally independent** (given the rest of x) if

$$(\Sigma^{-1})_{ij} = 0$$

modeled as undirected graph with n nodes; arc i, j is absent if $(\Sigma^{-1})_{ij} = 0$



$$\Sigma^{-1} = \begin{bmatrix} \bullet & \bullet & 0 & \bullet & \bullet \\ \bullet & \bullet & \bullet & 0 & \bullet \\ 0 & \bullet & \bullet & \bullet & 0 \\ \bullet & 0 & \bullet & \bullet & 0 \\ \bullet & \bullet & 0 & 0 & \bullet \end{bmatrix}$$

Maximum likelihood estimation

Log-likelihood function for N independent samples of x

$$\frac{N}{2}\left(\log\det\Sigma^{-1} - \mathbf{tr}(C\Sigma^{-1})\right)$$

C is sample covariance

ML estimation of Σ, for given topology

$$\begin{array}{ll} \text{minimize} & -\log\det X + \mathbf{tr}(CX) \\ \text{subject to} & \text{given sparsity pattern of } X \end{array}$$

a convex problem in $X = \Sigma^{-1}$

known as covariance selection (Dempster 1972)

Topology selection via model selection criteria

- enumerate topologies and for each topology, solve ML problem

$$\begin{array}{ll} \text{minimize} & -\mathcal{L}(X) = -\log\det X + \mathbf{tr}(CX) \\ \text{subject to} & \text{sparsity pattern of } X \end{array}$$

- rank ML estimates $X_{\rm ml} = \Sigma_{\rm ml}^{-1}$ using an information criterion

$$\begin{array}{rcll} \mathrm{AIC} & = & -2\mathcal{L}(X_{\rm ml}) + 2k & \quad (\textit{Akaike}) \\ \mathrm{AIC_c} & = & -2\mathcal{L}(X_{\rm ml}) + \dfrac{2Nk}{N-k-1} & \quad (\textit{second order Akaike}) \\ \mathrm{BIC} & = & -2\mathcal{L}(X_{\rm ml}) + k\log N & \quad (\textit{Bayes}) \end{array}$$

k is number of parameters ($\propto$ nonzeros in X); N is sample size

this approach is only feasible for small graphs

Topology selection via 1-norm regularization

Regularized ML problem

$$\text{minimize} \quad -\log\det X + \mathbf{tr}(CX) + \gamma \sum_{i,j} |X_{ij}|$$

Dual problem

$$\begin{array}{ll} \text{maximize} & \log\det(C+Z) \\ \text{subject to} & |Z_{ij}| \leq \gamma, \quad i,j = 1,\ldots,n \end{array}$$

- convex; primal or dual can be solved by first-order methods

  Yuan & Lin 2007; Banerjee, El Ghaoui & d'Aspremont 2008; Friedman, Hastie & Tibshirani 2008; Lu 2008, 2009; Scheinberg & Rish 2009, . . .

- choice of γ: rank topologies on trade-off curve by AIC, $\mathrm{AIC_c}$ or BIC

  other methods: Banerjee et al. 2008, Friedman et al. 2007, Ravikumar et al. 2008; Meinshausen & Bühlmann 2006

Outline

- Gaussian graphical models
- **Graphical model of autoregressive time series**
- Algorithms

Graphical models of time series

n-dimensional stationary Gaussian time series $x(t)$, $t \in \mathbf{Z}$

x_i and x_j are **conditionally independent** (given the rest of $x(t)$) if

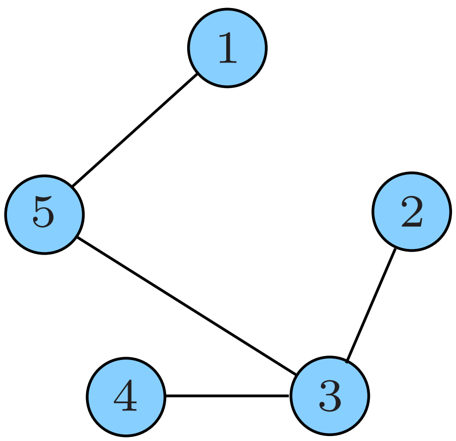
$$\left(S(\omega)^{-1}\right)_{ij} = 0$$

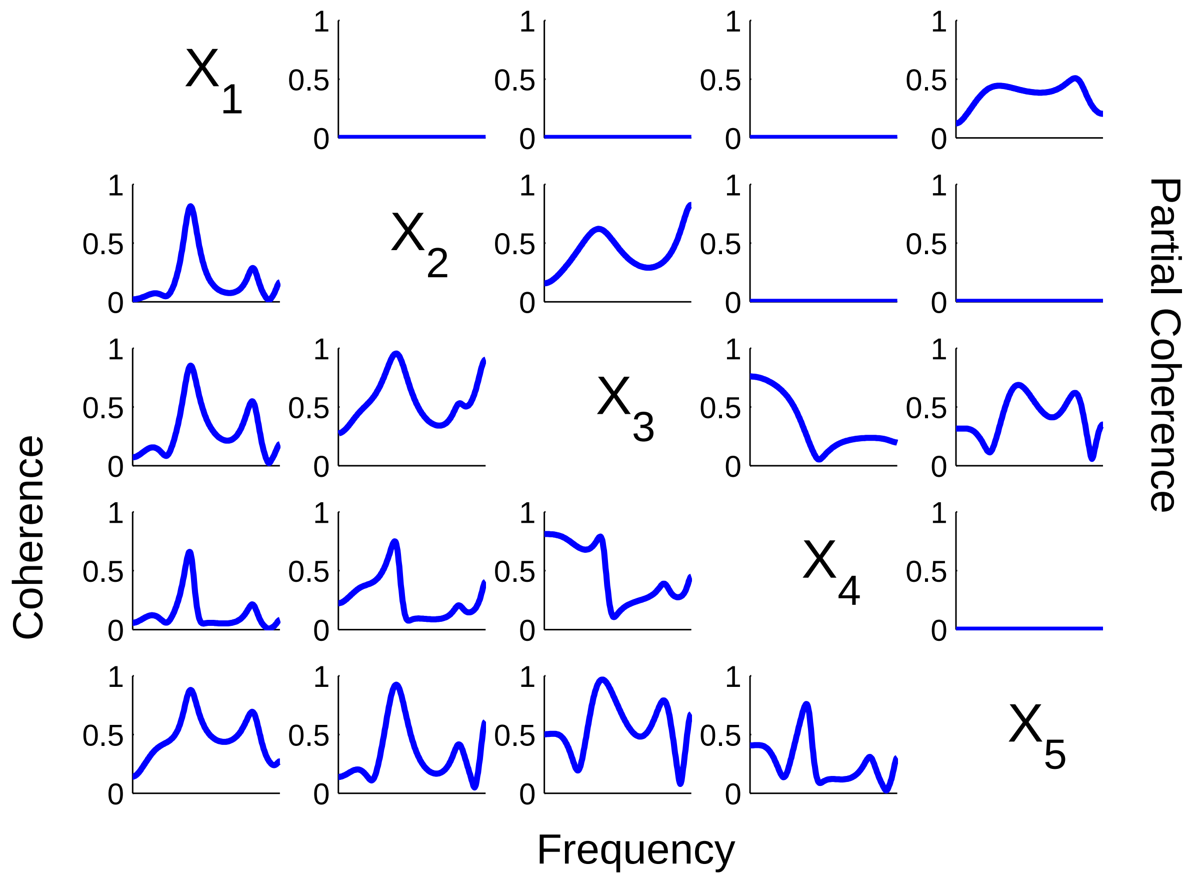
$S(\omega)$ is **spectral density matrix:**

$$S(\omega) = \sum_{k=-\infty}^{\infty} R_k e^{-\mathrm{j}k\omega}, \qquad R_k = \mathbf{E}\, x(t+k)x(t)^T \qquad (\mathrm{j} = \sqrt{-1})$$

Brillinger 1981, Dahlhaus 2000

Example (autoregressive model of order 4)





- coherence spectrum: $S(\omega)$ normalized to have diagonal one
- partial coherence: $S(\omega)^{-1}$ normalized to have diagonal one

Autoregressive time series

$$B_0 x(t) = -\sum_{k=1}^{p} B_k^T x(t-k) + w(t), \qquad w(t) \sim \mathcal{N}(0, I)$$

without loss of generality, assume B_0 symmetric, positive definite

Inverse spectrum

$$S(\omega)^{-1} = Y_0 + \sum_{k=1}^{p} \left(Y_k e^{-\mathrm{j}k\omega} + Y_k^T e^{\mathrm{j}k\omega} \right), \qquad Y_k = \sum_{l=0}^{p-k} B_l B_{l+k}^T$$

Conditional independence relations

$$\left(S(\omega)^{-1} \right)_{ij} = 0 \qquad \Longleftrightarrow \qquad (Y_k)_{ij} = (Y_k)_{ji} = 0, \quad k = 0, 1, \ldots, p$$

ML estimate with conditional independence constraints

$$
\begin{array}{ll}
\text{minimize} & -\log\det X_{00} + \mathbf{tr}(CX) \\
\text{subject to} & X = \begin{bmatrix} X_{00} & \cdots & X_{0p} \\ \vdots & \ddots & \vdots \\ X_{p0} & \cdots & X_{pp} \end{bmatrix} = \begin{bmatrix} B_0 \\ \vdots \\ B_p \end{bmatrix} \begin{bmatrix} B_0 \\ \vdots \\ B_p \end{bmatrix}^T \\
 & \text{given sparsity pattern of } \sum_l X_{l,l+k} \text{ for } k = 0, \ldots, p
\end{array}
$$

- maximizes conditional likelihood (conditioned on first p values)
- C is sample covariance estimate from observations of $x(t)$ (see later)
- variables are X, B_0, . . . , B_p
- equality constraints $X_{ij} = B_i B_j^T$ make problem nonconvex

Convex relaxation

$$
\begin{array}{ll}
\text{minimize} & -\log\det X_{00} + \mathbf{tr}(CX) \\
\text{subject to} & X = \begin{bmatrix} X_{00} & \cdots & X_{0p} \\ \vdots & \ddots & \vdots \\ X_{p0} & \cdots & X_{pp} \end{bmatrix} \succeq 0 \\
& \text{given sparsity pattern of } \sum\limits_{l} X_{l,l+k} \text{ for } k = 0, \ldots, p
\end{array}
$$

- exact if optimal X has rank n, $i.e.$, can be factored as $X_{ij} = B_i B_j^T$
- from duality: relaxation is exact if C is pos. definite and **block-Toeplitz**
- in practice: often exact even for non-Toeplitz sample covariances C

Duality

$$\begin{array}{ll} \text{minimize} & -\log\det X_{00} + \mathbf{tr}(CX) \\ \text{subject to} & X \succeq 0, \quad \mathsf{P}\left(\sum_l X_{l,l+k}\right) = 0 \end{array}$$

- X is $(p+1)\times(p+1)$ symmetric block matrix with blocks X_{ij} of order n
- $\mathsf{P}(U)$ is projection on zero pattern

Dual problem

$$\begin{array}{ll} \text{maximize} & \log\det W + n \\ \text{subject to} & \left[\begin{array}{cc} W & 0 \\ 0 & 0 \end{array}\right] \preceq C + \mathsf{T}(\mathsf{P}(Z_0), \ldots, \mathsf{P}(Z_p)) \end{array}$$

- variables W, Z_0, . . . , Z_p are $n \times n$, with W and Z_0 symmetric
- $\mathsf{T}(U_0, \ldots, U_p)$ is block-Toepliz matrix with first row U_0, . . . , U_p

Optimality condition

Property of block-Toeplitz matrices: if $W \succ 0$ and

$$\mathsf{T}(U_0, \ldots, U_p) = \left[\begin{array}{cccc} U_0 & U_1 & \cdots & U_p \\ U_1^T & U_0 & \cdots & U_{p-1} \\ \vdots & \vdots & \ddots & \vdots \\ U_p^T & U_{p-1}^T & \cdots & U_0 \end{array}\right] \succeq \left[\begin{array}{cccc} W & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{array}\right]$$

then $\mathsf{T}(U_0, \ldots, U_p) \succ 0$

Complementary slackness for ML problem $(Z = \mathsf{T}(\mathsf{P}(Z_0), \ldots, \mathsf{P}(Z_p)))$

$$W = X_{00}^{-1}, \qquad \mathbf{tr}\left(X\left(C + Z - \left[\begin{array}{cc} W & 0 \\ 0 & 0 \end{array}\right]\right)\right) = 0$$

hence, if C is block-Toeplitz, optimal X has rank n

Sample covariance matrix

from measurements $\hat{x}(1), \ldots, \hat{x}(N)$, estimate $R_k = \mathbf{E}\, x(t)x(t+k)^T$

$$C = \frac{1}{M}\sum_{t=t_1}^{t_2} \begin{bmatrix} \hat{x}(t) \\ \hat{x}(t-1) \\ \vdots \\ \hat{x}(t-p) \end{bmatrix} \begin{bmatrix} \hat{x}(t) \\ \hat{x}(t-1) \\ \vdots \\ \hat{x}(t-p) \end{bmatrix}^T \approx \begin{bmatrix} R_0 & R_1 & \cdots & R_p \\ R_1^T & R_0 & \cdots & R_{p-1} \\ \vdots & \vdots & \ddots & \vdots \\ R_p^T & R_{p-1}^T & \cdots & R_0 \end{bmatrix}$$

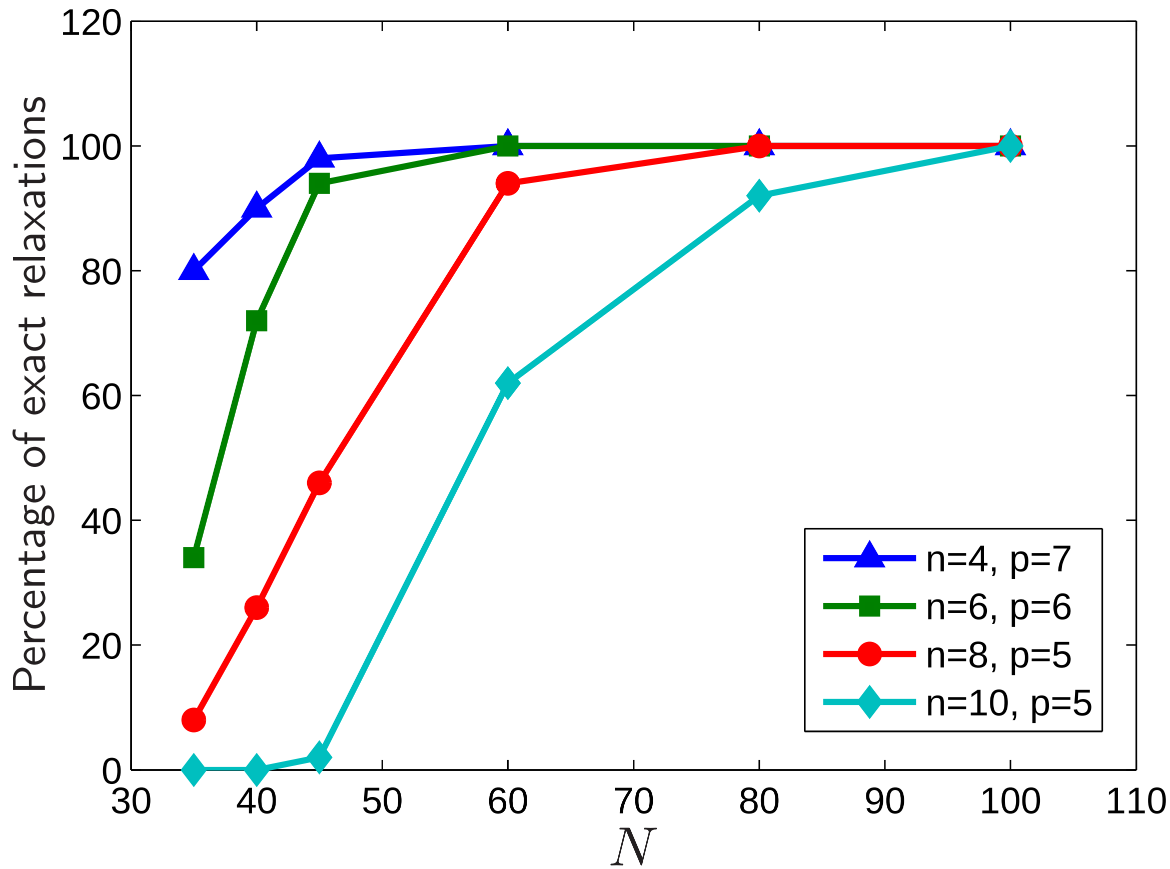
Non-windowed estimate: $t_1 = p+1$, $t_2 = N$, $M = N - p$

- arises in conditional ML/LS estimation
- generally not block-Toeplitz but approaches block-Toeplitz for large N

Windowed estimate: $t_1 = 1$, $t_2 = N + p$, $M = N$

- assume $\hat{x}(t) = 0$ for $t < 1$ and $t > N$
- block-Toeplitz

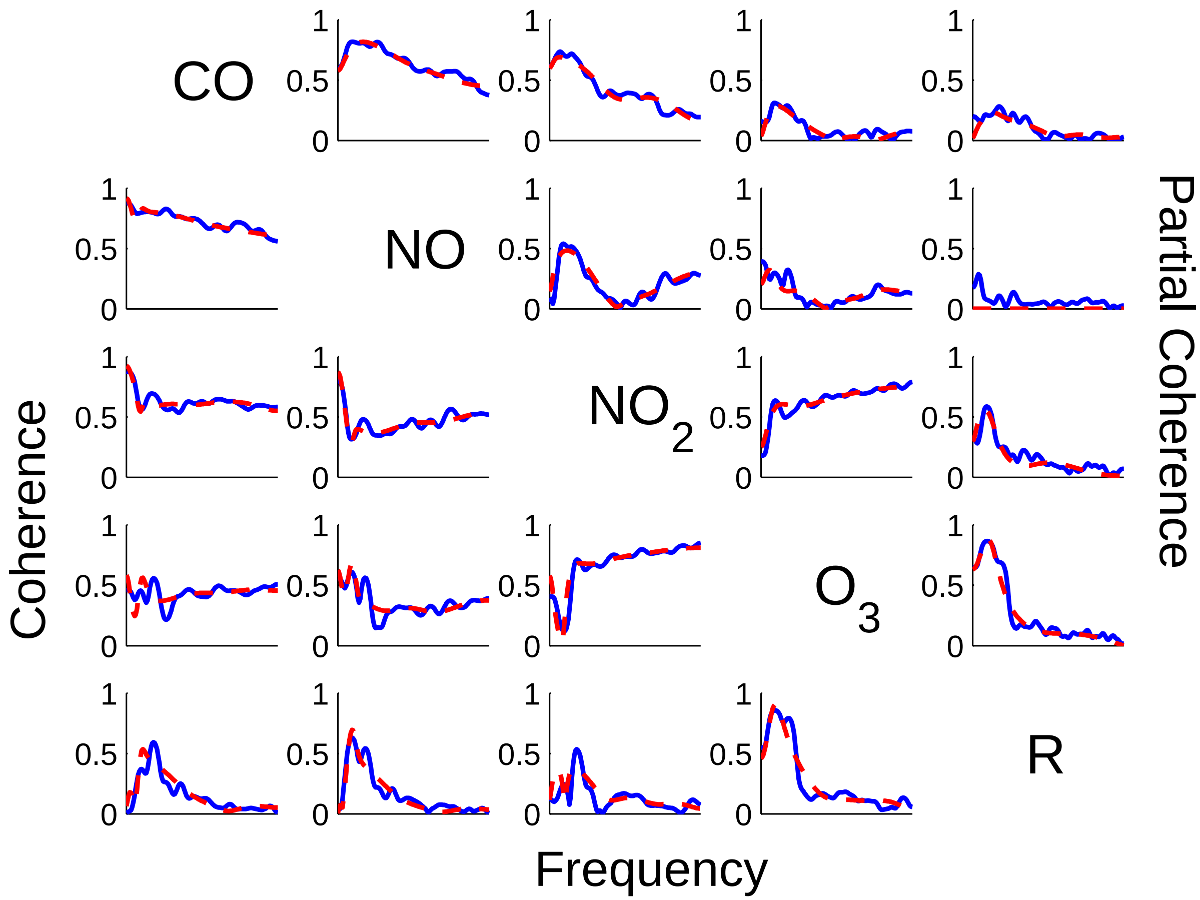
Example: Exactness of relaxation



- generate 50 trials of C for each model
- relaxation for (non-Toeplitz) non-windowed sample covariance matrix C

Example: Air pollution data

hourly values of CO, NO, NO_2, O_3, solar radiation at Azusa ($N = 8370$)



blue: empirical spectrum; red: optimal AR model for BIC ($p = 4$)

Topology selection via nonsmooth regularization

Estimation with known topology

$$
\begin{array}{ll}
\text{minimize} & -\log\det X_{00} + \mathbf{tr}(CX) \\
\text{subject to} & X \succeq 0 \\
& \text{given sparsity pattern of } Y_k = \displaystyle\sum_{l=0}^{p-k} X_{l,l+k}, \quad k = 0, \ldots, p
\end{array}
$$

Nonsmooth regularization

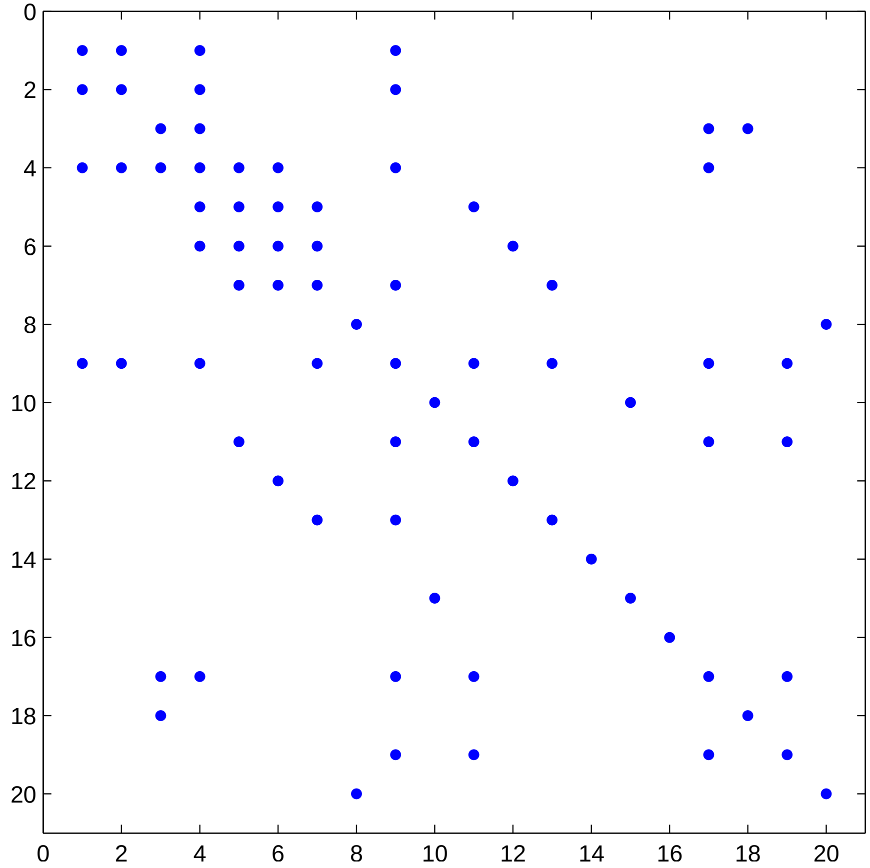
$$
\begin{array}{ll}
\text{minimize} & -\log\det X_{00} + \mathbf{tr}(CX) + \gamma h(Y_0, \ldots, Y_p) \\
\text{subject to} & X \succeq 0
\end{array}
$$

convex penalty h promotes common, symmetric sparsity pattern of Y_k:

$$
h(Y_0, Y_1, \ldots, Y_p) = \sum_{i>j} \max_k \max\{|(Y_k)_{ij}|, |(Y_k)_{ji}|\}
$$

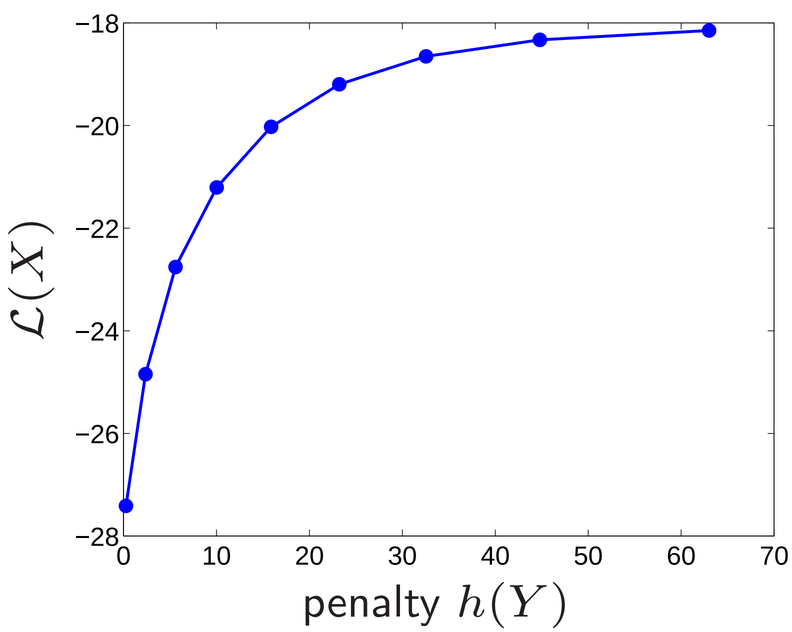
Example

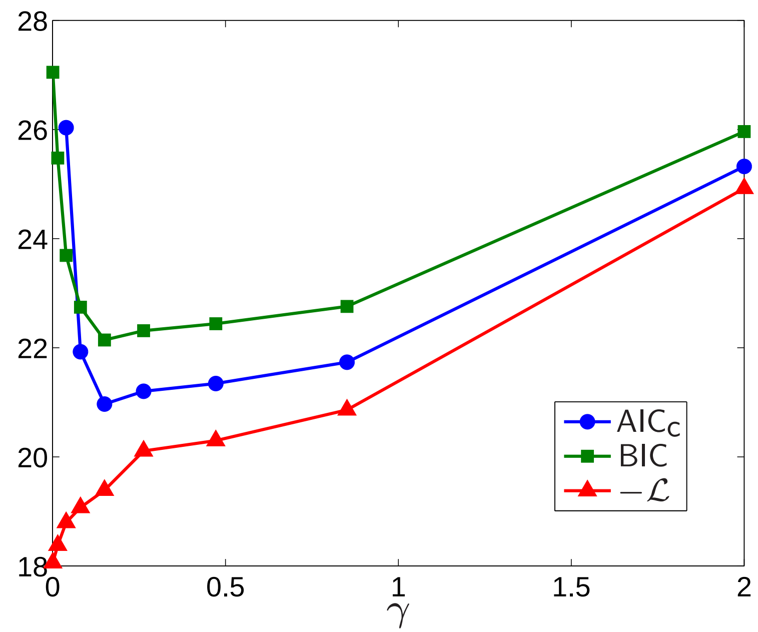
$n = 20$, $p = 2$, exact $S(\omega)^{-1}$ has 76 nonzeros



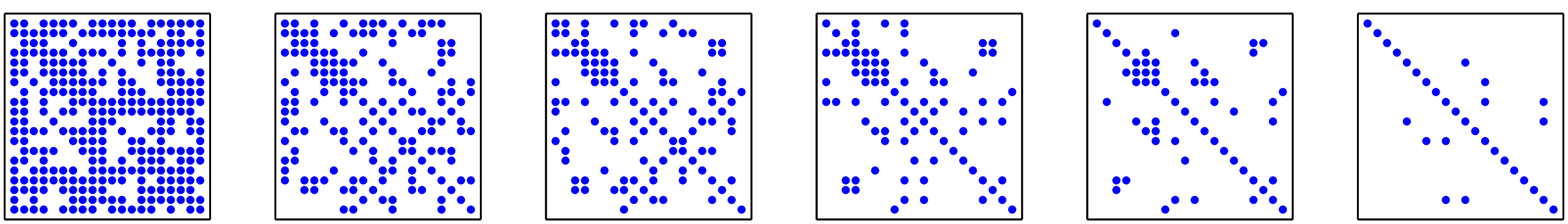
Regularized maximum likelihood problem (512 samples)

$$\text{minimize} \quad -\mathcal{L}(X) + \gamma h(Y) \qquad (\mathcal{L}(X) = \log\det X_{00} - \mathbf{tr}(CX))$$

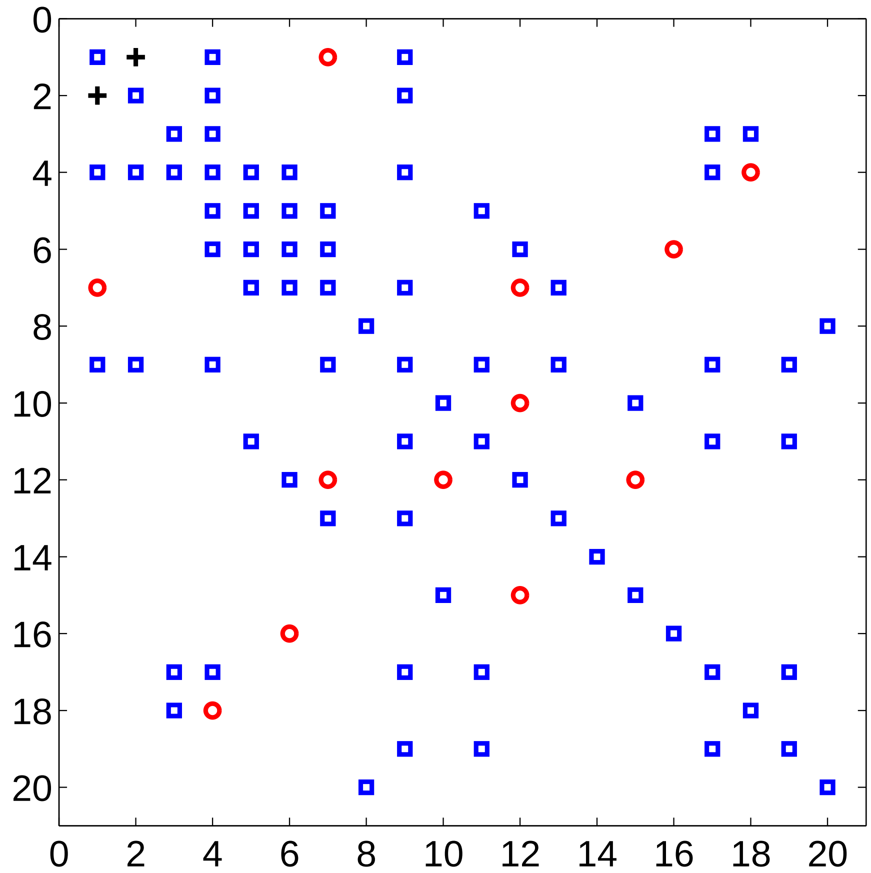




$\gamma = 0$ $\gamma = 0.04$ $\gamma = 0.08$ $\gamma = 0.15$ $\gamma = 0.85$ $\gamma = 2$



Topology for $\gamma = 0.15$



- blue squares: correctly classified
- red circles: incorrectly classified as nonzero
- plus signs: incorrectly classified as zero

Comparison with other estimates

threshold the inverse spectrum from one of three estimation methods

- **ML estimation** (a.k.a. LS estimation): solve Yule-Walker equations

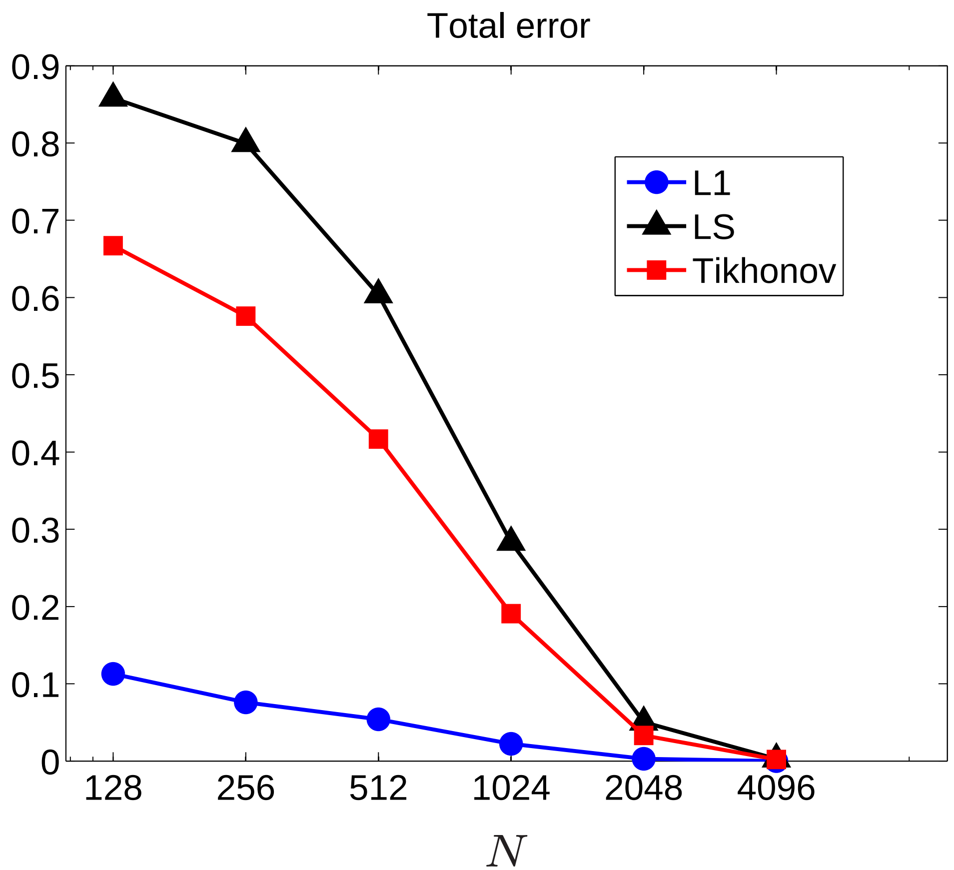
- **ML estimation with added Tikhonov regularization term**

$$\gamma \sum \|B_k\|_F^2 = \gamma \, \mathbf{tr}\, X$$

  equivalent to ML estimate using $C := C + \gamma I$

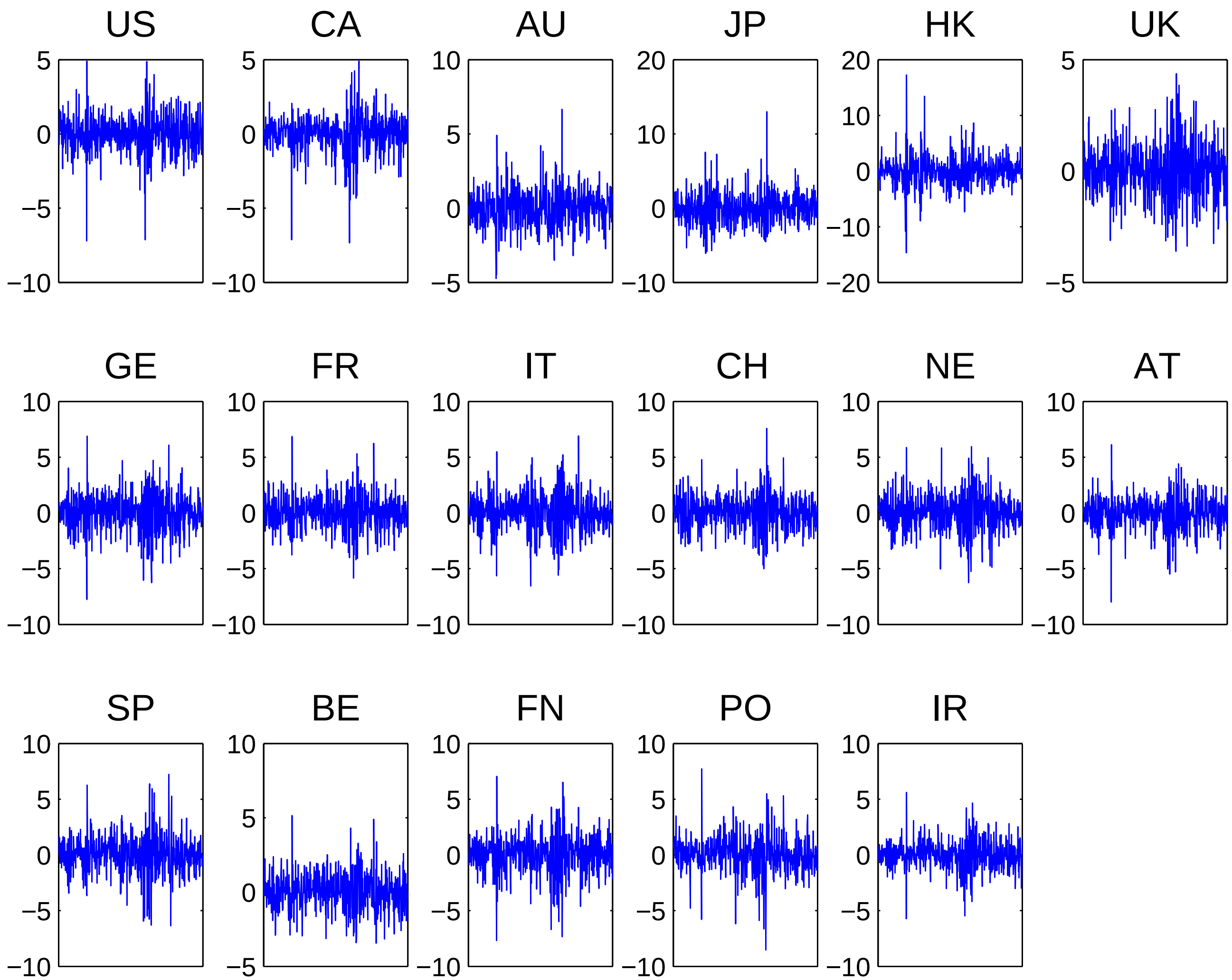
- **ML estimation with added nonsmooth regularization term** $h(Y)$

Error in topology as function of sample size



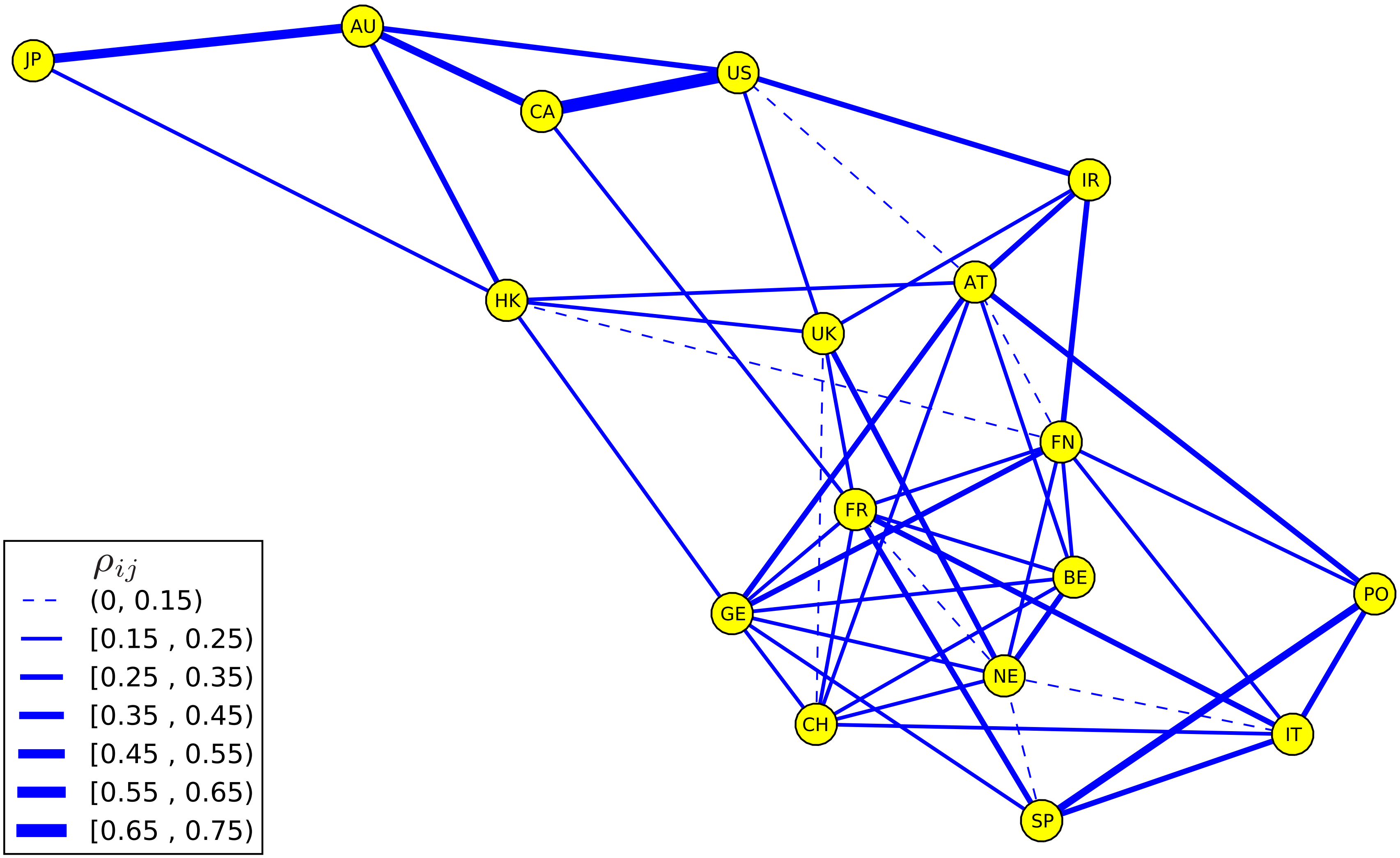
- topology estimated by applying threshold to ML estimate (LS), Tikhonov-regularized ML estimate, and l_1-regularized ML estimate (L1)
- graphs show fraction of entries misclassified as zero/nonzero

Example: 17 stock market indices



detrended daily returns of 17 market indices during 1997-1999; $N = 540$

Selected model ($p = 1$)



link widths show $\rho_{ij} = \max |R_{ij}(\omega)|$, $R(\omega)$ is normalized inverse spectrum

Example: fMRI time series

Dataset (Mitchell, 2004)

- time series of length $N = 640$
- 1718 voxels in ROI, reduced to $n = 7$, 50, 100, 190 by averaging groups
- two inputs: 'picture', 'sentence'

Experiment

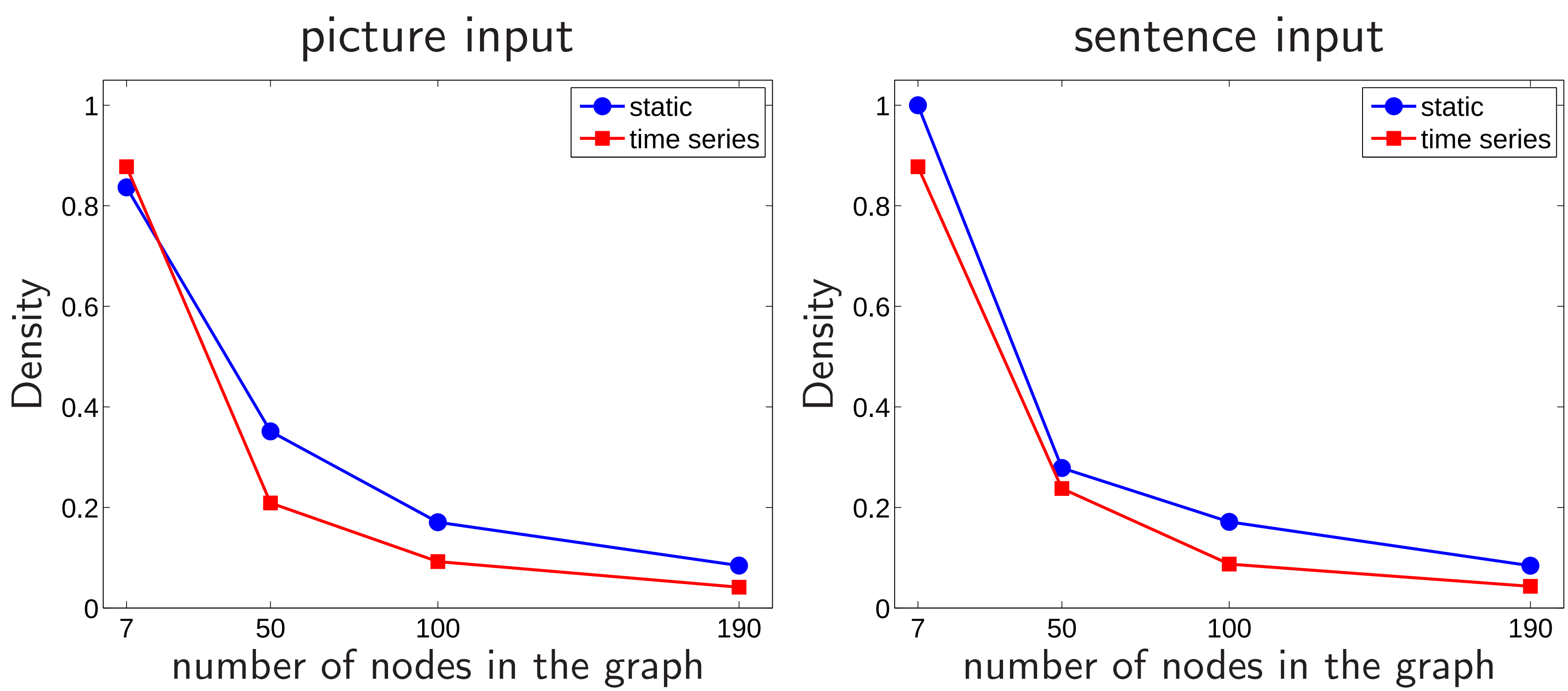
- estimate sparse models for the two inputs by regularized ML estimation
- validate models by an input classification experiment

Model selection from regularized ML estimation plus BIC

- model order

input	$n = 7$	$n = 50$	$n = 100$	$n = 190$
picture	$p = 1$	$p = 1$	$p = 0$	$p = 0$
sentence	$p = 1$	$p = 1$	$p = 0$	$p = 0$

- sparsity



Model validation via input classification

- use selected models to classify the two inputs from unseen data
- select the input with the highest likelihood

Classification error versus model size

model order	$n = 7$	$n = 50$	$n = 100$	$n = 190$
$p = 0$	21%	16%	**11%**	**6%**
$p = 1$	**20%**	**16%**	16%	11%

Outline

- Gaussian graphical models
- Graphical models of autoregressive time series
- **Algorithms**

Regularized ML problem

$$
\begin{array}{ll}
\text{minimize} & -\log\det X_{00} + \mathbf{tr}(CX) + \gamma h(Y) \\
\text{subject to} & X = \left[\begin{array}{ccc} X_{00} & \cdots & X_{0p} \\ \vdots & \ddots & \vdots \\ X_{p0} & \cdots & X_{pp} \end{array}\right] \succeq 0 \\
& Y_k = \sum_{l=0}^{p-k} X_{l,l+k}, \quad k = 0,\ldots,p
\end{array}
$$

- variables X and $Y = (Y_0, \ldots, Y_p)$ with Y_k and X_{ij} square of order n
- $h(Y_0, Y_1, \ldots, Y_p)$ is nonsmooth penalty (sum of infinity norms)

$$
h(Y) = \sum_{j>i} \max_k \max\{|Y_{k,ij}|, |Y_{k,ij}|\}
$$

Dual problem

$$\begin{array}{ll} \text{minimize} & f(C + \mathsf{T}(Z)) \\ \text{subject to} & \sum_{k=0}^{p} (|Z_{k,ij}| + |Z_{k,ji}|) \leq \gamma, \quad i \neq j \end{array}$$

variable $Z = (Z_0, Z_1, \ldots, Z_p)$, Z_k square with zero diagonal

- constraints are independent 1-norm constraints
- $\mathsf{T}(Z)$ is block-Toeplitz matrix with first row Z
- $f(V) = -\log\det(V_{00} - V_{1:p,0}^T V_{1:p,1:p}^\dagger V_{1:p,0})$ if

$$V = \begin{bmatrix} V_{00} & \cdots & V_{p0}^T \\ \vdots & \ddots & \vdots \\ V_{p0} & \cdots & V_{pp} \end{bmatrix} = \begin{bmatrix} V_{00} & V_{1:p,0}^T \\ V_{1:p,0} & V_{1:p,1:p} \end{bmatrix} \succeq 0$$

- objective is convex differentiable, and **closed** if C is block-Toeplitz

Gradient projection

$$\begin{array}{ll} \text{minimize} & g(x) \\ \text{subject to} & x \in \mathcal{C} \end{array}$$

g convex, differentiable; $\mathcal{C}$ a 'simple' convex set ($e.g.$, 1-norm ball)

Basic algorithm

$$x^{(k+1)} = \mathcal{P}\left(x^{(k)} - \alpha \nabla g(x^{(k)})\right)$$

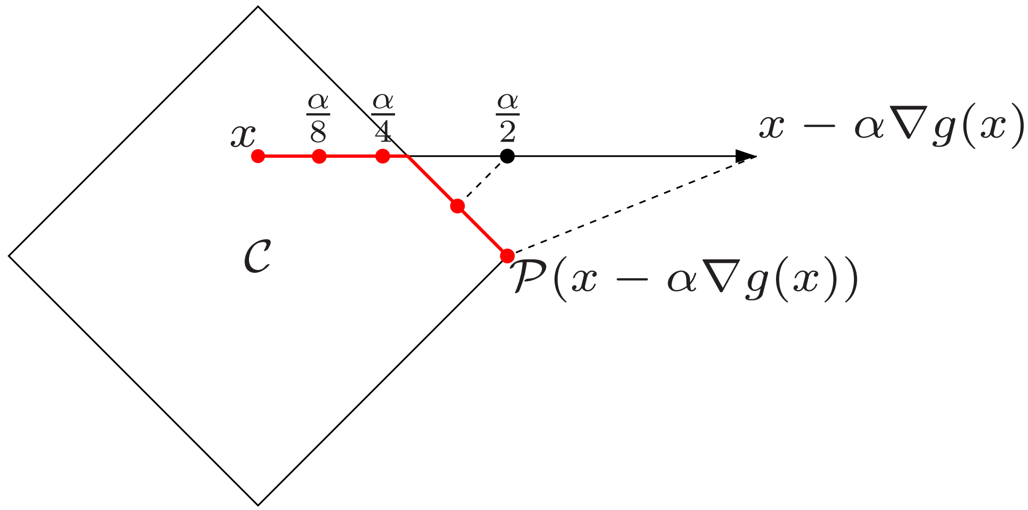
$\mathcal{P}$ is (inexpensive) projection on $\mathcal{C}$

Accelerated algorithms (Nesterov, Beck and Teboulle, Tseng, . . .)

- same complexity per step; faster ('optimal') convergence in theory
- known convergence theory does not apply to our problem

Step size

'Arc search': backtracking search for α



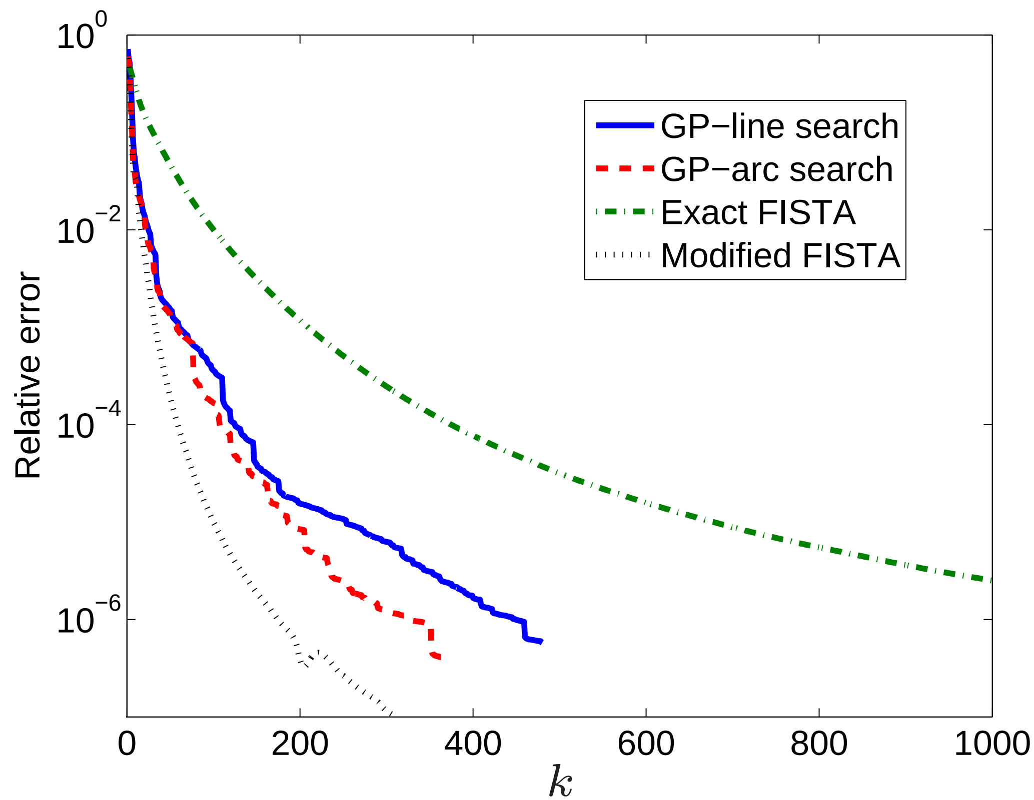
initialize α by Barzilai-Borwein stepsize

(Straight) line search

$$x := (1-\alpha)x + \alpha\mathcal{P}\left(x - t\nabla g(x)\right)$$

α determined by backtracking

Numerical example



- example with $n = 300$, $p = 2$ $(225150$ variables)
- exact FISTA uses decreasing stepsize (required in convergence proof)
- modified FISTA is FISTA with non-monotone stepsize

Summary: Graphical models of AR processes

Estimation with known topology

- convex relaxation of constrained ML estimation problem
- relaxation is exact if sample covariance matrix C is block-Toeplitz
- in practice, it is often exact for almost-Toeplitz C

Topology selection

- convex nonsmooth (ℓ_1-type) regularization of ML problem
- useful heuristic for reducing #models ranked by information criteria
- efficient solution via first-order methods applied to the dual

References

- J. Songsiri, J. Dahl, L. Vandenberghe, *Graphical models of autoregressive processes*, in: Y. Eldar, D. Palomar (Eds.), *Convex Optimization in Signal Processing and Communications*, 2010.

- J. Songsiri and L. Vandenberghe, *Topology selection in graphical models of autoregressive processes*, Journal of Machine Learning Research, 2010.