

## Some Questions of Interest

- Covariance matrix estimation
- Structural learning & clustering via covariance selection
- Incorporate in a Bayesian hierarchical model: (i) uncertainty about and (ii) time-variation of the sought-after graph structure
- Online (as opposed to offline) tracking & prediction

2

## Graphical Gaussian Models

- Undirected graph  $g = (V, E)$ . Vertex set  $V = \{1, \dots, d\}$
- Multivariate Normal  $N_d(0, \Sigma)$  density, with  $\mathbf{K} = \Sigma^{-1}$ :

$$f(\mathbf{x}) = f(x_1, \dots, x_d) \propto \exp\left\{-\frac{1}{2} \sum_{i \in V} \left[ k_{ii} x_i^2 + \sum_{j \neq i} k_{ij} x_i x_j \right]\right\}$$

- $g$  constrains  $\Sigma$  by zeroing entries of  $\mathbf{K} = \Sigma^{-1}$  :

$$k_{ij} = (\Sigma^{-1})_{ij} = 0 \text{ iff } \{i, j\} \notin E, i \neq j$$

- Similar to variable selection in regression (projection):

$$f(x_i | x_j, j \neq i) \propto \exp\left\{-\frac{1}{2} k_{ii} \left[ x_i - \sum_{\{i,j\} \in E} \frac{-k_{ij}}{k_{ii}} x_j \right]^2\right\}$$

3

# Online Structural Learning of Dynamic Graphical Models via Particle Filters

Makram Talih

Department of Mathematics and Statistics  
Hunter College (City University of New York)

e-mail: makram.talih@hunter.cuny.edu

August 9, 2005

## Some Motivating Problems

- Asset pricing & clustering (Talih & Hengartner, 2005; Aguilar & West, 2000)
- Medical monitoring of IC patients (Fried & Didelez, 2003; Gather et al., 2002)
- Telecommunications networks (Cortes et al., 2003)

1

## Parameterization

- For simplicity, assume standardized mean zero variables
- For graph  $g_t = (V, E_t)$ , let

$$k_{iit} = 1 \quad \text{and} \quad k_{ijt} = -\theta_{ijt}$$

where

$$\theta_{ijt} \in [-1, 1]$$

with the understanding that

$$\theta_{ijt} = 0 \quad \text{whenever} \quad \{i, j\} \notin E_t$$

and that  $\mathbf{K}_t > 0$

6

## Parameter evolution

- In our model, the parameters change only with the graph:

$$\theta_{ij,t+1} = \begin{cases} \theta_{ijt}, & \text{if } d(g_t, g_{t+1}) = 0 \\ 0, & \text{if } d(g_t, g_{t+1}) \neq 0 \text{ and } \{i, j\} \notin E_{t+1} \\ u_{ij,t+1}, & \text{if } d(g_t, g_{t+1}) \neq 0 \text{ and } \{i, j\} \in E_{t+1} \end{cases}$$

- If  $\{i, j\} \in E_t$ ,  $u_{ij,t+1}$  is a small perturbation of  $\theta_{ijt}$ . Otherwise, it is drawn from a U-shaped density on  $[-1, 1]$ , such as  $\omega(u) = 1/(\pi\sqrt{1-u^2})$

- Let  $\iota(\boldsymbol{\theta}_{t+1} | \boldsymbol{\theta}_t, g_{t+1}, g_t)$  be the resulting transition density

7

## A Model for Time-Varying Graphs

- Data conceptualized as stream of  $d$ -dimensional observations

$$\underbrace{\mathbf{X}_{1,1}, \dots, \mathbf{X}_{1,N_1}}_{N_1}, \underbrace{\mathbf{X}_{2,1}, \dots, \mathbf{X}_{2,N_2}}_{N_2}, \dots$$

- Block  $b$  of observations  $\mathbf{X}_{b1}, \dots, \mathbf{X}_{bN_b}$  IID from  $N_d(0, \Sigma_b)$
- $\mathbf{K}_b = \Sigma_b^{-1} \in \mathcal{M}^+(g_b)$  for an undirected graph  $g_b = (V, E_b)$
- $\mathcal{M}(g) = \{\mathbf{A} : \mathbf{A} = \mathbf{A}^T, A_{ij} = 0, \{i, j\} \notin E, i \neq j\}$
- $\mathcal{M}^+(g)$  : all positive definite matrices in  $\mathcal{M}(g)$

4

## Underlying: Slowly Varying Sequence of Graphs

- Successive graphs differ by:  $d(G_t, G_{t+1}) = \#(E_t \Delta E_{t+1})$

- Transitions are locally uniform:

$$\mathbf{P}[G_{t+1} = h | G_t = g, d(g, h) = r] = \frac{1}{\#R(g, r)} \mathbf{1}_{\{h \in R(g, r)\}}$$

- Make  $d(G_t, G_{t+1})$  independent of  $G_t$ , with, say,

$$\mathbf{P}[d(G_t, G_{t+1}) = r | \lambda] \propto \frac{\lambda^r e^{-\lambda}}{r!}, \quad r = 0, 1, \dots, d(d-1)/2$$

- Obtain, eg.

$$P_\lambda(g, h) \propto \frac{\lambda^{d(g,h)} e^{-\lambda}}{d(g, h)!} \times \frac{1}{\#R(g, d(g, h))}$$

5

### Particle Filtering: Extend Step

- Maintain a (large) collection  $\{\Delta_{0:t,m}\}_{m=1}^M$  of sample paths (called particles) leading up to time  $t$
- Just after time  $t$ , sample  $\tilde{\lambda}_m$  from  $\rho_t(\lambda_m | D_{t,m})$
- Using  $\tilde{\lambda}_m$  and  $\Delta_{t,m}$ , generate  $\tilde{\Delta}_{t+1,m}$  according to  $Q_{\tilde{\lambda}_m}$
- This gives an approximate sample from the posterior

$$q_{t+1}(\Delta_{1:t+1}, \lambda | \mathbf{x}_{1:t+1})$$

except that data at time  $t + 1$  hasn't yet been used!

10

### Particle Filtering: Weight and Resample Step

- For each  $m$ , compute the importance weights

$$\tilde{w}_m = f_{t+1}(\mathbf{x}_{t+1} | \tilde{\Delta}_{t+1})$$

- For each  $m$ , normalize the weights so they add up to 1:

$$w_m^* = \frac{\tilde{w}_m}{\sum_{m=1}^M \tilde{w}_m}$$

- Now, let the data speak (or forever hold its peace...):  
**Sample with replacement** from the set of particles according to the weights  $w_m^*$  !! [cf. Kitagawa's stratified sampling]
- Obtain a new (re-weighted) set of particles  $\{\Delta_{0:t+1,m}^*\}_{m=1}^M$

11

### Recursive Updating of Evidence...

- $\Delta_t = \{g_t, \theta_t\}$  is a dynamic target that we are trying to track/learn

- The transition from  $\Delta_t$  to  $\Delta_{t+1}$  is governed by

$$Q_\lambda(\Delta_t, \Delta_{t+1}) = P_\lambda(g_t, g_{t+1}) \times \iota(\theta_{t+1} | \theta_t, g_{t+1}, g_t)$$

- Also, have a static parameter  $\lambda > 0$  that controls the rate of structural change along the sequence of graphs

- A priori,  $\lambda \sim \text{Gamma}(\alpha, \beta)$

8

### ... Recursive Updating of Evidence

- At time  $t$ , we update our evidence about  $\lambda$  via the total graph distance travelled along the sequence:

$$D_t = \sum_{s=0}^{t-1} d(g_s, g_{s+1}),$$

$$\rho_t(\lambda | D_t) \propto \text{Gamma}(\alpha + D_t, \beta + t)$$

- For the dynamic component  $\Delta_t$ , key recursion is:

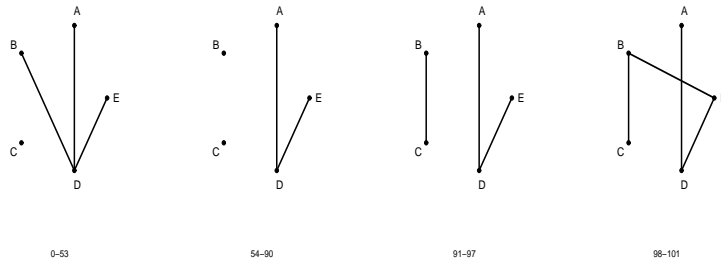
$$q_{t+1}(\Delta_{1:t+1}, \lambda | \mathbf{x}_{1:t+1}) \propto f_{t+1}(\mathbf{x}_{t+1} | \Delta_{t+1}) \times Q_\lambda(\Delta_t, \Delta_{t+1}) \times q_t(\Delta_{1:t}, \lambda | \mathbf{x}_{1:t})$$

where  $q_t(\Delta_{1:t}, \lambda | \mathbf{x}_{1:t}) = q_t(\Delta_{1:t} | \mathbf{x}_{1:t}) \rho_t(\lambda | D_t)$

9

### True Graph Sequence, stopped at T=101

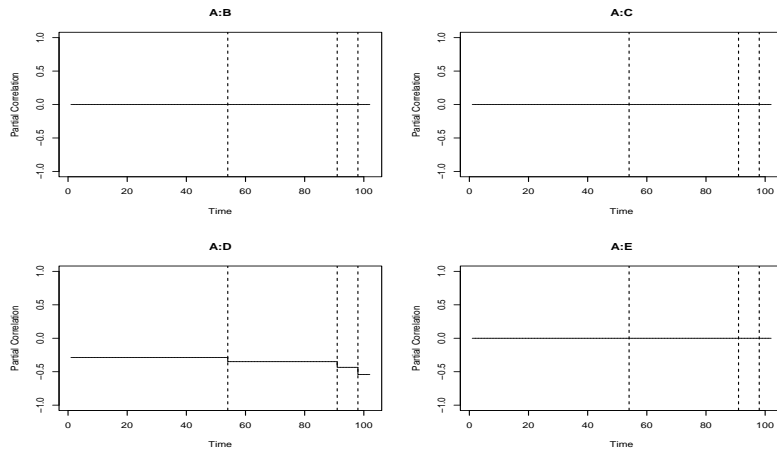
[Change Points at times 54, 91 & 98]



### Moving to Avoid Impoverishment

- In theory,  $M \rightarrow \infty$
- In practice, repeated resampling can lead to getting rid of all but one particle!
- Remedy is to slightly perturb each particle, one at a time
- MCMC algorithms are typically needed here
- Computationally expensive step. Nontrivial proposal design
- As long as such moves are reversible, we still have an adequate sample from the posterior  $q_{t+1}$

### Partial Correlation Sequence...



### Simulation Example

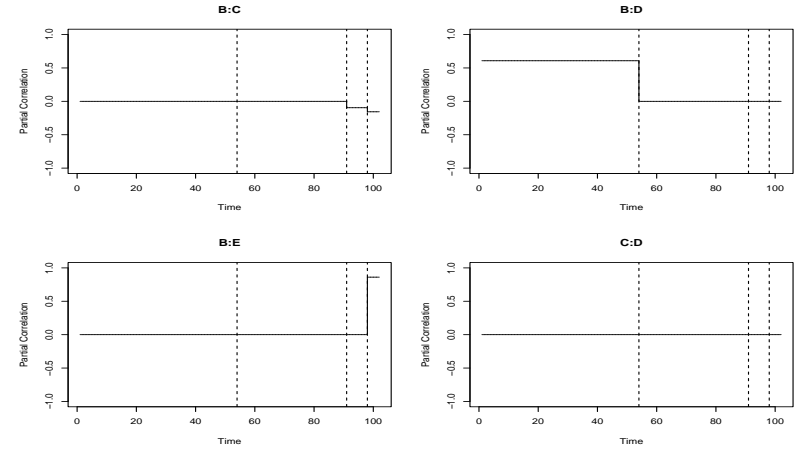
[lambda=0.085 – from Gamma(scale=1, rate=2)]

## Posterior Inference via filter of size 1000

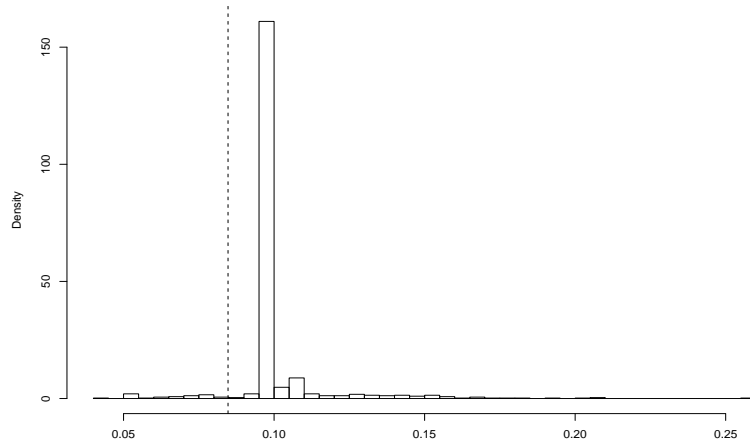
Average acceptance rates for MOVE step:

- Lambda: 14%
- Graph Sequence: 7%
- Partial correlations: 66%

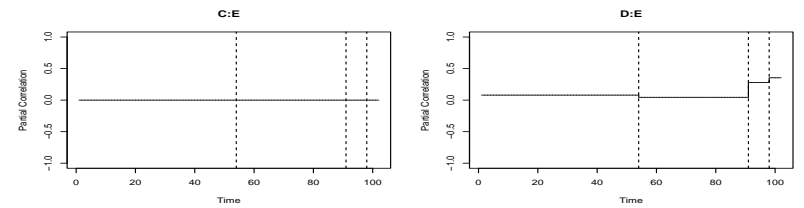
### ...Partial Correlation Sequence...



Posterior Distribution: Lambda

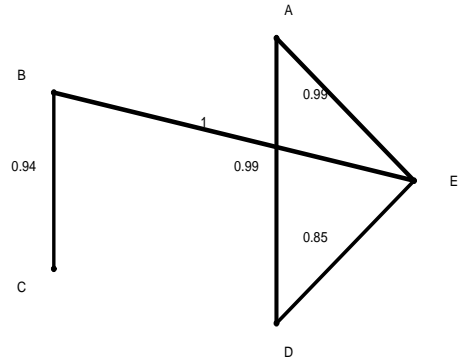


### ...Partial Correlation Sequence

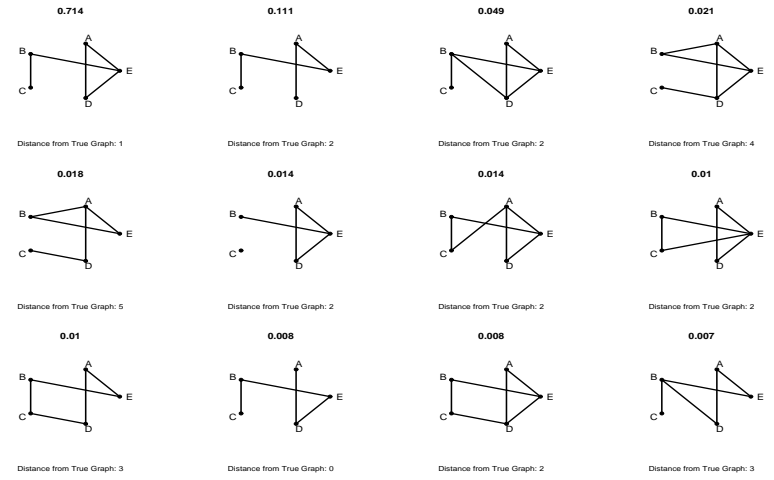


### Posterior "Median" Graph at time T

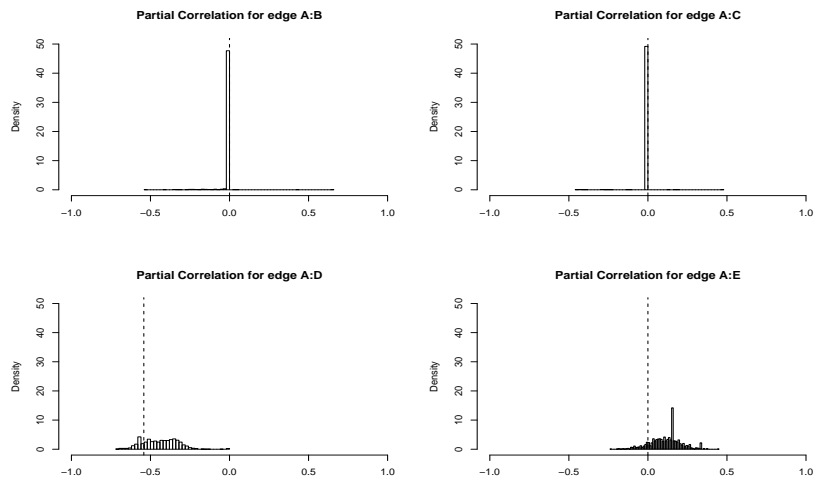
	A:B	A:C	A:D	A:E	B:C	B:D	B:E	C:D	C:E	D:E
true	0	0	1	0	1	0	1	0	0	1
median	0	0	1	1	1	0	1	0	0	1
mean	0.048	0.016	0.995	0.992	0.942	0.061	1	0.061	0.01	0.853



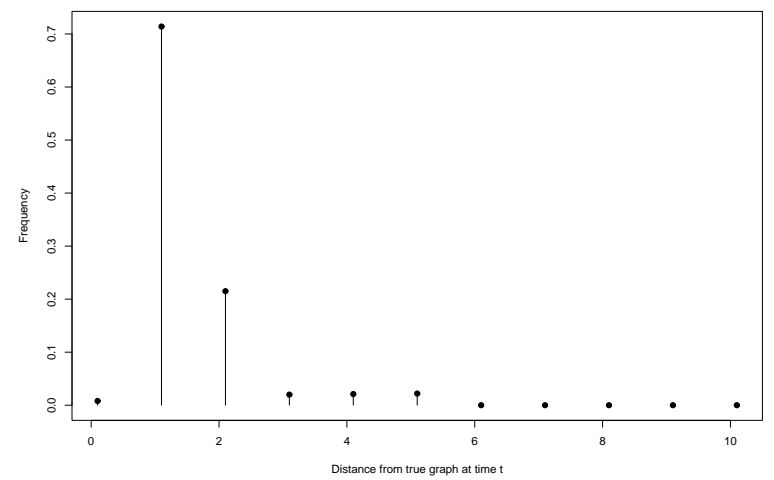
### Posterior Dist. of Graph at time T (First 12 only)



### Posterior Partial Correlations at time T...



### Posterior Distance from True Graph at time T



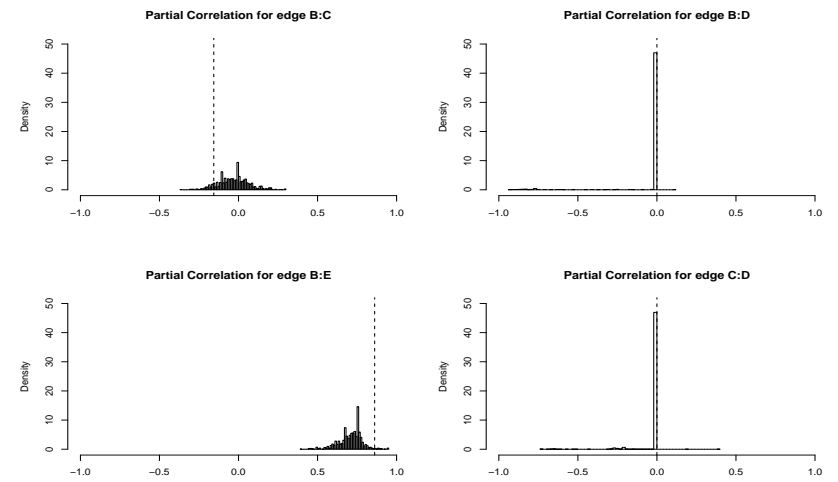
## Conclusions

- Methodology for inference and structural learning for time-varying graphical models
- Learning rate of structural change  $\lambda$  a posteriori
- Online prediction
- Extend to allow vertex set to change via:

$$d(g, h) = \#(V \Delta V') + \#(\bar{E} \Delta \bar{E}'),$$

identifying  $V$  with vectors in  $\{0, 1\}^{|\mathcal{V}|}$  for some over-arching  $\mathcal{V}$

### ...Posterior Partial Correlations at time T...



### ...Posterior Partial Correlations at time T...

