Learning linear Bayes networks with sparse Bayesian models

Statistical Modeling and Machine Learning in Computational Systems Biology June 22-26, 2009, Tampere, Finland

Ole Winther

Technical University of Denmark (DTU) & University of Copenhagen (KU)

June 24, 2009

All lectures

1. Introduction to graphical models and Bayesian networks
2. Estimating the size of the transcriptome
3. Using biological prior information in motif discovery
4. Learning linear Bayes networks with sparse Bayesian models

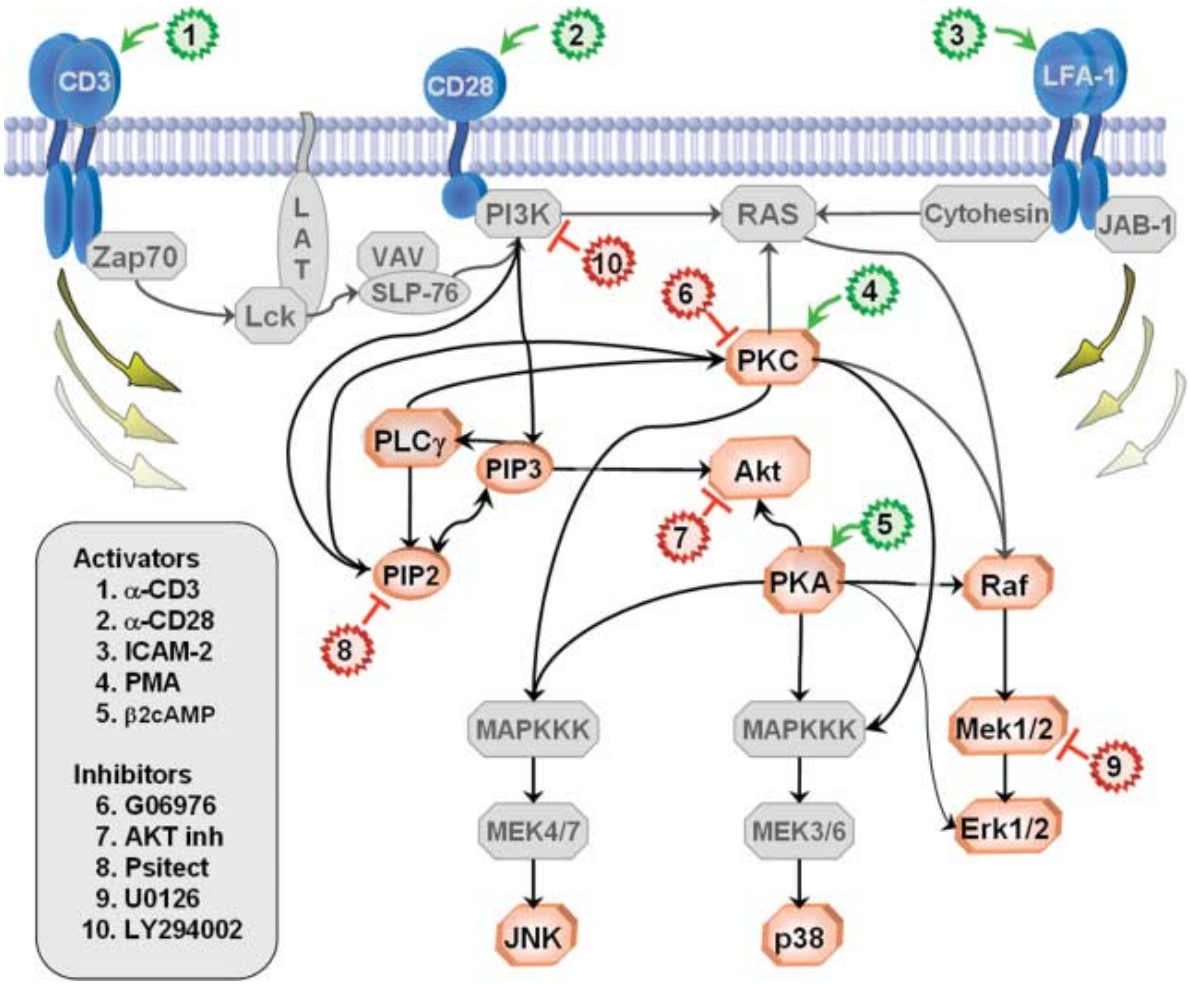
Common theme:

- Complex Bayesian model building possible and advantageous
- Model checking – prediction, marginal- and test-likelihood

Lecture 4

- Motivation - regulatory networks from multivariate data
- Learning identifiable and sparse factor models
- From factor models to DAGs - learn variable order.
- Model selection and comparison with test likelihood
- Extension to temporal processes

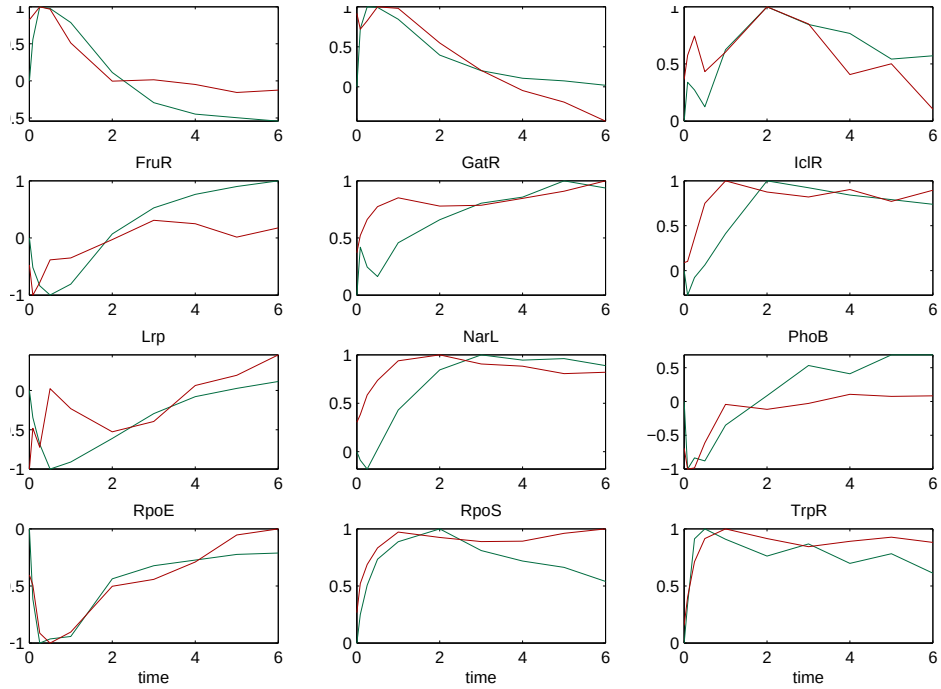
Protein signalling network textbook



- Single cell flow cytometry measurements of 11 phosphorylated proteins and phospholipids.
- Data was generated from a series of stimulatory cues and inhibitory interventions.
- Observational data: 1755 general stimulatory conditions,
- Experimental data $\sim$ 80% not used in our approach.
- Not “small n large d”!

E.Coli Transcription Factor network

- gene expression levels from 100 genes taken at 5, 15, 30 and 60 min, and every hour until 6 hours after transition from glucose to acetate (100×10).
- Objective is to find underlying transcription factor driving signal with or without *ground truth* regulatory networks (RegulonDB).



- A probabilistic model of $\mathbf{x}$ can be represented by a DAG

$$p(\mathbf{x}) = \prod_i p(x_i|\mathrm{Pa}(x_i))$$

- Linear DAG - $\mathbf{P}$ is an unknown permutation (order)

$$\mathbf{Px} = \mathbf{BPx} + \mathbf{Pz}\ , \quad \text{(DAG model)}$$

- $\mathbf{B}$ strictly lower triangular square matrix.
- Non-zero element of $\mathbf{B}$ corresponds to a link in the DAG.
- Noise-free factor model

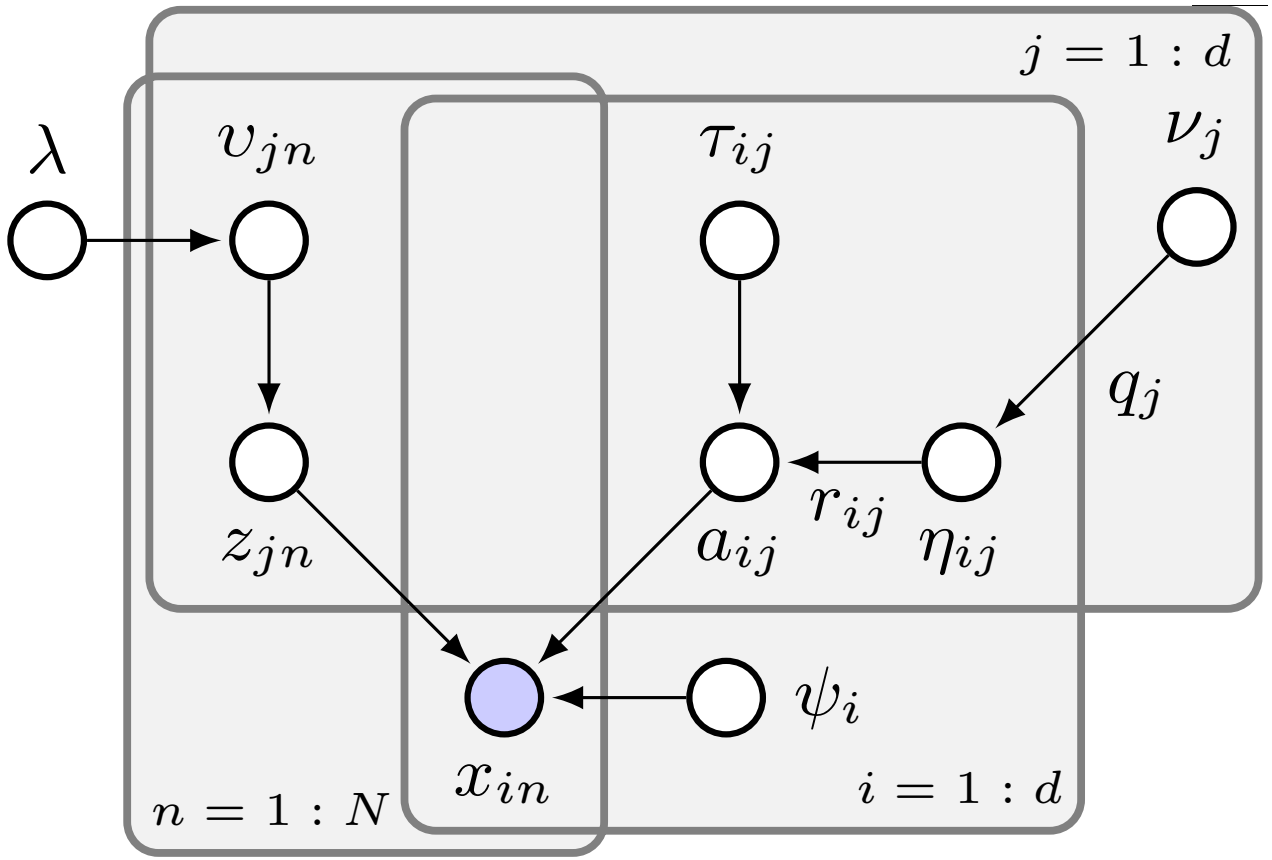
$$\mathbf{x} = \mathbf{P}^{-1}\mathbf{APz} = \mathbf{P}^{-1}(\mathbf{I} - \mathbf{B})^{-1}\mathbf{Pz}, \quad \text{(Factor model)}$$

- Idea: Learn sparse factor model

$$\mathbf{x} = \mathbf{P}_r \mathbf{A} \mathbf{P}_c \mathbf{z} + \epsilon,$$

- with row and column permutations $\mathbf{P}_r = \mathbf{P}$ and $\mathbf{P}_c = \mathbf{P}_f \mathbf{P}_r$ of $\mathbf{A}$
- such that the mixing matrix $\mathbf{A}$ is close to be triangular.
- $\mathbf{M}$ triangular mask: $\mathbf{A} \approx \mathbf{M} \odot \mathbf{A}$
- Learn sparse DAG model for fixed $\mathbf{P}$.
- $\mathbf{Px}$ is a DAG with ordering inferred by factor model.

DAGs & factor models



- Sparsity: spike and slab:

$$a_{ij}|r_{ij},\psi_i,\tau_{ij} \sim (1-r_{ij})\delta_0(\cdot) + r_{ij}\mathcal{N}(a_{ij}|0,\psi_i\tau_{ij})$$

  Plus more complications for the hierarchy for r_{ij}.

- Identifiability non-Gaussian

$$z_{jn}|\mu,\lambda \sim \text{Laplace}(z_{jn}|\mu,\lambda), \quad z_{jn}|\mu,\sigma^2,\theta \sim t(z_{jn}|\mu,\theta,\sigma^2)$$

- Infinite mixture representation:

$$\text{Laplace}(z|\mu,\lambda) = \int_0^\infty \mathcal{N}(z|\mu,\upsilon)\text{Exponential}(\upsilon|\lambda^2)d\upsilon$$

- Order search - no preference for any order

- Sparse prior $\rho(\mathbf{A}|\cdot)$ measure able to produce exact zeros in $\mathbf{A}$.
- Discrete spike and slab prior (West 2003, Lucas et. al. 2006),

$$\begin{aligned}
a_{ij}|r_{ij},\psi_i,\tau_{ij} &\sim (1-r_{ij})\delta_0(\cdot) + r_{ij}\mathcal{N}(a_{ij}|0,\psi_i\tau_{ij}),\\
r_{ij}|\eta_{ij} &\sim \text{Bernoulli}(r_{ij}|\eta_{ij}),\\
\eta_{ij}|q_j,\alpha_p,\alpha_m &\sim (1-q_j)\delta_0(\cdot) + q_j\text{Beta}(\eta_{ij}|\alpha_p\alpha_m,\alpha_p(1-\alpha_m)),\\
q_j|\nu_j &\sim \text{Bernoulli}(q_j|\nu_j),\\
\tau_{ij}^{-1}|t_s,t_r &\sim \text{Gamma}(\tau_{ij}^{-1}|t_s,t_r),\\
\nu_j|\beta_m,\beta_p &\sim \text{Beta}(\nu_j|\beta_p\beta_m,\beta_p(1-\beta_m)).
\end{aligned} \tag{1}$$

- No go for identifiability for Gaussian model $\overline{\mathbf{z}\mathbf{z}^T} = \mathbf{I}$:

$$\mathbf{X} = \mathbf{A}\mathbf{Z} = \mathbf{A}\mathbf{U}^{-1}\mathbf{U}\mathbf{Z} = \hat{A}\hat{Z}$$

- Second order statistics unchanged $\hat{\mathbf{z}} = \mathbf{U}\mathbf{z}$:

$$\overline{\hat{\mathbf{z}}\hat{\mathbf{z}}^T} = \mathbf{U}\overline{\mathbf{z}\mathbf{z}^T}\mathbf{U}^T = \mathbf{U}\mathbf{U}^T = \mathbf{I}\,.$$

- Non-Gaussianity is enough (Comon 1994). We use

$$\begin{aligned} z_{jn}|\mu,\lambda &\sim \text{Laplace}(z_{jn}|\mu,\lambda) \\ z_{jn}|\mu,\sigma^2,\theta &\sim t(z_{jn}|\mu,\theta,\sigma^2) \end{aligned}$$

- Process priors (temporal or spatial smoothness)
- Gaussian process is enough (more about that later)

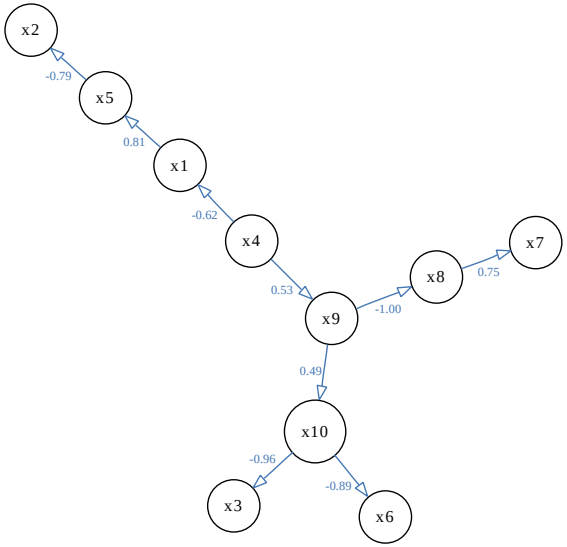
- All parameters apart from $\mathbf{P}$ standard Gibbs sampling!
- Order search - stochastic search over $\mathbf{P}_\mathrm{r}$ and $\mathbf{P}_\mathrm{c}$:
  - Proposal: $q(\mathbf{P}_\mathrm{r}^\star|\mathbf{P}_\mathrm{r})$ swaps two random rows and $q(\mathbf{P}_\mathrm{c}^\star|\mathbf{P}_\mathrm{c})$ swaps two random columns.
  - Metropolis-Hastings acceptance probability

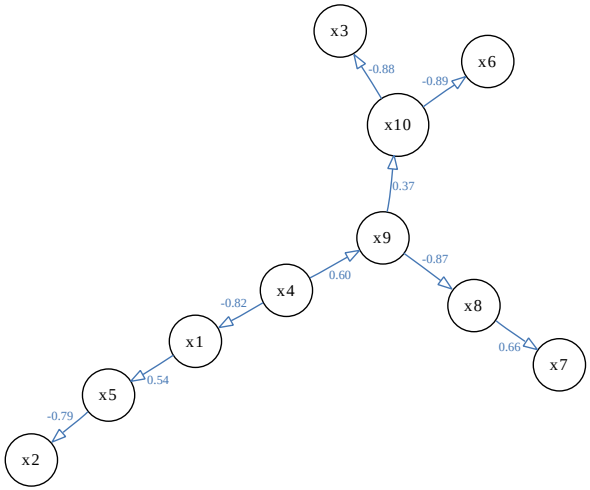
$$\min(1, \xi_{\rightarrow\star}) \quad \xi_{\rightarrow\star} = \frac{\mathcal{N}(\mathbf{X}|\mathbf{P}_\mathrm{r}^\star(\mathbf{M}\odot\mathbf{A})\mathbf{P}_\mathrm{c}^\star\mathbf{Z}, \Psi)}{\mathcal{N}(\mathbf{X}|\mathbf{P}_\mathrm{r}(\mathbf{M}\odot\mathbf{A})\mathbf{P}_\mathrm{c}\mathbf{Z}, \Psi)}.$$

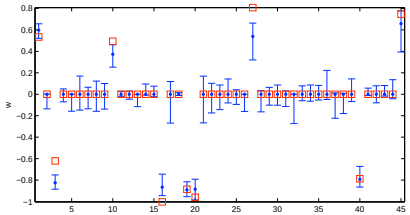
  - A lower triangular mask $\mathbf{M}$ breaks permutation symmetry.
- DAG - Gibbs sampling with $\mathbf{P}_\mathrm{r}$ top candidates:

$$\mathbf{X}|\mathbf{P}_\mathrm{r}, \mathbf{B}, \mathbf{X}, \cdot \sim \pi(\mathbf{X} - \mathbf{P}_\mathrm{r}^{-1}\mathbf{B}|\cdot), \quad \mathbf{B} \sim \rho(\mathbf{B}|\cdot),$$

Artificial data

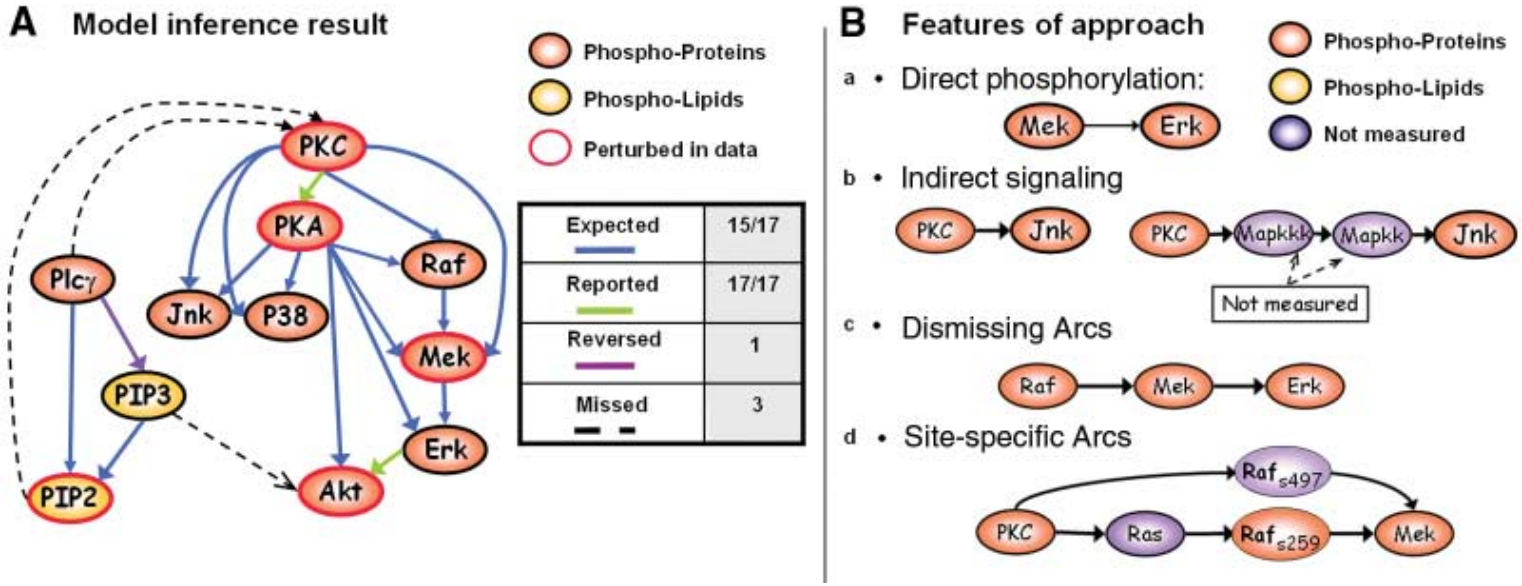






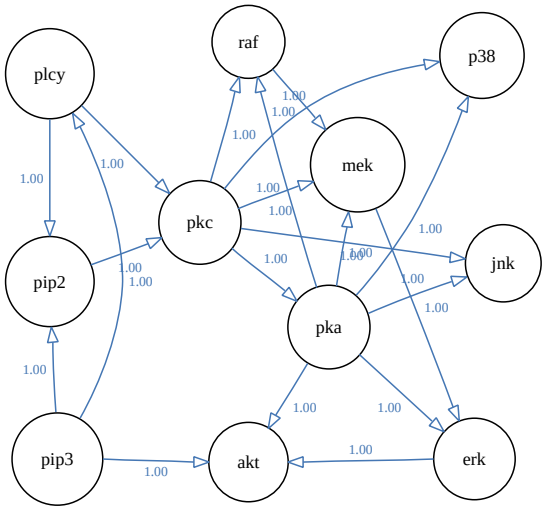
d	N	Method	TP (%)	TN (%)	AUC ($Q_{5\%}$)	OE
5	200	LINGAM	7294 (80.7%)	714 (73.7%)	500 (0.6)	42
		sFA	7428 (82.2%)	719 (74.2%)	608 (0.66)	42
5	500	LINGAM	7807 (86.4%)	607 (62.7%)	770 (0.0)	288
		sFA	7914 (87.6%)	775 (80.1%)	716 (0.7)	17
5	1000	LINGAM	8281 (90.9%)	765 (79.0%)	845 (0.2)	183
		sFA	8361 (92.5%)	654 (67.5%)	756 (0.7)	16
10	500	LINGAM	25836 (75,4%)	6566 (60.9%)	845 (0.06)	183
		sFA	28763 (84.0%)	7454 (69.2%)	179 (0.6)	462
10	1000	LINGAM	28281 (82.6%)	8012 (74.4%)	222 (0.00)	667
		sFA	31335 (87.4%)	8573 (79.6%)	261 (0.7)	265

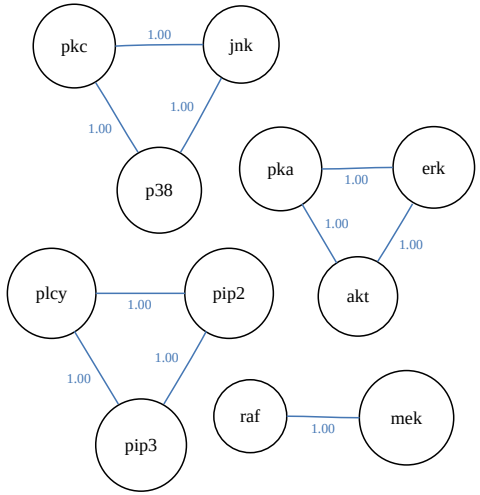
Protein signalling network learned

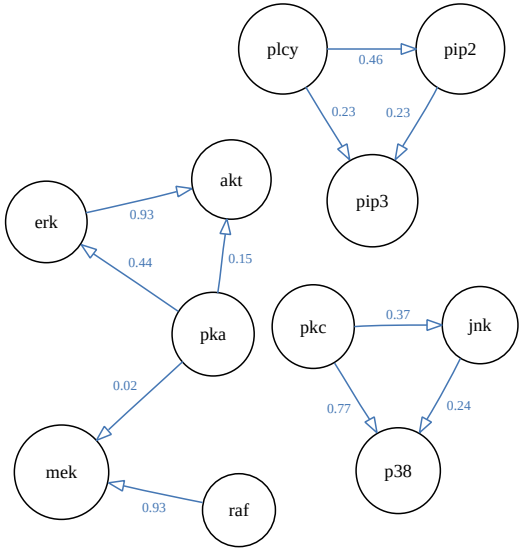


Sachs et. al. Science **308**, 523, (2005).

Protein signalling network learned







- Using textbook as ref: we found 10 true links (TP), one falsely added link (FP) and only two reversed links (RL)
- RL: $\mathrm{PIP}_2 \rightarrow \mathrm{PIP}_3$ is bidirectional (textbook) and $\mathrm{PLC}_\gamma \rightarrow \mathrm{PIP}_3$ also found reversed by Sachs et. al.

Model comparison with test likelihoods

- The likelihood of *intensive* variables **A** and Ψ on new data $\mathbf{X}^\star$.
- **Factor model:** Use scale mixture representation and integrate out

$$\begin{aligned} p(\mathbf{X}^\star|\mathbf{A},\Psi,\mathbf{X}) &= \int p(\mathbf{X}^\star|\mathbf{A},\mathbf{Z},\Psi)p(\mathbf{Z}|\cdot)d\mathbf{Z} \\ &\approx \frac{1}{\text{rep}}\prod_n \sum_r^{\text{rep}} \mathcal{N}(\mathbf{x}_n^\star|\mathbf{0},\mathbf{A}^T\mathbf{U}_n\mathbf{A}+\Psi), \end{aligned}$$

  where $\mathbf{U}_n = \text{diag}(\upsilon_{1n},\ldots,\upsilon_{dn})$ with υ_{jn} from the *prior*.
- **DAG model** Analytical integrate out **Z**:

$$p(\mathbf{X}^\star|\mathbf{B},\mathbf{X}) = \int p(\mathbf{X}^\star|\mathbf{B},\mathbf{X},\mathbf{Z})p(\mathbf{Z}|\cdot)d\mathbf{Z} = \prod_{i,n} \text{Laplace}(\mathbf{x}_n^\star|\mathbf{BX},\cdot)$$

- Artificial data – generate 500 random DAGs and 500 factor models with $d = 5$ and $N = 500, 1000$.
- Use 20% of data as test set.
- For $N = 500$ selects true DAGs 91.5% of the times and true factor models 89.2%.
- For $N = 1000$ the numbers are 98.5% and 94.6%
- Protein signalling network – factor model preferred - could be explained by the presence of non-measured components.

- Gaussian process (GP) $\mathbf{z}_j^T \sim \mathrm{GP}(\mathbf{z}_j^T|\mathbf{0}, \mathbf{K}_j)$.
- $\mathbf{K}_j$ covariance function of factor j:

$$k_j(t_n, t_{n'}) = \exp(-\upsilon_j(t_n - t_{n'})^2) \quad \mathbf{K} = \mathrm{block}(\mathbf{K}_1, \ldots, \mathbf{K}_m)$$

- Inverse squared length-scale υ:
  $\upsilon_j|u_s, \kappa \sim \mathrm{Gamma}(\upsilon_j|u_s, \kappa)$.
- t-process (Yu et. al. 2007) $\mathbf{z}_j^T \sim \mathrm{TP}(\mathbf{z}_j^T|\mathbf{0}, \mathbf{K}_j, \theta_j)$.
- Scale mixture representation - Just one parameter needed!

$$\mathbf{z}_j^T \sim \mathcal{N}(\mathbf{z}_j^T|\mathbf{0}, \frac{1}{\tau_j}\mathbf{K}_j) \quad \tau_j \sim \mathrm{Gamma}(\tau|\frac{\theta}{2}, \frac{\theta}{2})$$

- Equivalent to a GP with a Gamma-prior over the inverse scale of the kernel $k_j(t_n, t_{n'}) = \exp(-\upsilon_j(t_n - t_{n'})^2)/\tau_j$

No go for identifiability for Gaussian model $\overline{\mathbf{z}\mathbf{z}^T} = \mathbf{I}$:

$$\mathbf{X} = \mathbf{A}\mathbf{Z} = \mathbf{A}\mathbf{U}^{-1}\mathbf{U}\mathbf{Z} = \widehat{A}\widehat{Z}$$

Second order statistics unchanged $\widehat{\mathbf{z}} = \mathbf{U}\mathbf{z}$:

$$\overline{\widehat{\mathbf{z}}\widehat{\mathbf{z}}^T} = \mathbf{U}\overline{\mathbf{z}\mathbf{z}^T}\mathbf{U}^T = \mathbf{U}\mathbf{U}^T = \mathbf{I}\,.$$

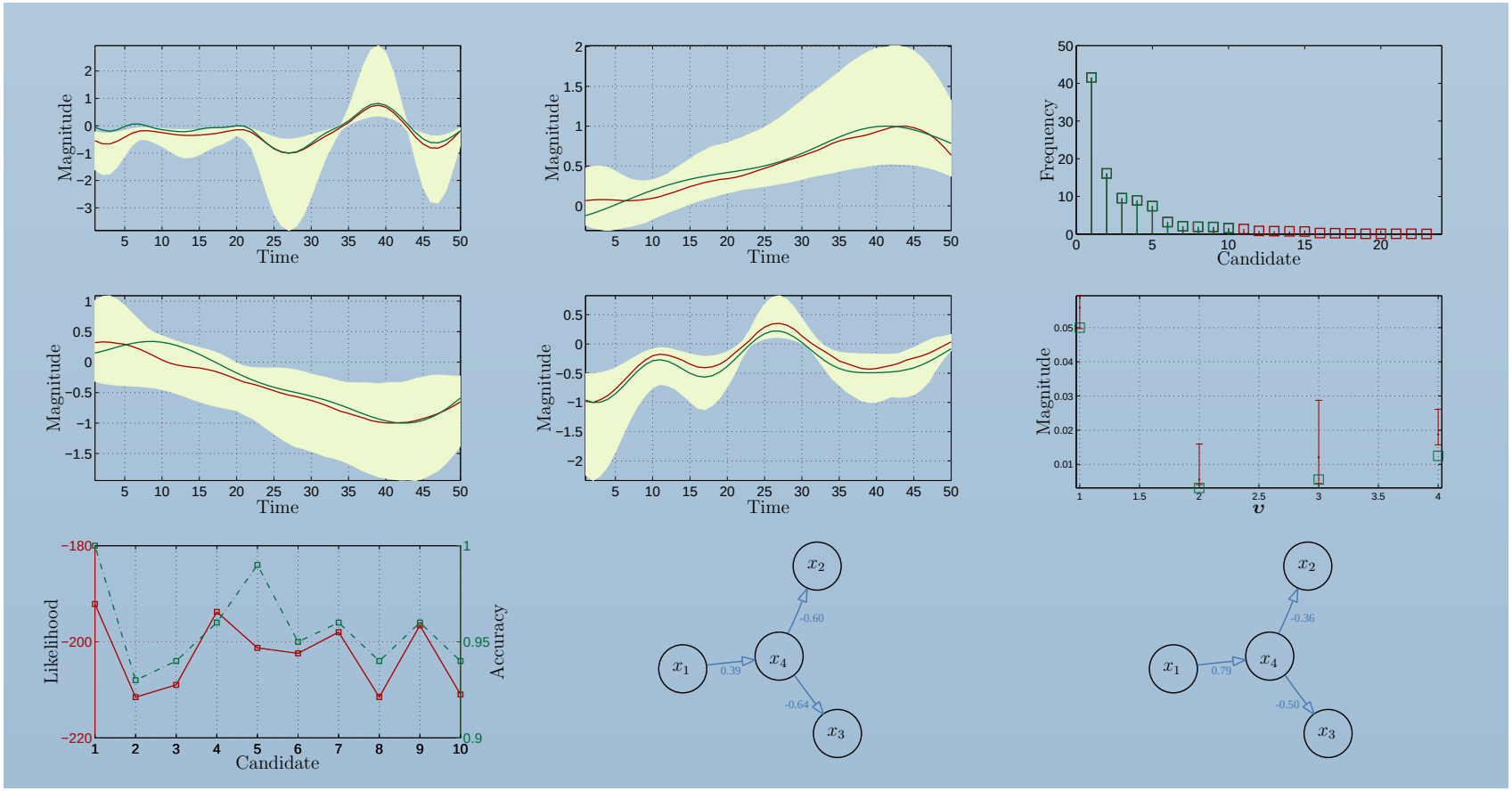
Enter Gaussian process: $\overline{z_{jn}z_{j'n'}} = \delta_{jj'}K_{j,nn'}$

$$\overline{z_{jn}z_{j'n'}} = \sum_{kk'} u_{jk}u_{j'k'}\overline{z_{kn}z_{k'n'}} = \sum_{k} u_{jk}u_{j'k}K_{k,nn'} \neq \delta_{jj'}K_{j,nn'}$$

if all kernels are different

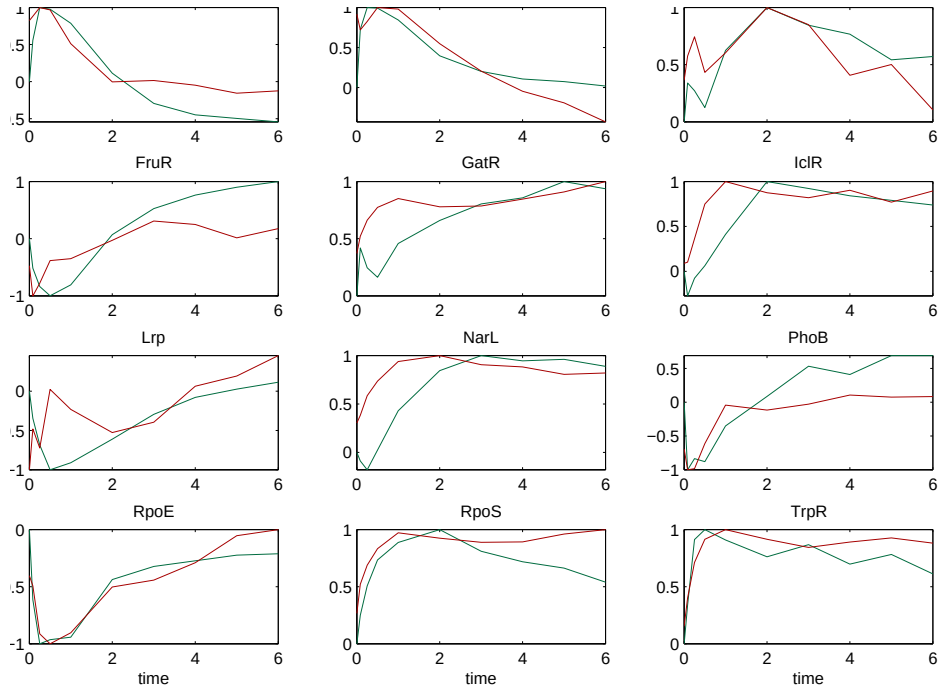
$$K_{j,nn'} \neq K_{j,nn'} \qquad \forall j, j'$$

Temporal processes artificial data



E.Coli Transcription Factor network

- Objective is to find underlying transcription factor driving signal with or without *ground truth* regulatory networks (RegulonDB).
- Our method with learned and fixed **A** give similar activities. But learned and "true" **A** somewhat different. Use model selection to decide which one is the best one.



- Hybrid model

$$\mathbf{x} = \mathbf{A}\mathbf{z} + \mathbf{B}\mathbf{x} + \epsilon$$

- Interventions = experimental data: easy in DAG and difficult in factor model!

Summary

- Sparse Bayesian linear models for structure learning (w Ricardo Henao, DTU and KU)
- Rich and flexible framework modeling linear latent and DAG structure
- Model comparison and checking - very important in biology. Not at all fully developed yet:
  - Compare models with inferred structure to “ground truth”.
  - Compare models with temporal smoothness (with different kernels robust) to iid (with different priors).

References

1. Pierre Comon. Independent component analysis, a new concept? Signal Processing, 36, 287-314, December 1994.
2. Mike West. Bayesian factor regression models in the “large p, small n” paradigm. In J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, and M. West, editors, Bayesian Statistics 7, pages 723-732. Oxford University Press, 2003.
3. J. Lucas, C. Carvalho, Q. Wang, A. Bild, J.R. Nevins, and M. West. Bayesian Inference for Gene Expression and Proteomics, chapter Sparse Statistical Modeling in Gene Expression Genomics, pages 155-176. Cambridge University Press, 2006.
4. Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-Gaussian acyclic model for causal discovery. Journal of Machine Learning Research, 7, 2003-2030, October 2006.
5. K.C. Kao, Y-L. Yang, R. Boscolo, C. Sabatti, V. Roychowdhury, and J.C. Liao. Transcriptome-based determination of multiple transcription regulator activities in escherichia coli by using network component analysis. PNAS, 101(2): 641-646, January 2004.
6. S. Yu, V. Tresp, and K. Yu. Robust multi-task learning with t -processes. In Proceedings of the 24th International Conference on Machine Learning, volume 227, pages 1103-1110, 2007.