Bayesian Sparse Factor Models and DAGs Inference and Comparison

Ricardo Henao and Ole Winther

DTU Compute · Technical University of Denmark, Kgs. Lyngby, Denmark

Bioinformatics Centre · University of Copenhagen, Copenhagen N, Denmark

Technical University of Denmark

DTU Compute
Department of Applied Mathematics and Computer Science

Introduction

- We present a novel approach to learn directed acyclic graphs (DAGs) and factor models within the same framework while also allowing for model comparison between them.
- We exploit the connection between factor models and DAGs to propose Bayesian hierarchies based on spike and slab priors to promote sparsity, heavy-tailed priors to ensure identifiability and predictive densities to perform the model comparison.

Model Specification

From DAGs to Factor Models

We assume that an ordered d-dimensional data vector $\mathbf{Px}$ can be represented as a DAG with only observed nodes, where $\mathbf{P}$ is an unknown permutation matrix, thus

$$\mathbf{x} = \mathbf{P}^{-1}\mathbf{BPx} + \mathbf{z} \ , \ \text{(DAG model)} \tag{1}$$

where $\mathbf{B}$ is a strictly lower triangular square matrix and $\mathbf{z}$ is a driving signal. In this setting, each non-zero element of $\mathbf{B}$ corresponds to a link in the DAG. Solving for $\mathbf{B}$ we can rewrite

$$\mathbf{x} = \mathbf{P}^{-1}\mathbf{APz} = \mathbf{P}^{-1}(\mathbf{I}-\mathbf{B})^{-1}\mathbf{Pz} \ , \ \text{(Noiseless factor model)} \tag{2}$$

- $\mathbf{P}^{-1}\mathbf{AP}$ must be sparse so it can be permuted to lower triangular since $(\mathbf{I}-\mathbf{B})^{-1}$ is triangular.
- $\mathbf{z}$ must be non-Gaussian to ensure identifiability [**?**].
- $\mathbf{P}$ is unknown, we can estimate $\mathbf{P}^{-1}\mathbf{AP}$ and then stochastically find $\mathbf{P}$.

From Factor Models to DAGs

Instead of using the noise-free factor model of equation (**??**) we allow for additive noise

$$\mathbf{x} = \mathbf{P}_{\mathrm{r}}^{-1}\mathbf{AP}_{\mathrm{c}}\mathbf{z} + \boldsymbol{\epsilon} \ , \ \text{(Factor model)}$$

where $\boldsymbol{\epsilon}$ is uncorrelated Gaussian noise, $\mathbf{P}_{\mathrm{r}} = \mathbf{P}$ is the permutation matrix for the rows of $\mathbf{A}$ and $\mathbf{P}_{\mathrm{c}} = \mathbf{P}_{\mathrm{f}}\mathbf{P}_{\mathrm{r}}$ another permutation for the columns of $\mathbf{A}$ with $\mathbf{P}_{\mathrm{f}}$ accounting for the permutation freedom of the factors. The Bayesian model is specified as follows

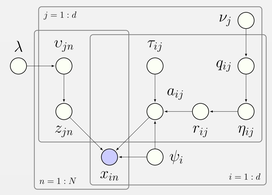
$$\begin{aligned}
\mathbf{X}|\mathbf{P}_{\mathrm{r}},\mathbf{A},\mathbf{P}_{\mathrm{c}},\mathbf{Z},\boldsymbol{\Psi} &\sim \mathcal{N}(\mathbf{X}|\mathbf{P}_{\mathrm{r}}^{-1}\mathbf{AP}_{\mathrm{c}}\mathbf{Z},\boldsymbol{\Psi}) \ , \quad \mathbf{Z} \sim \pi(\mathbf{Z}|\cdot) \ , \ \text{(Heavy-tailed factor prior)} \\
\psi_i^{-1}|s_s,s_r &\sim \mathrm{Gamma}(\psi_i^{-1}|s_s,s_r) \ , \quad \mathbf{A} \sim \rho(\mathbf{A}|\cdot) \ , \ \text{(Sparse mixing prior)}
\end{aligned}$$

Identifiability: We are restricted to use non-Gaussian distributions $\pi(\mathbf{Z}|\cdot)$ for the factors $\mathbf{z}_n$, here we use Laplace distributions parameterized as scale mixtures of Gaussians [**?**]

$$\begin{aligned}
z_{jn}|\mu,\lambda &\sim \mathrm{Laplace}(z_{jn}|\mu,\lambda) = \int_0^\infty \mathcal{N}(z_{jn}|\mu,v)\mathrm{Exponential}(v_{jn}|\lambda^2)dv_{jn} \ , \\
\lambda^2|\ell_s,\ell_r &\sim \mathrm{Gamma}(\lambda^2|\ell_s,\ell_r) \ ,
\end{aligned}$$

Sparsity: We require a sparse prior $\rho(\mathbf{A}|\cdot)$ able to produce exact zeros in $\mathbf{A}$. Here we adopt a two-layer discrete spike and slab prior for the elements a_{ij} of $\mathbf{A}$ similar to the one in [**?**]

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



We make the following Bayesian specification of linear DAG model of equation (**??**) as

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

where $\pi(\cdot)$ and $\rho(\cdot)$ are given above, $\mathbf{B}$ is a strictly lower triangular matrix and we use $\lambda_1,\ldots,\lambda_d$ to compensate for the fixed scaling of $\mathbf{X}$.

Permutation Search, $\mathbf{P}_{\mathrm{r}}$ and $\mathbf{P}_{\mathrm{c}}$: We perform a stochastic search over the space of all possible $d!$ orderings in the form of a Metropolis-Hastings (MH) algorithm.

- Acceptance probability $\min(1,\xi_{\to\star})$ where $\xi_{\to\star} = \dfrac{\mathcal{N}(\mathbf{X}|(\mathbf{P}_{\mathrm{r}}^\star)^{-1}(\mathbf{M}\circ\mathbf{P}_{\mathrm{r}}^\star\mathbf{A}(\mathbf{P}_{\mathrm{c}}^\star)^{-1})\mathbf{P}_{\mathrm{c}}^\star,\boldsymbol{\Psi})}{\mathcal{N}(\mathbf{X}|\mathbf{P}_{\mathrm{r}}^{-1}(\mathbf{M}\circ\mathbf{P}_{\mathrm{r}}\mathbf{AP}_{\mathrm{c}}^{-1})\mathbf{P}_{\mathrm{c}},\boldsymbol{\Psi})}$.
- Symmetric proposal consisting on a single uniform random transposition of $\mathbf{P}_{\mathrm{r}}$ and $\mathbf{P}_{\mathrm{c}}$.
- $\mathbf{M}$ is lower triangular and binary, to break the invariability of the model to permutations.

Predictive distributions: we use $p(\mathbf{X}^\star|\mathbf{X},\mathcal{M})$ with $\mathcal{M} = \{\mathcal{M}_{\mathrm{FA}},\mathcal{M}_{\mathrm{DAG}}\}$ instead of marginal likelihoods. With Gibbs sampling, we draw samples from $p(\mathbf{A},\boldsymbol{\Psi},\lambda|\mathbf{X},\cdot)$ and $p(\mathbf{B},\lambda_1,\ldots,\lambda_m|\mathbf{X},\cdot)$. Then we average over $p(\mathbf{Z}^\star|\cdot)$ for a test set $\mathbf{X}^\star$ using (permutation matrices are omitted for clarity)

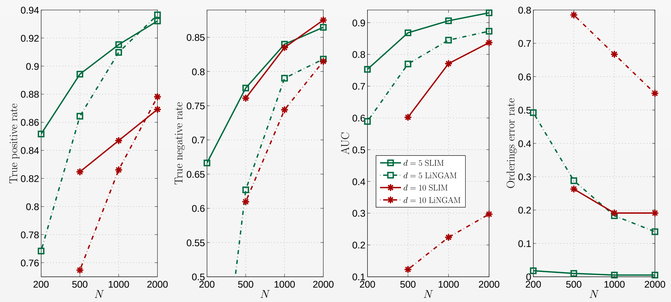
$$\begin{aligned}
p(\mathbf{X}^\star|\mathbf{A},\boldsymbol{\Psi},\cdot) &= \int p(\mathbf{X}^\star|\mathbf{A},\mathbf{Z},\boldsymbol{\Psi})p(\mathbf{Z}|\cdot)d\mathbf{Z} \approx \frac{1}{\mathrm{rep}}\prod_n\sum_r^{\mathrm{rep}}\mathcal{N}(\mathbf{x}_n^\star|\mathbf{0},\mathbf{A}^\top\mathbf{U}_n\mathbf{A}+\boldsymbol{\Psi}) \ , \ \text{(factor model)} \\
p(\mathbf{X}^\star|\mathbf{B},\cdot) &= \int p(\mathbf{X}^\star|\mathbf{B},\mathbf{X},\mathbf{Z})p(\mathbf{Z}|\cdot)d\mathbf{Z} = \prod_{i,n}\mathrm{Laplace}(x_{ij}|[\mathbf{BX}]_{in},\lambda_i) \ , \ \text{(DAG)}
\end{aligned}$$

where $\mathbf{U}_n = \mathrm{diag}(\upsilon_{1n},\ldots,\upsilon_{dn})$, the υ_{jn} are sampled from the prior and $[\mathbf{BX}]_{ij}$ is element of $\mathbf{BX}$.

Experiments

LiNGAM suite

- We compare against LiNGAM using the artificial model generator presented with LiNGAM [**?**].
- Both dense and sparse non-Gaussian networks with different degree of sparsity.
- The variables are randomly permuted to hide the correct order, $\mathbf{P}$.
- We consider $d = \{5,10\}$ and $N = \{200,500,1000,2000\}$.

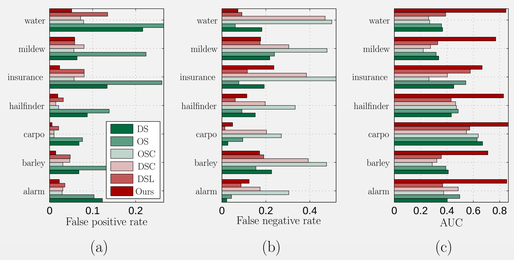


(a) (b) (c) (d)

FIGURE 1: Performance measures for LiNGAM suite. (a) True positive rate. (b) True negative rate. (c) Frequency of AUC being greater than 0.9. (d) Number of estimated correct orderings.

Bayesian networks repository

- 7 structures: alarm ($d = 37$), barley (48), carpo (61), hailfinder (56), insurance (27), mildew (35) and water (32).
- A single dataset of size $N = 1000$ is generated from each network.
- Comparison against: L1MB then DAG-search (DSL) [**?**].



(a) (b) (c)

FIGURE 2: Performance measures for Bayesian networks repository experiments.

Model comparison

- 1000 different datasets with $d = 5$ and $N = \{500,1000\}$.
- Approximately half of the datasets were generated using DAGs.
- We kept 20% of the data to compute the predictive densities to then select between DAGs and factor models.

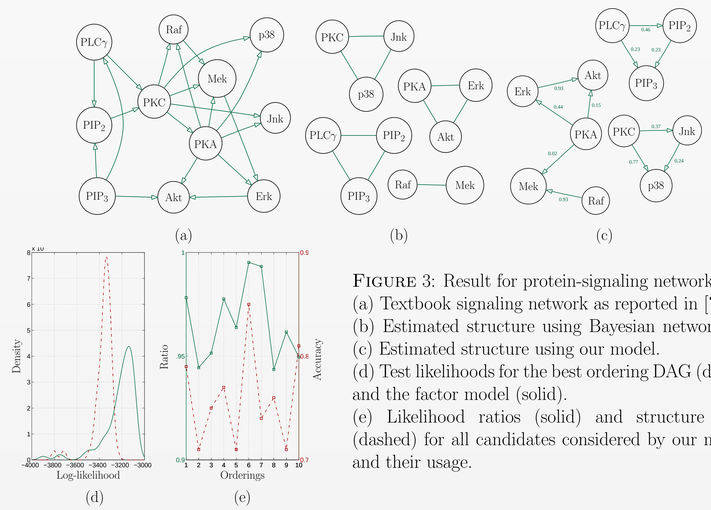
N	True DAG	True factor model	Error
500	91.5%	89.2%	9.6%
1000	98.5%	94.6%	5.0%

TABLE 1: Model selection accuracies and overall error rates.

Protein-signaling network

The dataset introduced by [**?**] consists on flow cytometry measurements of 11 different proteins.

- Observations are vectors of quantitative amounts measured from single cells.
- Data generated from a series of stimulatory cues and inhibitory interventions.
- Observational data only, 1755 observations corresponding to general stimulatory conditions.
- Our method found 10 true links (TP), one falsely added link (FP).
- Our method found two reversed links (RL). $\mathrm{PIP}_2 \to \mathrm{PIP}_3$ is bidirectional and $\mathrm{PLC}\gamma \to \mathrm{PIP}_3$ was also found reversed in [**?**] using interventional data.
- We also tried the methods above. Results were: TP$\approx$ 9, TN$\approx$ 32 and RL$\geq$ 6.



(a) (b) (c) (d) (e)

FIGURE 3: Result for protein-signaling network. (a) Textbook signaling network as reported in [**?**]. (b) Estimated structure using Bayesian networks [**?**]. (c) Estimated structure using our model. (d) Test likelihoods for the best ordering DAG (dashed) and the factor model (solid). (e) Likelihood ratios (solid) and structure errors (dashed) for all candidates considered by our method and their usage.

Conclusions & Outlook

- Novel approach to perform inference and model comparison of sparse factor models and DAGs within the same framework.
- First time that a method for comparing such a closely related linear models is proposed.
- Results on artificial and real data showed that our method significantly outperforms state-of-the-art techniques for structure learning.
- Currently investigating extensions to other source distributions (non-parametric Dirichlet process, temporal Gaussian processes and discrete).

References

[1] P. Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314, December 1994.

[2] D. F. Andrews and C. L. Mallows. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society: Series B (Methodology)*, 36(1):99–102, 1974.

[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] S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030, October 2006.

[5] M. W. Schmidt, A. Niculescu-Mizil, and K. P. Murphy. Learning graphical model structure using L1-regularization paths. In *AAAI*, pages 1278–1283, 2007.

[6] K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, April 2005.