FROSTY: a high-dimensional scale-free Bayesian network learning method

Joshua Bang[1], Sang-Yun Oh[1,2]

University of California, Santa Barbara[1]
Lawrence Berkeley National Lab[2]

Table of contents

Introduction

Graphical models

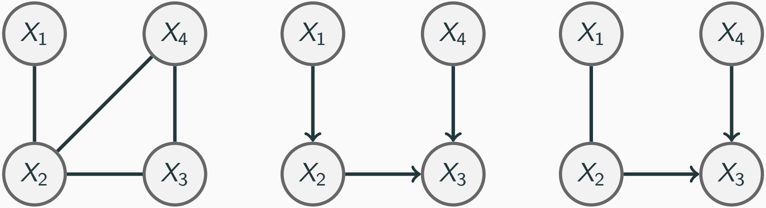

Figure 1: Undirected Graph (UG), Directed Acyclic Graph (DAG), Completed Partially Directed Acyclic Graph (CPDAG)

Graphical models

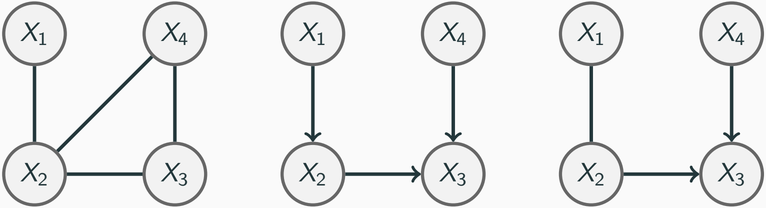

Figure 1: Undirected Graph (UG), Directed Acyclic Graph (DAG), Completed Partially Directed Acyclic Graph (CPDAG)

$$\text{UG}: \; P(X) = \frac{1}{Z}\psi_1(X_1, X_2)\,\psi_2(X_2, X_3, X_4)$$

$$\begin{aligned}\text{DAG}: \; P(X) &= \prod_{X_v \in \mathcal{V}} P(X_v \mid X_{pa(X_v)}) \\ &= P(X_1)\,P(X_2 \mid X_1)\,P(X_3 \mid X_2, X_4)\,P(X_4)\end{aligned}$$

Bayesian networks

- Indistinguishablility (Markov equivalence class)
  e.g., $V = \{X_1, X_2, X_3\}, \;\; I(P) = \{X_1 \perp X_3 \mid X_2\}$

Figure 2: Markov Equivalence Class

Bayesian networks

- Indistinguishablility (Markov equivalence class)
  e.g., $V = \{X_1, X_2, X_3\}, \;\; I(P) = \{X_1 \perp X_3 \mid X_2\}$

Figure 2: Markov Equivalence Class

- Large non-convex search space (NP-hard)
  $|B_p| = \sum_{i=1}^{p} (-1)^{i+1} \binom{p}{i} 2^{i(p-i)} |B_{p-i}|$
  $|B_{14}| = 1,439,428,141,044,398,334,941,790,719,839,535,103$

Recent works

- Annealing on Regularized Cholesky Score (ARCS)
  - developed by Ye, Amini, and Zhou (2020) from UCLA
  - $f(L; P) = -\log|L| + tr(PSP^T LL^T) + \sum_{i>j} \rho_\theta(L_{ij})$
  - alternately minimize f with respect to Cholesky factor L and variable ordering P with simulated annealing

Recent works

- Annealing on Regularized Cholesky Score (ARCS)
  - developed by Ye, Amini, and Zhou (2020) from UCLA
  - $f(L; P) = -\log|L| + tr(PSP^T LL^T) + \sum_{i>j} \rho_\theta(L_{ij})$
  - alternately minimize f with respect to Cholesky factor L and variable ordering P with simulated annealing
- Sparsest Permutation (SP)
  - developed by Raskutti and Uhler (2018) from UW-Madison & MIT
  - lay down theoretical result that connects variable ordering and sparsest Bayesian network

Recent works

- Annealing on Regularized Cholesky Score (ARCS)
  - developed by Ye, Amini, and Zhou (2020) from UCLA
  - $f(L; P) = -\log|L| + tr(PSP^T LL^T) + \sum_{i>j} \rho_\theta(L_{ij})$
  - alternately minimize f with respect to Cholesky factor L and variable ordering P with simulated annealing
- Sparsest Permutation (SP)
  - developed by Raskutti and Uhler (2018) from UW-Madison & MIT
  - lay down theoretical result that connects variable ordering and sparsest Bayesian network
- Removal-Fill-Degree (RFD)
  - developed by Squires, Amaniampong, and Uhler (2020) from MIT
  - given undirected graph, recover the variable ordering based on removal/fill scores

Our Method: FROSTY

FROSTY utilizes Robust Selection (RobSel), (Cisneros, Petersen, and Oh, 2020) and sparse Choleksy algorithm called Approximate Minimum Degree ordering (AMD), (Amestoy, Davis, and Duff, 1996).

Algorithm 1 Pseudocode for FROSTY

Input : Dataset X, confidence level α

Output: Bayesian network B

Step 1: Estimate input undirected graph Θ

1: Set $\lambda = \text{RobSel}(X, \alpha)$

2: Estimate Θ_λ with Graphical Lasso

3: Prune Θ_λ with conditional independence test

Step 2: Recover DAG B

4: $L_\pi, P_\pi = \text{AMD}(\Theta_\lambda)$

5: $L = P_\pi^T L_\pi P_\pi$ and $B = (L_D - L)L_D^{-1}$

6: return B

Robust selection

Let $\ell(\Theta)$ be the Gaussian negative log-likelihood. Cisneros, Petersen, and Oh (2020) show that

$$\underbrace{\min_{\Theta} \sup_{\mathcal{P}:D_c(\mathcal{P},\mathcal{P}_n)\leq\lambda} E_{\mathcal{P}}[\ell(\Theta)]}_{\substack{\text{Distributionally Robust} \\ \text{Optimization (DRO) formulation}}} = \underbrace{\min_{\Theta}\{\ell(\Theta) + \lambda\|\Theta\|_1\}}_{\text{Graphical Lasso formulation}}, \tag{1}$$

Robust selection

Let $\ell(\Theta)$ be the Gaussian negative log-likelihood. Cisneros, Petersen, and Oh (2020) show that

$$\underbrace{\min_{\Theta} \sup_{\mathcal{P}:D_c(\mathcal{P},\mathcal{P}_n)\leq\lambda} E_{\mathcal{P}}[\ell(\Theta)]}_{\substack{\text{Distributionally Robust}\\ \text{Optimization (DRO) formulation}}} = \underbrace{\min_{\Theta}\{\ell(\Theta) + \lambda\|\Theta\|_1\}}_{\text{Graphical Lasso formulation}}, \tag{1}$$

Under the DRO formulation, the *optimal* regularization parameter can be chosen by

$$\lambda = \inf\{\lambda > 0 : P_0(\Theta \in \mathcal{C}_n(\lambda)) \geq 1 - \alpha\}. \tag{2}$$

This can be efficiently computed by bootstrapping dataset X only.

Robust selection

The interpretation of α in RobSel is recently shown in Tran et al. (2022):

$$P(\text{having at least one false positive edge in } \hat{\Theta}) \leq \alpha \;\text{ as }\; n \to \infty \qquad (3)$$

Robust selection

The interpretation of α in RobSel is recently shown in Tran et al. (2022):

$$P(\text{having at least one false positive edge in } \hat{\Theta}) \leq \alpha \quad \text{as} \quad n \to \infty \qquad (3)$$

In other words, the tuning parameter α in RobSel has a direct connection to the asymptotic family-wise error rate of the zero-nonzero patterns.

Topological ordering

Under multivariate Gaussian assumption, $\Theta = \Sigma^{-1}$ corresponds to the undirected graph, which can be expressed as a function of B and $\Omega = diag(\omega_1^2, \ldots, \omega_p^2)$, the error variances of $X_1, \ldots, X_p$:

$$\Theta(B, \Omega) = (I - B)\Omega^{-1}(I - B)^T \tag{4}$$

Topological ordering

Under multivariate Gaussian assumption, $\Theta = \Sigma^{-1}$ corresponds to the undirected graph, which can be expressed as a function of B and $\Omega = diag(\omega_1^2, \ldots, \omega_p^2)$, the error variances of $X_1, \ldots, X_p$:

$$\Theta(B, \Omega) = (I - B)\Omega^{-1}(I - B)^T \tag{4}$$

If we appropriately order the variables, say $\pi = (\pi_{(1)}, \ldots, \pi_{(p)})$,

$$\begin{aligned} \Theta_\pi(B_\pi, \Omega_\pi) &= (I - B_\pi)\Omega_\pi^{-1}(I - B_\pi)^T \\ &= LDL^T \\ &= \widetilde{L}\widetilde{L}^T \end{aligned} \tag{5}$$

Approximate minimum degree ordering

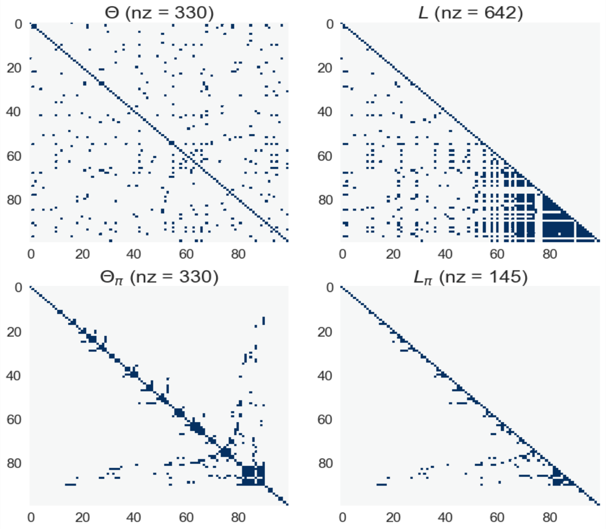

Figure 3: The impact of variable ordering.

Computational complexity

- RobSel: $O(bnp^2)$
- Graphical lasso: $O(p^3)$
- CI-testing: $O(ns^2 + s^3)$
- AMD: approximately $O(|L|) \ll O(p^2)$

where b is the number of boostrap samples, s is the average number of nonzeros per row in Θ, and L is the Cholesky factor.

Computational complexity

- RobSel: $O(bnp^2)$
- Graphical lasso: $O(p^3)$
- CI-testing: $O(ns^2 + s^3)$
- AMD: approximately $O(|L|) \ll O(p^2)$

where b is the number of boostrap samples, s is the average number of nonzeros per row in Θ, and L is the Cholesky factor.

<u>Note</u>: one only needs to fit Graphical Lasso *once*.

Simulation: empirical results

Performance metrics

All metrics are computed based on CPDAG, instead of DAG.

Performance metrics

All metrics are computed based on CPDAG, instead of DAG.

- Structural Hamming Distance, SHD ≥ 0

  SHD $=$ the number of edge changes (addition, deletion, reverse) required to turn an estimated graph into the true graph

Performance metrics

All metrics are computed based on CPDAG, instead of DAG.

- Structural Hamming Distance, SHD ≥ 0

  SHD $=$ the number of edge changes (addition, deletion, reverse) required to turn an estimated graph into the true graph

- Matthew's Correlation Coefficient, $-1 \leq \text{MCC} \leq 1$

  $$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

Simulation results

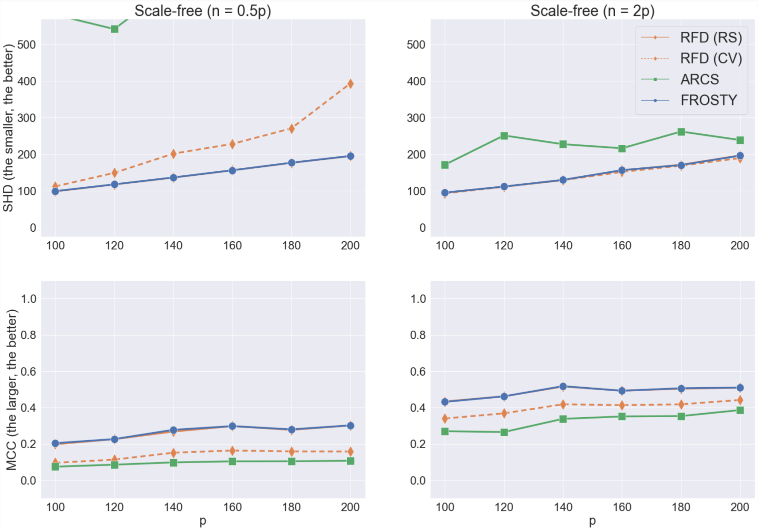

Figure 4: Performance comparison for scale-free graphs.

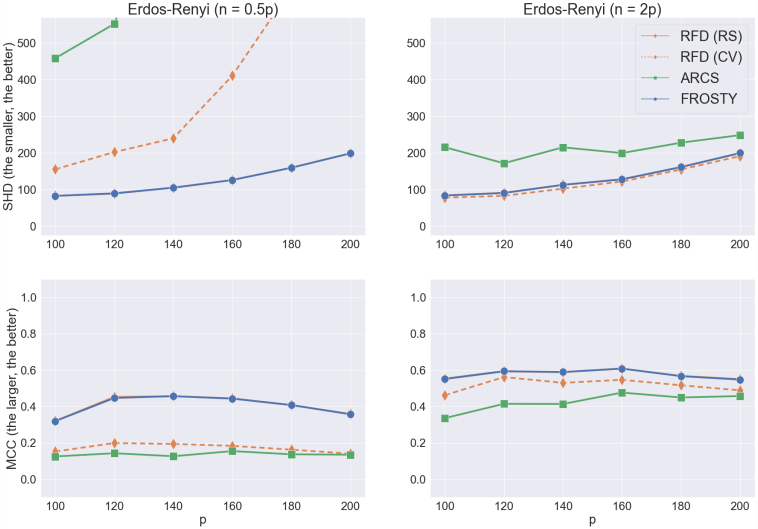

Figure 5: Performance comparison for Erdos-Renyi graphs.

Simulation results

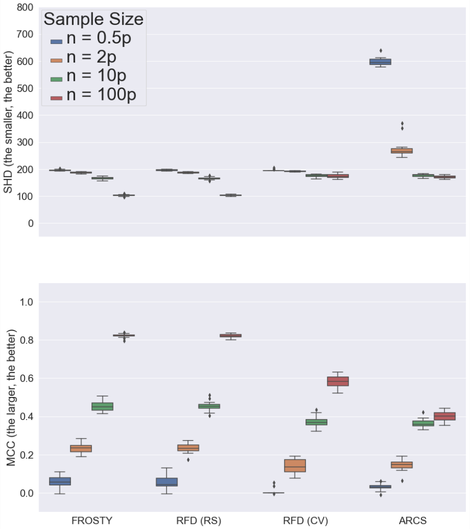

Figure 6: Performance comparison for Pathfinder network.

Simulation results

Method	Runtime (sec.)				
	p=100	p=200	p=500	p=1000	p=2000
FROSTY	0.12	0.28	1.39	5.52	33.69
RFD (RS)	0.86	6.52	168.62	2180.04	–
RFD (CV)	33.59	177.88	1817.85	14250.32	–
ARCS	57.52	101.47	562.56	5721.50	14581.00

Note: For $p = 2000$, RFD took up too much memory that it was not feasible to run within our available memory space of 60GB.

Table 1: Runtime analysis

Conclusion

Conclusion

- FROSTY is extremely scalable.

Conclusion

- FROSTY is extremely scalable.
  - For a large graph size of 2000 vertices, it estimates a Bayesian network less than a minute.

Conclusion

- FROSTY is extremely scalable.
  - For a large graph size of 2000 vertices, it estimates a Bayesian network less than a minute.
- FROSTY is simple.

Conclusion

- FROSTY is extremely scalable.
  - For a large graph size of 2000 vertices, it estimates a Bayesian network less than a minute.
- FROSTY is simple.
  - There is only one tuning parameter α, as in $(1-\alpha)$-confidence level for undirected graph, which has a direct relation to the asymptotic family-wise error rate of the zero-nonzero patterns.

Conclusion

- FROSTY is extremely scalable.
  - For a large graph size of 2000 vertices, it estimates a Bayesian network less than a minute.
- FROSTY is simple.
  - There is only one tuning parameter α, as in $(1-\alpha)$-confidence level for undirected graph, which has a direct relation to the asymptotic family-wise error rate of the zero-nonzero patterns.
- First step of FROSTY can improve other methods.

Conclusion

- FROSTY is extremely scalable.
  - For a large graph size of 2000 vertices, it estimates a Bayesian network less than a minute.
- FROSTY is simple.
  - There is only one tuning parameter α, as in $(1-\alpha)$-confidence level for undirected graph, which has a direct relation to the asymptotic family-wise error rate of the zero-nonzero patterns.
- First step of FROSTY can improve other methods.
  - Methods that take an undirected graph as an input can be significantly improved by adapting FROSTY's undirected graph estimation step.

References

- Amestoy, Patrick R, Timothy A Davis, and Iain S Duff (1996). "An approximate minimum degree ordering algorithm". In: *SIAM Journal on Matrix Analysis and Applications* 17.4, pp. 886–905.
- Cisneros, Pedro, Alexander Petersen, and Sang-Yun Oh (2020). "Distributionally Robust Formulation and Model Selection for the Graphical Lasso". In: *International Conference on Artificial Intelligence and Statistics*, pp. 756–765.
- Raskutti, Garvesh and Caroline Uhler (2018). "Learning directed acyclic graph models based on sparsest permutations". In: *Stat* 7.1, e183.

Reference ii

- Squires, Chandler, Joshua Amaniampong, and Caroline Uhler (2020). "Efficient Permutation Discovery in Causal DAGs". In: *arXiv preprint arXiv:2011.03610*.
- Tran, Chau et al. (2022). "Family-wise error rate control in Gaussian graphical model selection via Distributionally Robust Optimization". In: *arXiv preprint arXiv:2201.12441*.
- Ye, Qiaoling, Arash Amini, and Qing Zhou (2020). "Optimizing regularized cholesky score for order-based learning of bayesian networks". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Appendix

Robust selection

$$
\begin{aligned}
\lambda &= \inf\{\lambda > 0 : P_0(\Theta \in \mathcal{C}_n(\lambda)) \geq 1-\alpha\} \\
&= \inf\{\lambda > 0 : P_0(R_n(\Theta) \leq \lambda) \geq 1-\alpha\} \\
&= \inf\{\lambda > 0 : P_0(\|\text{vec}(S-\Theta^{-1})\|_\infty \leq \lambda) \geq 1-\alpha\}
\end{aligned} \tag{6}
$$

where $\mathcal{C}_n(\lambda)$ is the confidence region for Θ and

$$
\begin{aligned}
R_n(\Theta) = \inf\{&D_c(\mathcal{P}, \mathcal{P}_n) : \\
&E_{\mathcal{P}}[\frac{\partial}{\partial\Theta'}(tr(S\Theta') - \log|\Theta'|)\Big|_{\Theta'=\Theta}] = \mathbf{0}\}
\end{aligned} \tag{7}
$$

is called the Robust Wasserstein Profile (RWP) function, which represents the minimum distance between the empirical distribution and any *plausible* distribution that satisfies the first order optimality condition for the precision matrix Θ.

Undirected graph estimation

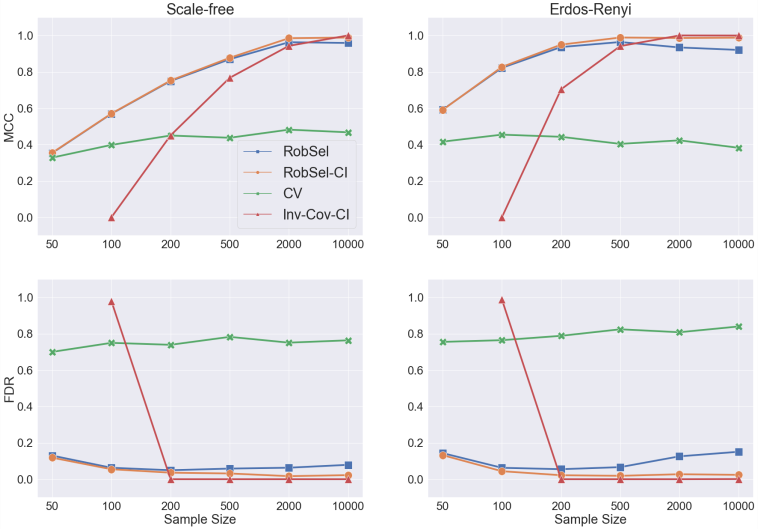

Figure 7: Undirected graph estimation comparison.