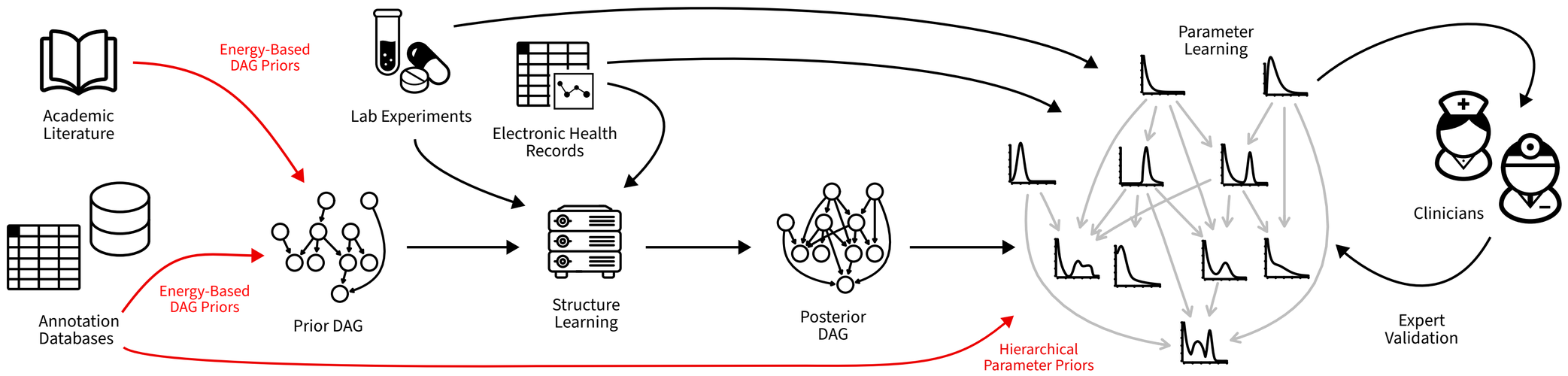
Bayesian posterior estimation of Gene Regulatory Networks from Zero-Inflated Single-Cell Data

Marco Scutari, Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA)
Noriaki Sato and Seiya Imoto, The Institute of Medical Science, The University of Tokyo
Shuichi Kawano, Faculty of Mathematics, Kyushu University
Rui Yamaguchi, Aichi Cancer Center & Nagoya University Graduate School of Medicine

Informative Priors to Augment Molecular Network Learning



Energy-Based Priors & Structure Learning

KEGG, DoRothEA and other annotation databases [1] contain enough information to build an *a priori* DAG $\mathcal{G}_0$ to prime gene regulatory network learning. With energy-based priors [2], the energy function

$$E(\mathcal{G}, \mathcal{G}_0; w) = w\widetilde{\text{SHD}}(\mathcal{G}, \mathcal{G}_0) + (1-w)\widetilde{\text{SID}}(\mathcal{G}, \mathcal{G}_0), \quad w \in [0,1]$$

and the associated graphical prior

$$\text{P}(\mathcal{G}; \mathcal{G}_0, \beta, w) = \exp\left\{-\beta E(\mathcal{G}, \mathcal{G}_0; w)\right\}, \qquad \beta \geqslant 0,$$

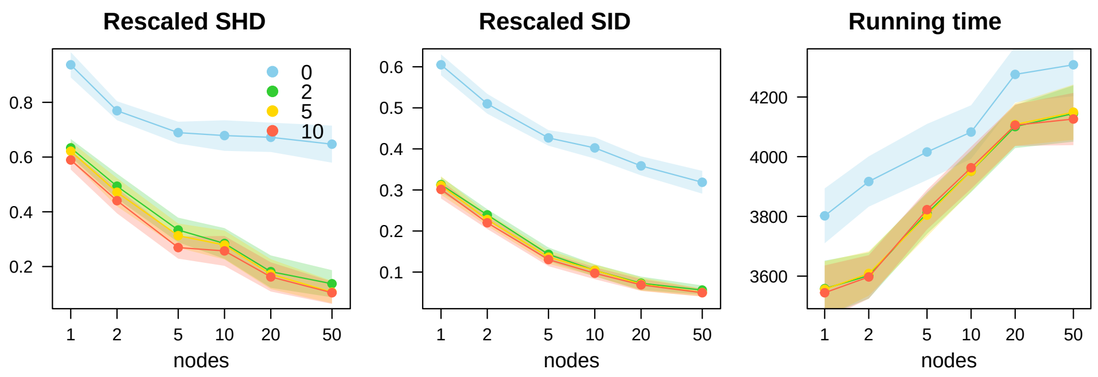
encourage both local (SHD) and global (SID) consistency.

- Structural Hamming Distance ($\widetilde{\text{SHD}}$, scaled [3]): individual-arc differences in presence and direction.
- Structural Interventional Distance (SID, scaled [4]): discrepancies along causal pathways after interventions (say, knock-outs).

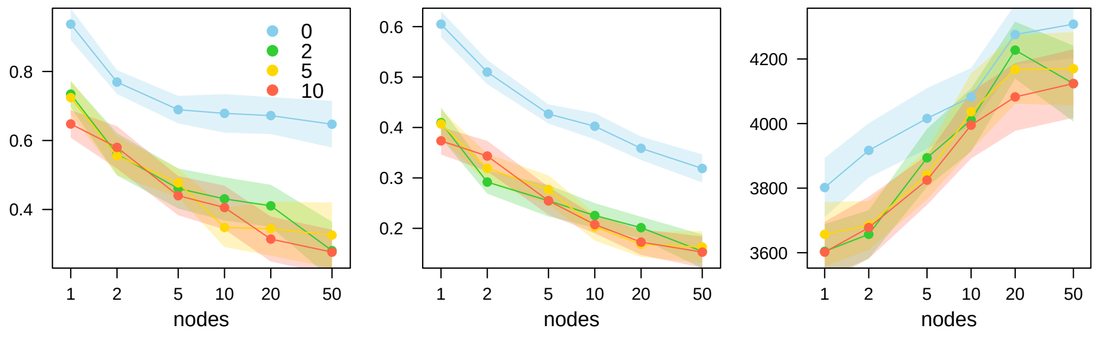
Practical Performance

Benchmarked with the gene expression (ECOLI70) and SNP + phenotype (MAGIC-NIAB, MAGIC-IRRI) reference networks [5].

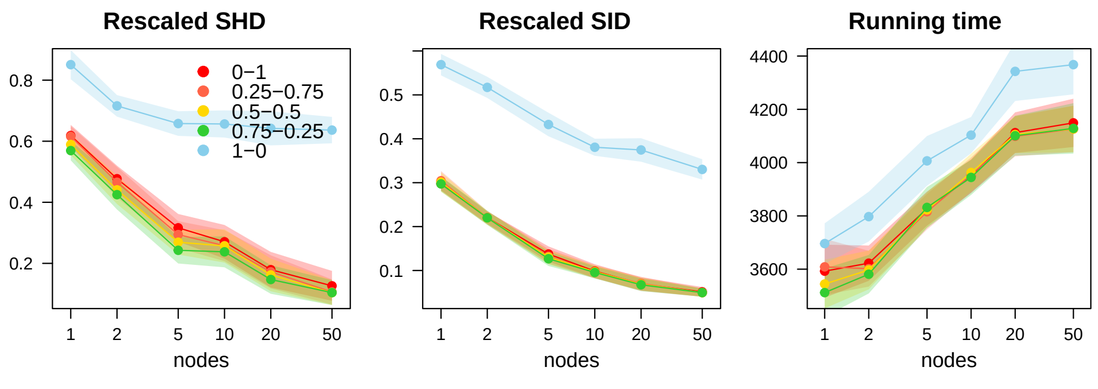
- More accurate and faster ($\beta > 0$) than baseline ($\beta = 0$).



- Robust against even to 20% incorrect arcs in $\mathcal{G}_0$.



- Not sensitive to the choice of w as long as $w < 1$ (here $\beta = 10$).



Zero-Inflated DAGs

Zero-inflated generalised hypergeometric (GHP) DAGs [ZiG-DAG; 6],

$$\text{P}(X_i \mid \text{pa}(X_i)) = \begin{cases} \pi_i + (1-\pi_i)\text{GHP}(X_i, \lambda_i) & \text{(structural zeros)} \\ (1-\pi_i)\text{GHP}(X_i, \lambda_i) & \text{(pseudocounts)} \end{cases},$$

$\pi_i = \pi(\text{pa}(X_i)), \lambda_i = \lambda(\text{pa}(X_i))$, allow causal identifiability, but:

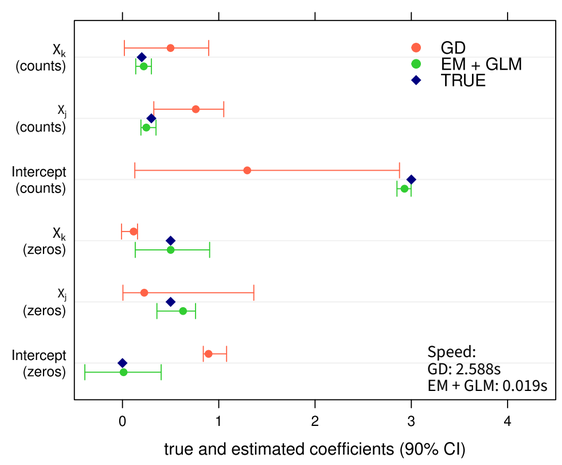
- estimating both components simultaneously by gradient descent is numerically unstable.
- estimating the parameters independently for each $X_i \mid pa(X_i)$ disregards pathway status (active/inactive).

Hierarchical, Stable Parameter Estimation

EM + GLM parameter estimation [7] beats gradient descent even in $X_i \mid \text{pa}(X_i)$ with few parameters and large samples.

1. *E-Step:* Estimate the structural zero indicators.
2. *M-Step:* Estimate the coefficients of
   - the logistic regression modelling the zero-inflation probability;
   - the GLM modelling the pseudocounts.

Consider:



$X_i \sim ZiPois(\pi_i, \lambda_i)$ with $\text{pa}(X_i) = \{X_j, X_k\}$ and:

$$\text{logit}(\pi_i) = 0.5X_j + 0.5X_k$$
$$\log(\lambda_i) = 3 + 0.3X_j + 0.2X_k$$

200 replicates, random samples with $n = 200$.

Estimates can improve even further with a hierarchical prior that uses the hierarchical grouping of regulatory network nodes (TODO).

References

[1] D. J. Rigden and X. M. Fernández. The 2025 Nucleic Acids Research Database Issue and the Online Molecular Biology Database Collection. *Nucleic Acids Research*, 53(D1):D1–D9, 2025.

[2] S. Imoto, T. Higuchi, T. Goto, K. Tashiro, S. Kuhara, and S. Miyano. Combining Microarrays and Biological Knowledge for Estimating Gene Networks via Bayesian Networks. *Journal of Bioinformatics and Computational Biology*, 2(1):77–98, 2004.

[3] I. Tsamardinos, L. E. Brown, and C. F. Aliferis. The Max-Min Hill-Climbing Bayesian Network Structure Learning Algorithm. *Machine Learning*, 65(1):31–78, 2006.

[4] J. Peters and P. Bühlmann. Structural Intervention Distance (SID) for Evaluating Causal Graphs. *Neural Computation*, 27:771–799, 2015.

[5] M. Scutari. Bayesian Network Repository, 2012. `http://www.bnlearn.com/bnrepository`.

[6] J. Choi and Y. Ni. Model-Based Causal Discovery for Zero-Inflated Count Data. *Journal of Machine Learning Research*, 24(200):1–32, 2023.

[7] Z. Wang, S. Ma, C. Wang, M. Zappitelli, P. Devarajan, and C. Parikh. EM for Regularized Zero-Inflated Regression Models with Applications to Postoperative Morbidity After Cardiac Surgery in Children. *Statistics in Medicine*, 33(29):5192–5208, 2014.

email: scutari@bnlearn.com