Avoiding pitfalls in L_1-regularised inference of gene networks

Andreas Tjärnberg,[ab] Torbjörn E. M. Nordling,[acd] Matthew Studham,[a]
Sven Nelander,[cd] and Erik L.L. Sonnhammer,[abe]

[a]Stockholm Bioinformatics Centre, Science for Life Laboratory, Box 1031, 17121 Solna, Sweden
[b]Department of Biochemistry and Biophysics, Stockholm University
[c]Department of Immunology, Genetics and Pathology, Uppsala University, Rudbeck laboratory, 75185 Uppsala, Sweden
[d]Science for Life Laboratory, Uppsala University
[e]Swedish eScience Research Center

The Problem

Statistical regularisation methods such as LASSO and related L_1 regularised regression methods are commonly used to construct models of gene regulatory networks.

Although they theoretically can infer the correct network structure under the linear systems approximation assumption, they have been shown in practice to make errors, *i.e.* leave out existing links and include non-existing links.

Linear systems representation

$$\boldsymbol{Y} = -\boldsymbol{A}^{-1}\boldsymbol{P} + \boldsymbol{A}^{-1}\boldsymbol{F} + \boldsymbol{E}$$

L_1 regularisation (LASSO)

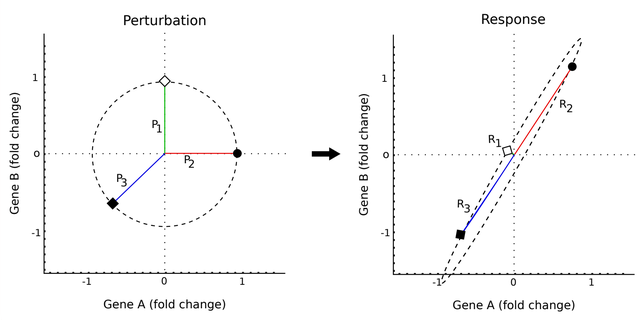
$$\hat{\boldsymbol{A}}_{\text{reg}}(\tilde{\zeta}) = \arg\min_{\boldsymbol{A}} ||\boldsymbol{AY} + \boldsymbol{P}||_{L_2}^2 + \tilde{\zeta}||\boldsymbol{A}||_{L_1}.$$

L_1 regularisation pitfall

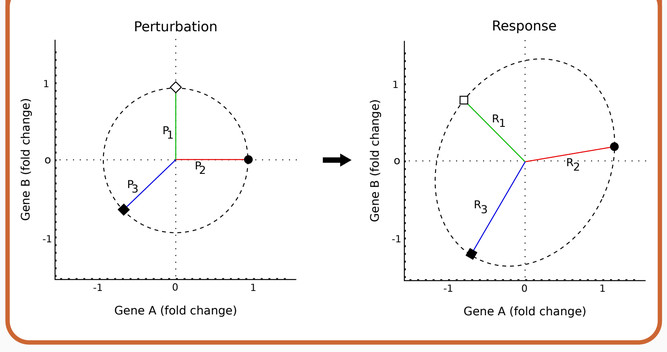
- L_1 GRN inference methods fail even though the data is informative enough for inferring the correct network structure.
  - Inference methods performance is dependent on data properties
  - Underlying system structure enforces response patterns and defines data properties
  - Design of experiments dictates final data structure

Data properties

ill-conditioned response



well-conditioned response



The condition number, κ indicates how much the input gets skewed when translated by the system. High number means ill-conditioned, low numbers $\kappa \to 1$, means well-conditioned.
In biological systems this means some responses get amplified while others get attenuated[2].

Signal to Noise Ratio

We applied noise to each data set with a variance λ selected to give the desired Signal to Noise Ratio (SNR)

$$\text{SNR} \triangleq \frac{\sigma_N(\check{\boldsymbol{Y}})}{\sqrt{\chi^{-2}(\alpha, NM)\lambda}}$$

Results

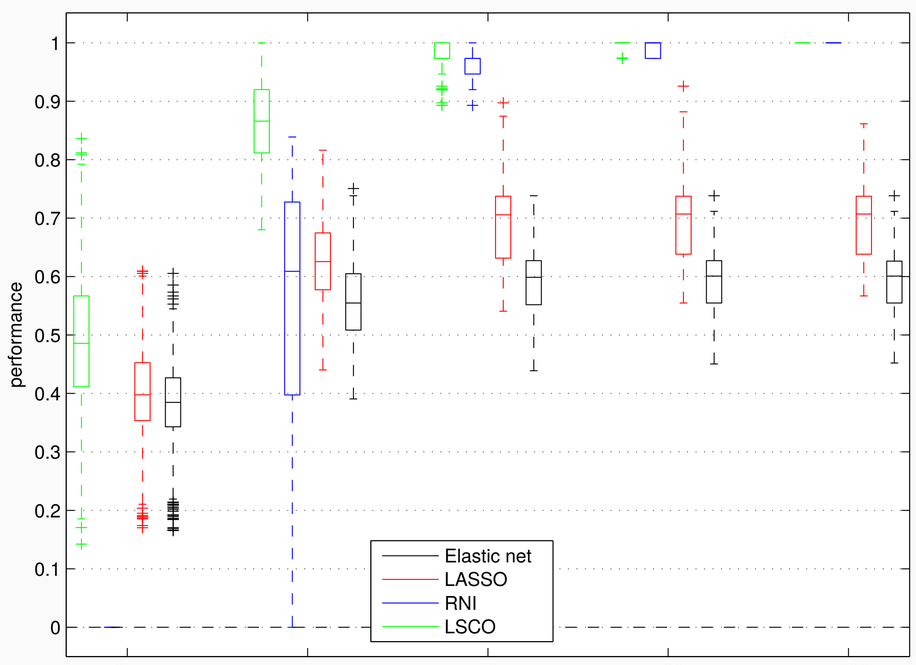


Figure 1: Performance on high κ data sets as a function of the Signal to Noise Ratio

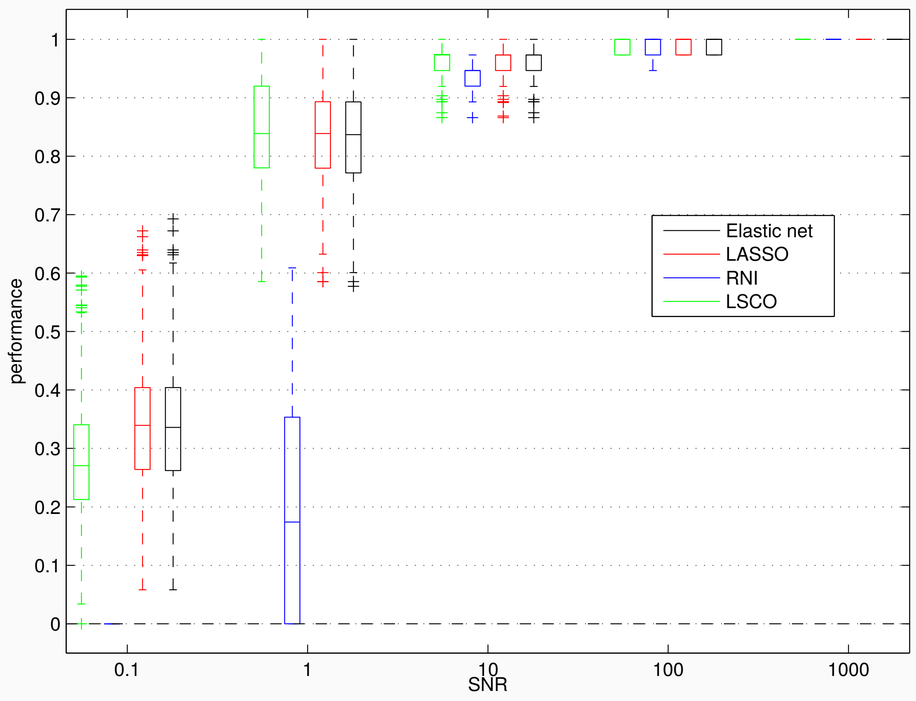


Figure 2: Performance on low κ data sets as a function of the Signal to Noise Ratio

Discussion

The most striking result on the data set with high response matrix condition number is that all the L_1 regularisation methods fail to recover the true network model even when the SNR is so high that the data is informative enough for network inference and all existing links can be proven to exist

The indicators SIC and WIC (Strong and Weak Irrepresentable Conditions) are fulfilled, only for the data sets with a low condition number and SNR of one or higher. An alternative representation of SIC can be seen in Figure 3.

SIC and WIK cannot in practice be used to evaluate performance. Until a better testable criterion for failure of L_1 regularisation is presented, we recommend all users to check the condition number of the response matrix. The condition number has the advantage of being a classical tool in linear algebra that is easy to calculate.

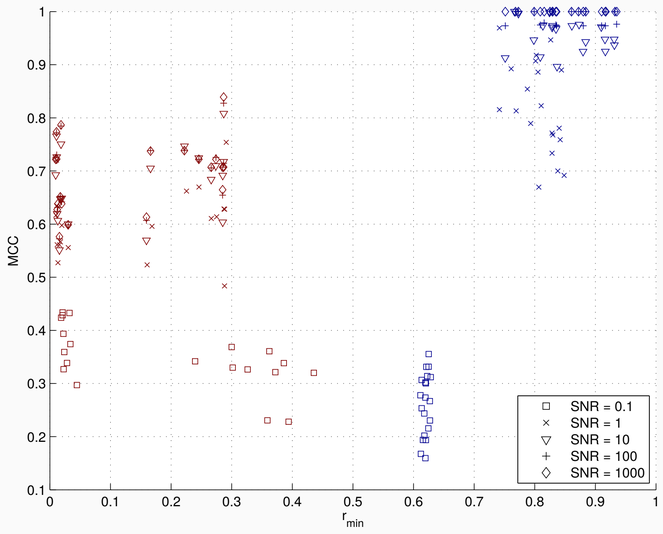


Figure 3: Performance of LASSO measured in r_{min} = minimum ratio between the shortest regressor corresponding to a nonzero element and the longest corresponding to a zero. Blue marks represent low κ data sets, while Red represents high κ data sets. A clear separation is seen based on the condition number of the data

Biological Data

We demonstrate how our simulations can be used to estimate the optimal performance based on the properties of the ten gene network of the Snf1 signalling pathway in *S. cerevisiae* and the *in vivo* data collected by Lorenz *et. al.*[1]. We identified the SNR 0.1 data point in Figure 1 as the closest one in our simulations. We thus expect the optimal performance, in terms of MCC, of LASSO, Elastic Net, LSCO, and RNI to be far below 0.5.

Performance	S10	S19	S9
LASSO	0.18	0.22	0.36
LSCO	0.21	0.20	0.32
Elastic net	0.18	0.27	0.40
NIR (S9)	0.25	0.28	1.00

Conclusions

- L_1 regularisation methods–LASSO and Elastic Net–typically perform poorly in GRN inference when using data as ill-conditioned as typical experimental data.
- For both well-conditioned and ill-conditioned data, we found an SNR, of 10 to be sufficient for LSCO and RNI to achieve maximum accuracy close to one.
- For data with a SNR below one the accuracy of all methods was in general low.

References

[1] David R. Lorenz, Charles R. Cantor, and James J. Collins. A network biology approach to aging in yeast. *Proceedings of the National Academy of Sciences*, 106(4):1145–50, January 2009.

[2] Torbjörn E M Nordling. *Robust inference of gene regulatory networks*. PhD thesis, KTH School of Electrical Engineering, Automatic Control Lab, 2013.