Semi-parametric neighborhood selection for estimating linear and non-linear gene co-expression networks

Juho A. J. Kontio, Marko J. Rinta-aho and Mikko J. Sillanpää

Supplement A

1 The role of parameters γ and α on the estimated non-linear neighborhood structures

Let us evaluate the robustness of the SNS procedure with respect to the values of the elastic net estimator parameter α (Zou and Hastie 2005) and the kernel parameter γ in the PH-E model (see the model (14) in the article). We go through all possible (γ, α) combinations using the following sparse grids:

$$\alpha \in \{1/3, 2/3, 1\} \quad \text{and} \quad \gamma \in \{1/5, 1, 2\}.$$

The same simulated datasets and replicates are used as in the second example analysis in the article (replicates are available in the supplementary materials B). The simulation was proceeded such that a real protein expression dataset provided by DREAM9 challenge (Noren *et al.* 2016) was used as a basis dataset. This dataset consists of measurements of 231 proteins $(X_1, X_2, \ldots, X_{231})$ over 191 individuals diagnosed with acute myeloid leukemia. Subsequently, ten replicates of six new proteins $(Z_1, Z_2, Z_3, Z_4, Z_5)$ were simulated on top of these 231 proteins such that

$$Z_1 = X_5 + \cos(X_{100}) + \mathcal{N}(0, 0.6^2), \quad Z_2 = X_{15} + X_{47}^2 + \mathcal{N}(0, 1^2),$$
$$Z_3 = X_{25} + X_{80}X_{110} + \mathcal{N}(0, 1.5^2), \quad Z_4 = X_{30} + X_{200}^3 + \mathcal{N}(0, 3^2),$$
$$Z_5 = X_{35} + X_{40} + X_{50} + X_8X_{150}X_{220} + \mathcal{N}(0, 0.6^2).$$

As reported in the article, the average proportions of signal variations on the overall variations were 0.73, 0.62, 0.41, 0.36 and 0.90 for the new simulated proteins Z_1, Z_2, Z_3, Z_4 and Z_5 over the replicates (calculated as $(\text{var}(Z_k) - \varepsilon_k^2)/\text{var}(Z_k)$ separately for each $k = 1, \ldots, 5$).

In the forthcoming examples, we only consider and estimate the neighborhoods of simulated proteins $(Z_1, Z_2, Z_3, Z_4, Z_5)$ for which the true underlying structures are fully known. The results of the analyses are averaged over the replicates. Moreover, despite some linear regulatory effects were simulated we

are considering only the non-linear relationships since identifying the linear neighborhoods in these examples would be a relatively simple task. The linear relationships are only included to show how different outcomes in the linear step (e.g. resulted by different (γ, α) pairs) affect to the results of the non-linear neighborhood selection. Thus, we compute the receiver operating characteristics (ROC) curves with respect to the neighborhoods of simulated proteins $(Z_1, Z_2, Z_3, Z_4, Z_5)$ and compare the areas under the ROC curves (AUCs) resulted by different parameter pairs (γ, α).

The estimation of linear and non-linear neighborhoods is done with the *glmnet* R-package (Friedman *et al.* 2010) in accordance to the article. For each pair (γ, α), the penalty parameter $\lambda_{1,k}$ and $\lambda_{2,k}$ values are chosen for the linear $(\lambda_{1,k})$ and non-linear $(\lambda_{2,k})$ neighborhood selection steps by the cross-validation (CV) criterion in each replicate.

A hard-thresholding procedure is applied to calculate truncated and non-truncated areas AUCs from the estimated non-linear neighborhoods (given by CV based values $\lambda_{1,k}$ and $\lambda_{2,k}$). Table S1 displays the non-truncated and truncated (at 0.10 and 0.40 false-positive rates) AUCs averaged over the replications for each parameter pair (γ, α). Note that the α parameter can be different for linear and non-linear neighborhood selection steps that are later separated with symbols α_1 (linear) and α_2 (non-linear). However, the results presented in Table S1 are produced using the same values for α_1 and α_2 that are simultaneously denoted by α.

Table S1: Averaged areas under the truncated and non-truncated ROC curves (AUCs) for the SNS procedure with different parameter pairs (γ, α) over ten replicates in the simulated scenarios. Truncation points are set to 0.10 and 0.40 false-positive rates and shown in brackets above each corresponding column. The same α parameter value is used for linear and non-linear neighborhood selection steps. The CV based penalty parameters $\lambda_{1,k}$ in linear neighborhood selection and $\lambda_{2,k}$ in non-linear neighborhood selection were used for each replicate.

(γ, α)	Truncated AUC (0.10)	Truncated AUC (0.40)	Non-truncated AUC
(0.20, 0.33)	0.832	0.908	0.928
(0.20, 0.66)	0.850	0.911	0.929
(0.20, 1.00)	0.847	0.916	0.934
(1.00, 0.33)	0.802	0.881	0.910
(1.00, 0.66)	0.811	0.876	0.900
(1.00, 1.00)	0.810	0.880	0.901
(2.00, 0.33)	0.747	0.848	0.873
(2.00, 0.66)	0.748	0.837	0.864
(2.00, 1.00)	0.762	0.840	0.869

It appears, that all parameter pairs (γ, α) are capable of identifying correct non-linear relationships efficiently – non-truncated AUCs are ranging between 0.864 and 0.928 (see the right column in Table S1). In particular, these results suggest that the performance of the SNS procedure becomes slightly better with

smaller parameter γ values. We therefore recommend to use the value 0.20 for γ which is the smallest possible value that can be used in the exponential kernel function (see e.g. Shi and Choi, 2011). The same conclusion has been also made in Kontio and Sillanpää (2019) that focused on finding higher-order gene-by-gene interaction terms using the exponential kernel function. However, for each fixed value of $\gamma \in \{1/5, 1, 2\}$ the SNS procedure shows extremely robust performance with respect to the changes in α parameter values. In fact, the truncated and non-truncated AUCs produced by different α values with fixed γ values are practically identical.

We have also found that for a fixed value of $\gamma \in \{1/5, 1, 2\}$ the SNS procedure is not sensitive to the choices of α_1 and α_2 even if they are different. For instance, few examples are given in Table S2 with the γ parameter fixed to 0.20.

Table S2: Averaged areas under the truncated and non-truncated ROC curves (AUCs) for the SNS procedure with fixed γ value 0.20 and different parameter α_1 and α_2 values over ten replicates in the simulated scenarios. Here α parameters used for linear and non-linear neighborhood selection steps are separated with symbols α_1 (linear) and α_2 (non-linear). Truncation points are set to 0.10 and 0.40 false-positive rates and shown in brackets above each corresponding column. The CV-based parameter $\lambda_{1,k}$ and $\lambda_{2,k}$ values were used for each replicate in both linear and non-linear estimation steps.

$(\gamma, \alpha_1, \alpha_2)$	Truncated AUC (0.10)	Truncated AUC (0.40)	Non-truncated AUC
(0.20, 0.33, 1.00)	0.835	0.908	0.931
(0.20, 0.33, 0.66)	0.840	0.905	0.927
(0.20, 0.66, 1.00)	0.837	0.905	0.926

Note that the selection of optimal tuning parameters is extremely challenging task in general and is highly depending on the context. Especially, to same extent that the optimal choice of (γ, α) in some case is not necessarily the optimal one in another case, neither might be the criteria by which they are chosen. Considering this issue would be beyond the scope of this paper. In particular, the above examples only illustrate that we would get decent results even if the most optimal pair of (γ, α) is not used.

2 Robustness with respect to the subsamples

Let us now evaluate the robustness of the SNS method under a series of different subsamples to test if the results are sensitive to the small changes in data. We select ten different subsamples among the original 191 individuals each of which consisting of m randomly chosen individuals. Then we analyze the same simulated dataset separately over each subsample to see how much the results are differing from each other. The test dataset is chosen to be the first replicate among the simulated datasets described in the previous example (this specific dataset can be found in the supplementary materials B).

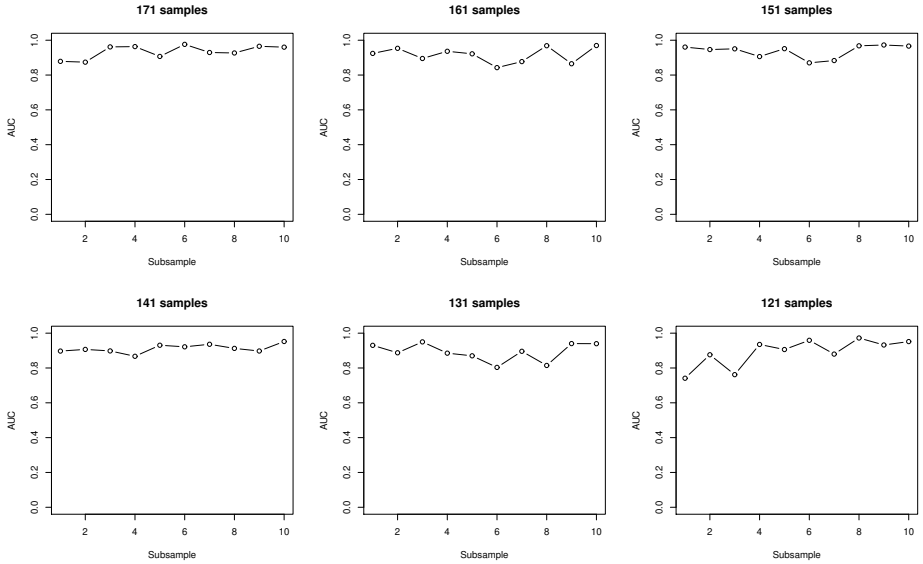

Figure S1: Non-truncated areas under the ROC curves (AUCs) over ten different randomly chosen subsamples for each subsample size 121, 131, 141, 151, 161 and 171. The non-linear neighborhoods were estimated with the SNS method using the parameter values $\gamma = 0.20$, $\alpha_1 = 1/3$ and $\alpha_2 = 1$. Here α parameters used for linear and non-linear neighborhood selection steps are separated with symbols α_1 (linear) and α_2 (non-linear). The CV based penalty parameters $\lambda_{1,k}$ in linear neighborhood selection and $\lambda_{2,k}$ in non-linear neighborhood selection were used in each case.

This subsampling analysis is repeated for different subsample sizes $m \in \{121, 131, 141, 151, 161, 171\}$. In each case, the non-linear neighborhoods are estimated with the SNS method using the parameters $\gamma = 0.20$, $\alpha_1 = 1/3$ and $\alpha_2 = 1$ as in the article. Similarly to the previous example, we only consider and estimate the neighborhoods around simulated proteins $(Z_1, Z_2, Z_3, Z_4, Z_5)$ as the corresponding true neighborhoods are known.

The non-truncated AUCs over ten randomly chosen subsamples are displayed in the Figure S1. Each plot corresponds to a different subsample size $m \in \{121, 131, 141, 151, 161, 171\}$. As we can see, the non-truncated AUCs are scattered around 0.90 with small variances in the majority of cases. Averaged AUCs over different subsamples are 0.891, 0.892, 0.912, 0.938, 0.915 and 0.934 for sample sizes 121, 131, 141, 151, 161 and 171, respectively. It is evident that the SNS procedure captures the underlying signals concordantly even though randomness in data changes extensively. Only with the subsample size of 121, AUCs begin to be slightly more dispersed over subsamples. However, while the sample size of 121 is extremely low to capture non-linear relationships the performance of the SNS method is surprisingly robust over the subsamples.

3 Additional figures

Here we provide additional information regarding the results shown in Table 2 in the article. In Figure S2, we have plotted sensitivies against specificities separately in each replicate (ten replicates) for the SNS method and the best performing versions of both DC and MI based methods (DC-REL and MI-REL).

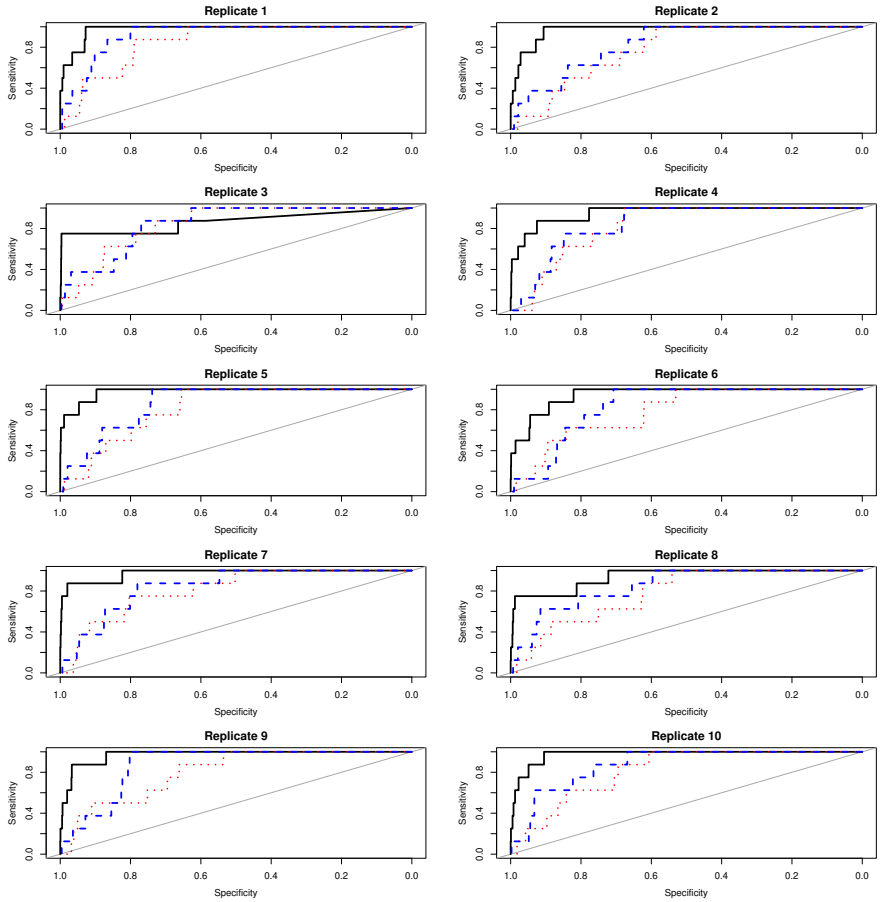

Figure S2: The receiver operator characteristic curves plotted for the SNS (black solid lines), DC-REL (blue dashed lines) and MI-REL (red dotted lines) methods in each ten replicate. In the SNS method, the CV-based parameter $\lambda_{1,k}$ and $\lambda_{2,k}$ values were used for each replicate in both linear (with $\alpha_1 = 1/3$) and non-linear (with $\alpha_2 = 1$) estimation steps.

It is evident that in each replicate the proposed SNS method consistently implies better accuracy than the best versions of the MI and DC based methods.

3.1 Yeast knock-out gene expression data

Figure S3 shows the GRNs estimated using the REL-MI and REL-DC methods on the DREAM3 yeast knock-out data we are referring in the article. These results are also presented in the paper of Guo *et al.* (2014).

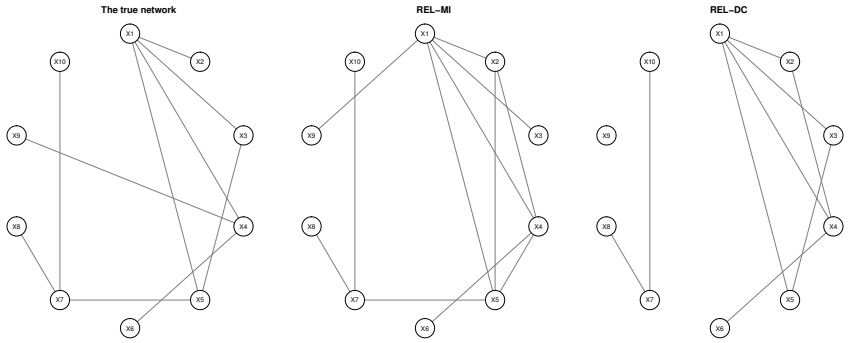

Figure S3: The true network and the network structures estimated with the REL-MI and REL-DC methods.

References

[1] Friedman J., T. Hastie, and R. Tibshirani, 2010 Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. **33:** 1-22.

[2] Guo, X., Y. Zhang, W. Hu, H. Tan, and X. Wang, 2014 Inferring nonlinear gene regulatory networks from gene expression data based on distance correlation. PLOS ONE **9:** 1–7.

[3] Kontio J. A. J. and M. J. Sillanpää, 2019 Scalable nonparametric prescreening method for searching higher-order genetic interactions underlying quantitative traits. Genetics **213:** 1209-1224.

[4] Noren, D. P., B. L. Long, R. Norel, K. Rrhissorrakrai, K. Hess *et al.*, 2016 A crowdsourcing approach to developing and assessing prediction algorithms for AML prognosis. PLoS Comput. Biol. **12:** e1004890 https://doi.org/10.1371/journal.pcbi.1004890.

[5] Shi, J. Q., and T. Choi, 2011 *Gaussian Process Regression Analysis for Functional Data*. Chapman Hall/CRC, London.

[6] Zou, H., and T. Hastie, 2005 Regularization and variable selection via the elastic net. J. R. Stat. Soc. B. **67:** 301-320.