SM VII: Comparison by simulation

The proposed ridge estimator is compared by means of simulation to (as far as we are aware) the only other method that estimates the VAR(1) model and its associated time series chain graph from high-dimensional data: SparseTSCGM proposed by [**?**]. SparseTSCGM too estimates parameters of the VAR(1) model in penalized fashion, but employs the smoothly clipped absolute deviation (SCAD, Fan2001) penalty on the autoregressive coefficient matrix $\mathbf{A}$ and precision matrix $\mathbf{\Omega}_\varepsilon$. As the estimation methods estimate the same VAR(1) model and only differ in the employed penalties, they can readily compared in terms of loss of the estimates and selection of the edges of the time series chain graph.

The ridge and 'SCAD' estimators of the VAR(1) model are compared under various choices of the model parameters $\mathbf{A}$ and $\mathbf{\Omega}_\varepsilon$, while the number of variates p, time points $\mathcal{T}$ and the number of samples n are varied in accordance with a factorial design. We first describe the levels of the factors of the factorial design. The number of variates equals either $p = 25$ or $p = 50$[1], representing the size of non-trivial but well-defined pathways. Simultaneously, we set $n \in \{5, 15\}$ and $\mathcal{T} \in \{10, 20\}$. We stress that the particular case of $(n = 5, \mathcal{T} = 10)$ is the most challenging but also most relevant as its design is closest to that which is most prevalent in practice. [**?**] observed that more than 80% of time-course gene expression microarray data sets have 8 or fewer number of time points. This is confirmed by a more recent survey of our own (confer SM III).

We now outline parameter choices of the various VAR(1) models employed in the simulation. Autoregression matrix $\mathbf{A}$ with a hub, cluster, random and complete network structure. Two sparsity levels are used for $\mathbf{A}$, more sparse with around 10% of edges and sparse with 25% of edges. These matrices $\mathbf{A}$ are used in combination with a banded and full precision matrix $\mathbf{\Omega}_\varepsilon$. We first detail the various choices of $\mathbf{\Omega}_\varepsilon$:

- Full $\mathbf{\Omega}_\varepsilon$: In the precision matrix $\mathbf{\Omega}_\varepsilon$ existing contemporaneous conditional independences is represented with $\rho = 0.5$.
- Banded $\mathbf{\Omega}_\varepsilon$: A banded precision matrix is generated with $(\mathbf{\Omega}_\varepsilon)_{j_1,j_2} = \rho^{|j_1 - j_2|}$ with $\rho = 0.5$ for $|j_1 - j_2| < 3$ and $(\mathbf{\Omega}_\varepsilon)_{j_1,j_2} = 0$ otherwise.
- Full but data-driven $\mathbf{\Omega}_\varepsilon$: The precision matrix is obtained from a VAR(1) model fitted to data from a time-course gene experiment experiment. The data of this experiment are from the aforementioned experiment. The data are limited to genes mapping to the p53 signaling pathway as defined by KEGG [**?**]. The VAR(1) model is fitted to these pathway data by means of the proposed ridge estimation procedure using penalty parameter values $\lambda_a = 0.3, \lambda_\omega = 0.1$. The thus estimated $\mathbf{\Omega}_\varepsilon$ is used in the simulation. For the various choice of p, first p genes from the pathway are used in the comparison. **WHICH GENES FOR THE VARIOUS CHOICES OF P?**

For the regression coefficient matrix $\mathbf{A}$ the following variants are employed:

- Hub $\mathbf{A}$: All elements of $\mathbf{A}$ are set to zero, except for $(\mathbf{A})_{j_1,j_2} = 0.3$ for $j_1 = \{d, 2d, 3d, ...\}$ with $j_1 > j_2$, while for more sparse $d = 10$ and for sparse $d = 2$ case. Hence, only a part of the variates affect (in varying degree) the temporal variation of the others.
- Cluster $\mathbf{A}$: Eight for more sparse and two for sparse equally-sized lower-triangular blocks filled with the value 0.3 are aligned along the diagonal. The remaining elements of $\mathbf{A}$ are zero.
- Random $\mathbf{A}$: A random network is generated, selecting 10% for more sparse and 25% for sparse of the total possible edges, with the additional constraint that $(\mathbf{A})_{j_1,j_2} = 0$ if $j_1 < j_2$. Elements of $\mathbf{A}$ are set equal to 0.3 if the corresponding element in the adjacency matrix of the random network is nonzero.

[1] Initially, we also included $p = 100$, but the estimation with the SCAD procedure fails to converge for $n = 5$ and $T = 10$, as is often the case for $p = 100$ with larger n and $\mathcal{T}$.

- Full data-driven $\mathbf{A}$: The same data set as for the generation of the data-driven $\mathbf{\Omega}_\varepsilon$ is used to estimate $\mathbf{A}$. Using these data $\mathbf{A}$ is estimated by the ridge procedure with penalty parameters $\lambda_a = 0.3$ and $\lambda_\omega = 0.1$. The resulting $\mathbf{A}$ is used in the simulation.

Note that the more sparse choice of $\mathbf{A}$ in principle favors the 'SCAD' estimation procedure SparseTSCGM.

For (combination of) these choices for the regression parameter matrix $\mathbf{A}$ and precision matrix $\mathbf{\Omega}_\varepsilon$, data are simulated in accordance with the VAR(1) model. Data for the first time point $t = 1$ are drawn from $\mathcal{N}(\mathbf{0}_{p\times 1}, \mathbf{\Omega}_\varepsilon^{-1})$ and for the following time points $t = 2, ..., T$ data are sampled from $\mathcal{N}(\mathbf{A}\mathbf{Y}_{*,i,t-1}, \mathbf{\Sigma}_\varepsilon)$. In total for each employed combination of the model parameters and design parameters p, n, $\mathcal{T}$, fifty time series data sets are generated.

For each simulated data set, the VAR(1) model is fitted using both the proposed ridge procedure and the 'SCAD'-based method SparseTSCGM. The parameters estimates obtained from both methods are chosen in an identical fashion: using maximization of the leave K-fold-out cross-validated log-likelihood. Instead of removing single element from the set of design points (as discussed in Section ?? of the main document), whole cell line is removed at each ford, due to the fact that the SparseTSCGM cannot deal with unbalanced design. From the set of design points $\mathcal{D} = \{(t, i) : t = 1, \ldots, \mathcal{T}, i = 1, \ldots, n\}$, it is removed one sample: $\mathcal{D} = \{(i) : i = 1, \ldots, n\}$. Resulting parameter estimates of $\mathbf{A}$ and $\mathbf{\Omega}_\varepsilon$ obtained by SparseTSCGM are sparse, whereas their ridge counterparts are not. When comparing the methods with respect to the ability to reconstruct the time series chain graph, the ridge estimates undergo post-estimation sparsification. This is done by means of the local false discovery (lFDR) approach detailed in Section **INSERT??** of the main document.

The ridge and SCAD estimators of $\hat{\mathbf{A}}(\lambda_a)$ and $\widehat{\mathbf{\Omega}}_\varepsilon(\lambda_a, \lambda_\omega)$ are evaluated in terms of their Frobenius loss and their ability to select edges of time series chain graph. The Frobenius loss for (e.g.) a precision matrix estimate is:

$$\left\|\widehat{\mathbf{\Omega}}_\varepsilon(\lambda_a, \lambda_\omega) - \mathbf{\Omega}_\varepsilon\right\|_F^2 \quad = \quad \sum_{j_1, j_2} \left\{ \left[\widehat{\mathbf{\Omega}}_\varepsilon(\lambda_a, \lambda_\omega)\right]_{j_1, j_2} - \left(\mathbf{\Omega}_\varepsilon\right)_{j_1, j_2} \right\}^2.$$

The Frobenius loss for the estimate of $\mathbf{A}$ is defined similarly.

The edge selection ability of the ridge (augment with the post-estimation lFDR-based selection procedure) and SCAD estimation methods are compared by means of sensitivity and specificity. The two methods may yield differing number of edges. This may hamper the comparison of their sensitivity and specificity. To facilitate a better comparison of the two methods in this respect, a simple but different post-estimation edge selection procedure is applied to the ridge estimate. It comprises of selecting the same number of edges as the SCAD method, thus favouring the latter. Having fixed the number of to-be-selected edges, they are selected on the basis of the (absolute) size of the statistic derived from the elements of $\mathbf{A}$ (as pointed out in Section 4.1 the main document) and (standardized) $\widehat{\mathbf{\Omega}}_\varepsilon(\lambda_a, \lambda_\omega)$. This thus means that in each data set for the number of edges selected by the SCAD estimator, we select the same number of largest elements of from the estimate of $\mathbf{A}$.

We first discuss the Frobenius loss comparison. Figure **??** shows the Frobenius loss (as boxplots) of ridge and SCAD estimates of $\mathbf{A}$ (upper panel) and $\mathbf{\Omega}_\varepsilon$ (lower panel) from fifty data sets generated with $p = 25$, $T = 10$, $n = 5$ (the most relevant case empirically), and the combinations of $\mathbf{\Omega}_\varepsilon$ and (both sparse and dense) $\mathbf{A}$. The panels of Figure **??** reveal that the Frobenius loss of the SCAD estimates of both $\hat{\mathbf{A}}(\lambda_a)$ and $\widehat{\mathbf{\Omega}}_\varepsilon(\lambda_\omega)$ exceed that of its ridge counterpart ,for both sparsity levels. This observation is consistent over the data sets generated from different choices of the regression coefficient and precision matrices. This is confirmed when the VAR(1) model is increased to include $p = 50$ variates. The difference in Frobenius loss between the ridge and SCAD estimators grows substantially in favour of the former (confer Figure **??**). When the number of time points $\mathcal{T}$ is increased, the ridge procedure outperforms its SCAD counterpart (as can be witnessed from Figure **??** and Figure **??** representing the '$T = 20$, $n = 5$'-case with $p = 25$ and $p = 50$, respectively,except

for the hub sparsest hub structure where p=25. A similar conclusion holds for the '$T = 10$, $n = 15$'-case with $p = 25$ and $p = 50$ in the case of the sparse $\mathbf{A}$, while for more sparse choice of $\mathbf{A}$ SCAD perform better, except for data driven $\mathbf{A}$ and $\mathbf{\Omega}_\varepsilon$., as can be witnessed from Figure **??** and Figure **??**, respectively.

Finally, Figures **??** and **??** gives the comparison for the '$T = 20$, $n = 15$'-case. The ridge estimator performs better in the simulations that employ data-driven $\mathbf{A}$ and $\mathbf{\Omega}_\varepsilon$, as well as in the case with less sparse choice of $\mathbf{A}$, when p=50. On the other hand, the SCAD estimator performs better on remaining types of networks

The superior Frobenius loss of the SCAD estimator is consequence of the sparse $\hat{\mathbf{A}}(\lambda_a)$ and $\hat{\mathbf{\Omega}}_\varepsilon(\lambda_\omega)$ which favours the SCAD estimator due to the sparsity of the estimates. This is verified by comparing the ridge and SCAD estimates separately for zero and non-zero elements of $\mathbf{A}$. The SCAD estimator performs better for the zero elements, while the ridge estimator outperforms on non-zero elements of $\mathbf{A}$. Only for the hub network does the SCAD estimator generally outperforms the ridge estimator.

Figures **??** to **??** present the edge selection sensitivity and specificity of both VAR(1) estimation methods for the various models. The lower panels of the Figure **??** summarize the comparison of the methods where edge selection for the ridge method is done by means of the post-estimation local FDR procedure , where the number of parameters is determined based on the SCAD selection, favouring this method. In the selection comparison methods are on a par for the data set with a set up closest to the empirically most prevalent ($T = 10$ and $n = 5$) for more sparse choice of the $\mathbf{A}$, while in less sparse case ridge perform slightly better. For data sets with $T = 20$ and $n = 15$ the methods are more less on the par, however in the sparse set up ridge perform slightly better, while in the more sparse SCAD estimator for hub structure and p=50. The ridge procedure augmented with a post-estimation selection step has the potential to be equally good (or in some cases even better) as the in-built selection of the SCAD method.