Time-Varying Dynamic Bayesian Networks Appendix

Le Song, Mladen Kolar and Eric Xing
School of Computer Science, Carnegie Mellon University
{lesong, mkolar, epxing}@cs.cmu.edu

1 Proof Sketch of Theorem 1

1.1 Assumptions

1. The model given in the equation

$$\boldsymbol{X}^t = \boldsymbol{A}^t \cdot \boldsymbol{X}^{t-1} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}) \tag{1}$$

   is a locally stationary model (see [4] for a rigorous definition).

2. Element of the matrix $\boldsymbol{A}^t$ are smooth functions with bounded second derivatives, *i.e.*, there exists a constant $L > 0$ such that

$$|\frac{\partial}{\partial t} \boldsymbol{A}^t_{ij}| < L \quad \text{and} \quad |\frac{\partial^2}{\partial t^2} \boldsymbol{A}^t_{ij}| < L \tag{2}$$

3. The minimum absolute value of non-zero element of the matrix $\boldsymbol{A}^t$ is bounded away from zero at observation points, and this bound tends to zero as we observe more and more samples, *i.e.*,

$$a_{\min} := \min_{t \in \{1/T, 2/T, \ldots, 1\}} \min_{i \in [p], j \in S^t_i} |A^t_{ij}| > 0. \tag{3}$$

4. Let $\boldsymbol{\Sigma}^t = \mathbb{E}[\boldsymbol{X}^t (\boldsymbol{X}^t)^T] = [\sigma_{ij}(t)]_{i,j=1,\ldots,p}$ and let S^t_i denote the set of non-zero elements of the i-th row of the matrix $\boldsymbol{A}^t$, *i.e.*, $S^t_i = \{j \in [p] \ : \ \boldsymbol{A}^t_{ij} \neq 0\}$. We assume that there exist a constant $d \in (0, 1]$ such that

$$\max_{j \in S^t_i, k \neq j} |\sigma_{jk}(t)| \leq \frac{d}{s}, \quad \forall i \in [p], t \in [0, 1],$$

   where s is an upper bound on the number of non-zero elements, *i.e.*, $s = \max_{t \in [0,1]} \max_{i \in [p]} |S^t_i|$.

5. The kernel $K(\cdot) : \mathbb{R} \mapsto \mathbb{R}$ is a symmetric function and has bounded support on $[0, 1]$. There exist a constant M_K that upper bounds the following quantities $\max_{x \in \mathbb{R}} |K(x)|$ and $\max_{x \in \mathbb{R}} K(x)^2$.

1.2 Theorem

Theorem 1 *Assume that the conditions made above hold. Let the penalty scale as*

$$\lambda = \mathcal{O}(\sqrt{\frac{\log p}{Th}})$$

and let the minimum non-zero value be sufficiently large

$$a_{\min} \geq 2\lambda.$$

If we assume

$$h = \mathcal{O}(T^{1/3}) \quad \textit{and} \quad \frac{s \log p}{Th} = o(1)$$

then

$$\mathbb{P}[\operatorname{supp}(\hat{\boldsymbol{A}}^{t^*}) = \operatorname{supp}(\boldsymbol{A}^{t^*})] \to 1, \quad t^* \in [0,1]. \tag{4}$$

To prove the above theorem we follow the proof strategy of [3], however, there are a lot of technical details that need to be addressed. First, in the case of the auto-regressive model, the observations are not *i.i.d.*. Second, the process is not stationary, but can only be approximated by a locally stationary process. For this reason, we additionally need to deal with the bias term that arises due to non-stationarity. We leave the detailed proof of this theorem for a full version of the paper, and outline the approach below.

1.3 Proof Sketch

We outline the proof strategy, which follows [3] with necessary modifications to allow for dependent data. Since the estimation problem decomposes across different rows of matrix $\boldsymbol{A}$, it is enough to show that one row is estimated sparsistently with high probability and then apply the union bound to show that the whole matrix $\boldsymbol{A}$ is estimated sparsistently. We proceed by characterizing the optimum

$$\hat{\boldsymbol{A}}^{t^*}_{i\cdot} = \operatorname*{argmin}_{\boldsymbol{a} \in \mathbb{R}^{1\times n}} \frac{1}{T} \sum_{t=1}^{T} w^{t^*}(t)(x_i^t - \boldsymbol{a}\boldsymbol{x}^{t-1})^2 + 2\lambda ||\boldsymbol{a}||_1. \tag{5}$$

From the estimator $\hat{\boldsymbol{A}}^{t^*}_{i\cdot}$ we obtain the estimator $\hat{S}^{t^*}_i$. We show that $\mathbb{P}[\hat{S}^{t^*}_i = S^{t^*}_i] \to 1$ in two steps, showing that $\mathbb{P}[\hat{S}^{t^*}_i \not\subseteq S^{t^*}_i] \to 0$ and that $\mathbb{P}[S^{t^*}_i \not\subseteq \hat{S}^{t^*}_i] \to 0$. From the analysis of the KKT conditions, we have that if $\hat{\boldsymbol{A}}^{t^*}_{ik} = 0$, then

$$|\frac{1}{T} \sum_{t=1}^{T} w^{t^*}(t)(x_i^t - \hat{\boldsymbol{A}}^{t^*}_{i\cdot}\boldsymbol{x}^{t-1})x_k^{t-1}| \leq \lambda, \tag{6}$$

which we use to show that

$$\begin{aligned}
&\mathbb{P}[S^{t^*}_i \not\subseteq \hat{S}^{t^*}_i] \\
&\quad \leq s \max_{k \in S^{t^*}_i} \mathbb{P}[|\frac{1}{T} \sum_{t=1}^{T} w^{t^*}(t)(x_i^t - \hat{\boldsymbol{A}}^{t^*}_{i\cdot}\boldsymbol{x}^{t-1})\boldsymbol{x}_k^{t-1}| \leq \lambda; \quad \boldsymbol{A}^{t^*}_{ik} \neq 0] \\
&\quad \leq s \max_{k \in S^{t^*}_i} \mathbb{P}[|\frac{1}{T} \sum_{t=1}^{T} w^{t^*}(t)((t - t^*)\frac{\partial}{\partial t}\boldsymbol{A}^{t^*}_{i\cdot}\boldsymbol{x}^{t-1} + \boldsymbol{\epsilon}_i^t)\boldsymbol{x}_k^{t-1}| \geq \frac{\lambda}{2}] + \\
&\qquad s \max_{k \in S^{t^*}_i} \mathbb{P}[|\frac{1}{T} \sum_{t=1}^{T} w^{t^*}(t)((\boldsymbol{A}^{t^*}_{i\cdot} - \hat{\boldsymbol{A}}^{t^*}_{i\cdot})\boldsymbol{x}^{t-1})\boldsymbol{x}_k^{t-1}| \geq \frac{\lambda}{2}].
\end{aligned} \tag{7}$$

Next, we describe how to bound the two probabilities above. It is not hard to show that $\frac{1}{T}\sum_{t=1}^{T} w^{t^*}(t)\boldsymbol{x}_l^{t-1}\boldsymbol{x}_k^{t-1} \to_p \sigma_{lk}(t^*)$, sufficiently fast, combing the standard results on stationary time series [2] and Lemma 4 in [1]. Combining this result, together with exponential inequality for martingales [5], one obtains that the first term converges to 0, exponentially fast. The convergence of the second term to 0, can be shown in a similar way as Theorem 2.2 in [3], utilizing the fact that $\frac{1}{T}\sum_{t=1}^{T} w^{t^*}(t)\boldsymbol{x}_l^{t-1}\boldsymbol{x}_k^{t-1} < \sigma_{lk}(t^*) + \delta$ with high probability and using exponential inequality for martingales in place of Bernstein's inequality. Combining these results, we have

$$\mathbb{P}[\hat{S}^{t^*}_i \not\subseteq S^{t^*}_i] \to 0. \tag{8}$$

To finalize the proof of the theorem, we show that

$$\begin{aligned}
&\mathbb{P}[\hat{S}_i^{t^*} \not\subseteq S_i^{t^*}] \\
&\leq \sum_{k \notin S_i^{t^*}} \mathbb{P}[|\frac{1}{T}\sum_{t=1}^{T} w^{t^*}(t)(x_i^t - \hat{\boldsymbol{A}}_{i\cdot}^{t^*}\boldsymbol{x}^{t-1})\boldsymbol{x}_k^{t-1}| \geq \lambda] \\
&\leq s \max_{k \in S_i^{t^*}} \mathbb{P}[|\frac{1}{T}\sum_{t=1}^{T} w^{t^*}(t)((t-t^*)\frac{\partial}{\partial t}\boldsymbol{A}_{i\cdot}^{t^*}\boldsymbol{x}^{t-1} + \boldsymbol{\epsilon}_i^t)\boldsymbol{x}_k^{t-1}| \geq \frac{\lambda}{2}] + \\
&\quad \sum_{k \notin S_i^{t^*}} \mathbb{P}[|\frac{1}{T}\sum_{t=1}^{T} w^{t^*}(t) \sum_{j \in S_i^{t^*}} (\boldsymbol{A}_{ij}^{t^*} - \hat{\boldsymbol{A}}_{ij}^{t^*})\boldsymbol{x}_j^{t-1}\boldsymbol{x}_k^{t-1}| \geq \frac{\lambda}{2}],
\end{aligned} \tag{9}$$

which can be again argued to converge to 0, exponentially fast.

2 Additional Figures

Table 1: (a) Timeline of the 23 enriched transcriptional factor target gene sets, and (b) the 26 enriched gene knockout signature gene sets. Each cell in the plot corresponds to one gene set at one specific time point. The cells in each row are ordered according to their time point across two cell cycles. Cells colored blue indicate the corresponding gene set listed in the right column is detected in the estimated network; blank color indicates the gene set is not detected. It can be seen that many of them are dynamic and transient, and can not be captured by the static network.

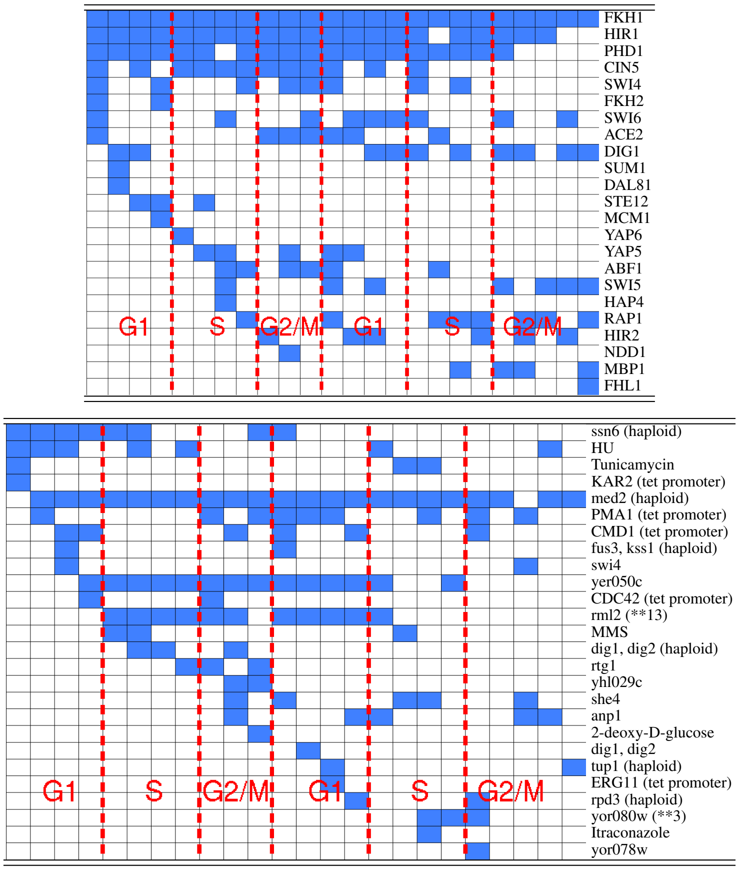
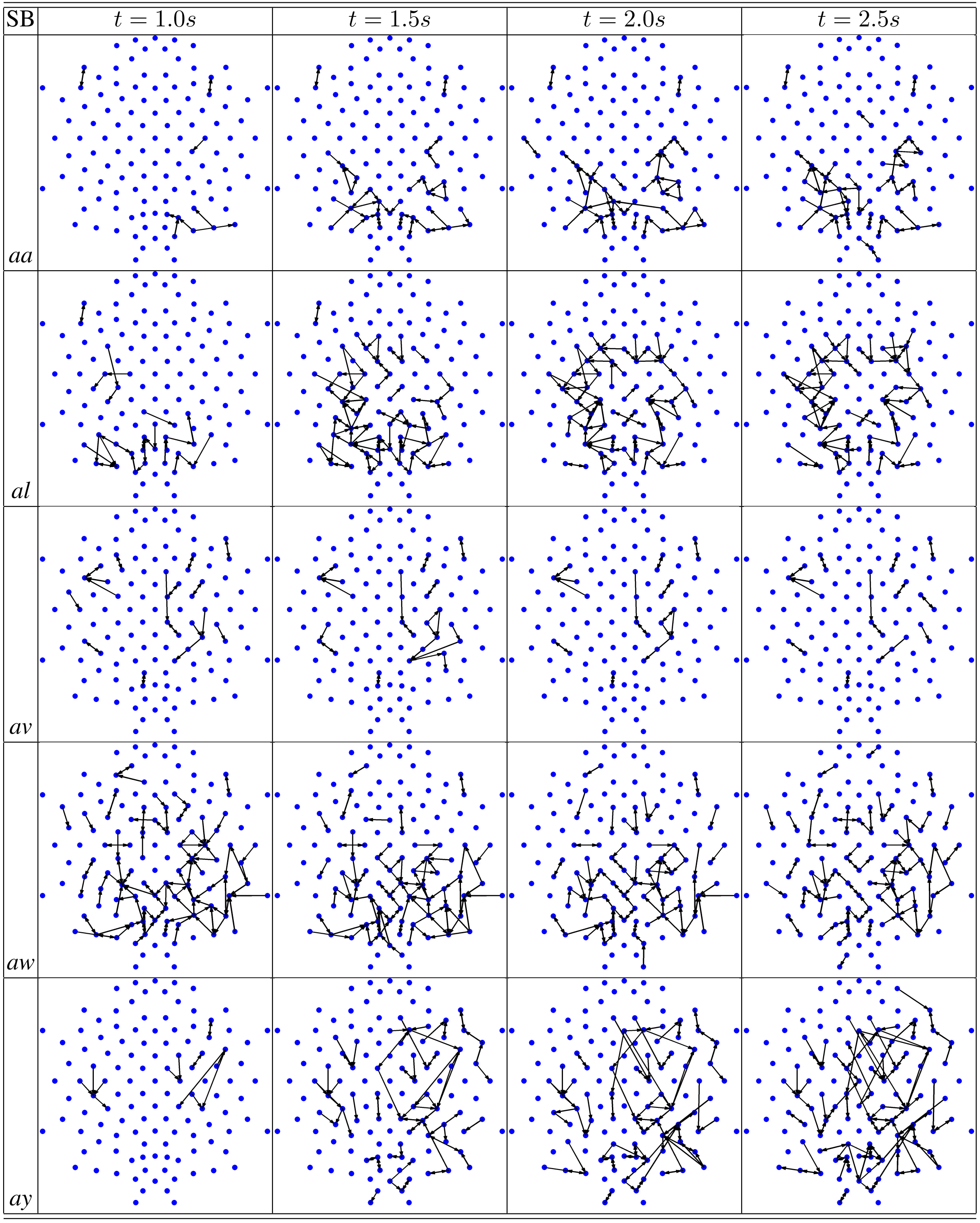


Figure 1: Temporal progressions of brain interactions for five BCI subjects. The dots correspond to EEG electrode positions in 10-5 system. The estimated TV-DBN reveal that the brain interaction of subject 'av' is particularly weak and the brain connectivity actually decreases as the experiment proceeds. In contrast, all other four subjects show an increased brain interaction as they engage in active imagination. Particularly, these increased interactions occur between visual and motor cortex. This dynamic coherence between visual and motor cortex corresponds nicely to the fact that subjects are consciously transforming visual stimuli into motor imaginations which involve the motor cortex. It seems that subject 'av' fails to perform such integration due to the disruption of brain interactions.



References

[1] Peter J. Bickel and Elizaveta Levina. Some theory for fisher's linear discriminant function, 'naive bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, 10(6):989–1010, 2004.

[2] Peter J. Brockwell and Richard A. Davis. *Time Series: Theory and Methods (Springer Series in Statistics)*. Springer, September 1991.

[3] Florentina Bunea. Honest variable selection in linear and logistic regression models via ℓ_1 and $\ell_1 + \ell_2$ penalization. *Electronic Journal of Statistics*, 2:1153, 2008.

[4] R. Dahlhaus. Fitting time series models to nonstationary processes. *Ann. Statist*, (25):1–37, 1997.

[5] V.H. de la Pena. A general class of exponential inequalities for martingales and ratios, 1999.