Inferring network statistics from high-dimensional undersampled time-course data

Dominik Linzner[⋆] **and Heinz Koeppl**[⋆†]

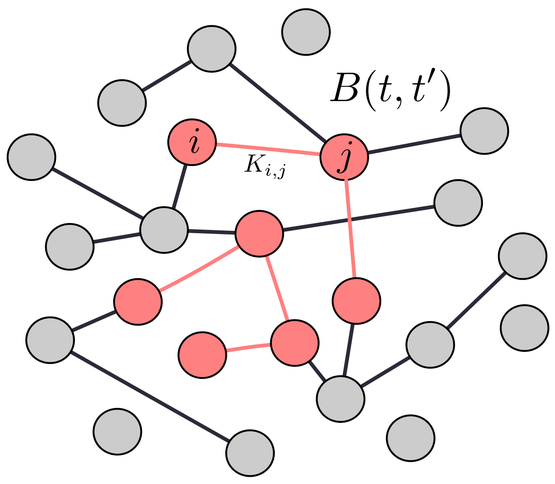
[⋆]**Department of Electrical Engineering and** [†]**Department of Biology, Technische Universität Darmstadt, Germany**

1. Motivation

Reconstructing networks from current high-dimensional datasets is a notoriously ill-posed problem. For this reason we want to focus on more general statistical properties, as finding the degree distribution or sparsity.

We consider sub-systems of continuously valued random variables embedded in a larger hidden system. In this scenario, also referred to as undersampling, the dynamics of the sub-systems are influenced by the hidden system. With increasing number of connections between sub-system and hidden system these effects can be treated statistically. Analysing these effects then allows for inference of statistical properties of the hidden system.

2. Setting

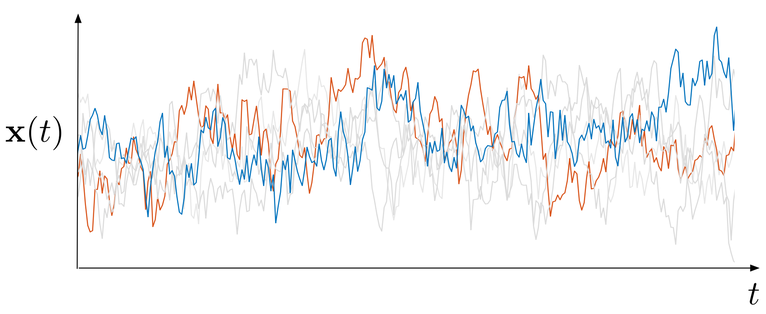


An observed network (red) with couplings between nodes K_{ij}. Artificial couplings are mediated through a hidden network (grey) which we will model by mean-field coloured noise $B(t,t')$.

Consider a dynamic network of random variables with continuous degrees of freedom. We now assume that only some parts of the network have been observed. This defines a sub-network and a bulk. We consider dynamics described by Langevin dynamics:

$$\dot{\mathbf{x}} = f(\mathbf{x}) + \xi.$$

$\mathbf{x}$ is a vector of the continuous variables of interest, which for example can describe a concentration. These are then coupled by $f(\mathbf{x})$ and driven by external gaussian noise ξ which is assumed to be delta-correlated.



Only a fraction of nodes can actually be observed(blue, red). The others are hidden (grey).

3. Second order mean-field approximation

Mean-field approximation for dynamical equations with continuous degrees of freedom subject to stochastic noise.

- Probability of this system realising a specific trajectory as a a so-called Martin-Siggia-Rose-Janssen-De Dominicis (MSRJD) functional integral

$$p(\mathbf{x}) = (2\pi)^{-n/2} \int_{-\infty}^{\infty} \prod_t \mathrm{d}\hat{\mathbf{x}}(t) \langle e^{\mathscr{H}[\mathbf{x},\hat{\mathbf{x}}]} \rangle_\xi,$$

with the action

$$\mathscr{H}[\mathbf{x},\hat{\mathbf{x}}] = \sum_t i\hat{\mathbf{x}}(t)\mathscr{D}\left\{\dot{\mathbf{x}}(t) - f_{\text{local}}(\mathbf{x}(t)) - \alpha f_{\text{int}}(\mathbf{x}(t)) - \xi(t)\right\}.$$

- Expanding w.r.t. α and simultaneously fixing the moments

$$\begin{aligned}
\mu(t) &= \langle x_i(t) \rangle_\alpha \\
\hat{\mu}(t) &= \langle \hat{x}_i(t) \rangle_\alpha \\
C_i(t,t') &= \langle x_i(t) x_i(t') \rangle_\alpha \\
R_i(t,t') &= -i\langle \hat{x}_i(t) x_i(t') \rangle_\alpha \\
B_i(t,t') &= -\langle \hat{x}_i(t) \hat{x}_i(t') \rangle_\alpha,
\end{aligned}$$

by extremizing with respect to the appropriate constraints. This yields an effective gaussian and local approximation.

For a more in-depth discussion we refer the reader to [1, 2].

4. Gaussian likelihood sub-network

Gaussian integral can then be solved casting the likelihood into a simple analytic form

$$P(\vec{x}_a|B,K) = \frac{(2\pi)^{|T|/2}}{|B|^{1/2}} \exp\left[-\frac{1}{2(\Delta\Sigma)^2}\mathrm{Tr}\{\mathrm{SB}\}\right],$$

- $S = \delta\vec{x}_a \delta\vec{x}_a^T$ and $|T|$ as the number of observed time points.
- $\delta\vec{x}_a = x_a(t+\Delta) - x_a(t) + \Delta\lambda x_a(t) - \Delta m_a(t) - \Delta\sum_b x_b(t) K_{ba}$
- $\partial_t m_a(t) = -\lambda m_a(t) + \langle K^2 \rangle x_a(t) + \mathscr{O}(\langle K \rangle^2)$

Closed system of equations entirely determined by data and network statistics!

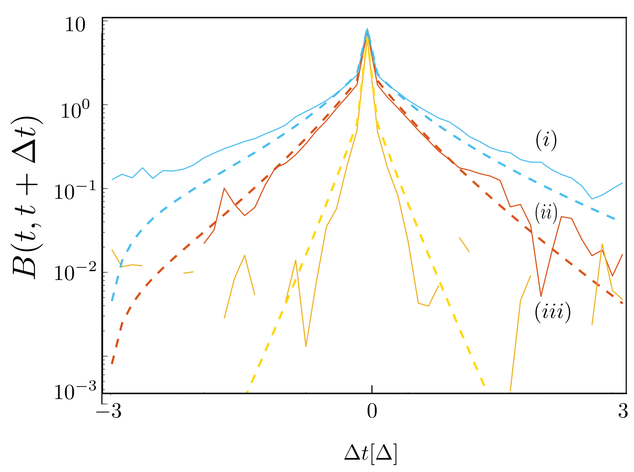
5. Mean-field noise

Model the hidden network as Erdos-Renyi graph (edge inclusion with probability p) with gaussian couplings and zero mean.

- $B(t,t')$ only depends on the networks spectral density $\rho(k)$.
- Spectral density can then be expressed through a semi-circle law [3]

$$\rho(k) = \begin{cases} \frac{\sqrt{4-k^2/\gamma}}{\sqrt{\gamma}} & \text{if } |k| \leq 2\sqrt{\gamma} \\ 0 & \text{else} \end{cases}.$$

with a sparsity parameter $\gamma = Np(1-p)\sigma^2$.



Comparison of mean-field coloured noise with coloured noise averaged from the data for different parameters $(i):[\lambda=1,\gamma=0.16]$, $(ii):[\lambda=1,\gamma=0.09]$, $(iii):[\lambda=2,\gamma=0.09]$.

6. MLE network

To make the mean-field approximation of close to infinite neighbours viable for sparse graphs, we have to consider large undersampled sub-systems. We thus have to first infer the sub-network first.

- Assume local relaxation rates λ are known.
- Maximisation of likelihood yields MLE $\hat{K}$:

$$\begin{aligned}
\hat{\mathbf{K}}_j &= D_j^{-1}\mathbf{q}_j. \\
D_j &= \Delta \sum_{t,t'} B(t,t') \Omega_{\bullet/j,\bullet/j}(t,t') \\
\mathbf{q}_j &= \frac{1}{2}\sum_{t,t'} B(t,t')[\mathbf{x}_{\bullet/j}(t) dx_j(t') + \mathbf{x}_{\bullet/j}(t') dx_j(t)]
\end{aligned}$$

- $\Omega_{i,j}(t,t')$ is the empirical covariance matrix in space and time $\Omega = \mathbf{x}(t)\mathbf{x}(t')^T$ and $dx_j(t) = x_j(t+\Delta) - x_j(t) + \Delta\lambda x_j(t)$.
- Note that for insufficient data a Laplace prior distribution still yields a maximum a-posteriori.

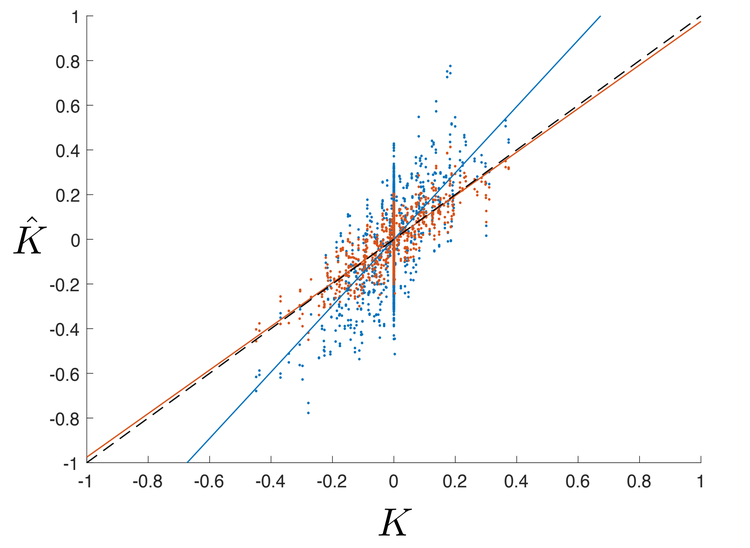
Fully observed networks

- Coloured noise is absent ($B(t,t') = \delta_{t,t'}$) we can, by comparing with the original linear model lower-bound the error

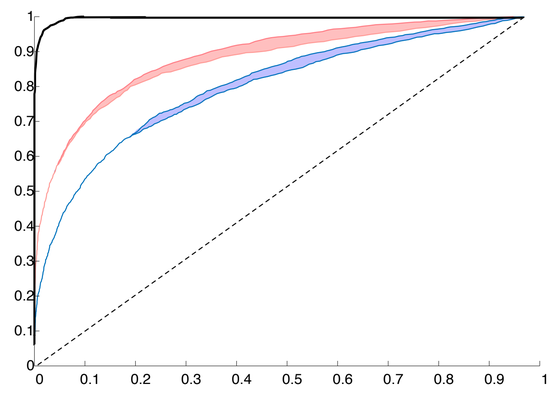
$$||\hat{\mathbf{K}}_j - \mathbf{K}_j|| = \sqrt{\Delta}\xi(t),$$

- We recover the expected simple regression model of the naive approach.

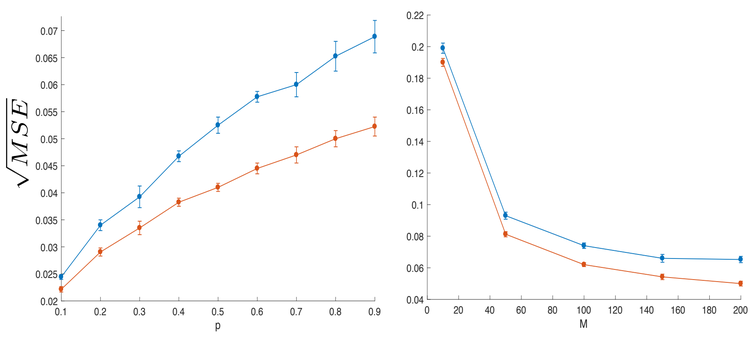
7. Simulation results



Comparison of naive MLE (blue) and our method (red) with the true graph, for 40 observed nodes in contact with 1000 hidden nodes. The graph is generated as an Erdos-Renyi graph with sparsity parameter $\gamma = 0.09$. We used 2000 trajectories each consisting of 10 time points with time steps $\Delta = .1/\Sigma$.



AUC for edge detection improves for conducted experiments (40 observed 400 hidden nodes).



Setting: 40 observed nodes in contact with 160 hidden nodes. Naive MLE(blue) and our method(red). Left: MSE for different sparsities. We used 400 trajectories each consisting of 10 time points with time steps $\Delta = .1/\Sigma$ Right:MSE for different number of samples for $p = 0.6$. The bias of the bulk can only partly be removed using our method.

8. Towards inference of network statistics

It is well-known that real networks are in general only badly described by Erdos-Renyi graphs[3]. To infer the true network statistics it is therefore necessary to use more sophisticated models for the bulk network. We are aiming at finding a general sampling scheme employing our sub-network likelihood over general statistical graph models and then identifying the best fit by scoring.

References

[1] Bravi, B., Sollich, P. and Opper, M. Extended Plefka expansion for stochastic dynamics. Journal of Physics A: Mathematical and Theoretical, 49(**19**), 194003

[2] Bravi, B., and Sollich, P. Inference for dynamics of continuous variables: the Extended Plefka Expansion with hidden nodes. J. Stat. Mech. (2017) 063404

[3] Albert, R., and Barabasi, A.-L. Statistical mechanics of complex networks. Reviews of Modern Physics, 74(**1**), 47-97.

This work was funded by the European Union's Horizon 2020 research and innovation programme under grant agreement 668858.