Imperial College London

25 April 2024

A gentle introduction to data story telling

The art and science of communicating via complex data

Daniel Salnikov
PhD retreat Goodenough College

Today's talk

- We are all story tellers.
- Statistical intuition and understanding a complex world.
- *Mathematical formalism*.
- Beyond mathematical formalism.
- How to think statistically.
- Case study: Glimpse of network statistical learning.
- Why you should not be afraid of "AI".
- Why you should be afraid of "AI"
- Factfulness: tell a data-based story.

Stories by the campfire

Statistics is a a set of tools for obtaining, transmitting and presenting *knowledge* based on empirical data, e.g., what are the best mammoth hunting grounds.

Communicating Facts

Acting based on inference

So Data

Figure 1: Reasoning with data.

Data are varied, yet we can study different presentations with analytical methods:

- Statistics
- Data Science
- Computer Science
- Augmented Intelligence and Machine Learning

The linear model and statistical intuition

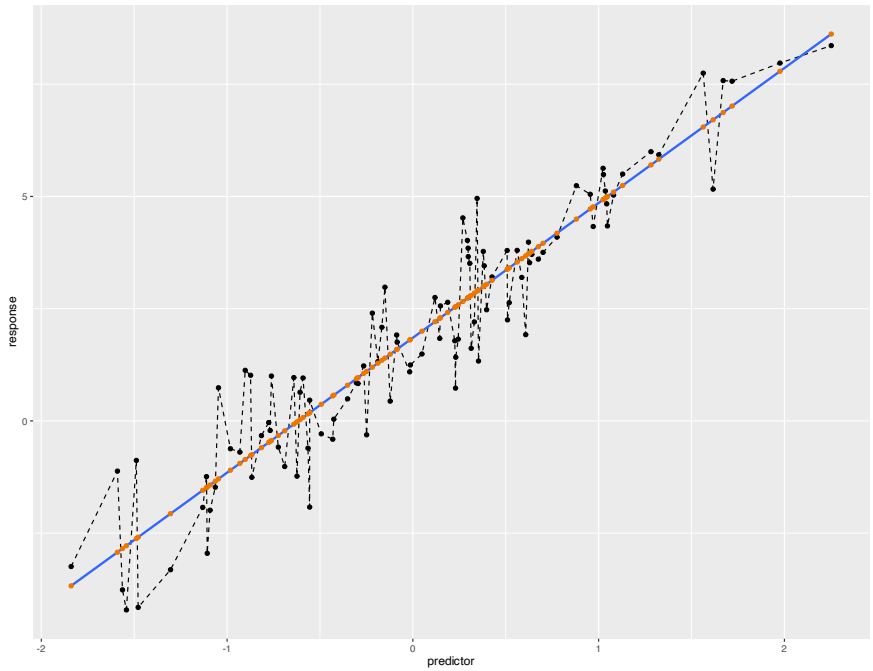


Figure 2: Simulated linear regression plot.

The curse of dimensionality

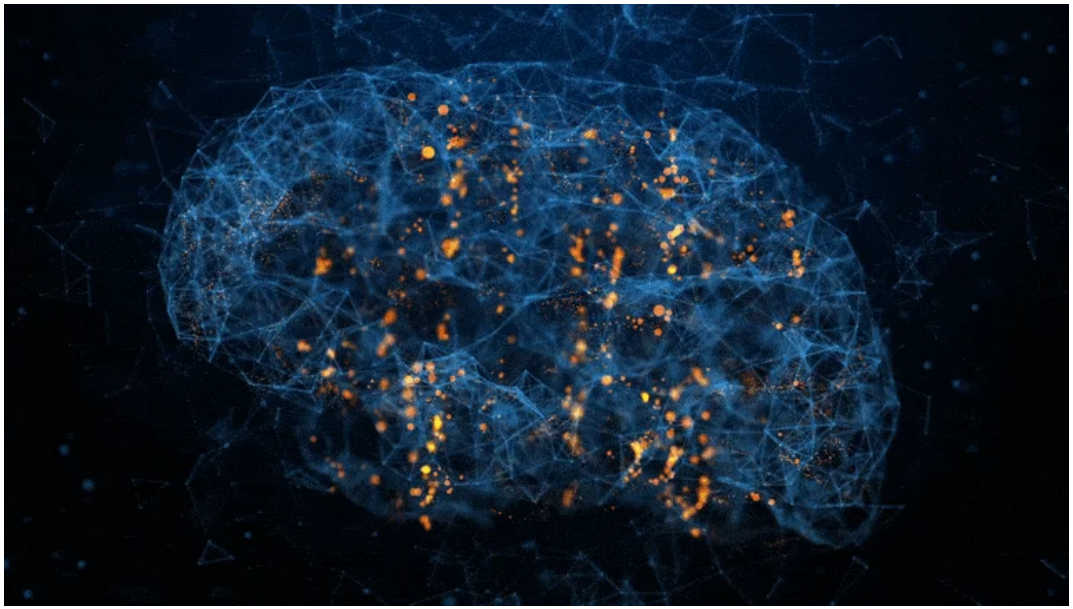
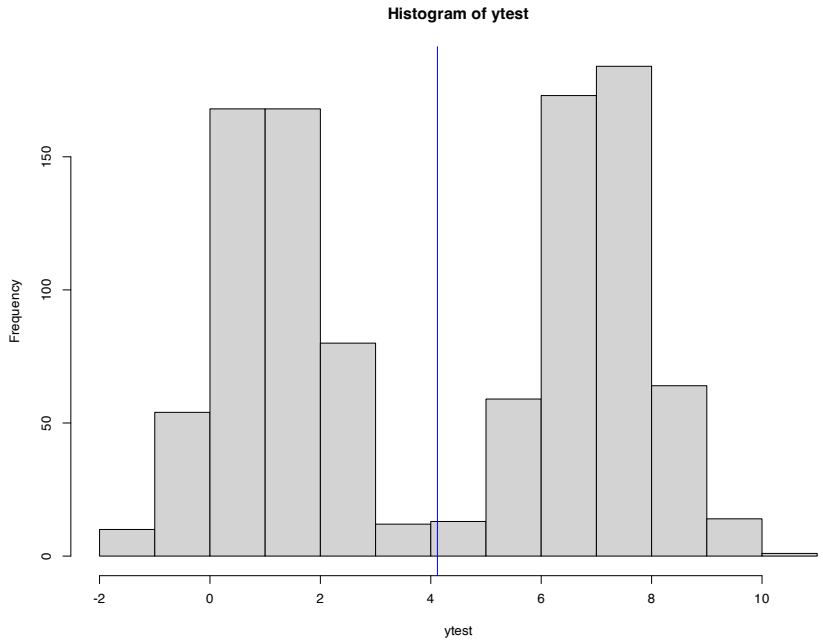


Figure 3: It is exponentially difficult to sample points in the required neighbourhoods.

The mean (may) hide spread



Mind over data and mathematical formalism

Figure 4: Hans Rosling

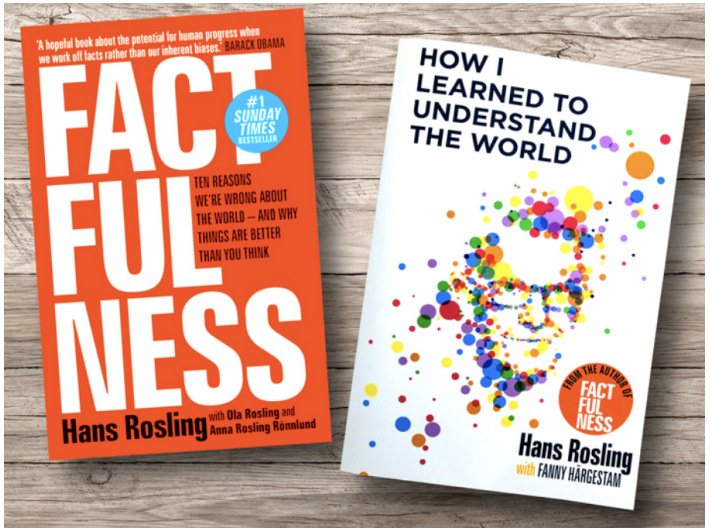


Figure 5: Books for reasoning with data in a good way.

Network statistical learning

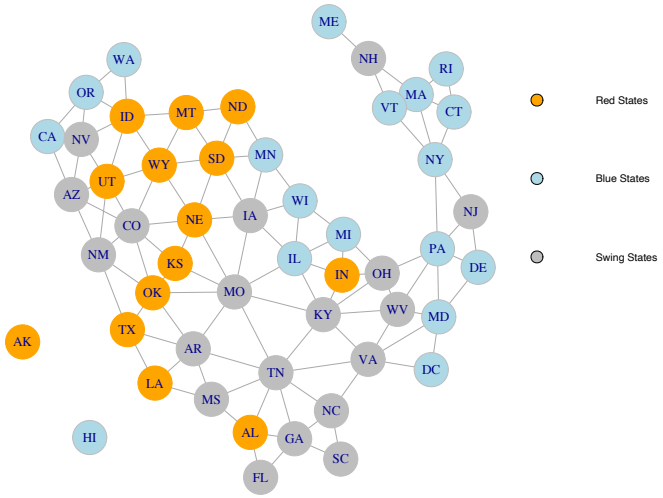


Figure 6: USA state-wise network. We aim to study signals polluted by noise linked to this type of structures.

Application to a COVID-19 Network Time Series

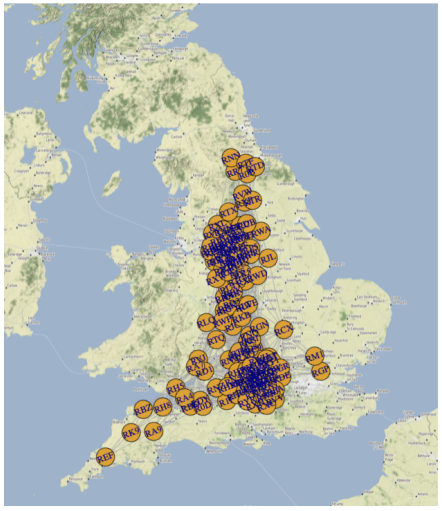
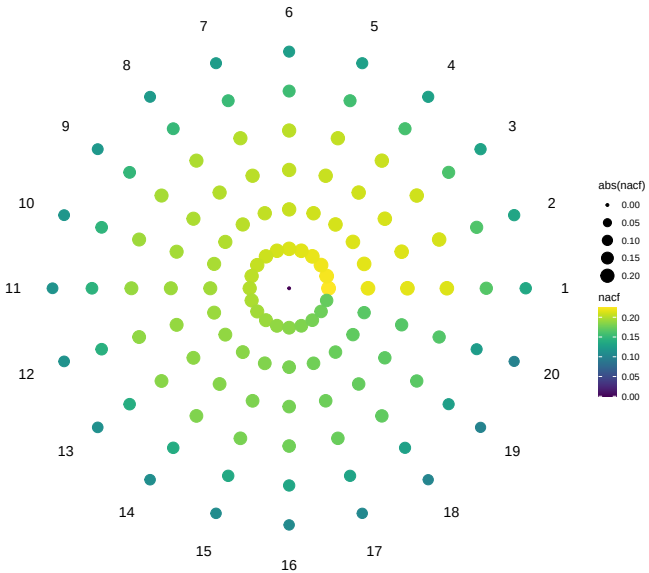


Figure 7: Network for the $\boldsymbol{X}_t$ COVID-19 network time series.

The data $\mathbf{X} := [\boldsymbol{X}_1, \cdots, \boldsymbol{X}_T]$ are obtained from the UK Government's Coronavirus Dashboard (`coronavirus.data.gov.uk`).

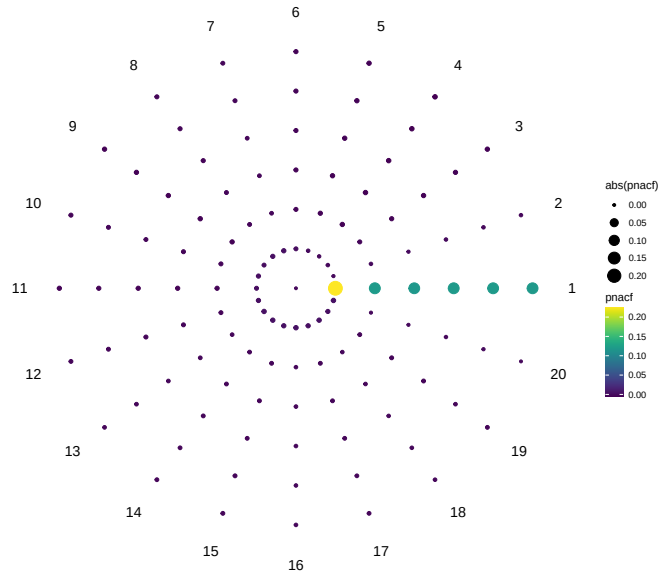
- Each $X_{i,t}$, $i \in \{1, \cdots, d\}$ is the logarithm of one plus the number of patients transferred to mechanical ventilation beds.
- Each i refers to an individual NHS Trust in England and there 140 trusts in the network.
- Consist of 452 days.

Understanding COVID-19 dynamics



nacf corbit plot with max lag 20 and max path length 6

(a) COVID-19 NACF Corbit plot.



pnacf corbit plot with max lag 20 and max path length 6

(b) COVID-19 PNACF Corbit plot.

Figure 8: Figure 8a shows the Corbit plot for the observed NACF, and Figure 8b shows the observed PNACF with respect to the COVID-19 network time series.

Conduit towns

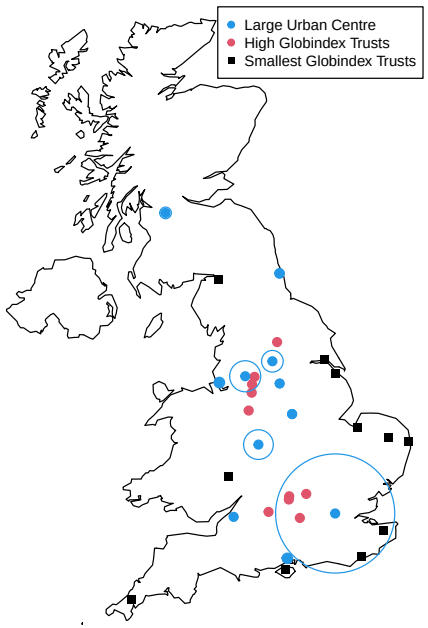


Figure 9: NHS trusts that influence network correlation.

Understanding voting dynamics

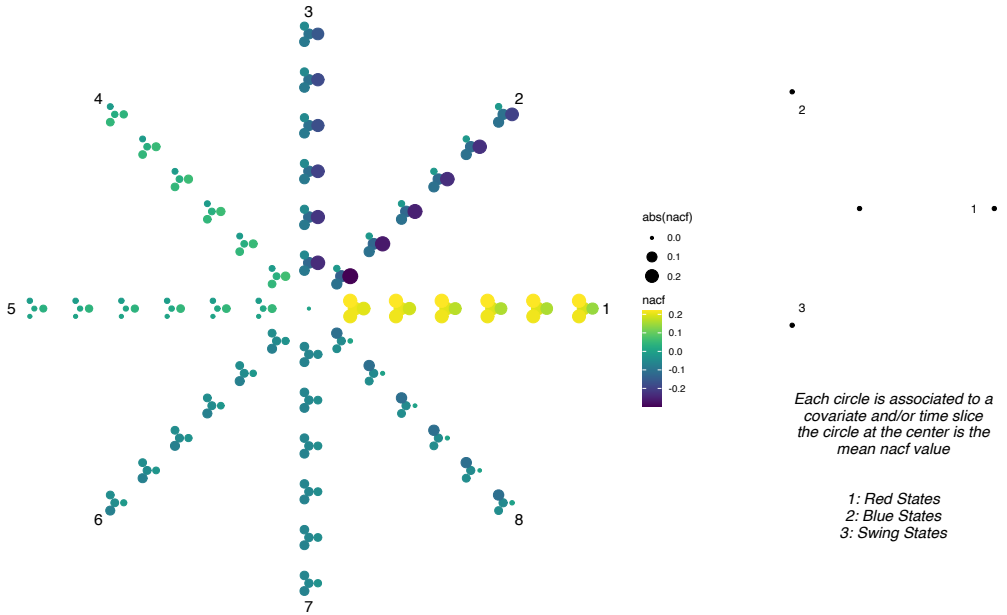


Figure 10: Wagner plot of voting dynamics in the USA (1976 to 2020).

Why you should fear "AI"

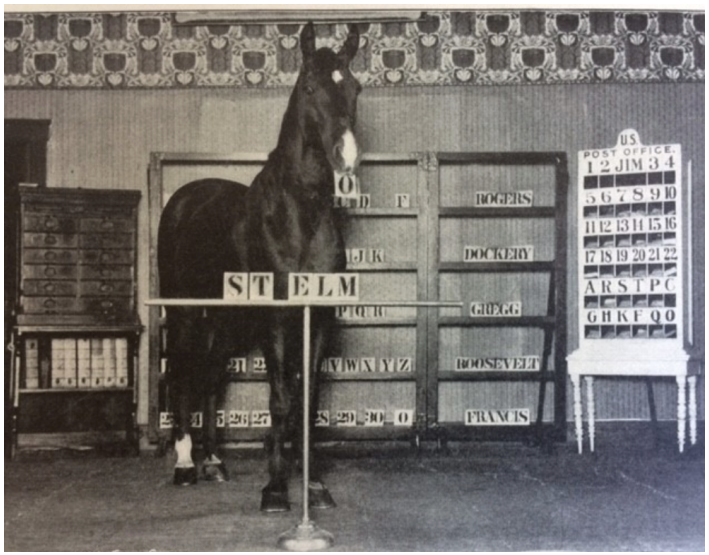


Figure 11: Clever Hans.

Figure 12: Correlation does not imply causation. Source: spurious correlations.

Why you should not fear "AI"

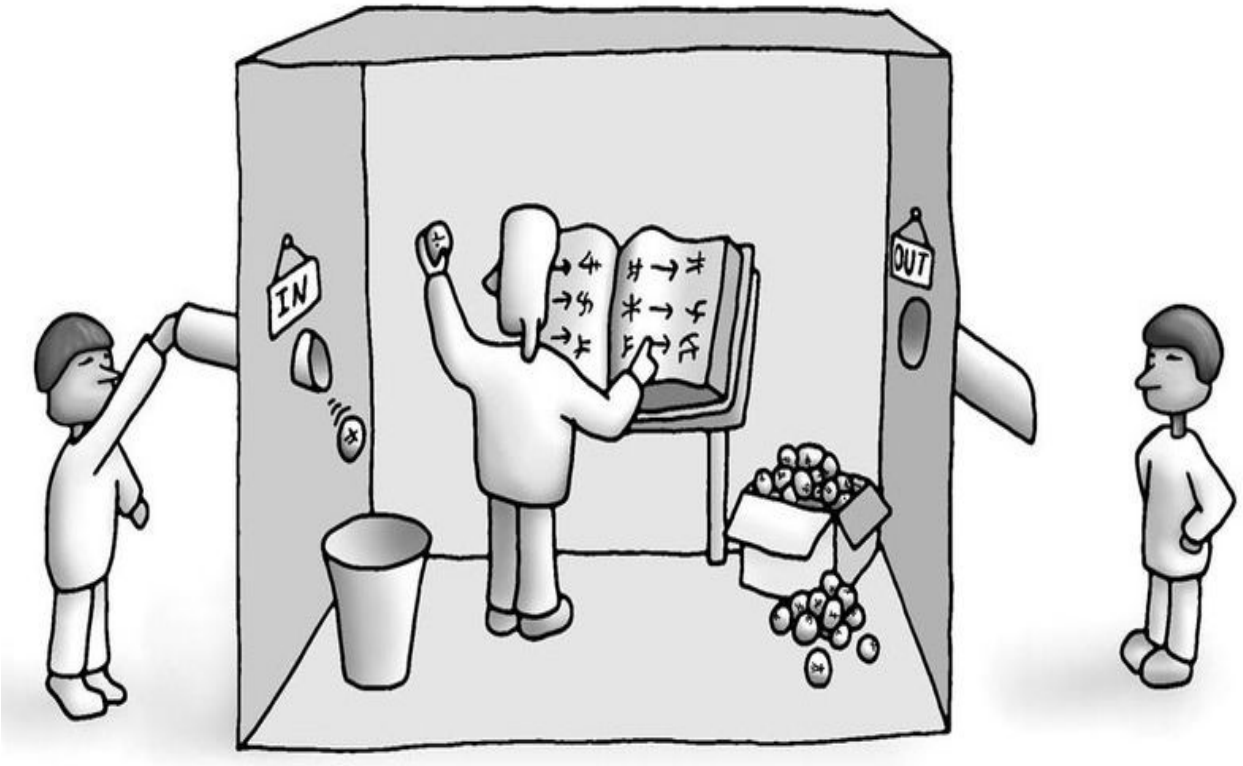
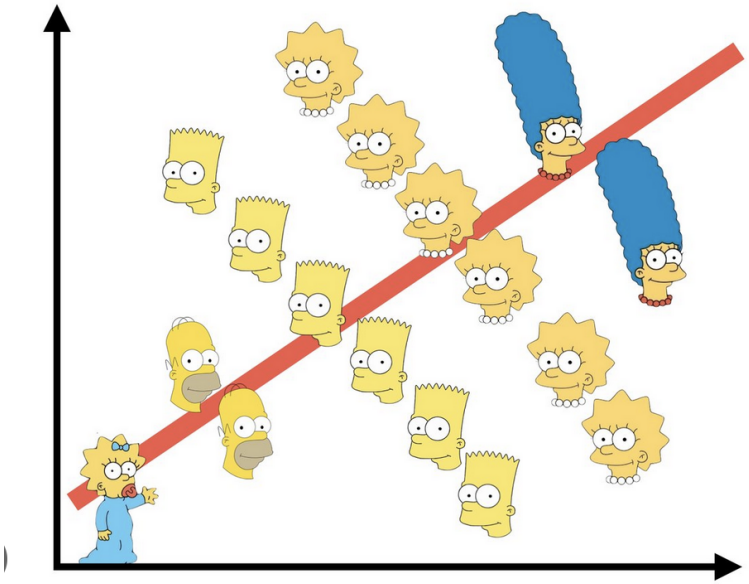


Figure 13: Chinese room thought experiment.

Simpson's paradox



It is a leap of faith

We can only be less wrong, remember,

I do not control the wind, I study it and act accordingly.

Summary

There is no true interpretation of anything; interpretation is a vehicle in the service of human comprehension. The value of interpretation is in enabling others to fruitfully think about an idea.

-Andreas Buja

To properly understand complex data, we must dig in, remain humble and curious. The data are there, it is up to us to make sense of it all.

Acknowledgements

Engineering and Physical Sciences Research Council

Questions

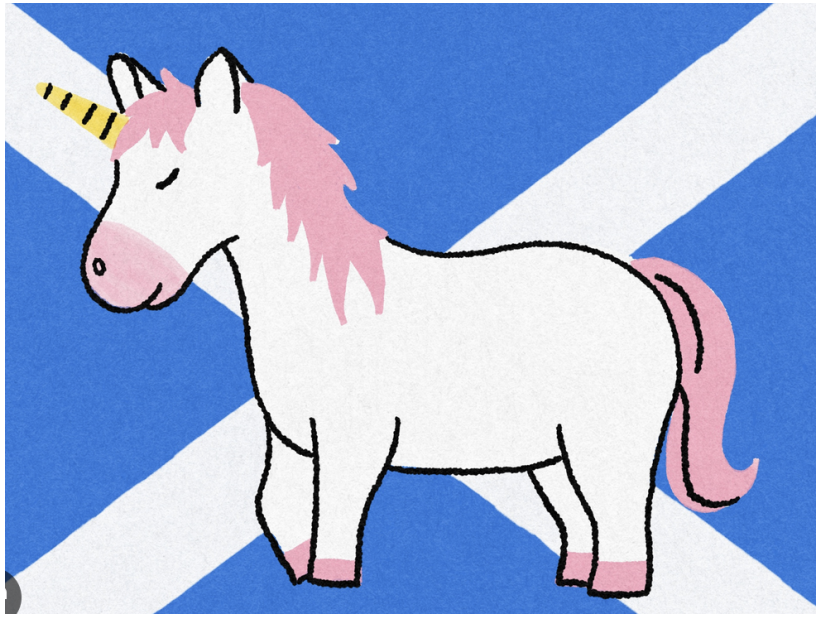


Figure 15: Thanks for listening :)

References

Hastie, T., Tibshirani, R., & Friedman, J. 2017, The Elements of Statistical Learning, 2nd edn. (New York: Springer), doi: 10.1007/b94608

Knight, M., Leeming, K., Nason, G., & Nunes, M. 2020, Journal of Statistical Software, 96, 1, doi: 10.18637/jss.v096.i05

Nason, G., Salnikov, D., & Cortina-Borja, M. 2023, New tools for network time series with an application to COVID-19 hospitalisations.

https://arxiv.org/abs/2312.00530

Wainwright, M. J. 2019, High-Dimensional Statistics: A Non-Asymptotic Viewpoint (Cambridge: Cambridge University Press), doi: 10.1017/9781108627771