Methods in Graphical Models with Missingness

BST245: Multivariate & Longitudinal Data

Dominic DiSanto
Department of Biostatistics, Harvard University

1 Graphical Models

Graphical models are a powerful tool for modelling of multivariate data, often with the goal of understanding conditional dependencies. Focusing on undirected graphs for the proposed project, a graph $\mathcal{G}$ with p nodes (read: random variables/vectors) is commonly represented by its vertex and edge sets V, E as such: $\mathcal{G} = (V, E), V = \{1, ..., p\}, E \subseteq V \times V$ [7]. Nodes correspond to some random vector $X = (X_1, ..., X_p) \sim F$ for some distribution F, where a common and convenient assumption is multivarate Gaussianity, i.e. $F \equiv N(\mu, \Sigma)$.

In the ideal setting, we completely observe n (often *iid*) realizations $X^{(1)}, ..., X^{(n)}$, resulting in $N = np$ total data points. In the high-dimensional setting ($p \gg n$), methods with regularization have been proposed to recover true, underlying graphical structure, such as the neighborhood selection [7] and graphical lasso algorithms [3]. However theoretical guarantees on consistency of graph recovery do not hold universally [6] and are greatly complicated by missingness, including both partially or completely unobserved nodes.

A natural, naïve approach involves simple mean-imputation for completely missing nodes, advantageous for its simplicity but almost certainly expected to be inferior to more sophisticated methods. The MissGLasso method was proposed as a natural extension to GLasso for missing data, with assumptions of missing at random (MAR) data [8]. A further extension proposed a method for Gaussiain graphical modelling with partially missing data with stronger theoretical guarantees on graph recovery, but requiring the assumption of missign completely at random (MCAR) data for theoretical guaranteess (although performance in simulations was shown to be preferable under the looser, MAR assumption) [5].

In their 2022 pre-print *Graphical Model Inference with Erosely Measured Data*, Lili Zheng and Genevera Allen propose a method beyond these restrictive assumptions on missingness, where their method is proposed to be consistent and adequately powered for data with intermittent and "drastically different" missingness (the authors' *erose* definition) throughout a graph [10].

2 Applications & Software

Current implementations of the Graphical lasso procedure are available in R. We currently plan to focus on implementation of the "typical" graphical lasso using the `glasso` R package [4] and the `mglasso()` function from the `classo` package [1] as an implementation of (Städler and Buühlmann)'s presented method for precision matrix estimation with missingness. Zheng and Allen provide an example implementation of their GI-JOE method which we aim to use/adapt [9]. We aim to compare use of the GI-JOE method to other, existing methods in the graphical modelling literature beyond those simulations provided in Zheng & Allen [10] using both simulated data and real data as provided by the authors in their simulations [9]. Of particular interest is the relative performance

of GI-JOE in non-erose settings (a sense of "robustness" for this new method compared to earlier developments specific to non-erose, MAR settings).

From our the discussion above, one can see that much of the existing literature and (implemented) methodology focuses on Guassian graphical models. There remains interest in extensions to graphical models of non-Gaussian and mixed (exponential family) distributions (see Introduction of Chen et al. for brief overview of applications in non-Guassian extensions for graphical models)[2]. Time and space pending, we hope to also explore the barriers in extending these methods for missing data to more generalized (i.e. non-Gaussian) graphical models. Lastly, again time-space pending, we hope to qualitatively compare methods in graphical modelling to non-graphical approaches in similar, multivariate data analysis problems with (both erose and non-erose) missingness.

3 References

References

[1] Luigi Augugliaro et al. *cglasso: Conditional Graphical LASSO for Gaussian Graphical Models with Censored and Missing Values*. Version 2.0.6. Jan. 17, 2023.

[2] Shizhe Chen, Daniela M. Witten, and Ali Shojaie. "Selection and estimation for mixed graphical models". In: *Biometrika* 102.1 (Mar. 1, 2015), pp. 47–64.

[3] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. "Sparse inverse covariance estimation with the graphical lasso". In: *Biostatistics* 9.3 (July 1, 2008), pp. 432–441.

[4] Jerome Friedman and Trevor Hastie {and} Rob Tibshirani. *glasso: Graphical Lasso: Estimation of Gaussian Graphical Models*. Version 1.11. Oct. 1, 2019.

[5] Mladen Kolar and Eric P Xing. "Estimating Sparse Precision Matrices from Data with Missing Values". In: *Proceedings of the 29 th International Conference on Machine Learning* (2012).

[6] Nicolai Meinshausen. "A note on the Lasso for Gaussian graphical model selection". In: *Statistics & Probability Letters* 78.7 (May 1, 2008), pp. 880–884.

[7] Nicolai Meinshausen and Peter Bühlmann. "High-dimensional graphs and variable selection with the Lasso". In: *The Annals of Statistics* 34.3 (June 2006). Publisher: Institute of Mathematical Statistics, pp. 1436–1462.

[8] Nicolas Städler and Peter Bühlmann. "Missing values: sparse inverse covariance estimation and an extension to sparse regression". In: *Statistics and Computing* 22.1 (2010), pp. 219–235.

[9] Lili Zheng. *GI-JOE: Graph Inference when Joint Observations are Erose*. Mar. 31, 2023.

[10] Lili Zheng and Genevera I. Allen. *Graphical Model Inference with Erosely Measured Data*. May 14, 2023. arXiv: 2210.11625[math,stat].