Proceedings 1st International Workshop on Advanced Analytics and Learning on Temporal Data
AALTD 2015

Estimating Dynamic Graphical Models from Multivariate Time-Series Data

Alexander J. Gibberd and James D.B. Nelson
Department of Statistical Science, University College London, Gower Street, London, WC1E 6BT

Abstract. We consider the problem of estimating dynamic graphical models that describe the time-evolving conditional dependency structure between a set of data-streams. The bulk of work in such graphical structure learning problems has focused in the stationary i.i.d setting. However, when one introduces dynamics to such models we are forced to make additional assumptions about how the estimated distributions may vary over time. In order to examine the effect of such assumptions we introduce two regularisation schemes that encourage piecewise constant structure within Gaussian graphical models. This article reviews previous work in the field and gives an introduction to our current research.

1 Introduction

As the current data explosion continues, governments, business, and academia are now not only harvesting more data points but also measuring an ever-increasing number of variables. The complex systems represented by such datasets arise in many socio-scientific domains, such as: cyber-security, neurology, genetics and economics. In order to understand such systems, we must focus our analytic and experimental resources on investigating the most important relationships. However, searching for significant relationships between variables is a complex task. The number of possible graphs that encode such dependencies between variables becomes exponentially large as the number of variables increase. Such computational issues are only compounded when such graphs vary over time.

From a statistical estimation viewpoint, the significance of a model component can often be viewed in terms of a model selection problem. Generally, one may construct an estimate of model fit (a lower score implies better fit) $L(M, \boldsymbol{\theta}, \boldsymbol{Y})$, relating a given model $M \in \mathcal{M}$ and parameters $\boldsymbol{\theta} \in \Theta(M)$ to some observed data $\boldsymbol{Y} \in \Omega$. Additionally, to account for differences in perceived model complexity one should penalise this by a measure of complexity $R(M, \boldsymbol{\theta})$ (larger is more *complex*). An optimal model and identification of parameters can be found through balancing the two terms, i.e:

$$(\hat{M}, \hat{\boldsymbol{\theta}}) = \underset{M \in \mathcal{M}, \boldsymbol{\theta} \in \Theta(M)}{\arg\min} \left[L(M, \boldsymbol{\theta}, \boldsymbol{Y}) + R(M, \boldsymbol{\theta}) \right]. \tag{1}$$

In statistics such a formulation is referred to as an M-estimator [15], however such frameworks are popular across all walks of science [2], for example, maximum-likelihood (ML), least-squares, robust (Huber loss), penalised ML estimators can

all be discussed in this context. The principle idea is to suggest a mathematical (and therefore can be communicated objectively) statement to the effect of Occam's Razor, whereby given similar model-fit, one should prefer the simpler model. Depending on the specification of the functions $L(\cdot)$ and $R(\cdot)$ and associated model/parameter spaces, the problem in (1) can be either very easy or difficult (for example, are the functions smooth, convex, etc).

In the next section we introduce the canonical *Gaussian graphical model (GGM)*, and study the estimation of such models within the M-estimation framework. This lays the foundations for our proposed dynamical extensions. We conclude with an example of an estimated dynamic GGM, some recovery properties of our estimators and discuss future research directions.

2 Gaussian graphical models

A Gaussian graphical model is a generative model which encodes the conditional dependency structure between a set of P variables $(Y_{1,\dots}Y_P) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ as a graph $G(V, E)$. For now we will discuss the traditional i.i.d setting, in Section (4) we will demonstrate ways in which we may relax the assumption of the distribution being identical over time.

In the standard case, the vertex set $V = \{1, \dots, P\}$ identifies variables and the edge set $E = \{(i, j), \dots (l, m)\}$ contains an edge if variables are conditionally dependent, specifically if $(i, j) \notin E$ we can decompose a joint distribution as $P(Y_i, Y_j | Y_{V\setminus\{i,j\}}) = P(Y_i | Y_{V\setminus\{i,j\}}) P(Y_j | Y_{V\setminus\{i,j\}})$. The aim of our work is to estimate an edge-set that appropriately represents a given data-set. Within the GGM setting, learning such representations does not only provide insight by suggesting key dependencies, but also specifies a robust probabilistic model which we can use for tasks such as anomaly detection.

It is well known that the edges in a GGM are encoded by non-zero off-diagonal entries within the precision matrix $\boldsymbol{\Theta} := \boldsymbol{\Sigma}^{-1}$, specifically $(i, j) \in E \iff \Theta_{i,j} \neq 0$ (see [12] for details). Learning the structure within the GGM can then be linked with the general framework of (1) through a ML or Maximum a-posteriori (MAP) paradigm. Assuming T observations $\boldsymbol{Y} \in \mathbb{R}^{P \times T}$ drawn as i.i.d samples the model fit function $L(\cdot)$ can be related to the likelihood specified by the multivariate normal. Typically, one prefers to work with the log-likelihood, which if we assume $\boldsymbol{\mu} = \mathbf{0}$ (we assume this throughout) is given by:

$$\log(P(\boldsymbol{Y}|\boldsymbol{\Theta}))/T = \frac{1}{2}\log\det(\boldsymbol{\Theta}) - \frac{1}{2}\mathrm{trace}(\hat{\boldsymbol{S}}\boldsymbol{\Theta}) - \frac{P}{2}\ln(\pi) \ ,$$

where $\hat{\boldsymbol{S}} = \boldsymbol{Y}\boldsymbol{Y}^\top/T$. Taking $L(\cdot) = -\log\det(\boldsymbol{\Theta}) + \mathrm{trace}(\hat{\boldsymbol{S}}\boldsymbol{\Theta})$ gives (in the setting where $T > P$) a well-behaved smooth, convex function describing how well a given parameterisation $\boldsymbol{\Sigma}$ represents the data $\boldsymbol{Y}$.

3 Penalising complexity

If one considers Eq. (1) with the function $R(\cdot) = 0$, i.e. no complexity penalty, then the precision matrix estimator $\hat{\boldsymbol{\Theta}} := \arg\min_{\{\boldsymbol{\Theta} \succeq 0\} \in \mathbb{R}^{P \times P}} \big[-\log\det(\boldsymbol{\Theta}) + \text{trace}(\hat{\boldsymbol{S}}\boldsymbol{\Theta}) \big]$ demonstrates some undesirable properties indicative of over-fitting:

- The estimator exhibits large variance when $T \approx P$ and is very sensitive to changes in observations leading to poor generalisation performance.
- In the high-dimensional setting ($P > T$), the sample estimator is rank deficient ($\text{rank}(\hat{\boldsymbol{S}}) < P$) and there is no unique estimator $\hat{\boldsymbol{\Theta}}$.

In order to avoid estimating a complete GGM graph (where all vertices's are connected to each other), one must actively select edges according to some criteria. In the asymptotic setting where $T \gg P$ we can test for the significance of edges by considering the asymptotic distribution of the empirical partial correlation coefficients ($\rho_{ij} = -\Theta_{ij}/\Theta_{ii}^{1/2}\Theta_{jj}^{1/2}$) [4]. However, such a procedure cannot be performed in the high-dimensional setting (this is important for the dynamical extensions, see Sec. 4) as we require that the empirical estimate be positive semi-definite.

An alternative approach to testing is to consider prior knowledge about the number of edges in the graph. If we assume a flat prior on the model $\mathcal{M}$ and parameters $\boldsymbol{\Theta}(\mathcal{M})$, maximising the approximate posterior probability over models $P(\mathcal{M}|\boldsymbol{Y})$, then leads to the Bayesian information criterion for GGM [5]: $BIC(\hat{\boldsymbol{\Theta}}_{ML}) = N(-\log\det(\hat{\boldsymbol{\Theta}}_{ML}) + \text{trace}(\hat{\boldsymbol{S}}\hat{\boldsymbol{\Theta}}_{ML})) + \hat{p}\log(N)$, where $\hat{p}$ is given by the number of unique non-zeros within the ML estimated precision matrix $\hat{\boldsymbol{\Theta}}_{ML}$. Unfortunately, interpreting BIC under the framework in Eq. (1), we find the complexity penalty $R() = \hat{p}\log(N)$ is non-convex ($\hat{p} \propto \|\boldsymbol{\Theta}\|_0$ it basically counts the number of estimated edges). In order to arrive at a global minima an exhaustive search over the model space (all possible graphs $\mathcal{O}(2^{P^2})$) is required.

Alternatively, one can place an informative prior on the parameterisation and model (i.e. the GGM sparsity pattern) to encourage a parsimonious representation. One popular approach [6,11,17,20] is to place a Laplace type prior on the precision matrix in an effort to directly shrink off-diagonal values. Whilst one could choose to perform full Bayesian inference for the posterior $P(\boldsymbol{\Theta}|\boldsymbol{Y}, \gamma)$ (as demonstrated in [17]), a computationally less demanding approach is to perform MAP estimation resulting in the *graphical lasso* problem [6]:

$$\hat{\boldsymbol{\Theta}}_{GL} := \arg\min_{\boldsymbol{\Theta} \succ 0} \big[-\log\det(\boldsymbol{\Theta}) + \text{trace}(\hat{\boldsymbol{S}}\boldsymbol{\Theta}) + (\gamma/N)\|\boldsymbol{\Theta}\|_1 \big] , \tag{2}$$

where $\|\boldsymbol{\Theta}\|_1 = \sum_{1 \leq i,j \leq P} |\Theta_{i,j}|$ is the ℓ_1 norm of $\boldsymbol{\Theta}$. The graphical lasso problem can yet again be interpreted within the general framework, except this time with $R(\cdot) = (\gamma/N)\|\boldsymbol{\Theta}\|_1$. Unlike BIC this complexity penalty is convex thus we can quickly find a global minima.

4 Introducing dynamics

In this section we extend the basic GGM model to a dynamic setting whereby the estimated graph is permitted to change as a function of time. Consider the P-variate time-series data $\boldsymbol{Y} \in \mathbb{R}^{P \times T}$ as before, however, we now permit the generative distribution to be a function of time, i.e:

$$(Y_1^t, \dots Y_P^t) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}^t) \,, \tag{3}$$

the challenge is now to learn a GGM via $(\boldsymbol{\Sigma}^t)^{-1}$ for each time point $t = 1, \dots, T$. Clearly such a model is far more flexible than the identically distributed version, instead of $\mathcal{O}(P^2)$ parameters we now have $\mathcal{O}(P^2 T)$. In such a semi-parametric model the potential complexity can scale with the amount of data we have available. Our aim is to harness this additional flexibility to identify potential changes within the graphical models which may shed insight onto dynamics of the data-generating system.

Local kernel/window estimation

Zhou et. al. [20] consider the dynamic GGM model in a continuous setting such that the underlying graphs are assumed to vary smoothly as a function of time. To provide a local estimate of the covariance they suggest the estimator $\hat{\boldsymbol{S}}(t) = \sum_s w_{st} \boldsymbol{y}_s \boldsymbol{y}_s^\top / \sum_s w_{st}$, where $w_{st} = K(|s-t|/h_T)$ are weights derived from a symmetric non-negative kernel (typically one may use a box-car/Gaussian function) with bandwidth h_T. The idea is that by replacing $\hat{\boldsymbol{S}}$ with $\hat{\boldsymbol{S}}(t)$ in the graphical lasso problem (Eq. 2) it is possible to obtain a temporally localized estimate of the graph $\hat{\boldsymbol{\Theta}}(t)_{GL}$. Given some smoothness conditions on the true covariance matrices one can demonstrate [20] that the estimator is consistent (estimator risk converges in probability $R(\hat{\boldsymbol{\Sigma}}(t)) - R(\boldsymbol{\Sigma}^*(t)) \xrightarrow{P} 0$) even in the dynamic (non-identically distributed) case.

Piecewise constant GGM

The seminal work by Zhou et al. [20] focused in the setting where graphs continuously and smoothly evolve over time. However, there are many situations where we might expect the smoothness assumptions to be broken. Our research [7,8,9] focuses on how we can incorporate different smoothness assumptions when estimating dynamic GGM. In particular we wish to study piecewise constant GGM where the generative distribution is strictly stationary within regions separated by a set of changepoints $\mathcal{T} = \{\tau_1, \dots, \tau_K\}$, $\tau_i \in \{1, \dots, T\}$, such that:

$$P(Y^t) = P(Y^{t+i}) \; \forall \, t, (t+i) \in \{\tau_k, \dots \tau_{k+1}\} \text{ for } k = 0, \dots, K-1 \,.$$

If we keep the Gaussian assumption of Eq. (3), then estimation relates to finding a set of $K-1$ GGM describing the distribution between changepoints. Such a definition extends the usual definition of a changepoint [13] to multivariate distributions, it is expected that the number of changepoints should be small relative to the total period of measurement, i.e. $K \ll T$ and that such points may lead to insight about changes within observed systems.

5 Structure learning with dynamic GGM

Our approach to searching for changepoints falls naturally into the M-estimation framework of Eq. (1). As has already been discussed, appropriate complexity penalties $R(\cdot)$ may act to induce sparsity in a given set of parameters. We propose two sparsity aware estimators that use such properties not only to estimate the graphical model, but also jointly extract a sparse set of changepoints.

Independent Fusing

Our first approach (see [7,9], also related to [14,3,18]) constructs a model fit function $L(\boldsymbol{\Theta}, \boldsymbol{Y}) = \sum_{t=1}^{T} \big(-\operatorname{logdet}(\boldsymbol{\Theta}^t) + \operatorname{tr}(\hat{\boldsymbol{S}}^t \boldsymbol{\Theta}^t) \big)$, where $\hat{\boldsymbol{S}}^t = \boldsymbol{y}^t (\boldsymbol{y}^t)^\top / 2$ is an estimate of the covariance for a specific time t. Clearly, there is not enough information within $\hat{\boldsymbol{S}}^t$ to recover a graph, as we are effectively trying to estimate with only one data point. To solve this problem we introduce an explicit prior on the smoothness of the graph via a complexity function

$$R_{IFGL}(\boldsymbol{\Theta}) = \lambda_1 \sum_{t=1}^{T} \|\boldsymbol{\Theta}^t\|_1 + \lambda_2 \sum_{t=2}^{T} \|\boldsymbol{\Theta}^t - \boldsymbol{\Theta}^{t-1}\|_1 \ , \tag{4}$$

where λ_1, λ_2 control the level of sparsity and number of changepoints in the model. Unlike in the work of Zhou et al. our prior encodes an assumption that the model has a piecewise constant parameterisation (this is similar to the fused lasso, see [16,10]). We refer to the problem $\{\hat{\boldsymbol{\Theta}}\}_{t=1}^{T} = \arg\min_{\boldsymbol{\Theta} \succeq 0} \big[L(\cdot) + R_{IFGL}(\cdot) \big]$ as defined above, as the *independently fused graphical lasso (IFGL)*, it estimates changepoints at an individual edge level such that changepoints do not necessarily coincide between edges.

Group Fusing

Sometimes we have a-priori knowledge that particular variables may change in a grouped manner, that is changepoints across the edges which connect variables may coincide. Examples, might include genes associated with a specific biological function (see the example in Fig. 1), or stocks within a given asset class. In order to encode such prior structure for changepoints one can adapt the smoothing prior to act over a group of edges, for such cases we suggest the *group-fused graphical lasso (GFGL)* penalty[9]:

$$R_{GFGL}(\boldsymbol{\Theta}) = \lambda_1 \sum_{t=1}^{T} \|\boldsymbol{\Theta}^t\|_1 + \lambda_2 \sum_{t=2}^{T} \|\boldsymbol{\Theta}^t - \boldsymbol{\Theta}^{t-1}\|_2 \ . \tag{5}$$

Optimisation

Both IFGL and GFGL form non-smooth convex optimisation problems which can be tackled within a variety of optimisation schemes. We have developed

an alternating direction method of multipliers (ADMM) algorithm that efficiently solves both the above problems by taking advantage of subtle separability properties of the estimators. Typically, one can solve for several changepoints $K = 1, \ldots, \sim 10$ on problems of size $T \approx 100 - 1000$, $P \approx 10 - 100$ in a few minutes. Due to the convex formulation scaling is linear in time $\mathcal{O}(TP^3K^2)$ for GFGL, which is a considerable advantage when compared to the quadratic time complexity of dynamic programming approaches [1].

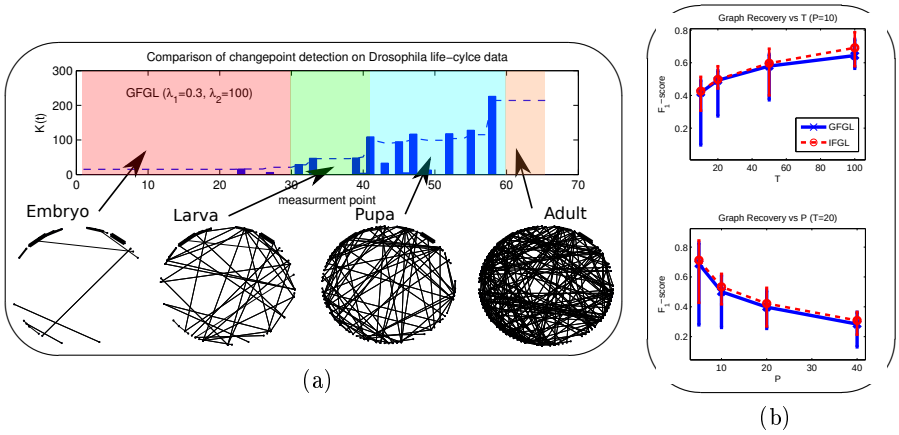

Fig. 1: a) Example of estimated graphical models (using GFGL) describing gene dependency through the life-cycle of *Drosophila melanogaster* (the common fruit fly). The generative distribution is assumed to be stationary between the blue bars which indicate changepoints. The dashed line indicates the number of active edges in the recovered graph. b) Results from an empirical study [9] considering how recovery of the graph changes with problem size.

6 Conclusion

To date, we have examined some properties of the IFGL and GFGL estimators in an empirical setting (see Fig. 1). Through the use of a wavelet framework our work [8] has also considered how one could allow for trends and changes in the mean parameter for dynamic GGM. Empirical results suggest some desirable properties for the proposed estimators (graph recovery improves when one increases the size and amount of data available within the stationary segments, see Fig. (1b), however, we have yet to examine the theoretical consistency properties. Theoretical analysis is complicated by the fact we regularise in multiple directions (the graph and over time), it is possible some insight in this direction can be gained from results in the regression setting [19].

References

1. D. Angelosante and G. B. Giannakis. Sparse graphical modeling of piecewise-stationary time series. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011.
2. S. Boyd and L. Vandenberghe. *Convex Optimization*. 2004.
3. P. Danaher, P. Wang, and D. M. Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2013.
4. M. Drton and M. D. Perlman. Model selection for Gaussian concentration graphs. *Biometrika*, 2004.
5. R. Foygel and M. Drton. Extended Bayesian information criteria for gaussian graphical models. In J.D. Lafferty, C.K.I. Williams, J. Shawe-Taylor, R.S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*. 2010.
6. J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics (Oxford, England)*, 2008.
7. A. J. Gibberd and J. D. B. Nelson. High dimensional changepoint detection with a dynamic graphical lasso. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.
8. A. J. Gibberd and J. D. B. Nelson. Estimating multi-resolution dependency graphs within a locally stationary wavelet framework. *In review*, 2015.
9. A. J. Gibberd and J. D. B. Nelson. Regularized Estimation of Piecewise Constant Gaussian Graphical Models: The Group-Fused Graphical Lasso. *In review*, 2015.
10. Z. Harchaoui and C. Lévy-Leduc. Multiple change-point estimation with a total variation penalty. *Journal of the American Statistical Association*, 2010.
11. J. Lafferty, H. Liu, and L. Wasserman. Sparse nonparametric graphical models. *Statistical Science*, 2012.
12. S. L. Lauritzen. *Graphical models*. Oxford, 1996.
13. M. A. Little and N. S. Jones. Generalized methods and solvers for noise removal from piecewise constant signals. II. New methods. *Proceedings. Mathematical, physical, and engineering sciences / the Royal Society*, 2011.
14. R. P. Monti, P. Hellyer, D. Sharp, R. Leech, C. Anagnostopoulos, and G. Montana. Estimating time-varying brain connectivity networks from functional MRI time series. *NeuroImage*, 2014.
15. S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 2012.
16. R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2005.
17. H. Wang. Bayesian graphical lasso models and efficient posterior computation. *Bayesian Analysis*, 2012.
18. S. Yang, Z. Pan, X. Shen, P. Wonka, and J. Ye. Fused multiple graphical lasso. *Arxiv*, 2012.
19. B. Zhang, J. Geng, and L. Lai. Multiple change-points estimation in linear regression models via sparse group lasso. *IEEE Trans. Signal Processing*, 2015.
20. S. Zhou, J. Lafferty, and L. Wasserman. Time varying undirected graphs. *Machine Learning*, 2010.