HIGH DIMENSIONAL CHANGEPOINT DETECTION WITH A DYNAMIC GRAPHICAL LASSO

Alexander J. Gibberd and James D. B. Nelson

Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, UK

ABSTRACT

The use of sparsity to encourage parsimony in graphical models continues to attract much attention at the interface between multivariate Signal Processing and Statistics. We propose and investigate two approaches for the detection of changepoints in the correlation structure of evolving Gaussian graphical models. Both approaches employ two-stages; first estimating the dynamic graphical structure through regularising the precision matrix, before changepoints are selected via a group fused lasso. Experiments on simulated data illustrate the efficacy of the two approaches. Furthermore, results on real internet traffic flow data containing a Denial Of Service attack demonstrate that the proposed approaches have potential utility in information forensics and security.

Index Terms— Graphical models, Time-varying systems, Intrusion detection, Statistical learning, System identification

1. INTRODUCTION

The detection of changepoints in data where the distributional properties change abruptly is of interest across a wide variety of domains, such as network security [1, 2, 3], neurology [4] and seismic analysis [5].

As we gather more data, with greater numbers of variables, we gain the power to estimate increasingly complex relationships within our observations. This is particularly evident in the analysis of network traffic data, where transfer rates are increasing alongside the number of devices participating in these networks. Cyber-attacks and espionage are a growing concern for governments, businesses and consumers who depend on these networks. Current intrusion detection systems (IDS) used to detect cyber-threats are primarily supervised in nature, driven by hard-coded rule-sets and heuristic analysis of packet data [6]. These are increasingly being bypassed by so-called zero-day attacks which evolve to counter the very specific rules in place. We present a more data driven approach to the problem and consider unsupervised learning to detect anomalous regions and relationships that may be indicative of attacks.

Early changepoint research focused on univariate signals [7], however, much recent work has focused on expanding this into high-dimensional settings where the number of variables p is comparable or greater than the amount of data points T [8, 9, 3]. In this $p \geq T$ regime it is often hard to interpret the dependency structure between variables due to the large number of possible interactions and the parameters required to describe these.

Utilising the parameter shrinking property of ℓ_1-based regularisers, much recent work has looked at how to recover sparse graphical models [10, 11], more recently, still via time-evolving dynamic graphical models [12, 13]. In addition to work in structural estimation, fast $\mathcal{O}(T \log T)$ methods have been developed to infer changepoints by optimising a regularised likelihood function and fitting sparse piecewise approximations to signals [9, 14].

This paper aims to combine the above state-of-the-art convex regularisation approaches, detecting both correlation structure and changepoints in network traffic. We introduce a novel two-stage methodology that provides comparable computational performance with current dynamic programming methods that require computational cost $\mathcal{O}(T^2p^{4.5})$ [15]. This formulation lays the foundation for the development of future methods, which take full advantage of the convexity, potentially reducing computational cost from polynomial to linear complexity.

2. PRELIMINARIES - DYNAMIC GAUSSIAN GRAPHICAL MODELS

The methodology proposed here aims to extract piecewise edge structure within a time dependent undirected Gaussian graphical model. We focus on modeling continuous multivariate time-series data $\mathbf{y}_t \sim \mathcal{N}(\boldsymbol{\mu}_t, (\boldsymbol{\Theta}^t)^{-1})$ $t = 1, \ldots, T$, where $\mathbf{y}_t \in \mathbb{R}^{p\times 1}$, $\boldsymbol{\mu}_t \in \mathbb{R}^{p\times 1}$ is the mean and $\boldsymbol{\Theta}^t = (\theta^t_{i,j}) \in \mathbb{R}^{p\times p}$ is the precision matrix at time t. This matrix encodes the edge structure of the graph G_t, where zero entries $\theta^t_{i,j} = 0$ denote conditional independence between variables i and j [16].

When deriving a maximum likelihood estimator for the precision matrix in the static case, we usually estimate the

Alex Gibberd acknowledges support from Defence Science Technology Laboratory (Dstl).

James Nelson acknowledges support from Technology Strategy Board and an EPSRC D2U grant.

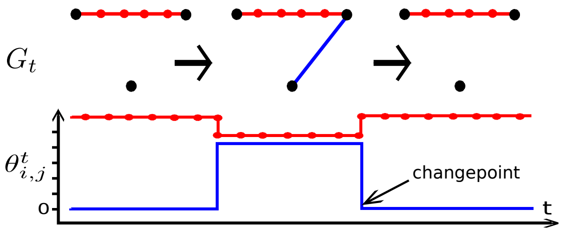

Fig. 1. We assume that the graph G_t is piecewise constant over time with the time points where edge values jump across the graph denoting changepoints. This piecewise model is perhaps the simplest way for us to allow for dynamics and relax the assumption of stationarity.

covariance using the sample mean $\bar{\boldsymbol{\mu}} = T^{-1}\sum_{t=1}^{T}\mathbf{y}_t$, such that: $\hat{\boldsymbol{\Theta}} = \arg\max_{\boldsymbol{X}\succeq 0}\{\mathcal{L}(\boldsymbol{X},\boldsymbol{S})\}$, where $\mathcal{L}(\boldsymbol{X},\boldsymbol{S}) := \log(\det(\boldsymbol{X})) - \mathrm{tr}(\boldsymbol{S}\boldsymbol{X})$ and $\boldsymbol{S} = T^{-1}\sum_{t=1}^{T}(\mathbf{y}_t - \bar{\boldsymbol{\mu}})(\mathbf{y}_t - \bar{\boldsymbol{\mu}})^\top$. However, in a non-stationary case, this approach with a static sample mean $\bar{\boldsymbol{\mu}}$ is no longer valid. If trends are present in the data, estimation of covariance $\mathbb{E}[(y_{i,t}-\mu_{i,t})(y_{j,t}-\mu_{j,t})]$ becomes difficult because, for example, the mean varies. As such, the sample mean (over time) no longer converges to the expectation at a given instance in time $T^{-1}\sum_{t=1}^{T} y_{i,t} \neq \mathbb{E}[y_{i,t}]$. If care is not taken, variation in the mean $\mu_{i,t}$ (or trends) in the data will give rise to erroneous components in the correlation structure, which is usually interpreted as a contemporaneous functional relationship between variables [17]. Throughout the rest of this paper we will assume data has been suitably de-trended. We are now only interested in modeling the Gaussian noise structure $\mathbf{y}_t \sim \mathcal{N}(\mathbf{0}, (\boldsymbol{\Theta}^t)^{-1})$.

The goal of detecting changepoints in GGM therefore boils down to enforcing sparsity on the precision matrix, whilst also encouraging piecewise constant edge structure in time with $K \ll T$ jump points.

3. GRAPHICAL CHANGEPOINT DETECTION

It is well known that addition of a regularisation factor constructed from an ℓ_1 norm can induce sparsity in the parameters of a linear regression problem [18]. More recently this property has been used to encourage sparsity in the estimation for GGM, resulting in the estimator: $\hat{\boldsymbol{\Theta}} = \arg\max_{\boldsymbol{X}\succeq 0}\{\mathcal{L}(\boldsymbol{X},\boldsymbol{S}) - \lambda_G\|\boldsymbol{X}\|_1\}$, where $\|\boldsymbol{X}\|_1 = \sum_{i,j}|x_{i,j}|$ is the element-wise matrix ℓ_1 norm [10, 19]. We consider the time-evolving variant of this model first proposed by Zhou et al. [12] which uses a weighted moving window to track changes in the graphical structure. The resultant estimator is now for a time dependent graph:

$$\hat{\boldsymbol{\Theta}}^t := \underset{\boldsymbol{X}^t\in\{\boldsymbol{X}^t_+\}_{t=T_s}^{T_e}}{\arg\max}\ \left\{\mathcal{L}(\boldsymbol{X}^t,\boldsymbol{S}^t) - \lambda_G\|\boldsymbol{X}^t\|_1\right\}, \qquad (1)$$

where we refer to the set of positive semi-definite matrices from time $t = T_s,\ldots,T_e$ as $\{\boldsymbol{X}^t_+\}_{t=T_s}^{T_e}$. To take into account the cropping due to the window of size w we set $T_s = \frac{w}{2}+1$, $T_e = T - \frac{w}{2}$. We now introduce a temporally local empirical covariance estimator:

$$\boldsymbol{S}^t = \Big(\sum_s f(s)\Big)^{-1}\sum_{k=t-w/2}^{t+w/2} f(k-t+w/2)\boldsymbol{y}^k(\boldsymbol{y}^k)^\top,$$

where $f(s)$ is a smooth positive kernel function defined over the window $s = 0,\ldots,w$. Whilst Zhou et al. consider the graph structure to be a function of time, they also consider the estimates of the graphs $\hat{\boldsymbol{\Theta}}^t$ to be independent (or at least assume no prior on this dependency/smoothness). Maintaining this independence in estimation comes at a cost due to the fact that we are only using a small portion of data (that within the window) to estimate the graphical structure.

3.1. Graph Dependency

In an effort to increase the accuracy of structural estimation and take into account dependency in the time series we now consider optimizing jointly across the data set, combining a smoothing and graphical regulariser. With this approach we construct the joint estimator for $\hat{\boldsymbol{\Theta}}^t$ making use of the dependency structure that we presume to exist between graphs. Reformulating and adapting the work of Danaher et al. [20] (and Ahmed et al. in the discrete case [13]) we propose the following joint graphical lasso (JGL) scheme:

$$\begin{aligned}\hat{\boldsymbol{\Theta}}^t : \;=\; & \underset{\boldsymbol{X}^t\in\{\boldsymbol{X}^t_+\}_{t=T_s}^{T_e}}{\arg\max}\sum_{t=T_s}^{T_e}\Big(\underbrace{\mathcal{L}(\boldsymbol{X}^t,\boldsymbol{S}^t) - \lambda_G\|\boldsymbol{X}^t\|_1}_{\text{independent estimate}}\Big)\\ & -\lambda_\nabla\underbrace{\sum_{t=T_s+1}^{T_e}\|\boldsymbol{X}^t - \boldsymbol{X}^{t-1}\|_1}_{\text{fused }\ell_1\text{ penalty}}. \qquad (2)\end{aligned}$$

The motivation for the above is to extend the fused lasso of Tibshirani [21] and apply it in the estimation of the precision matrix. In the time series case where we have meaningful ordering of the data, this fused ℓ_1 smoothing is consistent with attempting to enforce local temporal dependency in the graph structure. To optimize the objective in (2) we use an accelerated coordinate descent scheme based on Nesterov's method [22, 23].

3.2. Two-step regularisation

When estimating changepoints, we would like to highlight where the dependency of the graph changes taking into account all edges. We consider an approach to achieve this through further regularising the solution of Eqs(1,2).

Taking these estimated graphs we encourage a piecewise solution by smoothing $\hat{\boldsymbol{\Theta}}^t$ with a group lasso penalty combined with a least squares estimator. More specifically, we reparameterise the changepoint estimation problem according to the work of Bleakley et al. [9] who start with a multivariate (dimension q) piecewise model $\boldsymbol{U} \in \mathbb{R}^{T\times q}$ such that $u_{t,\cdot} = \boldsymbol{\gamma} + \sum_{i=1}^{t-1} d_i\phi_{i,\cdot}$ (note: we use $\phi_{i,\cdot}$ to denote all elements in row i of ϕ) for $t = 2,\ldots,T$, where $\boldsymbol{\gamma} = u_{1,\cdot}$ are the initial values of the time series and $\phi_{i,\cdot} = (u_{i+1,\cdot} - u_{i,\cdot})/d_i$ is a reparameterisation of the jumps in the model. In order to avoid boundary effects we weight the jumps setting $d_i = \sqrt{n/i(n-i)}$. This linear model can then be written in matrix form as $\boldsymbol{U} = \mathbf{1}_T\boldsymbol{\gamma} + \boldsymbol{X}\boldsymbol{\Phi}$, where $\boldsymbol{X} \in \mathbb{R}^{T\times(T-1)}$ is a lower diagonal design matrix with entries $x_{i,j} = d_j$ for $i > j$, $\boldsymbol{\Phi} = (\phi_{i,j}) \in \mathbb{R}^{(T-1)\times q}$ is the jump parameter matrix and $\mathbf{1}_T$ is a one vector of length T. Given this piecewise model a multivariate total variation smoothed least squares estimator takes the form of a group lasso problem:

$$\hat{\boldsymbol{\Phi}} := \underset{\boldsymbol{\Phi}\in\mathbb{R}^{(T-1)\times q}}{\arg\max} \; -\frac{1}{2}\|\bar{\boldsymbol{Y}} - \bar{\boldsymbol{X}}\boldsymbol{\Phi}\|_2^2 - \lambda\sum_{t=1}^{T-1}\|\phi_{t,\cdot}\|_2, \tag{3}$$

where $\bar{\boldsymbol{X}}, \bar{\boldsymbol{Y}}$ are the column centered design and data matrix $\boldsymbol{Y} = (y_{t,j}) \in \mathbb{R}^{T\times q}$ respectively [9]. Changepoints in this case are simply the time points where non-zero features are present $\hat{\mathcal{T}} = \{t \mid \phi_{t,i} \neq 0 \;\forall\; t = 1,..,T \;;\; i = 1,\ldots,q\}$. In this situation, the group lasso penalty (a sum over ℓ_2-norms) attempts to squeeze these features in the jump structure (columns within $\boldsymbol{\Phi}$) to zero [24]; increasing λ reduces the number changepoints. To apply this to the correlation structure estimated in Eqs(1,2) we rearrange the edge structure to form a new $p(p-1)/2$ dimensional time series. This is achieved by taking all upper triangular elements of $\hat{\boldsymbol{\Theta}}^t$ and concatenating them to form a vector for each point in time $\boldsymbol{z}^t := (\theta^t_{i,j} | \forall\; i = 1,\ldots,p-1,\; i < j \leq p)$ for $t = T_s,\ldots,T_e$ such that $\boldsymbol{Z} := (\boldsymbol{z}^t)_{t=T_s}^{T_e} \in \mathbb{R}^{(T-w)\times(p(p-1)/2)}$. Finally, to extract changepoints in the correlation structure we simply substitute $\boldsymbol{Z}$ for $\boldsymbol{Y}$ within Eq. (3) taking care to adjust the limits and sizes of the jump/design matrix such $T \rightarrow T - w$ and $q \rightarrow p(p-1)/2$.

The changepoint procedure described above is efficient due to its convexity with a theoretical complexity of $\mathcal{O}(KTq)$ for K jumps using an active-set strategy developed by Bleakley et al. [9]. Coupled with the required estimation for the graph at each point in time $\mathcal{O}(Tp^3)$ via (1) this gives us an overall complexity of $\sim \mathcal{O}(KT^2p^5)$.

4. EXPERIMENTS

4.1. Simulation

In order to examine the difference between independent and JGL graphical estimation we look at how well the methods recover graphical structure in some simulated cases. Data is simulated (without loss of generality) by randomly adding and subtracting four edges to an initial precision matrix at a changepoint $\tau = 300$, where the total length of the time series is $T = 600$. After checking that the resulting matrices are positive semi-definite we then generate multivariate Gaussian noise from this graph with $p = 10$ variables. For simplicity only a single ground truth changepoint is simulated here with $K^* = 1$, however there is no reason that one cannot perform this analysis for $K^* \geq 1$.

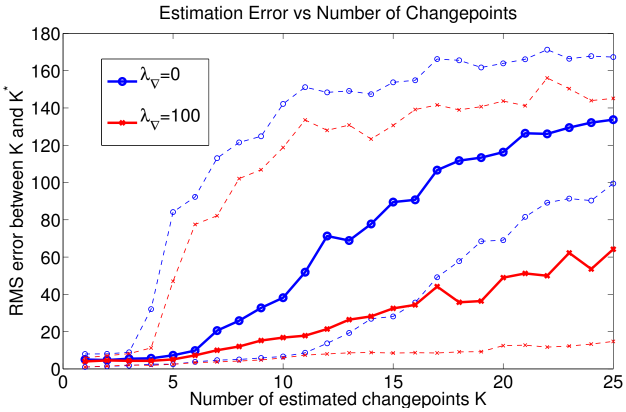

Fig. 2. RMS error as a function of number of K changepoints estimated. Dashed lines give 67% confidence intervals as approximated through the empirical cumulative distribution function over $n = 60$ experiments.

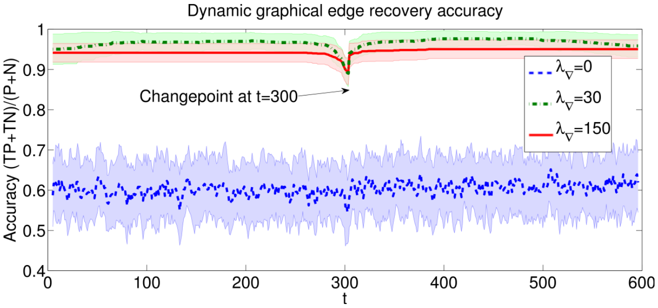

Fig. 3. The effect of smoothing via the JGL on edge recovery accuracy as calculated at each time step. Shaded areas give $\pm 1\sigma$ errors over $n = 60$ experiments.

Prior to changepoint estimation or smoothing we learn λ_G using a holdout data set. We choose λ_G such that when given no smoothing ($\lambda_\nabla = 0$) we recover a number of edges such that at least the ground truth edges are maintained in the solution. In the demonstrations, we fix $\lambda_G = 0.5$ and window width $w = 30$.

In Figure (2) we evaluate the performance of the two techniques by looking at the root mean square (RMS) error, $\sqrt{K^{-1}\sum_{i=1}^{K}(\tau - \hat{\tau}_i)^2}$, of the estimated changepoints $\hat{\tau}_i$ from the true changepoint τ. This plot demonstrates the benefit of smoothness when extracting $\hat{\boldsymbol{\Theta}}^t$ with the JGL where we see tighter clustering in the estimation of changepoints, this may be important for cases where $K^* > 1$.

The JGL estimation allows an extra degree of freedom (over the independent estimation of Eq. 1) which balances graph smoothness with sparsity. This is important, as applying large λ_∇ can overwhelm the sparsity enforcement of λ_G. This is such that some edges may be non-zero for the whole period $t = T_s, \ldots, T_e$, not just within their true classes. In selecting λ_∇, we need to balance changepoint detection accuracy in Fig. (2) against that of structural estimation within Fig. (3).

In Fig. (3) we note a region around the changepoint where more edges are estimated, this is due to the overlap of the kernel function with the two ground truth regimes either side of the changepoint and is particularly clear in the appropriately chosen $\lambda_\nabla = 30$ case. From comparing the analysis in both Figs. (2,3) we see that the JGL has an important role to play, increasing both changepoint detection accuracy and the stability of the edge structure beyond the independent graph estimation.

4.2. Changepoint detection within internet traffic

In this section, we give a demonstration of how the two-stage approach can be used to learn structure in network data and potentially identify cyber attacks. The first step is to construct a set of variables of interest. Ideally we want this set (of size p) to be as large as possible, whilst maintaining the normality assumptions required by likelihood function $\mathcal{L}(\boldsymbol{X}^t, \boldsymbol{S}^t)$. There are many variables we might consider aggregating against: source-destination, packet-type, protocol, packet size, etc. In this example, we use eight variables looking at both the protocol and size of packets. Variables are constructed by dis-aggregating packet flow at the edge of a local area network using the Wireshark packet inspection program. The data we use originates from attack simulations performed by DARPA to mimic cyber-attacks on a US airforce base [25]. We use data from the morning of Friday in week 7 which contains a so-called SYN-flood Denial of Service (DoS) attack.

Internet traffic is generally non-Gaussian as seen in Fig. (4), this may invalidate our methodology which is designed with Gaussian distributions in mind. To combat this we instead study the traffic as an auto-regressive process with a multivariate Gaussian noise term, such that $\boldsymbol{y}_t = \boldsymbol{y}_{t-1} + \boldsymbol{\epsilon}_t$ where $\boldsymbol{\epsilon}_t \sim (0, \mathcal{N}(\mathbf{0}, (\boldsymbol{\Theta}^t)^{-1})$.

In order to extract correlations within the data we first look to find the prominent correlation structure using the independent graph estimates of (1)—this allows us to set $\lambda_G = 0.1$. We then estimate the variation in this subset of edges via a smoothed solution using (2) with parameters $\lambda_\nabla = 1$, $\lambda_G = 5$ and $w = 10$ selected based on the timescale of typical DoS threats (>1 hour). Finally we estimate changepoints through the group lasso method of (3); we specify more changepoints than are known to exist allowing for false positives we set $K = 8$.

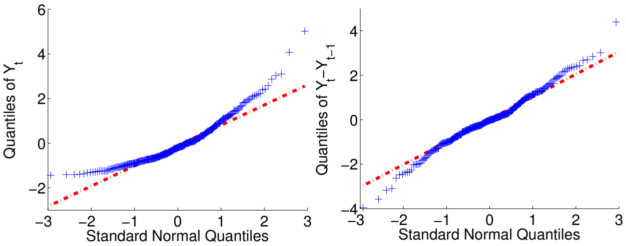

Fig. 4. QQ-Plots comparing best-fit Gaussian (red) and z-scored network traffic for packets between 90-120 bytes (Left) and differenced traffic (Right).

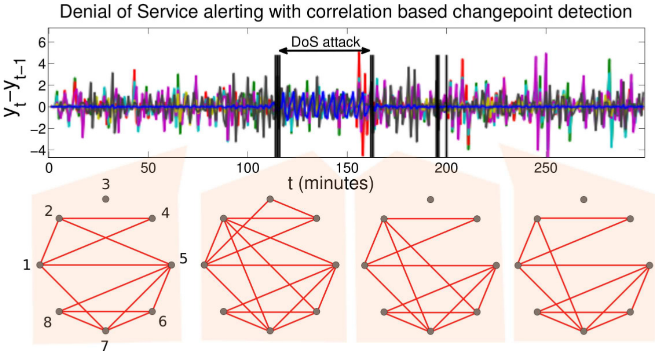

Fig. 5. Changepoints (vertical black lines) successfully identifying a DoS attack and recovered graph structure within network data. Variables identified in the graph are created according to packet protocol; 1 - (UDP), 2 - (IP), 3 - (HTTP), 4 - (SYN), and packet size (bytes); 5 - (<60), 6- (60-90), 7 - (90-120), 8 - (120-150).

5. CONCLUSION

We have presented a two-stage method for learning dynamic GGM alongside changepoints in the graphical structure. This method, based on two-stage regularisation has comparable computational efficiency with existing dynamic programming based methods. One computational limitation of the method is the need for a subsequent optimisation step (3). It may be possible to incorporate a fused group ℓ_2 penalty within (2) performing joint optimisation to enhance computational performance.

From an application point of view we have demonstrated how relational structures can be uncovered when modeling network traffic data which may have potential uses to build improved attack filters for IDS. However, there is not necessarily any clear reason that the largest changepoints we detect should be correlated with attacks; many cyber attacks do not even contribute to network traffic. Future work may look at learning an association between attacks and particular edges or recovered structure, this could act to reduce false positives, effectively using this work as an unsupervised feature extraction method.

6. REFERENCES

[1] Veronica Montes De Oca, Daniel R. Jeske, Qi Zhang, Carlos Rendon, and Mazda Marvasti, "A cusum change-point detection algorithm for non-stationary sequences with application to data network surveillance," *Journal of Systems and Software*, vol. 83, no. 7, pp. 1288–1297, July 2010.

[2] AS Polunchenko, "Nearly optimal change-point detection with an application to cybersecurity," *Sequential* . . . , vol. 1, no. x, pp. 1–23, 2012.

[3] Céline Lévy-Leduc and François Roueff, "Detection and localization of change-points in high-dimensional network traffic data," *The Annals of Applied Statistics*, vol. 3, no. 2, pp. 637–662, June 2009.

[4] P Xu, H Xu, and PJ Ramadge, "Detecting stimulus driven changes in functional brain connectivity," *Acoustics, Speech and Signal* . . . , pp. 3507–3511, 2013.

[5] Antonio Pievatolo and Renata Rotondi, "Analysing the interevent time distribution to identify seismicity phases: a Bayesian nonparametric approach to the multiplechangepoint problem," *Journal of the Royal Statistical Society:* . . . , vol. 49, no. 4, pp. 543–562, 2000.

[6] GC Tjhai and Maria Papadaki, "The problem of false alarms: Evaluation with snort and DARPA 1999 dataset," *Trust, privacy and Security* . . . , pp. 139–150, 2008.

[7] ES Page, "Continuous inspection schemes," *Biometrika*, pp. 100–115, 1954.

[8] Xiang Xuan and Kevin Murphy, "Modeling changing dependency structure in multivariate time series," *Proceedings of the 24th International Conference on Machine Learning (2007)*, vol. 227, no. m, pp. 1055–1062, 2007.

[9] K Bleakley and J P Vert, "The group fused Lasso for multiple change-point detection," Tech. Rep. HAL-00602121, HAL, 2011.

[10] L E Ghaoui O. Banerjee and A D'Aspremont, "Model Selection Through Sparse Maximum Likelihood Estimation for Multivariate Gaussian or Binary Data," *Journal of Machine Learning Research*, 2007.

[11] Jerome Friedman, Trevor Hastie, and Robert Tibshirani, "Sparse inverse covariance estimation with the graphical lasso.," *Biostatistics (Oxford, England)*, vol. 9, no. 3, pp. 432–41, July 2008.

[12] Shuheng Zhou, John Lafferty, and Larry Wasserman, "Time varying undirected graphs," *Machine Learning*, vol. 80, no. 2-3, pp. 295–319, Apr. 2010.

[13] Amr Ahmed and Eric P Xing, "Recovering time-varying networks of dependencies in social and biological studies.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 29, pp. 11878–83, July 2009.

[14] Z. Harchaoui and C. Lévy-Leduc, "Multiple Change-Point Estimation With a Total Variation Penalty," *Journal of the American Statistical Association*, vol. 105, no. 492, pp. 1480–1493, Dec. 2010.

[15] D Angelosante and G B Giannakis, "Sparse Graphical Modeling of Piecewise-Stationary Time Series," *International Conference on Acoustics, Speech and Signal Processing*, 2011.

[16] Michael I. Jordan, "Graphical Models," *Statistical Science*, vol. 19, no. 1, pp. 140–155, Feb. 2004.

[17] M Eichler, "Graphical modelling of multivariate time series," *Probability Theory*, 2012.

[18] R Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (* . . . , 1996.

[19] T Hastie J. Friedman and R Tibshirani, "Sparse Inverse Covariance Estimation with the Graphical Lasso," *Biostatistics*, 2007.

[20] Patrick Danaher, Pei Wang, and Daniela M Witten, "The joint graphical lasso for inverse covariance estimation across multiple classes," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2013.

[21] Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight, "Sparsity and smoothness via the fused lasso," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 1, pp. 91–108, Feb. 2005.

[22] Yu Nesterov, "Gradient methods for minimizing composite objective function," *ECORE*, 2007.

[23] Sen Yang, Zhisong Pan, Xiaotong Shen, Peter Wonka, Jieping Ye, and Computer Science, "Fused Multiple Graphical Lasso," *Arxiv*, pp. 1–14, 2012.

[24] Junzhou Huang and Tong Zhang, "The benefit of group sparsity," *The Annals of Statistics*, vol. 38, no. 4, pp. 1978–2004, Aug. 2010.

[25] MIT Lincoln Lab, "DARPA 1998 Dataset," `http://www.ll.mit.edu/mission/communications/cyber/CSTcorpora/ideval/data/1998data.html`, Accessed 5/3/2014.