DISTRIBUTED TOPOLOGY IDENTIFICATION FOR POINT PROCESS DYNAMIC NETWORKS

Syed Ahmed Pasha

Department of Electrical Engineering
Air University
Islamabad 44000, Pakistan

Victor Solo

School of Electrical Engineering & Telecom.
University of New South Wales
Sydney NSW 2052 Australia

ABSTRACT

Recently, the availability of high-dimensional point process data in a growing number of application areas is driving the demand for analysis tools for such data. An extremely challenging yet important problem is the inference of causal relationships from network data or network topology identification. This problem has received little attention in the literature until very recently. Here we develop, perhaps for the first time, a distributed optimization algorithm for large-scale dynamic networks of interacting Hawkes processes. Genomic data are analyzed to construct a transcriptional regulatory network in embryonic stem cells.

***Index Terms*—** Point process, stochastic intensity, topology identification, sparse estimation, distributed optimization

1. INTRODUCTION

In recent years, network data have become ubiquitous in a range of application domains. An extremely challenging problem is to infer causal relationships between nodes in a network given only the network data. This is compounded by the huge volume of such data which poses considerable challenges both in terms of storage as well as processing since traditional statistical methods are not distributed optimization algorithms. Furthermore, they suffer from the curse of dimensionality problem and quickly become unwieldy in the high-dimensional setting.

In a growing number of applications the network data are point processes where information is contained in the random times at which events occur. Applications typically involve counting interactions such as phone calls, emails [1, 2] and neural spike trains [3, 4]. [5] develop methods for the estimation of time-varying Poisson networks that have some limited history dependence. [3] model the point process stochastic intensity with history dependence in a finite interval but the estimation procedure requires fitting parameters to fine time bins on the interval. [4] develop a sparsity approach to fitting a network model with limited history dependence. [2] model pairwise directed interactions between a sender and a receiver using a multiplicative intensity model. The data are assumed to include information of the sender and receiver.

In previous work [6, 7], the authors discussed a Hawkes-Laguerre formulation for the point process stochastic intensity function and developed a l_1 regularized least squares procedure for grouped variables to solve the topology identification problem for sparse dynamic networks [8]. Subsequent work [9] discussed a systematic procedure to rigorously choose the regularization parameter and presented data analysis of genomic data. It was shown that the optimization problem is separable in the problem dimension and reduces to a convex optimization for each node. But for large-scale problems such as genome-wide analysis of the human genome a fully decentralized capability for data storage as well as optimization is needed. For example, in its pilot phase, the Encyclopedia of DNA Elements (ENCODE) Project [10] provided 29,998 kilobases (kb) of the human genome. These roughly 30 Mb of data account for only about 1% of the human genome.

In this paper, we develop a distributed optimization algorithm via the alternating direction method of multipliers (ADMM) method suitable for such large-scale problems. The ADMM method was developed in the 70's [11, 12] but a recent survey paper [13] has renewed interest in the algorithm. The approach is elegant due to its simplicity and robustness. Given an optimization problem, the ADMM form is an equivalent optimization problem where the objective function is separable across a splitting of the optimization variable. This leads to a decentralized algorithm where the local problems are solved in parallel. The utility of the ADMM method is immediately realized for statistical problems that involve optimization of a smooth function together with a non-smooth regularization term where the method naturally decouples the smooth function from the non-smooth term.

In the remainder of the paper we review the Hawkes-Laguerre dynamic point process model and the ADMM method in section 2. The sparse estimation algorithm and implementation are discussed in section 3. Some data analysis is presented in section 4. The paper concludes in section 5.

Notation. $\|.\|$ is the l_2 norm. Given $x \in \mathbb{R}^n$ and $P \in$

$\mathbb{R}^{n\times n}$, $\|x\|_P \equiv (x^T P x)^{\frac{1}{2}}$, $(\alpha)_+ \equiv \max(0,\alpha)$ and $\mathcal{S}_\kappa(a) = \left(1 - \frac{\kappa}{\|a\|}\right)_+ a$.

2. BACKGROUND

2.1. Point Process Dynamic Network

Consider a network comprising d nodes where each node represents a dynamic process. Causal dependencies in the network are represented by directed links between the nodes which determine the direction of information flows in the network. The output at each node is a spike train and furthermore observable on the time interval $0 < t \leq T$. The assumption that the output at each node is observed is standard (see [14]) so that the estimation problem remains tractable.

The network can be modelled by a multivariate point process on the interval $(0, T]$, characterized through its vector counting process $N_t := N(t) = (N_{1,t}, ..., N_{d,t})^T$ where the nodal process is a scalar counting process $N_{k,t} := N_k(t) =$ # events of the k-th process in $(0, t]$. We denote the counting process increment as $N^\delta_{k,t} =$ # events of the k-th process in $(t, t+\delta]$ and the history of the vector counting process as $\mathcal{H}^t = \{N_s, 0 \leq s < t\}$.

We introduce a standard assumption in point process theory, the No-Simultaneity condition [15],

$$P(\Sigma_1^d N^\delta_{k,t} > 1 | \mathcal{H}^t) = o(\delta)$$

where $\lim_{\delta\to 0} \frac{o(\delta)}{\delta} = 0$. This simply means that in an infinitesimal time interval with high probability simultaneous events of any type are not allowed. In the literature this is also referred to as orderliness [16].

Then, the vector stochastic intensity $\mu_t = (\mu_{1,t}, ..., \mu_{d,t})^T$ can be defined as

$$P(N^\delta_{k,t} = 1 | \mathcal{H}^t) = \mu_{k,t}\delta + o(\delta), \quad k = 1, ..., d$$

The observations at node k can be modelled as

$$Y_{k,t} = \mu_{k,t} + e_{k,t}$$

where $Y_{k,t} := \frac{1}{\delta} N^\delta_{k,t}$, $e_{k,t}$ is a martingale increment noise and the stochastic intensity $\mu_{k,t}$ is modelled by the Hawkes-Laguerre process [6, 7],

$$\mu_{k,t} = c_k + \Sigma_1^d \int_0^t h_{k,j}(u) dN_{j,t-u},\ k = 1, ..., d \quad (1)$$

where $c_k > 0$ is the background firing rate and $h_{k,j}(.) \geq 0$ is the impulse response of the directed link to node k from node j. We follow [6, 7] and expand $h_{k,j}$ in Laguerre polynomials which is a causal basis [17] (see [6, 7] for further advantages),

$$h_{k,j}(u) = \Sigma_1^{p_{k,j}} \beta_{k,j,l} (u\beta_{k,j,o})^{l-1} e^{-\beta_{k,j,o} u} \beta_{k,j,o} \quad (2)$$

where $\frac{1}{\beta_{k,j,o}}$ is a user chosen time constant.

Substituting (2) in (1), the stochastic component due to node j is

$$\begin{aligned} & \int_0^t h_{k,j}(u) dN_{j,t-u} \\ = \ & \Sigma_1^{p_{k,j}} \beta_{k,j,l} \int_0^t (u\beta_{k,j,o})^{l-1} e^{-\beta_{k,j,o} u} \beta_{k,j,o} dN_{j,t-u} \\ = \ & \Sigma_1^{p_{k,j}} \beta_{k,j,l} \psi_{k,j,l;t} \\ \equiv \ & \beta_{k,j}^T \psi_{k,j;t} \end{aligned}$$

where $\beta_{k,j}$ is the $p_{k,j}$ vector of coefficients of the directed link to node k from node j.

We thus have a linear (in the parameters) model,

$$\mu_{k,t} = c_k + \beta_k^T \psi_{k;t},\ k = 1, ..., d \quad (3)$$

where β_k comprises the d vector-valued coefficients of the directed links to node k assuming a fully connected network.

2.2. Alternating Direction Method of Multipliers

Many optimization problems have the form

$$\min_x f(x) + g(x)$$

where f, g are not necessarily convex (or concave). They may separately have nice features with respect to optimization but not when they are added. The idea of the alternating direction method of multipliers (ADMM) is to separate f from g as follows

$$\min_{x,z;x=z} f(x) + g(z)$$

Now one simply adds a quadratic penalty to enforce equality. The primal variables x, z are obtained via iterative minimization in alternating manner. Using the scaled dual variable η [13] we get the following iteration

$$\begin{aligned} x^{(i+1)} &:= \arg\min_x (f(x) + \tfrac{\rho}{2}\|x - z^{(i)} + \eta^{(i)}\|^2) \\ z^{(i+1)} &:= \arg\min_z (g(z) + \tfrac{\rho}{2}\|x^{(i+1)} - z + \eta^{(i)}\|^2) \\ \eta^{(i+1)} &:= \eta^{(i)} + x^{(i+1)} - z^{(i+1)} \end{aligned}$$

where $\rho > 0$ is the penalty parameter and $x - z$ is the residual.

The quadratic penalty greatly improves convergence properties of the algorithm which have been studied for example in [18, 19].

3. DECENTRALIZED TOPOLOGY IDENTIFICATION

We begin by expressing the point process data together with the Laguerre model (3) as a standard linear regression model

$$y_k = X_k \beta_k + e_k$$

where $y_k = (Y_{k,0} - c_k, Y_{k,\delta} - c_k ..., Y_{k,T-\delta} - c_k)^T$ and $X_k = (X_{k,1}, ..., X_{k,d})$ with

$$X_{k,j} = \begin{pmatrix} \psi_{k,j,1;0} & \cdots & \psi_{k,j,p_{k,j};0} \\ \vdots & \ddots & \vdots \\ \psi_{k,j,1;T-\delta} & \cdots & \psi_{k,j,p_{k,j};T-\delta} \end{pmatrix}$$

Given the point process observations at each node, the network topology identification problem is to recover a sparse network by setting some of the vector-valued coefficients $\beta_{k,j}$ to zero where directed links do not exist. In the least squares setting, the problem can be formulated as a l_1 regularized least squares for grouped variables [8, 9],

$$\min_{\beta_1,...,\beta_d} \Sigma_{k=1}^{d} \left(\tfrac{1}{2}\|y_k - X_k\beta_k\|^2 + \lambda_k \Sigma_{j=1}^{d} \|\beta_{k,j}\|_P \right) \quad (4)$$

where $\lambda_k = \lambda_o \sqrt{c_k/\delta}$ is the scaled regularization parameter where the scaling $\sqrt{c_k/\delta}$ is introduced for dimensional consistency (see [9]) and λ_o is the only tuning parameter. Following [20, 21], we take $P = pI$ with $p = \max\{p_{k,j}, k, j = 1, ..., d\}$.

Equation (4) is separable in the dimension d and can be solved as d separate minimizations in parallel

$$\min_{\beta_k} \tfrac{1}{2}\|y_k - X_k\beta_k\|^2 + \lambda_k \Sigma_{j=1}^{d} \|\beta_{k,j}\|_P, \quad k = 1, ..., d \quad (5)$$

This problem was solved in [8, 9]. Here, we cast the k-th minimization (5) in the ADMM framework and develop a fully decentralized solution.

Using $\nu_k = \lambda_k \sqrt{p}$, the ADMM form is

$$\begin{aligned} \min_{\beta_k} \quad & \tfrac{1}{2}\|y_k - X_k\beta_k\|^2 + \nu_k \Sigma_{j=1}^{d} \|z_{k,j}\| \quad (6) \\ \text{s.t.} \quad & \beta_{k,j} - z_{k,j} = 0, \ j = 1, ..., d \end{aligned}$$

which for orthonormal $X_{k,j}$ yields the following updates

$$\beta_{k,j}^{(i+1)} := \alpha X_{k,j}^T (y_k - X_k \beta_{k,-j}^{(i)}) + (1-\alpha)(z_{k,j}^{(i)} - \eta_{k,j}^{(i)}) \quad (7)$$

$$z_{k,j}^{(i+1)} := \mathcal{S}_{\alpha\nu_k/(1-\alpha)}(\beta_{k,j}^{(i+1)} + \eta_{k,j}^{(i)}) \quad (8)$$

$$\eta_{k,j}^{(i+1)} := \eta_{k,j}^{(i)} + \beta_{k,j}^{(i+1)} - z_{k,j}^{(i+1)} \quad (9)$$

with $\alpha = (1+\rho)^{-1}$ so that $0 < \alpha < 1$. This gives a useful interpretation of (7) as a weighted average where small α values place a strong penalty on violation of the primal feasibility and therefore produce smaller primal residuals. $\beta_{k,-j} = (\beta_{k,1}^T, ..., \beta_{k,j-1}^T, \mathbf{0}^T, \beta_{k,j+1}^T, ..., \beta_{k,d}^T)^T$ and $\mathcal{S}_{(.)}$ is the vector soft thresholding operator.

While orthonormal $X_{k,j}$ is not necessary, the algorithm in such a case has better numerical properties. For this reason, orthonormalizing $X_{k,j}$ is preferred [20]. Note that $\beta_{k,j}$ in the penalty term will need to be scaled and transformed back after the estimation to obtain the original variables.

Given the estimate of β_k,

$$c_k = (T/\delta)^{-1} \mathbf{1}^T (y_k - X_k\beta_k) \quad (10)$$

3.1. Computational Details

$X_k, k = 1, ..., d$ are known and need to be computed just once. In the special case that $\beta_{k,j,o}$ does not depend on k, $X_1 = X_2 = ... = X_d$.

The minimization (6) for $k = 1, ..., d$ can be done concurrently on d systems. The k-th minimization can be distributed to d subsystems indexed (k, j). Subsystem (k, j) will store the $T \times p$ and $p \times pd$ matrices $X_{k,j}, X_{k,j}^T X_k$ respectively and the current T and pd vectors y_k, β_k respectively. The updates are computed locally in the order (9),(7),(8). The k-th system will gather $\beta_{k,j}, j = 1, ..., d$ to do the c_k update (10). The updated y_k and assembled β_k are broadcast back to the subsystems. The k-th optimization is halted if $\max(|c_k^{(i+1)} - c_k^{(i)}|, \|\beta_k^{(i+1)} - \beta_k^{(i)}\|) < \epsilon$ for a suitable tolerance level ϵ.

4. TRANSCRIPTIONAL REGULATORY NETWORKS

Transcription factors (TFs), typically proteins, are used by cells to control gene expression i.e. into various types of cells. TFs perform their function either alone or by recruiting transcriptional regulators (TRs) and are frequently encoded in different configurations forming a transcriptional regulatory network (TRN).

We analyze genomic data from a study [22] to identify causal interactions in TRNs in embryonic stem cells. The data comprise co-ordinates of the binding sites of 13 TFs and 2 TRs.

4.1. Genomic Data Analysis

We use the data from chromosome 1 and following the approach in [23, 24], transform it to point process data at a resolution of $\delta = 100$ base pairs. Nodes with fewer than 100 counts after ensuring No-Simultaneity were dropped. This led to 12 TFs and 1 TR suz12 with about 500 counts per node. A raster plot of the output at each node is shown in Fig. 1.

We use Laguerre basis expansion for the impulse response $h_{k,j}, k, j = 1, ..., d$ with $p = 3$ terms. We take $\alpha = .09$.

The regularization parameter λ_o and tuning parameter $\beta_{k,j,o} = \beta_o$ were determined from the Bayesian Information Criterion (BIC) heat map shown in Fig. 2. BIC $= -2\mathcal{L} + \Sigma_1^d r_k \ln(N_{k,T})$ where $\mathcal{L} = \Sigma_1^d \int_0^T (\ln \mu_{k,t} dN_{k,t} - \mu_{k,t} dt)$ is the multivariate point process log-likelihood [16] and $r_k = \#$ active parameters of the k-th node.

We find that for $1 \leq \beta_o \leq 10$, the BIC plot exhibits a sharp fall initially with increasing λ_o values and is nearly flat for $\lambda_o \geq .2$. For $\beta_o > 10$, the BIC values at $\lambda_o = 0$ are relatively low. The surface falls gradually with increasing λ_o values and is nearly flat for $\lambda_o \geq .2$.

We have found that for a given λ_o value, the penalized regression algorithm gives a more parsimonious representation with a smaller β_o value but takes longer to converge. We take $\lambda_o = .2$. As already mentioned, $\beta_{k,j,o}$ controls the extent of memory in the model (3). By choosing $\beta_{k,j,o} = \beta_o = 20$, we set the *time constant* $\approx$ 20,000 base pairs. Note that in [24] the history dependence is limited to 1,000 base pairs only.

The starting values of c_k, β_k were obtained from a standard least squares estimate by setting negative values to zero to ensure positivity of the intensity function. The relative magnitude $(J - \hat{J})/\hat{J}$ iterates of the cost function (4) are

shown in Fig. 3 where $\hat{J}$ is the optimal cost at convergence. Also shown is the relative magnitude of the cost using the starting values of c_k, β_k and with $\lambda_o = 0$ to give a comparison with standard least squares method. We find that the cost obtained from the penalized regression algorithm falls monotonically until convergence, after about 40 iterations.

4.2. Results

The TRN recovered by the sparse estimation method is shown in Fig. 4. The self-exciting link at each node is omitted to avoid clutter. There were 3 bi-directional interactions; Oct4↔Sox2, E2f1↔Klf4 and Zfx↔n-Myc. These are indicated by undirected links.

In comparison to the fully connected network with $d^2 = 169$ links and $d + d^2p = 507$ parameters, the sparse estimation method removed 114 links and required 165 parameters. The network in Fig. 4 shows strong similarity to the network of 10 TFs estimated in [9] and that of [22] constructed from experimental studies [25, 26, 22].

The algorithm in [24] estimates a fully connected network and therefore suffers from over-fitting. The choice of B-splines basis expansion in their log-intensity model may lead to spurious estimates since B-splines do not form a causal basis system. Furthermore, the centralized nature of their algorithm places huge memory constraints for implementation.

[26] reported that depleting embryonic stem cells of Nanog resulted in the downregulation of Sox2 gene expression which led to the conclusion that Nanog activates transcription of the Sox2 gene. Since they infer interactions in a pairwise fashion it is not sufficient to establish the direct link Nanog → Sox2. Our analysis suggests that the interaction is most likely mediated via Oct4.

In addition, the TRN in Fig. 4 also uncovered some links that have not been studied experimentally. For example, (Esrrb,E2f1) → Tcfcp211, (Esrrb,E2f1) → Zfx, Oct4 → Zfx, etc.

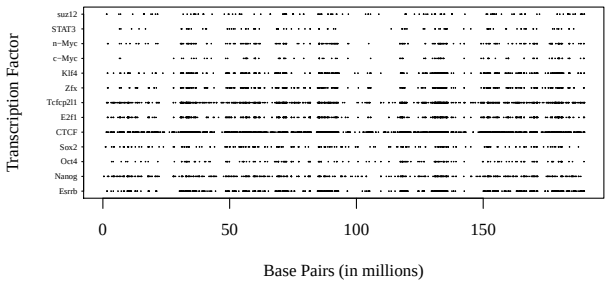

Fig. 1. Raster Plot of Output at the 13 Nodes.

5. CONCLUSIONS

In this paper we have developed a decentralized solution for topology identification in large-scale dynamic networks involving point processes. The sparse network estimation problem formulated as a l_1 regularized least squares optimization for grouped variables is cast in the ADMM form which leads to local optimization problems that can be solved in parallel using modest computing resources. We discussed a rigorous procedure for choosing the regularization parameter and the tuning parameter. The algorithm was tested on some genomic data relating to mouse embryonic stem cells.

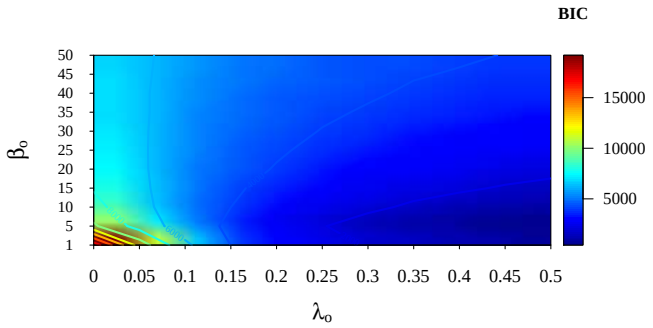

Fig. 2. Heat Map of the BIC Values.

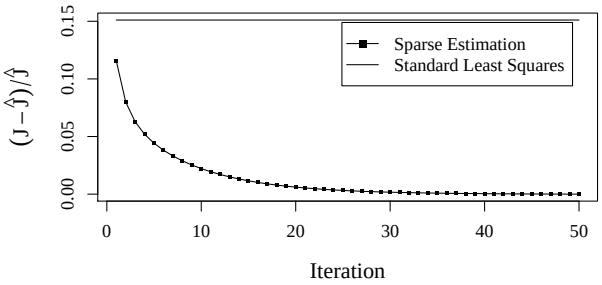

Fig. 3. Relative Magnitude Iterates of the Cost Function (4).

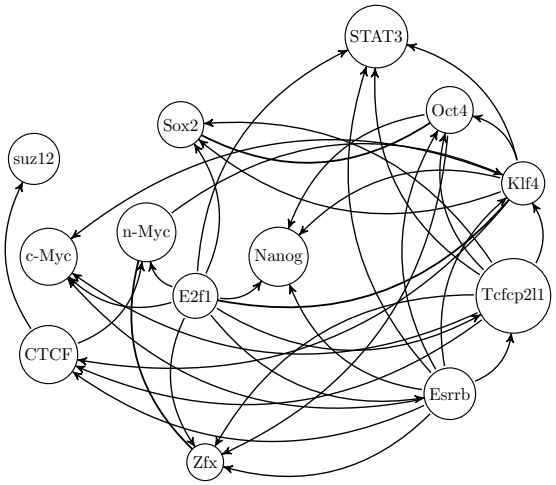

Fig. 4. Transcriptional Regulatory Network in the Mouse Embryonic Stem Cells.

6. REFERENCES

[1] J.R. Tyler, D.M. Wilkinson, and B.A. Huberman, "E-mail as spectroscopy: Automated discovery of community structure within organizations," *Inform. Soc.*, vol. 21, pp. 143–153, 2005.

[2] P.O. Perry and P.J. Wolfe, "Point process modeling for directed interaction networks," *Jl. Royal Stat. Soc.: Series B*, vol. 75, pp. 821–849, 2013.

[3] M. Okatan, M.A. Wilson, and E.N. Brown, "Analyzing functional connectivity using a network likelihood model of ensemble neural spiking activity," *Neural computation*, vol. 17(9), pp. 1927–1961, 2005.

[4] R.C. Kelly, R.E. Kass, M.A. Smith, and T.S. Lee, "Accounting for network effects in neuronal responses using l_1 regularized point process models," *Adv. Neural Inf. Process. Syst.*, vol. 23(2), pp. 1099–1107, 2010.

[5] R. Shyamsundar, T. Graepel, and R. Herbrich, "Poisson-networks: A model for structured point processes," in *Proc. 10th Intl. Workshop on AI and Stat.*, 2005.

[6] S.A. Pasha and V. Solo, "Hawkes-Laguerre reduced rank model for point processes," in *Proc. IEEE ICASSP*, 2013, pp. 6098–6102.

[7] S.A. Pasha and V. Solo, "Hawkes-Laguerre dynamic index models for point processes," in *Proc. IEEE CDC*, 2013, pp. 7028–7033.

[8] S.A. Pasha and V. Solo, "Topology identification of dynamic point process networks," in *Proc. IEEE ICASSP*, 2014, pp. 375–378.

[9] S.A. Pasha and V. Solo, "Topology identification for sparse dynamic point process networks," in *Proc. IEEE CDC, in press*, 2014.

[10] E. Birney et al., "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project," *Nature*, vol. 447, 2007.

[11] R. Glowinski and A. Marrocco, "Sur l'approximation, par elements finis d'ordre un, et la resolution, par penalisation-dualité, d'une classe de problems de dirichlet non lineares," *Revue Française d'Automatique, Informatique, et Recherche Opérationelle*, vol. 9, 1975.

[12] D. Gabay and B. Mercier, "A dual algorithm for the solution of nonlinear variational problems via finite element approximations," *Computers and Mathematics with Applications*, vol. 2, pp. 17–40, 1976.

[13] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine Learning*, vol. 3(1), pp. 1–122, 2011.

[14] A. Bolstad, B.D. Van Veen, and R. Nowak, "Causal network inference via group sparse regularization," *IEEE Trans. Sig. Proc.*, vol. 59, pp. 2628–2641, 2011.

[15] V. Solo, "High dimensional point process system identification: PCA and dynamic index models," in *Proc. IEEE CDC*, 2006, pp. 829–833.

[16] D.J. Daley and D. Vere-Jones, *An Introduction to the Theory of Point Processes, Volume I (2nd. ed.)*, Springer-Verlag, New York, 2003.

[17] B. Wahlberg, "System identification using Laguerre models," *IEEE Trans. Automatic Control*, vol. 36(5), pp. 551–562, 1991.

[18] D. Gabay, "Applications of the method of multipliers to variational inequalities," in *Augmented Lagrangian Methods: Applications to the Solution of Boundary-Value Problems*, M. Fortin and R. Glowinski, Eds. Amsterdam, North-Holland, 1983.

[19] J. Eckstein and D.P. Bertsekas, "On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators," *Mathematical Programming*, vol. 55, pp. 293–318, 1992.

[20] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 68(1), pp. 49–67, 2006.

[21] A.J. Seneviratne and V. Solo, "Topology identification of a sparse dynamic network," in *Proc. IEEE CDC*, 2012, pp. 1518–1523.

[22] X. Chen et al., "Integration of external signaling pathways with the core transcriptional network in embryonic stem cells," *Cell*, vol. 133, pp. 1106–1117, 2008.

[23] G. Gusto and S. Schbath, "FADO: a statistical method to detect favored or avoided distances between motif occurrences using the Hawkes' model," *Statist. Applications in Genetics and Molecular Bio.*, vol. 4(1), 2005.

[24] L. Carstensen, A. Sandelin, O. Winther, and N.R. Hansen, "Multivariate Hawkes process models of the occurrence of regulatory elements," *BMC bioinformatics*, vol. 11(1), 2010.

[25] N.B. Ivanova et al., "A stem cell molecular signature," *Science*, vol. 298, pp. 601–604, 2002.

[26] Y-H. Loh et al., "The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells," *Science*, vol. 38, pp. 431–440, 2006.