COMPRESSIVE KALMAN FILTERING FOR RECOVERING TEMPORALLY-REWIRING GENETIC NETWORKS

Jehandad Khan[1]*, Nidhal Bouaynaya*[2] *and Hassan M. Fathallah-Shaykh*[3]

[1] Department of Systems Engineering, University of Arkansas at Little Rock, Little Rock, AR
[2] Department of Electrical and Computer Engineering, Rowan University, Glassboro, NJ
[3] Department of Neurology, University of Alabama at Birmingham, Birmingham, AL

ABSTRACT

Genetic regulatory networks undergo rewiring over time in response to cellular developments and environmental stimuli. The main challenge in estimating time-varying genetic interactions is the limited number of observations at each time point; thus making the problem unidentifiable. We formulate the recovery of temporally-rewiring genetic networks as a tracking problem, where the target to be tracked over time consists of the set of genetic interactions. We circumvent the observability issue (due to the limited number of measurements) by taking into account the sparsity of genetic networks. With linear dynamics, we use a compressive Kalman filter to track the interactions as they evolve over time. Our simulation results show that the compressive Kalman filter achieves good tracking performance even with one measurement available at each time point; whereas the classical (unconstrained) Kalman filter completely fails in obtaining meaningful tracking.

Index Terms— Genetic networks; Kalman filtering; compressed sensing.

1. INTRODUCTION

Deciphering the complex dynamic nature of genetic regulatory networks holds the key to progressive therapeutic methods for many genetic ailments, including cancer. Much recent work has gone into identifying (or reverse-engineering) the structure of time-invariant gene regulatory networks from expression data (e.g., microarrays). Most popular methods include (probabilistic) Boolean networks, (dynamic) Bayesian networks, information-theoretic approaches and differential equations models [1, 2]. The DREAM (Dialogue on Reverse Engineering Assessment and Methods) project, which built a blind framework for performance assessment of methods for gene network inference, showed that there is no correlation between the inference methods used and the performance scores [3]. Rather, the success of a method is more related to the details of the implementation than the choice of the general methodology

These methods, however, estimate one single network from the available data, independently of the cellular "themes" or environmental conditions under which the measurements were collected. Collections of time-dependent genetic data from dynamic biological processes such as cancer progression, response to therapeutic compounds and developmental processes, are increasing with the new developments in high-throughput technologies. Clearly, the "static" or time-invariant view of these dynamical systems does not capture the temporal rewiring of genetic networks due to internal and external requirements and stimuli. The current understanding of the cell as a fixed network of genes and proteins is obsolete and we must derive new methods that unravel the dynamic nature of genetic networks by tracking genetic interactions as they undergo systematic rewiring in response to cellular development and environmental changes. These changes in network topology are imperceptible given current viewpoints and practices.

The inference of time-invariant genetic networks suffers from the limited number of measurements available to unambiguously estimate the network connectivity. The "large p small n" problem poses a challenge in estimation due to the identifiability problem, where a large class of network topologies is consistent with the measurements and no unique solution exists. This problem is even more severe for temporally-rewiring networks, where at a given time t, one or very few measurements or observations are available.

One way to ameliorate this data scarcity problem is to presegment the time-series into stationary epochs, and infer a static network for each epoch separately [4]. The main problem with the segmentation approach is the limited number of time points available in each stationary segment, which limits the resulting networks in terms of their temporal resolution and statistical power. Full resolution techniques, which allow a time-specific network topology to be inferred from samples measured over the entire time series, rely on model-based approaches [5]. However, these methods learn the structure (or skeleton) of the network but not the detailed strength of the interactions between the nodes. Dynamic Bayesian networks (DBNs) have been extended to the time varying case [6]. In

time-varying DBNs (TVDBN), the time-varying structure and parameters of the networks are treated as additional hidden nodes in the graph model [6].

In this paper, we formulate the problem of estimating time-varying genetic networks as a tracking problem, where the tracked state is the network connectivity matrix. The tracking is formulated within a state-space model with the network connectivity being the state vector. In order to improve the estimation accuracy with a limited number of observations, we consider the sparsity of the network. Empirical data indicate that biological gene networks are sparsely connected, and that the number of regulators per gene is only a small fraction of the total number of genes. Recent studies have shown that sparse signals can be recovered accurately using less observations than what is considered necessary by the Nyquist sampling theory [7]. This theory, known as compressed sensing, carries signal recovery and compression simultaneously; thus reducing the number of required observations. In general, the recovery of sparse signals is an NP-hard problem [7]. However, under some restrictions, one can relax the problem into a convex optimization problem by adopting the l_1 norm rather than the l_0 measure [7]. We adopt a linear state-space model, where the gene expressions vary over time following a linear differential equation model. Recovery of the time-varying sparse network connectivity is achieved using a Kalman filtering-based compressed sensing approach [8].

In this paper, scalars are denoted by lower case letters, vectors in $\mathbb{R}^n$ are denoted by lower case bold letters, e.g. $\mathbf{v}$ and matrices in $\mathbb{R}^{m\times n}$ are denoted by bold upper case letters, e.g. $\mathbf{A}$. $\mathbf{I}_p$ stands for the identity matrix of dimension $p \times p$ and $\mathbf{x}^t$ indicates the transpose of the vector $\mathbf{x}$.

2. MODEL DESCRIPTION

We model the concentrations of genes, proteins and other molecules using a time-varying ordinary differential equation (ODE) model, where the concentration of every molecule is modeled as a linear combination of the concentrations of the other molecules in the network. The rewiring nature of the network is captured by the time-dependent ODE coefficients. We have

$$\dot{x}_i(k) = -\lambda_i(k)x_i(k) + \sum_{j=1}^{p} a_{ij}(k)x_j(k) + v_i(k), \quad (1)$$

where $i = 1, ..., p$, with p being the total number of genes, $x_i(k)$ is the gene expression level at time k and $\dot{x}_i(k)$ is its rate of change, λ_i is the rate of self degradation of the i^{th} gene, $a_{ij}(k)$ is the influence of gene j on gene i at time k and $v_i(k)$ models the biological and measurement noise. Equation (1) can be written in matrix form as

$$\boldsymbol{z}(k) = \boldsymbol{A}(k)\boldsymbol{x}(k) + \boldsymbol{v}(k), \quad (2)$$

where $\boldsymbol{z}(k) = [\dot{x}_1(k), \cdots, \dot{x}_p(k)]^t$, $\boldsymbol{x}(k) = [x_1(k), \cdots, x_p(k)]^t$, $\boldsymbol{A}(k) = \{a_{ij}(k)\}$ is the matrix of time-varying interactions with $a_{ii} = -\lambda_i$ and $\boldsymbol{v}(k) = [v_1(k), \cdots, v_p(k)]^t$. Let $\boldsymbol{a}(k) \in \mathbb{R}^{p^2}$ be the vectorized form of the matrix $\boldsymbol{A}(k)$, i.e., $\boldsymbol{a}(k)$ is the vector composed of the concatenation of the columns of $\boldsymbol{A}(k)$,

$$\boldsymbol{a}(k) = [a_{11}(k), \cdots, a_{1p}(k), \cdots, a_{p1}(k), \cdots, a_{pp}(k)]^t. \quad (3)$$

It can be easily shown that

$$\boldsymbol{A}(k)\boldsymbol{x}(k) = [I_p \otimes \boldsymbol{x}(k)^t]\boldsymbol{a}(k) = \boldsymbol{H}(k)\boldsymbol{a}(k), \quad (4)$$

where $\otimes$ denotes the Kronecker product operator. Hence, Eq. (2) can be rewritten as

$$\boldsymbol{z}(k) = \boldsymbol{H}(k)\boldsymbol{a}(k) + \boldsymbol{v}(k), \quad (5)$$

where $\boldsymbol{H}(k) = I_p \otimes \boldsymbol{x}(k)^t$. Equation (5) is the observation equation of the state-space model with state vector $\boldsymbol{a}(k)$. The state equation models the prior knowledge on the dynamics of the state vector $\boldsymbol{a}(k)$. In this paper, we consider a random walk model, which reflects a lack of prior about the network connectivity dynamics. The state-space model of the network connectivity is given by

$$\begin{aligned} \boldsymbol{a}(k+1) &= \boldsymbol{a}(k) + \boldsymbol{w}(k) && (6)\\ \boldsymbol{z}(k) &= \boldsymbol{H}(k)\boldsymbol{a}(k) + \boldsymbol{v}(k), && (7)\\ \boldsymbol{a}(k) &\quad \text{is sparse.} \end{aligned}$$

with $\boldsymbol{H}$ is as defined in Eq. (5). The state and observation noise, $\mathbf{w}(\mathbf{k})$ and $\mathbf{v}(\mathbf{k})$, respectively, are zero mean Gaussian processes with covariances $\mathbf{Q_k}$ and $\mathbf{R_k}$, respectively. Since $\boldsymbol{H}(k) \in \mathbb{R}^{p\times p^2}$ is underdetermined, the state-space model in (6)-(7) is not observable, and hence the Kalman filter algorithm is senseless. However, this problem can be circumvented if we consider that the state vector $\boldsymbol{a}(k)$ is sparse.

3. COMPRESSIVE KALMAN FILTERING

The state-space estimation problem in (6)-(7), can be solved using constrained Kalman filtering, which provides the solution to the following constrained minimum-variance problem:

$$\begin{aligned} \min_{\hat{\boldsymbol{a}}_k} \quad & E_{\boldsymbol{a}_k|\boldsymbol{z}_k}[\|\boldsymbol{a}_k - \hat{\boldsymbol{a}}_k\|^2] \\ \text{subject to} \quad & \|\hat{\boldsymbol{a}}_k\|_1 \leq \epsilon, \end{aligned} \quad (8)$$

where the l_1-norm constraint imposes sparsity and ϵ controls the amount of sparsity of the vector $\boldsymbol{a}(k)$. The constrained Kalman filter exploits the additional information and gets better filtering performance than the (unconstrained) Kalman filter provides. There are various methods to incorporate state constraints in the Kalman filter [9]. If the state constraints are linear, then all of these different approaches result in the same

state estimate, which is the optimal constrained linear state estimate. If the constraints are nonlinear, then constrained filtering is, in general, not optimal, and different approaches give different results [10]. The constrained optimization problem in (8) can be solved using the pseudo-measurement technique (PM) [11]. PM generates a fictitious observation from the constraint function by writing the l_1 constraint as

$$\begin{aligned} &\|\hat{\boldsymbol{a}}_k\|_1 - \epsilon = 0 \Longleftrightarrow \boldsymbol{d}_k^t \hat{\boldsymbol{a}}_k - \epsilon = 0, \\ &\boldsymbol{d}_k = [\text{sign}(\hat{\boldsymbol{a}}_k(1)), \cdots, \text{sign}(\hat{\boldsymbol{a}}_k(p^2))]^t, \end{aligned} \tag{9}$$

where "sign" is the sign function. Observe that the pseudo observation matrix $\boldsymbol{d}_k$ depends upon the current state estimate. The complete algorithm is listed in Algorithm 1.The PM stage can be iterated to lessen the effect of base point and truncation errors associated with linearizing a non-linear constraint [11], [8], [12]. The method consists of repeating the PM stage multiple number of times to get closer and closer to the actual state estimate. The iteration is necessary because the constraint occurs around the state estimate rather than the actual estimate, which causes a shift in the estimate projection. Thus, repeating the constraint multiple number of times ensures that the error is reduced to a minimum [12]. We will now give the algorithm that describes the compressed sensing Kalman filter as proposed by Carmi *et al.* [8].

Algorithm 1 Compressive Kalman Filtering

1: *Prediction*

$$\hat{a}_{k+1|k} = \hat{a}_{k|k} \tag{10}$$
$$P_{k+1|k} = P_{k|k} + Q_k \tag{11}$$

2: *Measurement update*

$$K_k = P_{k+1|k} H^T (H P_{k+1|k} H^T + R_k)^{-1} \tag{12}$$
$$\hat{a}_{k+1|k+1} = \hat{a}_{k+1|k} + K_k (z_k - H \hat{a}_{k+1|k}) \tag{13}$$
$$P_{k+1|k+1} = (I - K_k H) P_{k+1|k} \tag{14}$$

3: *Pseudo-measurement*: Let $P^1 = P_{k+1|k+1}$ and $\hat{a}^1 = \hat{a}_{k+1|k+1}$.

4: **for** $\tau = 1, 2, \ldots, N_\tau - 1$ iterations **do**

$$\mathbf{d}_\tau = [\text{sign}(\hat{a}^\tau(1)), \ldots, \text{sign}(\hat{a}^\tau(p))]^T \tag{15}$$
$$K^\tau = P^\tau \mathbf{d}_\tau (\mathbf{d}_\tau^t P^\tau \mathbf{d}_\tau + \sigma_\epsilon^2)^{-1} \tag{16}$$
$$\hat{a}^{\tau+1} = (I - K^\tau \mathbf{d}_\tau^t) \hat{a}^\tau \tag{17}$$
$$P^{\tau+1} = (I - K^\tau \mathbf{d}_\tau^t) P^\tau \tag{18}$$

5: **end for**

6: Set $P_{k+1|k+1} = P^{N_\tau}$ and $\hat{a}_{k+1|k+1} = \hat{a}^{N_\tau}$.

4. SIMULATION RESULTS

We generate time-varying genetic networks obeying the dynamics in Eq. (2). We set the sparsity level at 7%. That is, the connectivity matrix has $0.07p^2$ non-zero entries with p being the number of genes or the dimension of the matrix. To assess and compare the performance of the algorithm we use the following error measure

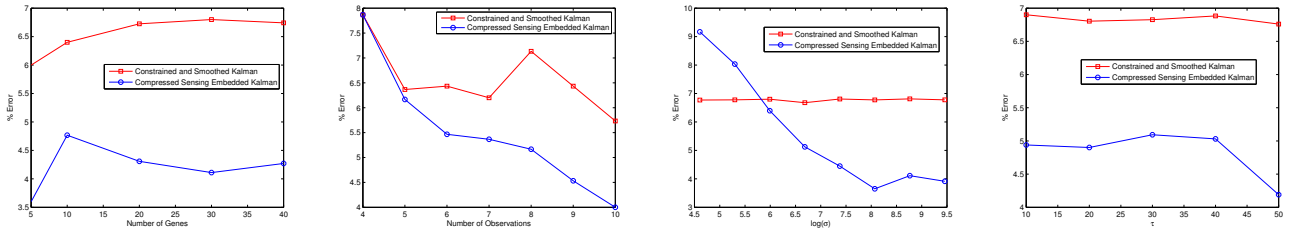
$$|a_{ij} - \hat{a}_{ij}| \leq \alpha a_{ij} \tag{19}$$

Where a_{ij} is the $(i,j)^{th}$ element of the true genetic interaction matrix and $\hat{a}_{ij}$ is the estimate of a_{ij}. α is a threshold parameter less than 1. Here, we fixed $\alpha = 0.2$. That is, we assume that the noise level (measurement errors, imperfection in the model and numerical errors) is about 20%. We count an error if the estimated interaction value, $\hat{a}_{ij}$, is not within 20% of the true value, a_{ij}. We compare the proposed compressive Kalman filter with the projection method adopt in [2] for estimating time-varying networks.

We first investigate the effect of the algorithm parameters, namely, the network size p, the number of measurements n, the sparsity parameter ϵ and the number of iterations τ. Figure 1(a) shows the error as a function of the number of genes. Observe that the state vector size increases exponentially with the number of genes. Nonetheless, the performance of the compressive Kalman filter does not seem to be affected by the large dimension of the state vector (for $p = 50$, there are $50^2 = 2500$ connections to be estimated). The number of measurements at each time instant are kept constant equal to $n = 5$. With the increase in the network size, the algorithm shows robustness and better prediction performance than the projection method adopted in [2], which leads to a uniform increase in the error with the network size. Figure 1(b) shows the performance of the algorithm when the number of measurements increases for a network of size $p = 50$. As expected, the estimation accuracy increases with the number of measurements or observations.

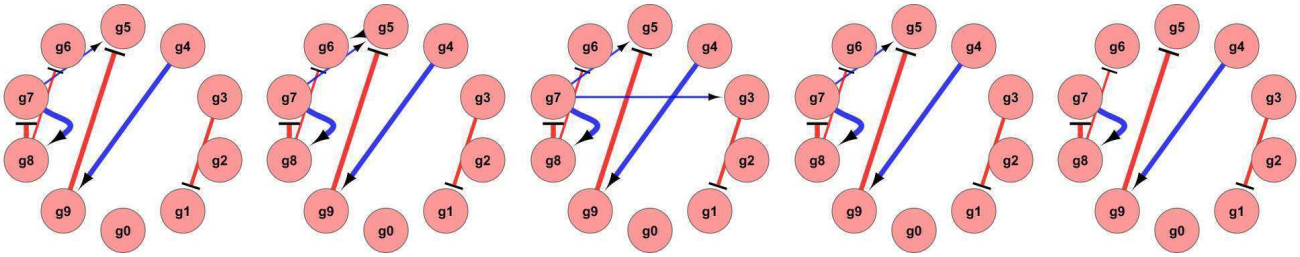
Figure 1(c) shows the effect of ϵ, the parameter that controls the sparsity in the estimated network. The network size, in this simulation, is $p = 30$ and the sparsity is 7%. It is evident that the choice of this parameter affects the estimation accuracy. ϵ is considered to be a prior knowledge on the degree of sparsity in the network. Figure 1(d) shows that the number of iterations τ has some effect on the accuracy error; Though, in our simulations, we found that, for a network of size $p = 50$, increasing τ from 10 to 50 decreases the error by less than 1%. However, we observed that an increase in the number of iterations τ enforces the constraint even more by making the network sparser; and thus may lead to increasing the false negative rate. Additionally, the number of iterations that are required to run for the constraint present a trade off between the accuracy of the estimate and the available computational power.

Figure 2 shows a ten-gene time-varying network evolv-

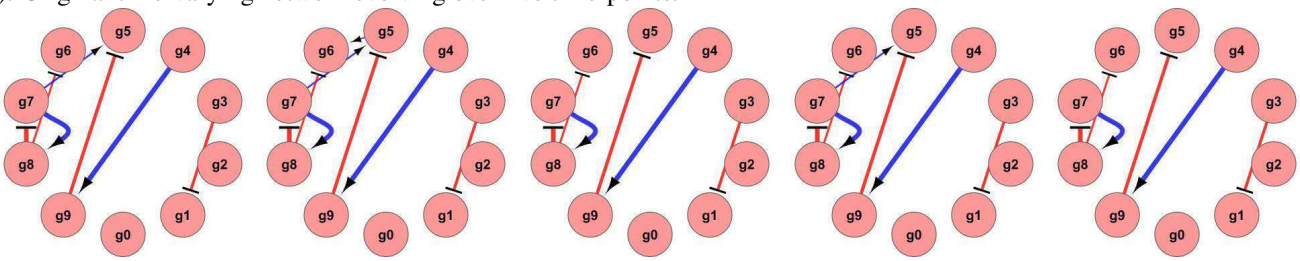


(a) (b) (c) (d)

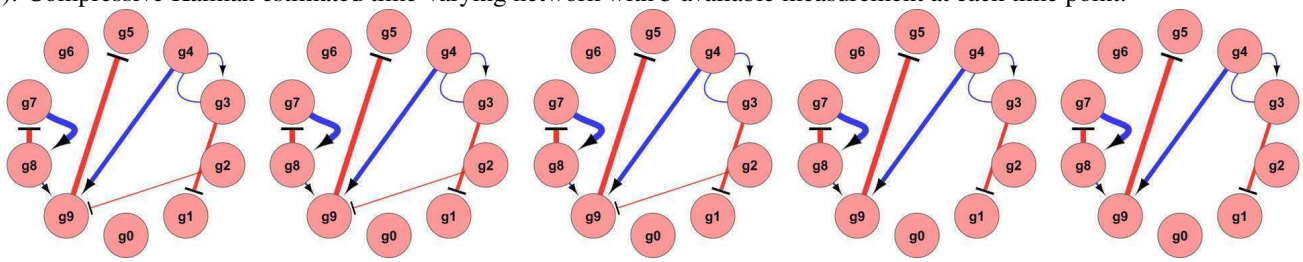
Fig. 1. Effect of algorithm parameters on the estimation accuracy and comparison with the projection method in [2]: (a) error vs. network size; (b) error vs. number of observations or measurements; (c) error vs. sparsity parameter ϵ; (d) error vs. number of iterations τ.



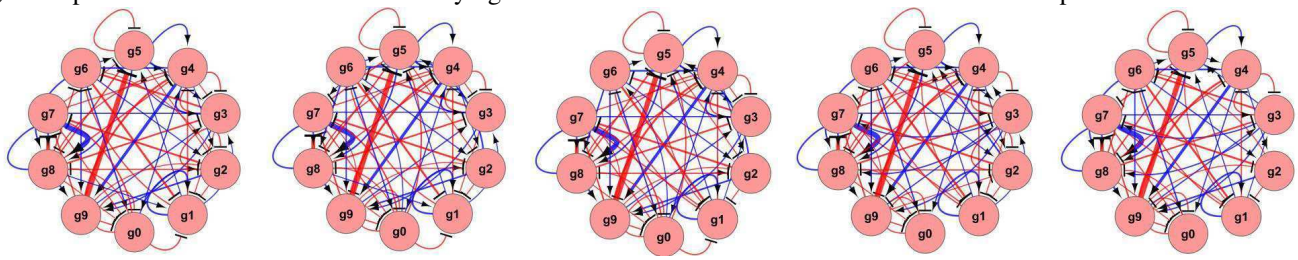
(a): Original time-varying network evolving over five time-points.



(b): Compressive Kalman estimated time-varying network with 5 available measurement at each time point.



(c): Compressive Kalman estimated time-varying network with 1 available measurement at each time point.



(d): Classical (unconstrained) Kalman estimated time-varying network.

Fig. 2. Tracking of a 10-gene time-varying network evolving over five time points.

ing over five time points. The compressive Kalman estimates of the network with five (resp. one) measurements at each time point is shown in the second (resp. third) row of Fig. 2. The classical (unconstrained) Kalman estimate is shown in the fourth row of Fig. 2. It is clear that compressive Kalman filtering is essential to obtain meaningful tracking of sparse time-varying genetic networks.

5. CONCLUSION

We formulated the problem of estimating genetic regulatory networks as a tracking problem of the network connectivity over time. It is well known that if the system is linear and observable, then the solution to this problem can be obtained using the Kalman filter. However, tracking of genetic networks is not an observable problem because the number of measurements is smaller than the number of genes. This issue is circumvented by taking into account the sparsity of the networks, and using a compressive sensing-based Kalman filter. We studied the effect of the algorithm on the estimation accuracy. We observed that the Kalman filter is robust to an increase in the number of genes, or equivalently an increase in the dimension of the state vector. The tracking results also depend on the sparsity parameter, which is a prior knowledge on the degree of sparsity of the network. Our simulations on synthetically generated time-varying networks show that the tracking performance is quite good even for one measurement at every time point. The performance also improves significantly with the number of measurements. At the same time, the unconstrained Kalman filter fails completely in giving any meaningful estimation or tracking of the network. Future research directions will explore tracking sparse genetic networks with non-linear system dynamics.

Acknowledgment

This project is supported by Award Number $R01GM096191$ from the National Institute Of General Medical Sciences (NIH/NIGMS). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute Of General Medical Sciences or the National Institutes of Health.

This work is also supported in part by the National Science Foundation under Grant CRI CNS-0855248, Grant EPS-0701890, Grant EPS-0918970, Grant MRI CNS-0619069, OISE-0729792, MRI 0722625, MRI-R2 0959124 and 0918970.

6. REFERENCES

[1] H M Fathallah-Shaykh, J L Bona, and S Kadener, "Mathematical model of the drosophila circadian clock: Loop regulation and transcriptional integration," *Biophysical Journal*, vol. 97, no. 9, pp. 2399–2408, November 2009.

[2] G Rasool, N Bouaynaya, H M Fathallah-Shaykh, and D Schonfeld, "Inference of genetic regulatory networks using regularized likelihood with covariance estimation," in *IEEE Statistical Signal Processing Workshop*, August 2012.

[3] D Marbach, R J Prill, T Schaffter, C Mattiussi, D Floreano, and G Stolovitzky, "Revealing strengths and weaknesses of methods for gene network inference," *Proceedings of the National Academy of Sciences (PNAS)*, vol. 107, no. 14, pp. 6286–6291, April 2010.

[4] A Rao, Alfred O Hero, David J States, and J D Engel, "Inferring time-varying network topologies from gene expression data," *EURASIP Journal on Bioinformatics and Systems Biology*, pp. 1–12, 2007.

[5] A Ahmed and E P Xing, "Recovering time-varying networks of dependencies in social and biological studies," *Proceedings of the National Academy of Sciences*, vol. 106, no. 29, pp. 11878–11883, July 2009.

[6] E E Kuruoğlu, X Yang, Y Xu, and T S Huang, "Time varying dynamic Bayesian network for nonstationary events modeling and online inference," *IEEE Transactions on Signal Processing*, vol. 59, no. 4, pp. 1553 – 1568, April 2011.

[7] E J Candes, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 489 – 509, February 2006.

[8] A Carmi, P Gurfil, and D Kanevsky, "Methods for sparse signal recovery using Kalman filtering with embedded pseudo-measurement norms and quasi-norms," *IEEE Transactions on Signal Processing*, vol. 58, no. 4, pp. 2405 – 2409, April 2010.

[9] Dan Simon, *Optimal State Estimation*, Wiley InterScience, 2006.

[10] D Simon, "Kalman filtering with state constraints: a survey of linear and nonlinear algorithms," *IET Control Theory & Applications*, vol. 4, no. 8, pp. 1303 – 1318, August 2010.

[11] Simon J Julier and J J LaViola Jr., "On Kalman filtering with nonlinear equality constraints," *IEEE Signal on Signal Processing*, vol. 55, no. 6, pp. 2774–2784, June 2007.

[12] J De Geeter, H Van Brussel, J De Schutter, and M Decreton, "A smoothly constrained Kalman filter," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 10, pp. 1171 – 1177, October 1997.