Imani and Braga-Neto *EURASIP Journal on Advances in Signal Processing* (2018) 2018:22
https://doi.org/10.1186/s13634-018-0543-y

EURASIP Journal on Advances in Signal Processing

RESEARCH Open Access

Gene regulatory network state estimation from arbitrary correlated measurements

Mahdi Imani* and Ulisses Braga-Neto

Abstract

Background: Advancements in gene expression technology allow acquiring cheap and abundant data for analyzing cell behavior. However, these technologies produce noisy, and often correlated, measurements on the transcriptional states of genes. The Boolean network model has been shown to be effective in capturing the complex dynamics of gene regulatory networks (GRNs). It is important in many applications, such as anomaly detection and optimal intervention, to be able to track the evolution of the Boolean states of a gene regulatory network using noisy time-series transcriptional measurements, which may be correlated in time.

Results: We propose efficient estimators for the Boolean states of GRNs using correlated time-series transcriptional measurements, where the nature of the correlation and of the measurements themselves are entirely arbitrary. More specifically, we propose new algorithms based on a hypothesis tree to compute optimal minimum mean square error (MMSE) filtering and smoothing state estimators for a Partially-Observed Boolean Dynamical System (POBDS) with correlated measurements. The algorithms are exact but may be computationally expensive for large state spaces or long time horizons, in which case a process for pruning the hypothesis tree is employed to obtain an approximation of the optimal MMSE estimators, while keeping computation tractable. Performance is assessed through a comprehensive set of numerical experiments based on the p53-MDM2 negative-feedback loop Boolean regulatory network, where the standard Boolean Kalman Filter (BKF) and Boolean Kalman Smoother (BKS) for uncorrelated measurements are compared to the corresponding new estimators for correlated measurements, called BKF-CORR and BKS-CORR, respectively.

Keywords: State estimation, Partially-observed Boolean dynamical system, Correlated measurement noise, Gene regulatory network, Boolean Kalman Filter and Smoother

1 Introduction

Gene regulatory networks (GRNs) govern the functioning of key cellular processes, such as the cell cycle, stress response, and DNA repair. Several mathematical models have been proposed to accurately capture the dynamical behavior of GRNs. These methods include Boolean networks [1–3], ordinary differential equations (OED) [4, 5], S-systems [6, 7], and Bayesian networks [8–10]. Boolean networks were first introduced as completely observable, deterministic models by Kauffman and collaborators [11, 12]. In a Boolean network, the transcriptional state of each gene is represented by 0 (OFF) or 1 (ON), and the relationship among genes is described by logical gates updated at discrete time intervals [13]. The Boolean network model has been successful in accurately modeling the dynamics of the cell cycle in the *Drosophila* fruit fly [14], the *Saccharomyces Cerevisiae* yeast [15], the mammalian cell cycle [16], and the switching behavior displayed by the p53 gene in tumor-suppressing pathways [17, 18]. Several variations of the original Boolean network models have been introduced in the literature to account the stochasticity in the behavior of gene regulatory networks. These models include Random Boolean Networks [1], Boolean Networks with perturbation (BNp) [19], Probabilistic Boolean Networks (PBN) [2], and Boolean Control Networks (BCN) [20, 21]. A key point is that all aforementioned models assume that the Boolean states of the system are directly observable. But, in practice, this is never the case. Modern transcriptional studies

*Correspondence: m.imani88@tamu.edu
Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843, USA

are based on technologies that produce noisy indirect measurements of gene activity, such as cDNA microarrays [22], RNA-seq [23], and cell imaging-based assays [24, 25].

The Partially-Observed Boolean Dynamical System (POBDS) model [3, 26] addresses the scenario encountered in practice in transcriptomic analysis by allowing for indirect and incomplete observation of gene states. The POBDS model is a special case of Hidden Markov Model (HMM) with Boolean state variables. The POBDS model unifies and generalizes most of the aforementioned Boolean network models. Several tools for the POBDS model have been developed in recent years, such as the optimal filter and smoother based on the minimum mean square error (MMSE) criterion, called the Boolean Kalman filter (BKF) [3, 26] and Boolean Kalman smoother (BKS) [3, 27], respectively; particle filter implementation of these filters [28]; fault detection [29]; optimal filter with correlated noise [30]; network inference [31]; sensor selection [32]; and control [33–38]. Most of these tools are freely available through an open-source R package called "BoolFilter" [39, 40].

All tools for estimation, identification, and control of POBDS have been built based on the assumption that the measurement noise is uncorrelated over time. However, this assumption may not hold in practice, due to the unavoidable measurement correlation existing in most real-world applications. The first development in this direction, for simple correlated binary measurement noise, was provided in [30]. However, in practice, the measurement space is never Boolean, but is in fact continuous-valued, such as in cDNA microarrays [22] and live cell imaging-based assays [24], or integer-valued, such as in RNA-seq data [41]. In this paper, we propose new algorithms based on a hypothesis tree to compute optimal MMSE filtering and smoothing state estimators for POBDS with arbitrary correlated measurements (Fig. 1). The proposed algorithms are exact, but, for large state spaces or long time horizons, computation is kept tractable by pruning the hypothesis tree, leading to an approximation of the optimal MMSE estimators. Performance is assessed through a comprehensive set of numerical experiments based on the p53-MDM2 negative-feedback loop Boolean regulatory network, where the standard Boolean Kalman Filter (BKF) and Boolean Kalman Smoother (BKS) for uncorrelated measurements are compared to the corresponding new estimators for correlated measurements, called BKF-CORR and BKS-CORR, respectively. In case there is no pruning, the BKF-CORR algorithm is equivalent to the filter estimator of [30] when the correlated observation noise is binary.

The article is organized as follows. In Section 2, the POBDS signal model with correlated observation noise is introduced. The proposed BKF-CORR and BKS-CORR estimators are developed in Sections 3.1 and 3.2, respectively. An instance of the POBDS model for gene regulatory networks observed through various sequencing technologies is discussed in Section 4. The performance of the proposed estimators is assessed in Section 5, through a comprehensive set of numerical experiments. Finally, Section 6 contains concluding remarks.

2 POBDS with correlated measurements

In this section, we introduce the model for a POBDS with correlated measurements. The model consists of a state model, which is the same as the one for an ordinary POBDS, and an observation model with general autoregressive measurement noise.

2.1 State model

The system is described by a state process $\mathbf{X}_k; k = 0, 1, \ldots$, where $\mathbf{X}_k \in \{0,1\}^d$ is a Boolean vector describing the activation/inactivation state of d genes at time k. The state is assumed to be updated at time k through the following nonlinear signal model

$$\mathbf{X}_k = \mathbf{f}\left(\mathbf{X}_{k-1}, \mathbf{u}_k\right) \oplus \mathbf{n}_k\,, \tag{1}$$

for $k = 1, 2, \ldots$, where $\mathbf{u}_k \in \{0,1\}^d$ is an input at time k, $\mathbf{f} : \{0,1\}^{2d} \rightarrow \{0,1\}^d$ is a Boolean function called

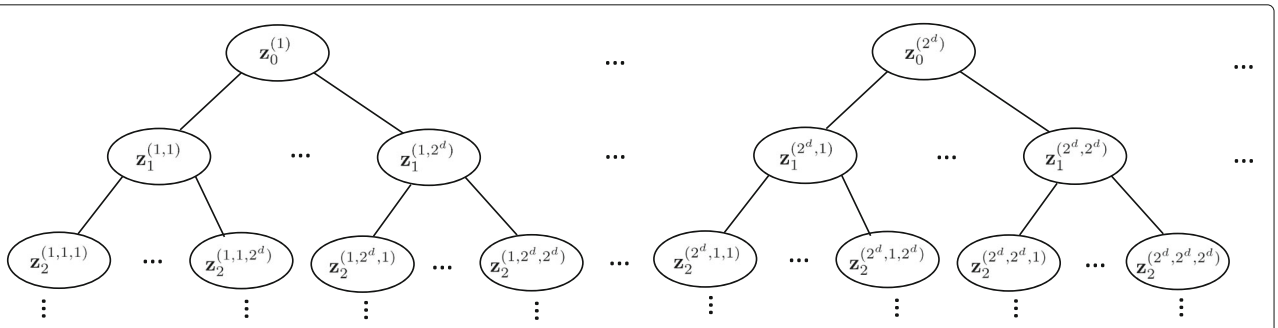

Fig. 1 Full hypothesis tree for a system with d Boolean variables

the network function, "$\oplus$" indicates the componentwise modulo-2 addition, and $\mathbf{n}_k \in \{0,1\}^d$ is the Boolean transition noise. The noise process $\{\mathbf{n}_k; k = 1,2,\ldots\}$ is assumed to be "white" in the sense that the noise at distinct time points is an independent random variable. We also assume that noise process is independent of the initial state $\mathbf{X}_0$ and the input sequence $\{\mathbf{u}_k; k = 1,2,\ldots\}$ is deterministic and known.

2.2 Observation model

Let $\mathbf{Y}_k$ be a vector containing the measurements at time k,

$$\mathbf{Y}_k = \mathbf{h}(\mathbf{X}_k, \mathbf{v}_k), \tag{2}$$

for $k = 1,2,\ldots$, where $\mathbf{v}_k$ is the measurement noise at time step k. We assume that $\{\mathbf{v}_k; k = 1,2,\ldots\}$ has a general autoregressive structure of the form

$$\mathbf{v}_k = \mathbf{g}(\mathbf{v}_{k-1}, \mathbf{w}_k), \tag{3}$$

where $\{\mathbf{w}_k; k = 1,2,\ldots\}$ is a white measurement noise process and $\mathbf{g}$ specifies the relationship between $\mathbf{v}_k$ and $\mathbf{v}_{k-1}$. The initial value of the noise is set to zero, i.e., $\mathbf{v}_0 = \mathbf{0}$.

For a given measurement $\mathbf{Y}_k$ and known Boolean state $\mathbf{X}_k$, we assume that there is a unique value of the measurement noise $\mathbf{v}_k$ that is accessible through a known mapping:

$$\mathbf{v}_k = \mathbf{r}(\mathbf{Y}_k, \mathbf{X}_k). \tag{4}$$

For example, in the case of simple additive noise, $\mathbf{Y}_k = \mathbf{X}_k + \mathbf{v}_k$, the inverse mapping would be $\mathbf{r}(\mathbf{Y}_k, \mathbf{X}_k) = \mathbf{Y}_k - \mathbf{X}_k$.

3 Proposed estimators

In this section, we describe the new algorithms for computing the optimal MMSE filter and smoother for a POBDS with correlated observations.

3.1 BKF-CORR

The optimal minimum mean square error (MMSE) filtering problem consists of, given observations $\mathbf{Y}_{1:k} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_k)$, finding an estimator $\hat{\mathbf{X}}_{k|k}$ of the state $\mathbf{X}_k$ that minimizes

$$\mathrm{MSE}\left(\hat{\mathbf{X}}_{k|k} \mid \mathbf{Y}_{1:k}\right) = E\left[\left|\left|\hat{\mathbf{X}}_{k|k} - \mathbf{X}_k\right|\right|^2 \mid \mathbf{Y}_{1:k}\right], \tag{5}$$

where $||.||$ denotes the usual L_2 vector norm. For a vector $\mathbf{v}$ of size d, define $\overline{\mathbf{v}} \in \{0,1\}^d$ via $\overline{\mathbf{v}}(i) = I_{\mathbf{v}(i)>1/2}$ for $i = 1,\ldots,d$. It has been shown ([3], Thm. 1) that

$$\hat{\mathbf{X}}^{\mathrm{MS}}_{k|k} = \overline{E\left[\mathbf{X}_k \mid \mathbf{Y}_{1:k}\right]} = \overline{\sum_{i_k \in I} P\left(\mathbf{X}_k = \mathbf{x}^{i_k} \mid \mathbf{Y}_{1:k}\right) \mathbf{x}^{i_k}}, \tag{6}$$

where $I = \{1,\ldots,2^d\}$ and $\left(\mathbf{x}^1,\ldots,\mathbf{x}^{2^d}\right)$ is an arbitrary enumeration of the possible Boolean state vectors.

For the standard POBDS model defined by (1)–(2) with *uncorrelated* observation noise ("white noise"), the previous estimator can be computed exactly by a recursive matrix-based algorithm, called the Boolean Kalman filter (BKF) [26]. It is our purpose in this section to derive an algorithm to accurately and efficiently compute this estimator in the case of the correlated noise model defined by (3)–(4). Computation is based on a hypothesis tree and is exact, but an approximate version of the estimator is also proposed for large state spaces or long time horizons, based on pruning the hypothesis tree.

Consider a new "state" vector $\mathbf{Z}_k = [\mathbf{X}_k, \mathbf{v}_k]^T$ consisting of the pair of state vector and observation noise and corresponding "transition" noise vector $\boldsymbol{\eta} = [\mathbf{n}_k, \mathbf{w}_k]$, which leads to the "state" model

$$\mathbf{Z}_k = \begin{bmatrix} \mathbf{X}_k \\ \mathbf{v}_k \end{bmatrix} = \begin{bmatrix} \mathbf{f}\left(\mathbf{X}_{k-1}, \mathbf{u}_k\right) \oplus \mathbf{n}_k \\ \mathbf{g}(\mathbf{v}_{k-1}, \mathbf{w}_k) \end{bmatrix} = \mathbf{q}(\mathbf{Z}_{k-1}, \boldsymbol{\eta}_{k-1}) \tag{7}$$

with observation model

$$\mathbf{Y}_k = \mathbf{h}(\mathbf{X}_k, \mathbf{v}_k) = \mathbf{h}(\mathbf{Z}_k). \tag{8}$$

Our approach is to compute $P(\mathbf{X}_k \mid \mathbf{Y}_{1:k})$ based on the probabilities of all possible realizations of the state trajectory $(\mathbf{Z}_0, \mathbf{Z}_1, \ldots, \mathbf{Z}_k)$ given the data $\mathbf{Y}_{1:k}$, which allows the computation of the optimal MMSE filter in (6).

The trajectories can be arranged in a hypothesis tree containing pairs. At time $k = 0$, using the fact that $\mathbf{v}_0 = \mathbf{0}$, there are 2^d possible realizations

$$\mathbf{z}_0^{(i)} = \left(\mathbf{X}_0 = \mathbf{x}^i, \mathbf{v}_0 = \mathbf{0}\right), \tag{9}$$

with probabilities

$$\pi_{0|0}^{(i)} = P\left(\mathbf{Z}_0 = \mathbf{z}_0^{(i)}\right) = P\left(\mathbf{X}_0 = \mathbf{x}^i\right), \tag{10}$$

for $i \in I = \{1,\ldots,2^d\}$. At time $k = 1$, each pair in (9) leads to 2^d additional pairs

$$\mathbf{z}_1^{(i,j)} = \left(\mathbf{X}_1 = \mathbf{x}^j, \mathbf{v}_1 = \mathbf{r}\left(\mathbf{Y}_1, \mathbf{x}^j\right)\right), \tag{11}$$

for $(i,j) \in I_2 = I \times I$, where we used the relationship (4). Each of these $2^d \times 2^d = 2^{2d}$ pairs corresponds to the terminal point of a unique trajectory $\left\{\mathbf{z}_0^{(i)}, \mathbf{z}_1^{(i,j)}\right\}$ through time $k = 1$. The probability of this trajectory is

$$\begin{aligned} \pi_{1|1}^{(i,j)} &= P\left(\mathbf{Z}_1 = \mathbf{z}_1^{(i,j)}, \mathbf{Z}_0 = \mathbf{z}_0^{(i)} \mid \mathbf{Y}_1\right) \\ &= P\left(\mathbf{Z}_1 = \mathbf{z}_1^{(i,j)} \mid \mathbf{Z}_0 = \mathbf{z}_0^{(i)}\right) P\left(\mathbf{Z}_0 = \mathbf{z}_0^{(i)}\right) \\ &= P\left(\mathbf{X}_1 = \mathbf{x}^j \mid \mathbf{X}_0 = \mathbf{x}^i\right) p\left(\mathbf{v}_1 = \mathbf{r}(\mathbf{Y}_1, \mathbf{x}^j)\right) \mid \mathbf{v}_0 = \mathbf{0}) \, \pi_{0|0}^{(i)}, \end{aligned} \tag{12}$$

for $(i,j) \in I_2$.

At time k, there are $2^{(k+1)d}$ pairs

$$\mathbf{z}_k^{(i_0,i_1\ldots,i_k)} = \left(\mathbf{X}_k = \mathbf{x}^{i_k}, \mathbf{v}_k = \mathbf{r}\left(\mathbf{Y}_k, \mathbf{x}^{i_k}\right)\right). \tag{13}$$

The probability of each of the unique $2^{(k+1)d}$ trajectories $\left\{\mathbf{z}_0^{(i_0)}, \mathbf{z}_1^{(i_0,i_1)}, \ldots, \mathbf{z}_k^{(i_0,i_1\ldots,i_k)}\right\}$ through time k can be computed recursively as

$$\begin{aligned}\pi_{k|k}^{(i_0,i_1\ldots,i_k)} &= P\left(\mathbf{Z}_k = \mathbf{z}_k^{(i_0,i_1\ldots,i_k)}, \ldots, \mathbf{Z}_0 = \mathbf{z}_0^{(i_0)} \mid \mathbf{Y}_{1:k}\right)\\ &= P\left(\mathbf{Z}_k = \mathbf{z}_k^{(i_0,i_1\ldots,i_k)} \mid \mathbf{Z}_{k-1} = \mathbf{z}_{k-1}^{(i_0,i_1\ldots,i_{k-1})}\right)\\ &\quad P\left(\mathbf{Z}_{k-1} = \mathbf{z}_{k-1}^{(i_0,i_1\ldots,i_{k-1})} \mid \mathbf{Y}_{1:k-1}\right)\\ &= P\left(\mathbf{X}_k = \mathbf{x}^{i_k} \mid \mathbf{X}_{k-1} = \mathbf{x}^{i_{k-1}}\right) p\left(\mathbf{v}_k = \mathbf{r}\left(\mathbf{Y}_k, \mathbf{x}^{i_k}\right) \mid \mathbf{v}_{k-1}\right.\\ &\quad \left. = \mathbf{r}\left(\mathbf{Y}_{k-1}, \mathbf{x}^{i_{k-1}}\right)\right) \pi_{k-1|k-1}^{(i_0,i_1\ldots,i_{k-1})}.\end{aligned} \tag{14}$$

for $(i_0, i_1 \ldots, i_k) \in I_{k+1}$, where $I_k = I \times \cdots \times I$ (k times). Since the state and noise transition probabilities, $P(\mathbf{X}_k \mid \mathbf{X}_{k-1})$ and $p(\mathbf{v}_k \mid \mathbf{v}_{k-1})$, are assumed to be known, this provides an efficient way to recursively compute the probability of all trajectories.

Now, since the event $\left[\mathbf{X}_k = \mathbf{x}^{i_k}\right]$ is equal to the disjoint union of all trajectories that end at $\mathbf{X}_k$ at time k, it is clear that the conditional probability $P\left(\mathbf{X}_k = \mathbf{x}^{i_k} \mid \mathbf{Y}_{1:k}\right)$ is equal to the sum of the conditional probabilities of those trajectories:

$$P\left(\mathbf{X}_k = \mathbf{x}^{i_k} \mid \mathbf{Y}_{1:k}\right) = \sum_{(i_0,\ldots,i_{k-1})\in I_k} \pi_{k|k}^{(i_0,i_1\ldots,i_k)} \tag{15}$$

for $i_k \in I$. Substituting this in (6) allows us to write the optimal MMSE estimator simply as

$$\hat{\mathbf{X}}_{k|k}^{\mathrm{MS}} = \overline{\sum_{(i_0,i_1\ldots,i_k)\in I_{k+1}} \pi_{k|k}^{(i_0,i_1\ldots,i_k)}\, \mathbf{x}^{i_k}}. \tag{16}$$

However, one can easily appreciate that the number of trajectories will quickly become intractable as the number of genes d and the horizon k increase. For example, for a network with eight genes, there will be $2^{40} = 1.1 \times 10^{12}$ trajectories after only $k = 4$ time points. To make the computation feasible, at each time k, we prune the trajectories with probability smaller than a threshold $\epsilon > 0$, by removing the corresponding pairs $(i_0, i_1 \ldots, i_k)$ from the index set I_{k+1}. The probabilities of the surviving trajectories are renormalized to add up to one, and the state estimator in (16) is computed on the reduced index set. Then, the surviving nodes are expanded, and the process is repeated. A larger value of ϵ results in more computational savings and a faster estimator, but at an increased loss of accuracy, and vice-versa. The resulting filter is called the BKF-CORR estimator. The effect of ϵ on the performance of the BKF-CORR estimator is investigated in Section 5.

3.2 BKS-CORR

The optimal filter uses the data $\mathbf{Y}_{1:k}$ observed up to the current time k to estimate the state at the current time k. By contrast, the (fixed-interval) smoother uses data $\mathbf{Y}_{1:T}$ that have been collected and stored "off-line" up to time T to estimate the states at any time point in the interval $0 \leq k \leq T$.

In Fig. 2a, it can be seen that the filtering process needs only a forward step for estimating the state at the last time point. In contrast, the smoothing process presented in Fig. 2b requires both forward and backward processes for state estimation over the fixed interval.

Given observations $\mathbf{Y}_{1:T}$, the optimal MMSE (fixed-interval) smoothing problem consists of finding an estimator $\hat{\mathbf{X}}_{k|T}$ of the state $\mathbf{X}_k$, for $0 < k < T$, which minimizes

$$\mathrm{MSE}(\hat{\mathbf{X}}_{k|T} \mid \mathbf{Y}_{1:T}) = E\left[\left\|\hat{\mathbf{X}}_{k|T} - \mathbf{X}_k\right\|^2 \mid \mathbf{Y}_{1:T}\right], \tag{17}$$

It can be shown that the solution is

$$\hat{\mathbf{X}}_{k|T}^{\mathrm{MS}} = \overline{E\left[\mathbf{X}_k \mid \mathbf{Y}_{1:T}\right]} = \overline{\sum_{i_k\in I} P\left(\mathbf{X}_k = \mathbf{x}^{i_k} \mid \mathbf{Y}_{1:T}\right)\, \mathbf{x}^{i_k}}, \tag{18}$$

It is instructive to compare the previous two equations to (5) and (6), respectively. For the standard POBDS model with uncorrelated observation noise, the estimator in (18)

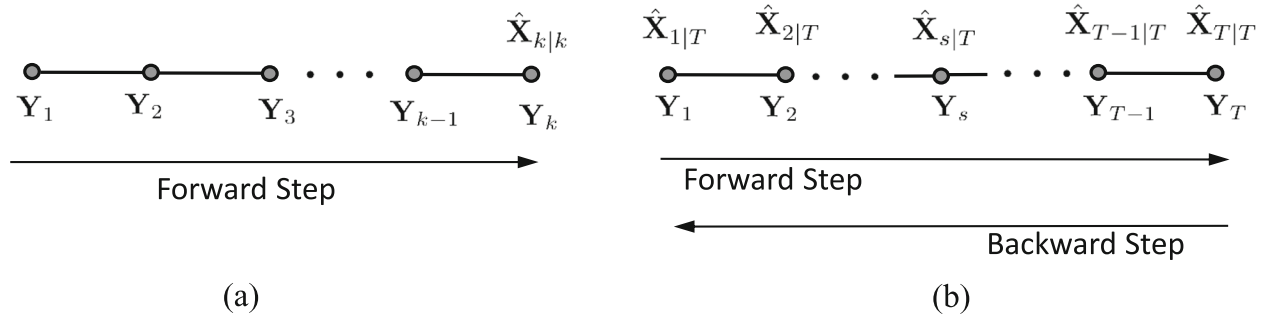

Fig. 2 Schematic diagram of filtering and smoothing processes. **a** Filter. **b** Fixed-interval smoother

can be computed exactly by a matrix-based algorithm, called the Boolean Kalman Smoother (BKS) [3, 27]. In this section, an exact MMSE smoother for a POBDS with correlated measurement defined by (3)–(4) is proposed.

The proposed smoother, called the BKS-CORR estimator, contains forward and backward steps. In the forward process, given a sequence of measurements $\mathbf{Y}_{1:T}$, one runs the proposed filter in Section 3.1 from time 0 to T to compute the filtering trajectories and their associated probabilities. Then, the backward process uses those values in a recursive fashion to compute the smoothed state estimate.

The filter at time step T creates $2^{(T+1)d}$ unique trajectories $\left\{\mathbf{z}_0^{(i_0)},\ldots,\mathbf{z}_T^{(i_0,i_1,\ldots,i_T)}\right\}$ with associated probabilities $\pi_{T|T}^{(i_0,i_1,\ldots,i_T)}$, for $(i_0,i_1,\ldots,i_T)\in I_{T+1}$. Clearly, the filtering and smoothing solutions in the last time step (at time step T) are the same. One can obtain the smoothed estimator by first computing the following smoothed posterior probabilities using the forward trajectories:

$$\begin{aligned}\pi_{T-1|T}^{(i_0,\ldots,i_{T-1})} &= P\left(\mathbf{Z}_{T-1}=\mathbf{z}_{T-1}^{(i_0,\ldots,i_{T-1})},\ldots,\mathbf{Z}_0=\mathbf{z}_0^{(i_0)}\mid\mathbf{Y}_{1:T}\right)\\ &=\sum_{i_T\in I}P\left(\mathbf{Z}_T=\mathbf{z}_T^{(i_0,\ldots,i_{T-1},i_T)},\mathbf{Z}_{T-1}=\mathbf{z}_{T-1}^{(i_0,\ldots,i_{T-1})}\right.\\ &\qquad\left.,\ldots,\mathbf{Z}_0=\mathbf{z}_0^{(i_0)}\mid\mathbf{Y}_{1:T}\right)\\ &=\sum_{i_T\in I}\pi_{T|T}^{(i_0,\ldots,i_{T-1},i_T)},\end{aligned}\tag{19}$$

for $(i_0,\ldots,i_{T-1})\in I_T$. The process can be repeated to compute the smoothed probability backwards to any desired time step via

$$\pi_{k-1|T}^{(i_0,\ldots,i_{k-1})}=\sum_{i_k\in I}\pi_{k|T}^{(i_0,\ldots,i_{k-1},i_k)},\tag{20}$$

for $(i_0,\ldots,i_{k-1})\in I_k$ and $k=1,\ldots,T$. The optimal MMSE smoother at time k can then be computed as

$$\hat{\mathbf{X}}_{k|T}^{\mathrm{MS}}=\overline{E[\mathbf{X}_k\mid\mathbf{Y}_{1:T}]}=\overline{\sum_{(i_0,i_1,\ldots,i_k)\in I_{k+1}}\pi_{k|T}^{(i_0,\ldots,i_k)}\mathbf{x}^{i_k}}.\tag{21}$$

The pruning process to make computation efficient is done in the forward process only, by using the same process described in the previous section.

4 Partially observed gene regulatory networks

In this section, we describe a specific instance of the POBDS model with correlated measurements in (1)–(3), which allows the application of the proposed BKF-CORR and BKS-CORR estimators to Boolean gene regulatory networks observed through noisy correlated gene-expression data.

4.1 Gene regulatory network state model

The state model adopted here is motivated by gene pathway diagrams commonly encountered in biomedical research, in which genes act to activate or inhibit the activity of other genes. The network function in (1) is expressed in component form as $\mathbf{f}=(f_1,\ldots,f_d)$, where each component $f_i:\{0,1\}^{2d}\to\{0,1\}$ is a Boolean function given by

$$f_i(\mathbf{x},\mathbf{u})=\begin{cases}1, & \sum_{j=1}^{d}a_{ij}\mathbf{x}(j)+b_i+\mathbf{u}(i)>0,\\ 0, & \sum_{j=1}^{d}a_{ij}\mathbf{x}(j)+b_i+\mathbf{u}(i)\le 0,\end{cases}\tag{22}$$

where a_{ij} and b_i are the system parameters. The former can take three values: $a_{ij}=+1$ if there is positive regulation (activation) from gene j to gene i; $a_{ij}=-1$ if there is negative regulation (inhibition) from gene j to gene i; and $a_{ij}=0$ if gene j is not an input to gene i. The latter specifies regulation *biases* and can take two values: $b_i=+1/2$ if gene i is positively biased, in the sense that an equal number of activation and inhibition inputs will produce activation, and the reverse being the case if $b_i=-1/2$. The proposed network function is depicted in Fig. 3, where the threshold units are step functions that output 1 if the input is nonnegative, and 0, otherwise.

The process noise $\mathbf{n}_k$ in (1) is assumed to have independent components distributed as Bernoulli(p), where the noise parameter p gives the amount of "perturbation" to the Boolean state process; the closer it is to $p=0.5$, the more chaotic the system will be, while a value of p close to zero means that the state trajectories are nearly deterministic, being governed tightly by the network function. From (1), the transition probabilities $P\left(\mathbf{X}_k=\mathbf{x}^i\mid\mathbf{X}_{k-1}=\mathbf{x}^j\right)$ of the state process, required for computation of the hypothesis tree probabilities in (14), take the form

$$\begin{aligned}P\left(\mathbf{X}_k=\mathbf{x}^i\mid\mathbf{X}_{k-1}=\mathbf{x}^j\right)&=P\left(\mathbf{n}_k=\mathbf{f}(\mathbf{x}^j,\mathbf{u})\oplus\mathbf{x}^i\right)\\ &=p^{||\mathbf{f}(\mathbf{x}^j,\mathbf{u})\oplus\mathbf{x}^i||_1}(1-p)^{d-||\mathbf{f}(\mathbf{x},\mathbf{u})\oplus\mathbf{x}^i||_1},\end{aligned}\tag{23}$$

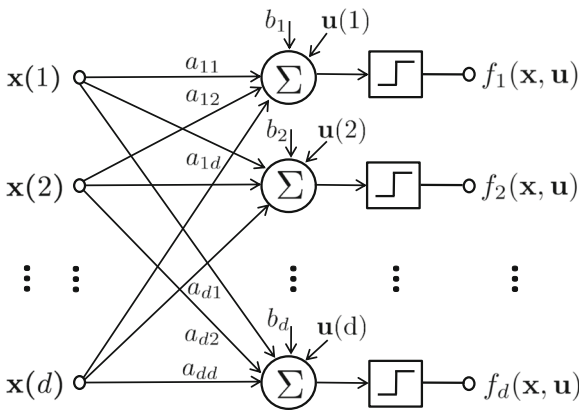

Fig. 3 Network model

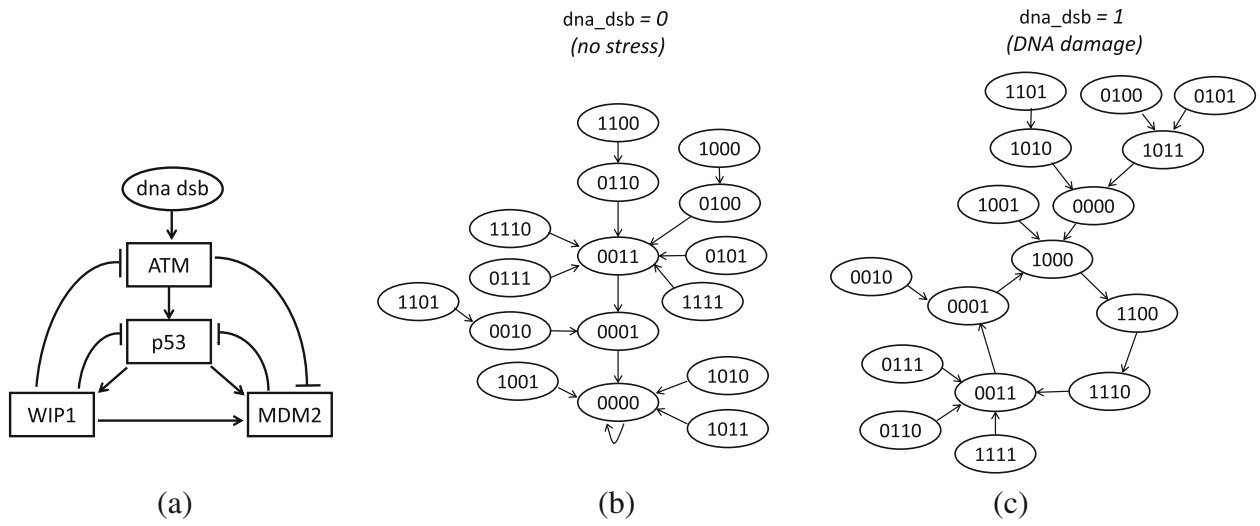

Fig. 4 a–c Activation/repression pathway diagram and state transition diagrams corresponding to a constant input dna_dsb = 0 (no-stress) and dna_dsb = 1 (DNA-damage) for the P53-MDM2 negative-feedback loop Boolean network model with negative regulation biases

for $i, j = 1, \ldots, 2^d$, where $||\mathbf{x}||_1$ denotes the number of 1's in the Boolean vector $\mathbf{x}$.

4.2 Gene-expression observation model

We employ here an additive Gaussian noise observation model even though the methodology proposed in the paper is entirely general and could be applied in principle to any observation model satisfying constraints (3) and (4). A Gaussian model is appropriate for modeling gene-expression data from technologies such as cDNA microarrays [22] and live cell imaging-based assays [24], in which gene expression measurements are continuous and unimodal (within a single population of interest) [42–45]. Let $\mathbf{Y}_k = (\mathbf{Y}_k(1), \ldots, \mathbf{Y}_k(d))$ be a vector containing the measurements at time k, for $k = 1, 2, \ldots$. The component $\mathbf{Y}_k(j) \in \mathbb{R}$ is the abundance measurement corresponding to transcript j, which is modeled as

$$\mathbf{Y}_k(j) = \mu_j^0 \left(1 - \mathbf{X}_k(j)\right) + \mu_j^1 \mathbf{X}_k(j) + \mathbf{v}_k(j), \qquad (24)$$

for $j = 1, \ldots, d$, where the parameters μ_j^0 and μ_j^1 specify the mean abundance of transcript j in the inactivated and activated states, respectively, and $\{\mathbf{v}_k; k = 1, 2, \ldots\}$ is the measurement noise process, with a standard AR(1) structure

$$\mathbf{v}_k = \eta\, \mathbf{v}_{k-1} + (1 - \eta)\, \mathbf{w}_k, \qquad (25)$$

where $0 \leq \eta \leq 1$ is a correlation parameter, and $\{\mathbf{w}_k; k = 1, 2, \ldots\}$ is a multivariate zero-mean white Gaussian noise process, with $\mathbf{w}_k \sim \mathcal{N}(0, \Sigma_k)$. The value $\eta = 0$ corresponds to uncorrelated observation noise, where as $\eta = 1$ corresponds to maximum correlation. Clearly, the conditional distribution $\mathbf{v}_k \mid \mathbf{v}_{k-1}$, required to compute the hypothesis tree probabilities in (14), is a multivariate Gaussian $N(\mathbf{v}_k, \Sigma_k)$.

Table 1 Parameter values used in the numerical experiments

Parameter	Value
Initial distribution $P(\mathbf{X}_0)$	$(1/16, \ldots, 1/16)$
Pruning parameter ϵ	0.01, 0.05, 0.0.1
Correlation parameter η	0.25, 0.50, 0.75
Mean in inactivated state μ^0	40
Mean in activated state μ^1	60
Standard deviation σ	10, 15

5 Results and discussion

In this section, we present the results of detailed numerical experiments to assess the performance of the proposed BKF-CORR and BKS-CORR estimators. We base our experiments on the well-known p53-MDM2 negative-feedback gene regulatory network [17, 18]. The *p53* gene codes for the tumor suppressor protein p53 in humans, and its activation plays a critical role in cellular responses to various stress signals that might cause genome instability. The gene regulatory network consists of four genes, *ATM*, *p53*, *Wip1*, and *MDM2*, and the input "dna_dsb," which indicates the presence of DNA double strand breaks.

The pathway diagram for this network is presented in Fig. 4a. We can see that *ATM* is the transductor gene for the DNA damage signal, which eventually activates *p53* through inactivation of *MDM2*. However, there is also a negative-feedback loop between *p53* and *ATM* through *Wip1*, so that *p53* is expected to display an oscillatory

Table 2 Average correct state estimation rates over 1000 independent runs for time series with length $T = 40$

			No-stress				DNA-damage			
p	σ	η	BKF	BKF-CORR	BKS	BKS-CORR	BKF	BKF-CORR	BKS	BKS-CORR
0.01	10	0.25	0.89	0.92	0.93	0.96	0.88	0.91	0.96	0.97
		0.50	0.86	0.91	0.88	0.95	0.78	0.91	0.87	0.97
		0.75	0.86	0.91	0.88	0.95	0.44	0.90	0.42	0.97
	15	0.25	0.87	0.89	0.90	0.92	0.83	0.86	0.93	0.95
		0.50	0.86	0.88	0.88	0.92	0.69	0.84	0.81	0.94
		0.75	0.86	0.88	0.88	0.92	0.43	0.85	0.47	0.94
0.05	10	0.25	0.73	0.78	0.79	0.84	0.71	0.76	0.83	0.88
		0.50	0.64	0.77	0.68	0.84	0.53	0.75	0.62	0.87
		0.75	0.59	0.77	0.62	0.84	0.35	0.76	0.35	0.87
	15	0.25	0.65	0.67	0.70	0.73	0.58	0.63	0.71	0.75
		0.50	0.61	0.67	0.64	0.73	0.48	0.62	0.58	0.75
		0.75	0.60	0.63	0.68	0.74	0.31	0.62	0.32	0.75

behavior under DNA damage [17]. On the other hand, under no stress, it is known that all four proteins are inactivated in the steady state [46].

These behaviors are captured nicely by the gene regulatory network model proposed in Section 4.1. Letting the state vector be $\mathbf{X}_k = (ATM, p53, Wip1, MDM2)$, the gene interaction parameters a_{ij} can be read off Fig. 4a:

$$\begin{aligned} &a_{11} = 0, \quad a_{12} = 0, \quad a_{13} = -1, \quad a_{14} = 0 \\ &a_{21} = +1, \quad a_{22} = 0, \quad a_{23} = -1, \quad a_{24} = -1 \\ &a_{31} = 0, \quad a_{32} = +1, \quad a_{33} = 0, \quad a_{34} = 0 \\ &a_{41} = -1, \quad a_{42} = +1, \quad a_{43} = +1, \quad a_{44} = 0 \end{aligned} \tag{26}$$

The input vector is $\mathbf{u}_k = (\text{dna\_dsb}, 0, 0, 0)$ and is assumed here to be held constant at one of its possible two values: DNA damage, $\mathbf{u}_k = (1, 0, 0, 0)$, or no stress, $\mathbf{u}_k = (0, 0, 0, 0)$, for $k = 1, 2, \ldots$. We assume negative regulation biases, $b_i = -1/2$, for $i = 1, \ldots, d$. This leads to two state transition diagrams, corresponding to each possible value of the input dna_dsb, which are depicted in Fig. 4b, c). We can see that under no-stress, "0000" is a singleton attractor state, while the other states are transient; on the other hand, under DNA damage, there is a cyclic attractor corresponding to an oscillation of p53 along with the other proteins in its regulatory pathway. This reproduces the known biological behavior described previously.

The mean expressions for activated and inactivated genes are assumed to be the same for all genes, with values μ_0 and μ_1, respectively, specified in Table 1. In addition, the covariance matrix for the noise $\mathbf{w}_k$ is assumed to be constant and equal to $\Sigma = \sigma^2 I_d^2$, with the value of σ specified in Table 1.

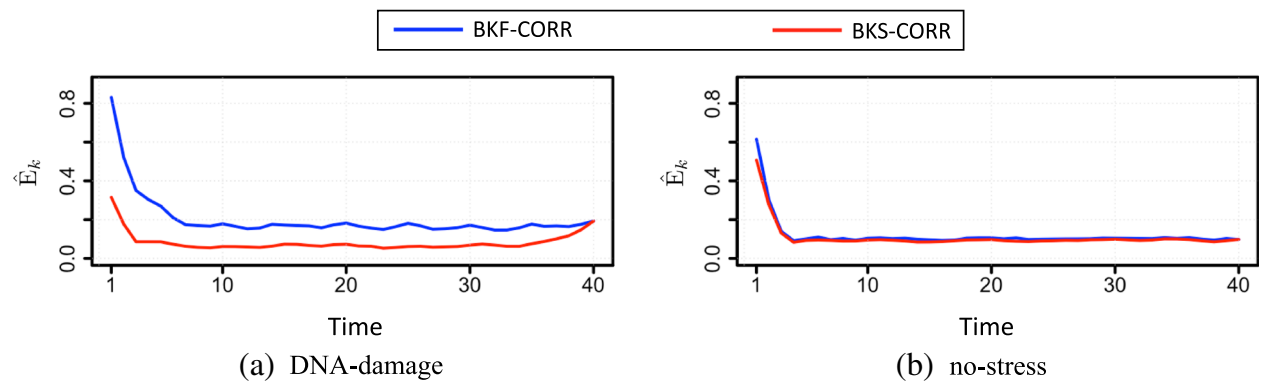

Fig. 5 Average correct state estimation rates for the BKF-CORR and BKS-CORR estimators under **a** DNA-damage and **b** no-stress conditions

Table 3 Effect of the pruning parameter on performance and running time

	BKF-CORR		BKS-CORR	
ϵ	Est. rate	Time	Est. rate	Time
0.01	0.76	10.33	0.88	15.76
0.05	0.72	5.41	0.84	8.56
0.10	0.62	1.72	0.73	2.34

Table 2 displays the average rate of correct state estimation for the standard BKF and BKS, which are optimal for uncorrelated noise but suboptimal in this case. The pruning parameter is set to be $\epsilon = 0.01$. As expected, the performance of the BKF-CORR and BKS-CORR estimators is better than that of the BKF and BKS estimators in various cases. As expected, the difference is more obvious for larger correlated noise.

Performance across the board is worse in the presence of large process and measurement noises. One can also see that better estimation is obtained in the "no-stress" condition in comparison to "DNA-damage" case. This can be explained by the attractor structure of each system, shown in Fig. 4b, c. Under no-stress, the system spends a significant amount of time in the rest state 0000, whereas under DNA damage, more states are visited due to the cyclic attractor, which makes the state estimation process more challenging.

Figure 5 displays the average correct state estimation rates E_k over 40 time steps using 1000 independent runs, defined as

$$\hat{\mathrm{E}}_k = \sum_{i=1}^{1000} ||\hat{\mathbf{X}}_{k,i} \oplus \mathbf{X}_{k,i}||_1 \,, \tag{27}$$

for $k = 1, \ldots, 40$, where $\hat{\mathbf{X}}_{k,i}$ is the estimate of the true state $\mathbf{X}_{k,i}$ in the ith iteration. The error $\hat{E}_k$ takes a value between 0 and 1. When $\hat{E}_k$ is close to 0, the proposed estimator has accurately estimated the transcriptional state of all genes at time step k over all independent runs. By contrast, a value of $\hat{E}_k$ close to 1 corresponds to the maximum possible estimation error at time step k. For the plot in Fig. 5, the process noise intensity, pruning parameter, and correlation rate are assumed to be $p = 0.01$, $\epsilon = 0.1$, and $\eta = 0.1$, respectively. The standard deviation of the measurement noise is also assumed to be $\sigma = 10$. In both cases, the BKF-CORR and BKS-CORR estimators have performed accurately, leading to small average estimation error. However, the BKS-CORR estimator has smaller error on average in comparison to the BKF-CORR estimator throughout the interval. This is due to the fact that the smoother uses future observations, but the filter uses only the observations up to the present time. The average estimation error is larger in the early

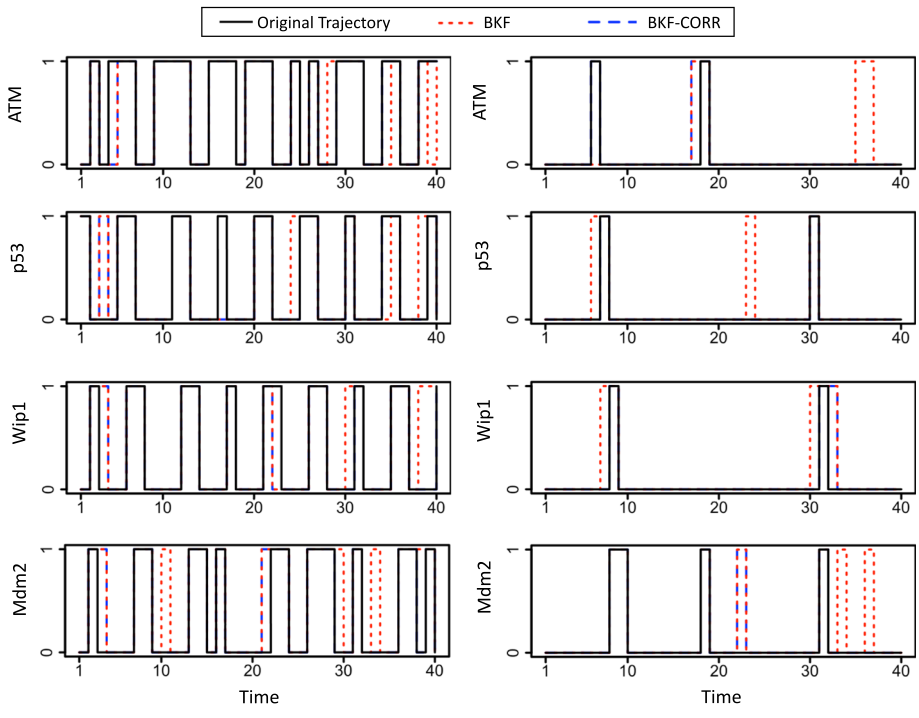

Fig. 6 Original and estimated state trajectories for **a** no-stress and **b** DNA-damage conditions

steps, due to the initial uniform distribution assumed over the Boolean states. However, as time goes on, the average error quickly becomes small. One can notice that the difference between the average estimation errors of the BKF-CORR and BKS-CORR estimators is larger in the presence of DNA damage. This can be justified by the fact that the p53-MDM2 network in the presence of the DNA damage has a cyclic attractor (see Fig. 4), as opposed to the no-stress condition in which "0000" is a singleton attractor. Clearly, the estimation in the presence of the cyclic attractor is more challenging than that of a singleton attractor. Thus, the use of future data in the smoothing process makes the estimation process more accurate in the middle of the interval. Finally, as expected, at the end of the horizon (i.e., $k = T$), the filter and smoother are equivalent (since no future data is available), and as a result, the same average error can be seen for both estimators in that case.

Next, the effect of the pruning parameter on performance and computational time is examined. Table 3 displays the average correct estimation rate and running time of the proposed methods for different pruning parameters, computed over 1000 independent runs for sequences of length $T = 40$. The process noise intensity and the standard deviation of measurements are assumed to be $p = 0.05$ and $\sigma = 10$, respectively. The system is assumed to be in the DNA-damage condition. As mentioned previously, as the pruning rate ϵ increases, running time decreases, but performance decreases. In this case, the performance of both the BKF-CORR and BKS-CORR estimators decreases significantly for $\epsilon = 0.10$, but it does not vary much by moving from $\epsilon = 0.01$ to $\epsilon = 0.05$. The choice of ϵ depends principally on the amount of available resources and time-limit constraints.

Figure 6 displays sample original and estimated state trajectories of all genes obtained by the BKF-CORR and the BKF estimators on a single time series of length 40, with correlation parameter $\eta = 0.2$, $p = 0.05$, $\epsilon = 0.1$, and $\sigma = 10$. It is clear that the gene states are better tracked by the BKF-CORR algorithm in comparison to the BKF. Notice that less gene activity can be observed in the case of no-stress condition due to the singleton rest attractor of the system, whereas several oscillations can be seen under DNA damage due to the existence of a cyclic attractor.

6 Conclusions

In practice, the existence of correlation between data points acquired from gene expression technologies should be expected, and there is a need for accurate estimation of transcriptional states of genes under these conditions. In this paper, gene regulatory networks observed through noisy correlated gene-expression data were modeled with a modified Partially-Observed Boolean Dynamical System (POBDS) model that accounts for measurement noise correlation. The BKF-CORR and BKS-CORR algorithms for state estimation from correlated measurements were proposed, which are built on a hypothesis tree and an efficient pruning process to keep the computation tractable. Numerical results demonstrated that the proposed BKF-CORR and BKS-CORR estimators achieve good state tracking performance under modest computational requirements.

Abbreviations

BKF: Boolean Kalman Filter; BKS: Boolean Kalman Smoother; BKF-CORR: Boolean Kalman Filter for correlated measurements; BKS-CORR: Boolean Kalman Smoother for correlated measurements; GRNs: Gene regulatory networks; MMSE: Minimum mean square error; POBDS: Partially-Observed Boolean Dynamical System

Funding

The authors would like to acknowledge the support of the National Science Foundation through NSF awards CCF-1320884 and CCF-1718924.

Authors' contributions

MI proposed the algorithms based on the hypothesis tree and carried out the numerical experiments. UB proposed the original idea of studying POBDS with correlated measurement noise. Both authors made significant contributions in the writing of the manuscript. Both authors read and approved the final manuscript.

Ethic approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Received: 25 October 2017 Accepted: 19 March 2018
Published online: 04 April 2018

References

1. SA Kauffman, Metabolic stability and epigenesis in randomly constructed genetic nets. J. Theor. Biol. **22**(3), 437–467 (1969)
2. I Shmulevich, ER Dougherty, W Zhang, From Boolean to probabilistic Boolean networks as models of genetic regulatory networks. Proc. IEEE. **90**(11), 1778–1792 (2002)
3. M Imani, U Braga-Neto, Maximum-likelihood adaptive filter for partially-observed Boolean dynamical systems. IEEE Trans. Signal Process. **65**, 359–371 (2017)
4. T Chen, HL He, GM Church, et al, in *Pacific Symposium on Biocomputing*. Modeling gene expression with differential equations. vol. 4, (1999), p. 40
5. MS Yeung, J Tegnér, JJ Collins, Reverse engineering gene networks using singular value decomposition and robust regression. Proc. Natl. Acad. Sci. **99**(9), 6163–6168 (2002)
6. S Kikuchi, D Tominaga, M Arita, K Takahashi, M Tomita, Dynamic modeling of genetic networks using genetic algorithm and S-system. Bioinformatics. **19**(5), 643–650 (2003)
7. S Kimura, K Ide, A Kashihara, M Kano, M Hatakeyama, R Masui, N Nakagawa, S Yokoyama, S Kuramitsu, A Konagaya, Inference of S-system models of genetic networks using a cooperative coevolutionary algorithm. Bioinformatics. **21**(7), 1154–1163 (2004)
8. N Friedman, M Linial, I Nachman, D Pe'er, Using Bayesian networks to analyze expression data. J. Comput. Biol. **7**(3-4), 601–620 (2000)
9. K Murphy, S Mian, et al, *Modelling gene expression data using dynamic Bayesian networks*. (Technical report, Computer Science Division, University of California, Berkeley, CA, 1999)

10. B-E Perrin, L Ralaivola, A Mazurie, S Bottani, J Mallet, F d'Alche–Buc, Gene networks inference using dynamic Bayesian networks. Bioinformatics. **19**(suppl_2), 138–148 (2003)
11. SA Kauffman, Metabolic stability and epigenesis in randomly constructed genetic nets. J. Theor. Biol. **22**, 437–467 (1969)
12. SA Kauffman, Homeostasis and differentiation in random genetic control networks. Nature. **224**, 177–178 (1969)
13. I Shmulevich, ER Dougherty, S Kim, W Zhang, Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks. Bioinformatics. **18**(2), 261–274 (2002)
14. R Albert, HG Othmer, The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in drosophila melanogaster. J. Theor. Biol. **223**(1), 1–18 (2003)
15. F Li, T Long, Y Lu, Q Ouyang, C Tang, The yeast cell-cycle network is robustly designed. Proc. Natl. Acad. Sci. U S A. **101**(14), 4781–6 (2004)
16. A Faure, A Naldi, C Chaouiya, D Thieffry, Dynamical analysis of a generic Boolean model for the control of the mammalian cell cycle. Bionformatics. **22**(14), 124–131 (2006)
17. E Batchelor, A Loewer, G Lahav, The ups and downs of p53: understanding protein dynamics in single cells. Nat. Rev. Cancer. **9**, 371–377 (2009)
18. R Layek, A Datta, Fault detection and intervention in biological feedback networks. J. Biol. Syst. **20**(4), 441–453 (2012)
19. I Shmulevich, ER Dougherty, *Probabilistic Boolean networks*. (SIAM, Philadelphia, 2009)
20. D Cheng, H Qi, A linear representation of dynamics of Boolean networks. IEEE Trans. Automatic Control. **55**(10), 2251–2258 (2010)
21. D Cheng, H Qi, Z Li, *Analysis and control of Boolean networks: a semi-tensor product approach*. (Springer, 2010)
22. Y Chen, ER Dougherty, ML Bittner, Ratio-based decisions and the quantitative analysis of cDNA microarray images. J. Biomed. Opt. **2**(4), 364–374 (1997)
23. A Mortazavi, BA Williams, K McCue, L Schaeffer, B Wold, Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods. **5**(7), 621–628 (2008)
24. J Hua, C Sima, M Cypert, GC Gooden, S Shack, L Alla, EA Smith, JM Trent, ER Dougherty, ML Bittner, Dynamical analysis of drug efficacy and mechanism of action using GFP reporters. J. Biol. Syst. **20**(04), 403–422 (2012)
25. SZ Dadaneh, X Qian, M Zhou, Bnp-seq: Bayesian nonparametric differential expression analysis of sequencing count data. J. Am. Stat. Assoc. (2017) just-accepted
26. U Braga-Neto, in *Signals, Systems and Computers (ASILOMAR), 2011 Conference Record of the Forty Fifth Asilomar Conference On*. Optimal state estimation for Boolean dynamical systems (IEEE, 2011), pp. 1050–1054
27. M Imani, U Braga-Neto, in *2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. Optimal state estimation for Boolean dynamical systems using a Boolean Kalman smoother (IEEE, 2015), pp. 972–976
28. M Imani, U Braga-Neto, Particle filters for partially-observed Boolean dynamical systems. Automatica. **87**, 238–250 (2018)
29. A Bahadorinejad, UM Braga-Neto, Optimal fault detection and diagnosis in transcriptional circuits using next-generation sequencing. IEEE/ACM Trans. Comput. Biol. Bioinform. (2015)
30. LD McClenny, M Imani, U Braga-Neto, in *the 42nd IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*. Boolean Kalman Filter with correlated observation noise (IEEE, 2017)
31. M Imani, U Braga-Neto, in *2015 49th Asilomar Conference on Signals, Systems and Computers*. Optimal gene regulatory network inference using the Boolean Kalman filter and multiple model adaptive estimation (IEEE, 2015), pp. 423–427
32. M Imani, U Braga-Neto, in *2017 51th Asilomar Conference on Signals, Systems and Computers*. Optimal finite-horizon sensor selection for Boolean Kalman filter (IEEE, 2017)
33. M Imani, U Braga-Neto, Control of gene regulatory networks with noisy measurements and uncertain inputs. IEEE Trans. Control Netw. Syst. (2018). https://doi.org/10.1109/TCNS.2017.2746341
34. M Imani, U Braga-Neto, Point-based methodology to monitor and control gene regulatory networks via noisy measurements. IEEE Trans. Control Syst. Technol. (2018). https://doi.org/10.1109/TCST.2017.2789191
35. M Imani, U Braga-Neto, in *American Control Conference (ACC), 2016*. State-feedback control of partially-observed Boolean dynamical systems using RNA-seq time series data (IEEE, 2016), pp. 227–232
36. M Imani, UM Braga-Neto, in *Proceedings of the 2017 American Control Conference (ACC 2017)*. Multiple model adaptive controller for partially-observed Boolean dynamical systems (IEEE, Seattle, 2017), pp. 1103–1108
37. M Imani, U Braga-Neto, in *Decision and Control (CDC), 2016 IEEE 55th Conference On*. Point-based value iteration for partially-observed Boolean dynamical systems with finite observation space (IEEE, 2016), pp. 4208–4213
38. M Imani, UM Braga-Neto, in *Proceedings of the 2018 American Control Conference (ACC 2018)*. Optimal Control of Gene Regulatory Networks with Unknown Cost Function (IEEE, 2018)
39. LD Mcclenny, M Imani, UM Braga-Neto, BoolFilter: an R package for estimation and identification of partially-observed Boolean dynamical systems. BMC Bioinformatics. **18**(1), 519 (2017)
40. LD McClenny, M Imani, U Braga-Neto, Boolfilter package vignette. The Comprehensive R Archive Network (CRAN) (2017)
41. N Ghaffari, MR Yousefi, CD Johnson, I Ivanov, ER Dougherty, Modeling the next generation sequencing sample processing pipeline for the purposes of classification. BMC Bioinformatics. **14**(1), 307 (2013)
42. S Boluki, M Shahrokh Esfahani, X Qian, ER Dougherty, Constructing pathway-based priors within a Gaussian mixture model for Bayesian regression and classification. IEEE/ACM Trans. Comput. Biol. Bioinformatics (2017). https://doi.org/10.1109/TCBB.2017.2778715
43. S Xie, M Imani, E Dougherty, U Braga-Neto, in *2017 51th Asilomar Conference on Signals, Systems and Computers*. Nonstationary linear discriminant analysis (IEEE, 2017)
44. S Boluki, M Shahrokh Esfahani, X Qian, ER Dougherty, Incorporating biological prior knowledge for Bayesian learning via maximal knowledge-driven information priors. BMC bioinformatics (2017)
45. A Karbalayghareh, U Braga-Neto, ER Dougherty, Classification of single-cell gene expression trajectories from incomplete and noisy data. IEEE/ACM Trans. Comput. Biol. Bioinformatics (2017). https://doi.org/10.1109/TCBB.2017.2763946
46. RA Weinberg, *The Biology of Cancer*. (Garland Science, Princeton, 2006)