Dynamic Inspection of Latent Variables in State-Space Systems

Tianshu Feng, Xiaoning Qian, *Senior Member, IEEE*, Kaibo Liu, *Member, IEEE*, and Shuai Huang, *Member, IEEE*

***Abstract*— The state-space models (SSMs) are widely used in a variety of areas where a set of observable variables are used to track some latent variables. While most existing works focus on the statistical modeling of the relationship between the latent variables and observable variables or statistical inferences of the latent variables based on the observable variables, it comes to our awareness that an important problem has been largely neglected. In many applications, although the latent variables cannot be routinely acquired, they can be occasionally acquired to enhance the monitoring of the state-space system. Therefore, in this paper, novel dynamic inspection (DI) methods under a general framework of SSMs are developed to identify and inspect the latent variables that are most uncertain. Extensive numeric studies are conducted to demonstrate the effectiveness of the proposed methods.**

***Note to Practitioners*—The SSM aims to estimate crucial latent variables that characterize the states of a system but cannot be measured routinely or directly. The conventional way has been solely based on a measurement capacity dedicated to observed variables. However, we realize there are situations that, although latent variables cannot be measured routinely, it is possible to inspect a small portion of latent variables at a given frequency. Thus, the problem is how to allocate the inspection resources to help monitor the latent variables of the state-space system optimally, conditioning on the established statistical machinery of the SSM for model estimation and inference. We propose a DI method to select and partially measure the latent variables and improve the estimation accuracy by combining the measured latent variables and observations.**

***Index Terms*— Dynamic inspection (DI), state-space models (SSMs).**

Manuscript received October 23, 2018; accepted November 20, 2018. This paper was recommended for publication by Associate Editor C. Hadjicostis and Editor S. Reveliotis upon evaluation of the reviewers' comments. This work was supported in part by the National Science Foundation Grants Division of Computing and Communication Foundations under Award 1718513 and Award 1715027 and in part by the Air Force Office of Scientific Research under Award FA9550-18-1-0145. *(Corresponding author: Shuai Huang.)*

T. Feng and S. Huang are with the Department of Industrial and Systems Engineering, University of Washington, Seattle, WA 98195 USA (e-mail: tsfeng@uw.edu; shuaih@uw.edu).

X. Qian is with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843 USA (e-mail: xqian@ece.tamu.edu).

K. Liu is with the Department of Industrial and Systems Engineering, University of Wisconsin-Madison, Madison, WI 53706 USA (e-mail: kliu8@wisc.edu).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the author. It contains discussion and description of the parameter tuning algorithm mentioned in the paper and provides experiments to verify the proposed algorithm. Meanwhile, it contains two additional real-world experiments to further demonstrate the proposed DI method. The total size of the file is 181 KB.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASE.2018.2884149

I. Introduction

RECENT advances in sensing and information technologies provide the unprecedented volume of measurements for complex systems in different applications. However, we still face the situations that some variables in complex systems cannot be directly measured, either due to technological limitations or economic reasons. For example, in healthcare applications, monitoring patients' health conditions is apparently a crucial task. However, a patient's health condition cannot be measured as the definition of health condition is often subjective or experience-based, characterized by multiple biological variables that can be challenging to acquire routinely outside of clinical settings. Examples can be found in engineering applications as well, such as the monitoring of the working conditions of Lithium-ion batteries [1] and wind turbines [2]. Indeed, these variables are considered as the states of interest of the underlying systems. They are not directly measurable and often considered as the latent variables. Often, we can measure some other observable variables associated with these latent variables to infer their corresponding status. Thus, in many real-world applications, we have to face the challenge of latent variable inference from the observed data, regardless of the available data volume from sensing technologies.

To tackle this challenge, the state-space model (SSM) has been one of the mathematical system models that characterize the internal behavior or system states by the latent variables, in addition to the observable variables as the input and output of the system models [3]. It is hoped that, with the statistical modeling of the relationship between the latent variables and observable variables, the latent variables can be inferred based on the measurements acquired on the observable variables. Applications of the SSMs have been ubiquitous in different areas, such as engineering [4], healthcare [5], bioinformatics [6]–[8], robotics [9], and even economics [10], [11] and climatology [12]. For example, in [1], an SSM is used to monitor the working condition of Lithium-ion batteries, where the latent variables are the open-circuit voltage, the voltage of a resistance-capacitance network and the state of charge, and the observable variable is the output voltage. In [13], an SSM is used to analyze the real-time urban traffic information and improve the short-term prediction accuracy of traffic parameters, in order to control traffic and reduce traffic congestion and delays, where the latent variable is a time-dependent linear transformation of the observable variables with the addition of exogenous regressors, such as roadway

capacity and weather, and the observable variables are the traffic volumes. While most of the SSMs mentioned above characterize linear systems, they can also be extended for nonlinear systems, e.g., as shown in a recent work that combines the SSM and deep learning methods [14] to characterize the brain dynamics measured by resting-state functional magnetic resonance imaging data.

Despite the popularity of the SSM, an important problem has been neglected in the literature. In many applications, although the latent variables cannot be routinely acquired due to economic reasons, they can be occasionally acquired. Actually, our work is motivated by our previous effort in two applications. One is the risk monitoring for disease onset of type 1 diabetes (T1D). It is important for T1D patients to monitor their immunological and metabolic variables regularly, based on which their doctors can analyze the patients' current health status and make proper medical plans. However, it can be challenging for patients to acquire these immunological and metabolic variables outside of clinical settings. Rather, it is more realistic for them to regularly measure their glucose level and body mass index (BMI). It is known that these "cheap" and measurable variables correlate with those "expensive" immunological and metabolic variables that are difficult to measure by patients but are measurable by their doctors. Another example is the emerging crowdsourcing applications that design online platforms to harvest expertise and idle labors all over the world. Briefly speaking, crowdsourcing employs anonymous workers who volunteer their labors to complete tasks. The quality of the workers is latent, which can only be acquired with testing tasks (answers known in advance). The responses to the tasks are the measurements that we use to infer the underlying quality.

For both applications, first of all, the latent variables are the most useful for decision making. Second, although these latent variables cannot be measured on a regular basis, it is possible to measure them with a higher cost and less frequent sampling patterns. In such scenarios, there is a dynamic inspection (DI) problem to decide on which latent variables to be measured given the observed data at each time epoch so that the uncertainty of the state estimation can be maximally reduced with the limited budget. The basic idea is that, based on the observed data, we can not only infer the status of the latent variables but also evaluate their conditional uncertainty conditioning on the data. Intuitively, estimation variance implies uncertainty, and a latent variable is more uncertain than others if its estimation variance is larger than those of other latent variables. Thus, the task is to dynamically identify and further inspect the most uncertain latent variables at every time epoch. However, a complication arises that the uncertainty of the latent variables is interrelated, encoded in the conditional covariance structure. To solve these problems, we propose a novel DI method under the general framework of SSMs. At each time epoch, DI selects and inspects latent variables that are most uncertain in a collective way, e.g., measured by their covariance matrix. Under the DI, the selection of the latent variables to be inspected integrating marginal variances and associations between the latent variables derived from the SSM formulation is achieved by solving a nonconvex optimization problem with a set of constraints at each time.

The paper is organized as follows. In Section II, we will review the related works. In Section III, we will introduce the basic framework of the SSM and formulate our DI framework. In Section IV, we will develop a simulation study to evaluate the performance of our DI method in comparison with other state-of-the-art methods. In Section V, we will demonstrate how our method can be applied to type 1 diabetes and crowdsourcing problems. Section VI concludes this paper.

II. Related Works

One type of related works focuses on sensor allocation under SSMs, i.e., the study of how to optimize the layout of a sensor network in a manufacturing process [15]–[19]. These works differ from ours as they aim to optimize the design of the observable variables in order to monitor the latent variables better. For example, the works in [15]–[17] proposed strategies to determine the locations of measurement stations and the minimum number of sensors by integrating sensing information into an SSM. Jin *et al.* [18] and Jin and Tsung [19] proposed to allocate control charts by considering the interrelationship between stages in serial-parallel multistage manufacturing processes described by an SSM. Another approach for dynamic sensor selection method is called the maximum mutual information principle [20] in which sensors were selected to query for tracking the target by maximizing the mutual information between the sensor output and target state. It was shown that maximizing the mutual information between the sensor output and target state was equivalent to minimizing the expected posterior uncertainty. This idea is further improved in terms of computational efficiency [21] and extended for prediction [22] and high-dimensional problems [23].

There have been considerable research efforts in extending the SSMs, such as the continuous time SSMs and nonlinear SSMs [24]. Much of the attention was given to the modeling of the system and the estimation of the parameters of the model. For example, Oud and Jansen [25] proposed the maximum likelihood estimation of the parameters of the continuous time SSMs on panel data using the structural equation modeling. Different approximation approaches are proposed for non-Gaussian/nonlinear SSMs, such as Monte Carlo methods [26], [27] and the particle filter [28]. However, our work is different from this line of research as well, since we focus more on how to monitor the latent variables better with a given knowledge of the SSM.

There exist several works that try to identify the structures of dynamic systems with partial observations [29]–[33]. For example, Yuan *et al.* [29] tried to infer the structure of the network based on partial exact observation of states. Our work is different from this line of works because we aim at selecting latent variables to be inspected and estimating latent variables according to previous observations which are linear combinations of latent variables with noise, and the structure of the system in our model is estimated from established methods which will be introduced in Section III.

One particularly related work that actually inspired this paper is [34]. This paper shares a similar motivation as ours, although it was mainly motivated to show that the ill-posed state-space problem, where the number of the observed variables was much less than the latent variables, could become well-posed with partial inspection of the latent variables. In [34], only a simple heuristic approach was developed to dynamically acquire measurements from the latent variables by selecting the latent variables with the largest variances conditioning on the observations without considering their covariance structure. Also, [34] only considered a simple SSM with fixed structure of process errors and no observation error. Our work extends this line of works to applications where there are observation errors and nontrivial process errors, and optimize the DI problem using a systematical formulation rather than heuristics to account for the covariance structure of the latent variables. Incorporation of the covariance between the latent variables provides a better evaluation of the true uncertainty of the latent variables. In other words, if a latent variable is closely associated with the latent variables with larger uncertainty, then the underlying true uncertainty of this latent variable is more likely to be large even though it has a relatively small estimated conditional variance.

III. Methodology

In this section, we will introduce a general framework of the SSM in Section III-A. Then, based on this framework, we will introduce our proposed DI method in Section III-B and its computational algorithm in Section III-C. As our method integrates statistical inference, model updating, and sensing simultaneously, a summary of the entire process is presented in Section III-D. Note that in this paper, we use lower case letters, e.g., x, to represent scalars, bold-face lower case letters, e.g., $\mathbf{x}$, to represent vectors, bold-face upper case letters, e.g., $\mathbf{X}$, to represent matrices, and bold-face upper case italic letters, e.g., $\boldsymbol{X}$, to represent random variables.

A. Review of the State-Space Model

The SSM is a big umbrella that includes many variants. Basically, it has two parts. One is the state equation that characterizes the dynamics of the latent variables. Another one is the measurement equation that characterizes the relationship between the latent variables and observable variables. In this paper, we use I and J as the number of the observable variables and the number of the latent variables, respectively. Also, we denote $\mathbf{x}^{(t)}$, a $J \times 1$ vector, as the realization of the latent variables at time t. The state equation is shown as follows:

$$\mathbf{x}^{(t+1)} = \mathbf{B}\mathbf{x}^{(t)} + \boldsymbol{\epsilon}^{(t)} \tag{1}$$

$$\boldsymbol{\epsilon}^{(t)} \sim N(0, \mathbf{Q}). \tag{2}$$

Here, the matrix $\mathbf{B}$, a $J \times J$ matrix, is commonly referred as the state transition matrix, that characterizes the autocorrelation between the latent variables. For example, in the example of T1D patient monitoring, the latent variables can be some expensive physiological variables such as HbA1c and glutamic acid decarboxylase. Then, the state transition matrix $\mathbf{B}$ characterizes how the previous states of these physiological variables impact the current states as J linear regression models. The error term, $\boldsymbol{\epsilon}^{(t)}$, corresponds to the variability of these physiological variables that cannot be explained by their temporal correlations. It is often assumed that $\boldsymbol{\epsilon}^{(t)}$ is from a multivariate normal distribution $N(0, \mathbf{Q})$, while $\mathbf{Q}$ is a $J \times J$ variance-covariance matrix that is stationary over time.

Similarly, we denote $\mathbf{y}^{(t)}$, an $I \times 1$ vector, as the realization of the observable variables at time t. The measurement equation is shown in the following:

$$\mathbf{y}^{(t+1)} = \mathbf{Z}\mathbf{x}^{(t+1)} + \mathbf{v}^{(t)} \tag{3}$$

$$\mathbf{v}^{(t)} \sim N(0, \mathbf{R}). \tag{4}$$

Here, $\mathbf{Z}$, an $I \times J$ matrix, is the measurement matrix. Each row of $\mathbf{Z}$ corresponds to an observable variable and encodes the linear coefficients of the J latent variables. In other words, this essentially suggests that there are I linear regression models, while each model represents how the J latent variables impact one of the observable variables. The error term, $\mathbf{v}^{(t)}$, corresponds to the variability of these observable variables that cannot be explained by the latent variables. It is often assumed that $\mathbf{v}^{(t)}$ is from a multivariate normal distribution $N(0, \mathbf{R})$, while $\mathbf{R}$ is an $I \times I$ variance-covariance matrix that is stationary over time.

As mentioned earlier in this paper, we focus on how to monitor the latent variables better with a given knowledge of the SSM. Thus, we assume that the model parameters, $\mathbf{B}$, $\mathbf{Z}$, $\mathbf{Q}$, have been known while $\mathbf{R}$ is unknown. This can be done in many applications given the availability of historical data (i.e., measurements of y) and domain knowledge for training the model. In particular, in our paper, we employ the expectation–maximization (EM) algorithm proposed in [35] to estimate the unknown parameter $\mathbf{R}$ of the model with observations, and solve for the expected values and estimation variances of the latent variables with the classic Kalman filter [36]. The log-likelihood function, as well as details of the derivative of the EM algorithm, can be found in [37]. Note that although $\mathbf{B}$, $\mathbf{Z}$, $\mathbf{Q}$ are assumed to be known in this paper, they can be estimated with the EM algorithm as well, as described in [37].

B. Dynamic Inspection Strategy

Inspecting some latent variables can greatly help us to estimate and monitor the latent variables. Suppose that if there are available resources allowing us to inspect a small number of latent variables at each time epoch, we then can estimate the other latent variables based on the inspections and the observations on the observable variables. To optimally decide on which latent variables to be inspected at each time point, we need to devise an optimization formulation first. The naive approach, denote as the random inspection (RI), is to choose and inspect the latent variables randomly. However, at each time epoch, there are unequal levels of uncertainty of the latent variables. In order to gain as much information about the latent variables as possible, it is better to inspect the most uncertain latent variables conditioning on previous observations and inspections, i.e., the latent variables with the largest $\mathbf{S}_{ii}^{(t)}$, where $\mathbf{S}^{(t)} = \mathrm{Var}(\boldsymbol{X}^{(t)} | \boldsymbol{Y}^{(1:t)}, \boldsymbol{X}_I^{(1:t-1)})$, $\boldsymbol{Y}^{(1:t)}$

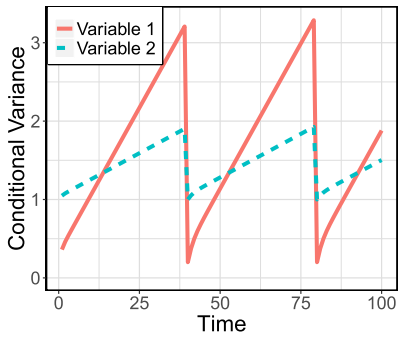


Fig. 1. Example of how marginal conditional variance changes over time.

represents the observations until time t, $\boldsymbol{X}_I^{(1:t-1)}$ denotes the inspections until time $t-1$, and $\boldsymbol{X}^{(t)}$ is the random vector of the latent variables at time t. Note that because the variance $\mathbf{S}^{(t)}$ is conditioning on previous inspections $\boldsymbol{X}_I^{(1:t-1)}$, the ranking of the uncertainty changes over time. This is from the fact that the conditional variance of $\boldsymbol{X}_i^{(t)}$ becomes small if we know the value of $\boldsymbol{X}_i^{(t-1)}$ and how latent variables transit, which, in turn, changes the ranking of the uncertainty of latent variables. One simple example with two related variables is shown in Fig. 1. The red solid line represents the marginal conditional variance of variable 1 while the blue dashed line denotes that of variable 2. Variable 1 is inspected at time points 40 and 80, and the ranking of the uncertainty of the two variables changes after each inspection. If an additional inspection is required at $t = 45$, then variable 2 should be chosen to reduce the overall uncertainty of the latent variables. Based on this motivation, we propose two approaches based on the conditional variance $\mathbf{S}^{(t)} = \mathrm{Var}(\boldsymbol{X}^{(t)}|\boldsymbol{Y}^{(1:t)}, \boldsymbol{X}_I^{(1:t-1)})$. The details of how $\mathbf{S}^{(t)}$ is derived can be found in the proof of Theorem 1, which involves the variance of noise in the states transition $\mathbf{Q}$ and the variance of noise in measurement $\mathbf{R}$.

The first approach is called the marginal inspection (MI) that inspects the latent variables with the largest marginal estimation variances conditioning on the observations and previous inspections at each time. Introducing a binary decision variable δ_{it} for the latent variable i such that $\delta_{it} = 1$ if and only if the latent variable i is inspected at time t, the MI approach is to find the optimal $\boldsymbol{\delta}_t^*$ that maximizes $\boldsymbol{\delta}_t^T \mathrm{diag}(|\hat{\mathbf{S}}^{(t)}|)\boldsymbol{\delta}_t$ with the constraint $\sum_{i=1}^J c_i\delta_{it} \le M$, where each element $|\hat{\mathbf{S}}_{ij}^{(t)}|$ of $|\hat{\mathbf{S}}^{(t)}|$ is the absolute value of $\hat{\mathbf{S}}_{ij}^{(t)}$, $\mathrm{diag}(\mathbf{S})$ is a diagonal matrix with the corresponding diagonal entries in $\mathbf{S}$, c_i is the cost of inspecting latent variable i and M is the cost budget. The MI strategy solves the following optimization problem:

$$\max_{\boldsymbol{\delta}_t \in \{0,1\}^J} \ \boldsymbol{\delta}_t^T \mathrm{diag}(|\hat{\mathbf{S}}^{(t)}|)\boldsymbol{\delta}_t$$
$$\text{s.t.} \sum_{i=1}^J c_i\delta_{it} \le M. \tag{5}$$

Another approach, called the DI, is to further account for the correlation between the latent variables. This may help us overcome the limitation of MI considering only the marginal variance of the latent variables. For example, suppose that the conditional estimation variance of the latent variable i is much higher than that of the latent variables j and k, and the conditional estimation variances of j and k are close to each other, then, with the consideration of the covariance between the latent variables, we prefer to inspect j rather than k if j is more correlated with i. This is because the underlying true variance of the latent variable j is more likely to be higher than that of the latent variable k with the knowledge from the latent variable i. Note that the uncertainty of the latent variables comes from $\mathbf{Q}, \mathbf{R}$, and thus the relation of the latent variables is influenced by the parameters $\mathbf{B}, \mathbf{Z}, \mathbf{Q}, \mathbf{R}$.

Thus, the DI is to find the optimal $\boldsymbol{\delta}_t^*$ that maximizes $\boldsymbol{\delta}_t^T \mathrm{diag}(|\hat{\mathbf{S}}^{(t)}|)\boldsymbol{\delta}_t$ with the constraint $\sum_{i=1}^J c_i\delta_{it} \le M$ and the regularization on the off-diagonal elements $|\hat{\mathbf{S}}_{ij}^{(t)}|$. Here, $|\hat{\mathbf{S}}_{ij}^{(t)}|$ refers to the association between latent variables i and j, conditioning on the observations and previous inspections. Formally, the DI strategy leads to the following optimization problem:

$$\max_{\boldsymbol{\delta}_t \in \{0,1\}^J} \ \boldsymbol{\delta}_t^T \mathrm{diag}(|\hat{\mathbf{S}}^{(t)}|)\boldsymbol{\delta}_t + \beta \sum_{i \ne j} \delta_{it}\delta_{jt}|\hat{\mathbf{S}}_{ij}^{(t)}|$$
$$\text{s.t.} \sum_{i=1}^J c_i\delta_{it} \le M. \tag{6}$$

The regularization term, $\beta \sum_{i \ne j} \delta_{it}\delta_{jt}|\hat{\mathbf{S}}_{ij}^{(t)}|$, on the decision variables in (6) controls the association of inspection decisions such that the larger the β is, the more associated the inspection decisions are. If $\beta = 0$, DI degenerates to MI. Now, the Lagrangian form of the optimization problem (6) is

$$\max_{\boldsymbol{\delta}_t \in \{0,1\}^J} \ \boldsymbol{\delta}_t^T \mathrm{diag}(|\hat{\mathbf{S}}^{(t)}|)\boldsymbol{\delta}_t - \eta \sum_{i=1}^J c_i\delta_{it} + \beta \sum_{i \ne j} \delta_{it}\delta_{jt}|\hat{\mathbf{S}}_{ij}^{(t)}|. \tag{7}$$

After solving (7) and getting the optimal set of indices of the selected latent variables at time t, $\Omega^{(t)} = \{i|\delta_{it}^* = 1\}$, we replace the estimated states of latent variables in $\Omega^{(t)}$ with observed values $\mathbf{x}_I^{(t)}$ from inspection.

Note that existing inference algorithms developed for the SSM have not considered the inference tasks when both data of the observable variables and partial data of some latent variables are given. Thus, there seems to be no readily available algorithm for us to derive $\mathbf{S}^{(t)} = \mathrm{Var}(\boldsymbol{X}^{(t)}|\boldsymbol{Y}^{(1:t)}, \boldsymbol{X}_I^{(1:t-1)})$. Fortunately, as shown in Theorem 1 (proof in Appendix VI), this problem can be readily derived based on existing results on the SSM. Lemmas 2 and 3 justify the correctness of the model construction in the proof of Theorem 1. If we can inspect all the latent variables at time t, the conditional variances after inspection become zeros, and the expected values of the latent variables are substituted by the observed values. By using the block matrices and Moore–Penrose inverse and following the deduction in the proofs, our computation also shows that any inspected latent variable has the marginal conditional variance being zero and the expected value being replaced with the observed value after inspection. Recall that, in [34], it has been shown that the ill-posed state-space problem, where the number of the observations I is much less than that of the hidden states J, can become well-posed with partial inspection of hidden states.

Theorem 1: The expected values and estimation variances of the latent variables conditioning on the observations and inspections, i.e., $E[\boldsymbol{X}^{(t)}|\boldsymbol{Y}^{(1:t)}, \boldsymbol{X}_I^{(1:t-1)}]$, and $\mathrm{Var}(\boldsymbol{X}^{(t)}|\boldsymbol{Y}^{(1:t)}, \boldsymbol{X}_I^{(1:t-1)})$, can be computed with the Kalman filter (proof in Appendix A).

Lemma 2: If all the latent variables are inspected at a time point, then under the model construction and notations from the proof of Theorem 1, we have the conditional covariance matrix of the latent variables $\mathbf{P}_t^{(t)} = 0$, and the states of the latent variables replaced by the inspections $\hat{\mathbf{x}}_t^{(t)} = \tilde{\mathbf{y}}_{(I+1):(I+J)}^{(t)}$ (proof in Appendix B).

Lemma 3: If none of the latent variables is inspected, then under the model construction and notations from the proof of Theorem 1, the inspection methods degenerate to the original SSM (proof in Appendix C).

C. Computational Algorithm

To solve the optimization problem in (7), because the decision variables are binary, first, we rewrite the objective function as

$$\min_{\boldsymbol{\delta}_t \in \{0,1\}^J} \sum_{i=1}^{J} \delta_{it}\left(c_i\eta - |\hat{\mathbf{S}}_{ii}^{(t)}|\right) - \beta \sum_{i\neq j} \delta_{it}\delta_{jt}|\hat{\mathbf{S}}_{ij}^{(t)}|. \tag{8}$$

Because

$$\begin{aligned}
&\sum_{i=1}^{J} \delta_{it}\left(c_i\eta - |\hat{\mathbf{S}}_{ii}^{(t)}|\right) - \beta \sum_{i\neq j} \delta_{it}\delta_{jt}|\hat{\mathbf{S}}_{ij}^{(t)}| \\
&\quad= \sum_{i=1}^{J} \delta_{it}\left(c_i\eta - |\hat{\mathbf{S}}_{ii}^{(t)}|\right) - \beta \sum_{i\neq j} \delta_{it}|\hat{\mathbf{S}}_{ij}^{(t)}| \\
&\qquad+\beta \sum_{i\neq j} \delta_{it}(1-\delta_{jt})|\hat{\mathbf{S}}_{ij}^{(t)}| \\
&\quad= \sum_{i=1}^{J} \delta_{it}\left(c_i\eta - \beta\sum_{j=1}^{J}|\hat{\mathbf{S}}_{ij}^{(t)}| - (1-\beta)|\hat{\mathbf{S}}_{ii}^{(t)}|\right) \\
&\qquad+\beta \sum_{i,j} \delta_{it}(1-\delta_{jt})|\hat{\mathbf{S}}_{ij}^{(t)}|
\end{aligned}$$

the problem can be further formulated as

$$\begin{aligned}
\min_{\boldsymbol{\delta}_t \in \{0,1\}^J} \sum_{i=1}^{J} \delta_{it}&\left(c_i\eta - \beta\sum_{j=1}^{J}|\hat{\mathbf{S}}_{ij}^{(t)}| - (1-\beta)|\hat{\mathbf{S}}_{ii}^{(t)}|\right) \\
&+\beta \sum_{i,j=1}^{J} |\hat{\mathbf{S}}_{ij}^{(t)}|\delta_{it}(1-\delta_{jt}).
\end{aligned} \tag{9}$$

Theorem 4: The optimization problem (9) is equivalent to an s-t min-cut problem

$$\begin{aligned}
\min_{\boldsymbol{\delta}_t \in \{0,1\}^J} \sum_{i=1}^{J} |\hat{\mathbf{S}}^{(t)}|_{u_1,i}\delta_{u_1 t}(1-\delta_{it}) + \sum_{i=1}^{J} |\hat{\mathbf{S}}^{(t)}|_{i,u_2}\delta_{it}(1-\delta_{u_2 t}) \\
+\beta \sum_{i,j=1}^{J} |\hat{\mathbf{S}}_{ij}^{(t)}|\delta_{it}(1-\delta_{jt})
\end{aligned} \tag{10}$$

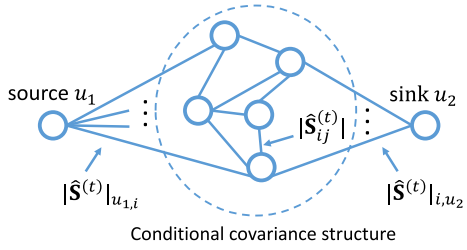


Fig. 2. Illustration of the s-t min-cut formulation for the DI strategy.

where $\boldsymbol{\delta}_{u_1,t} = 1$ and $\boldsymbol{\delta}_{u_2,t} = 0$

$$|\hat{\mathbf{S}}^{(t)}|_{u_1,i} = \left(\beta\sum_{j=1}^{J}|\hat{\mathbf{S}}_{ij}^{(t)}| + (1-\beta)|\hat{\mathbf{S}}_{ii}^{(t)}| - c_i\eta\right)_+$$

$$|\hat{\mathbf{S}}^{(t)}|_{i,u_2} = \left(c_i\eta - \beta\sum_{j=1}^{J}|\hat{\mathbf{S}}_{ij}^{(t)}| - (1-\beta)|\hat{\mathbf{S}}_{ii}^{(t)}|\right)_+$$

and $(x)_+ = \max\{x, 0\}$ (proof in Appendix D).

According to the max-flow min-cut theorem [38], the s-t min-cut problem in Theorem 4 can be efficiently solved with the maximum flow algorithm [39]. An illustration graph of the s-t min-cut problem for the DI strategy is shown in Fig. 2.

D. Summary of the DI Method

In this section, we introduce the framework of the DI. First, we construct an SSM for a given data set. Then, at each time t, we estimate the states and variances of latent variables with the Kalman filter based on the observations and previous inspections and inspect the states of selected latent variables with the DI. These latent variables are chosen by solving an s-t min-cut problem. Finally, we replace the estimated states of selected latent variables with the inspections of them. The parameter η controls the number of latent variables we choose, and β reflects how important we consider the covariance is in the model. A discussion about how to choose the parameters η and β can be found in the Supplementary file where we propose a search algorithm (as illustrated in Algorithm S.1 in the Supplementary file) to identify an interval of β using historical data within which β can achieve satisfying results. When $\beta = 0$, DI degenerates to MI. Note that the cost budget M at each time epoch is not a parameter to be tuned, but the resource constraint. Larger M tends to lead to better results, because it implies more latent variables can be inspected at each time epoch. Therefore, we should choose the largest M allowed. The algorithm for the DI method is given in Algorithm 1, and the corresponding flowchart is illustrated in Fig. 3.

IV. Simulations

A. Experiment Design

In this section, we conduct simulation experiments to compare the three inspection methods, the RI, the MI, and the DI. Furthermore, we include the state estimation by the Kalman filter for the SSM without inspection that solely relies

Algorithm 1 Dynamic Inspection Method

Initialize states of latent variables $\mathbf{x}^{(0)}$ and unknown parameters with historical observations of length T_h;
Find the approximate interval for β with Algorithm S.1;
for *each time epoch* t **do**
  Estimate unknown parameters with EM algorithm;
  Obtain expected values and variances $E[\boldsymbol{X}^{(t)}|\boldsymbol{Y}^{(1:t)}, \boldsymbol{X}_I^{(1:t-1)}]$ and $Var(\boldsymbol{X}^{(t)}|\boldsymbol{Y}^{(1:t)}, \boldsymbol{X}_I^{(1:t-1)})$ by applying the Kalman filter following Theorem 1;
  Select a β from the interval and find the η following the discussion in the Supplementary file;
  Solve the optimization problem (10) and get $\Omega^{(t)}$;
  Obtain $\mathbf{x}_I^{(t)}$, and replace the expected values of the latent variables in $\Omega^{(t)}$ with $\mathbf{x}_I^{(t)}$;
end
Output $E[\boldsymbol{X}^{(1:T)}|\boldsymbol{Y}^{(1:T)}, \boldsymbol{X}_I^{(1:T)}]$;

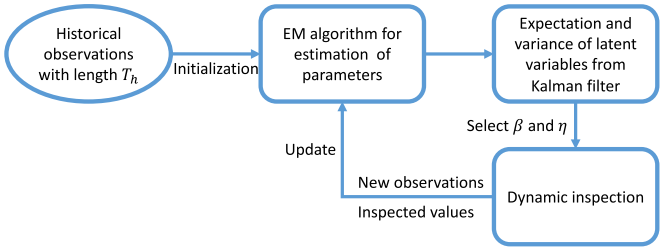


Fig. 3. Flowchart of the DI method.

on observations on the observable variables, to evaluate the impact of inspection. We evaluate the performance of the methods by the root-mean-square error (RMSE) of the conditional expectation $E[\boldsymbol{X}^{(t)}|\boldsymbol{Y}^{(1:t)}, \boldsymbol{X}_I^{(1:t-1)}]$, which is calculated as $r^{(t)} = (1/\sqrt{J})\|E[\boldsymbol{X}^{(t)}|\boldsymbol{Y}^{(1:t)}, \boldsymbol{X}_I^{(1:t-1)}] - \mathbf{x}^{(t)}\|_2$ at each time epoch t. The lower the RMSE is, the better the method is to estimate the latent variables.

We conduct the comparison across a number of scenarios with different numbers of observable variables and latent variables. Specifically, first, we randomly generate the initial vector $\mathbf{x}^{(0)}$ for the latent variables from a multivariate normal distribution $N(0, \mathbf{Q})$. The matrix $\mathbf{Q}$ is randomly generated in the following way to ensure that it is a positive-definite matrix as required for a eligible covariance matrix: let $\mathbf{P}$ be a matrix with the same size of $\mathbf{Q}$ whose elements are randomly sampled from a standard normal distribution, then, we construct $\mathbf{Q}$ as the inner product of $\mathbf{P}$, i.e., $\mathbf{Q} = \mathbf{P}'\mathbf{P}$. With the initial value $\mathbf{x}^{(0)}$, then, at each iteration t, the latent variable vector $\mathbf{x}^{(t)}$ is obtained via $\mathbf{x}^{(t)} = \mathbf{B}\mathbf{x}^{(t-1)} + \boldsymbol{\epsilon}^{(t)}$, and the observations vector $\mathbf{y}^{(t)}$ is obtained via $\mathbf{y}^{(t)} = \mathbf{Z}\mathbf{x}^{(t)} + \mathbf{v}^{(t)}$. Here, without loss of generality, we set $\mathbf{B} = \mathbf{I}$ where $\mathbf{I}$ is the identity matrix and draw the elements of $\mathbf{Z}$ from $\{0, 1, 2\}$ with equal probability and under the constraint that each row or column of $\mathbf{Z}$ is non-zero (i.e., to avoid the trivial scenario in which either some observable variables correspond to no latent variables or some latent variables cannot be observed). The process noise is simulated by $\boldsymbol{\epsilon}_t \sim N(0, \mathbf{Q})$, while the measurement error is simulated by $\mathbf{v}^{(t)} \sim N(0, \mathbf{I})$. After T iterations, we can obtain the matrix of states of the latent variables $\mathbf{X} = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(T)})$ and the matrix of the observations $\mathbf{Y} = (\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(T)})$. In this paper, because we focus on the inspection strategy rather than model estimation, we assume the availability of the matrices $\mathbf{B}$, $\mathbf{Z}$, $\mathbf{Q}$ and observations $\mathbf{y}$ that will be used in the three inspection methods. For simplicity, we assume the costs of inspections c_i, $i = 1, \ldots J$ are equal for all the latent variables. Since the methods are generic, weights can be applied if needed.

B. Results

The experiment is conducted with $T = 50$, $J = 10$, 12, 14 and $I = 4$, 6, 8. The length of historical data Th varies depending on the number of latent variables. Here, we choose $T_h = 50$; 100; 200 for $J = 10$; 12; 14, respectively. Under each combination of J and I, we follow the aforementioned experimental design, simulate the data, implement the three inspection methods, and obtain their performance evaluated by the RMSE criterion. Without loss of generality, here, we present results when the cost budget M is set to 2 and $c_i = 1$ for $i = 1, \ldots, J$.

The results are summarized in Table I and Fig. 4. Specifically, in Fig. 4, the points denote the RMSE of the estimations from different methods at different times, and each line represents the local regression of these points from a certain method. Table I summarizes Fig. 4 and reports the mean and standard derivation of the RMSE of each method in each combination of I and J across the simulated time points. Note that the results for SSM are omitted in Fig. 4 because the scale of the RMSE of SSM is very large that will distort the presentation of other methods. This can be seen from Table I that the RMSE values of SSM are much larger than those of the inspection methods. The first five time points are also removed so that we can focus on the converged results without distortion. Thus, a major observation is that with the inspection methods, even with the RI on only a few latent variables, the estimation of the latent variables can be much improved. This advantage is even bigger, as we can observe that, when the state-space systems have much less observable variables than the latent variables, i.e., $I = 4$, all the inspection models can dramatically improve the accuracy and robustness of estimations compared with SSM. Another major observation is that the DI method is overall the best approach that can achieve the lowest average RMSE and the smallest standard derivation of RMSE. This advantage will be smaller, however, if we increase the number of observable variables, as the difference between DI and MI decreases. One explanation can be that the estimations of latent variables and their uncertainty from the Kalman filter become more accurate with more observable variables; therefore, the MI model can correctly select the latent variables with high variances. Last but not least, as we can observe that the standard deviations of the RMSE from DI are smaller than those from MI and RI most of the time, incorporating the knowledge of the association between latent variables can improve the robustness of the inspection decision making.

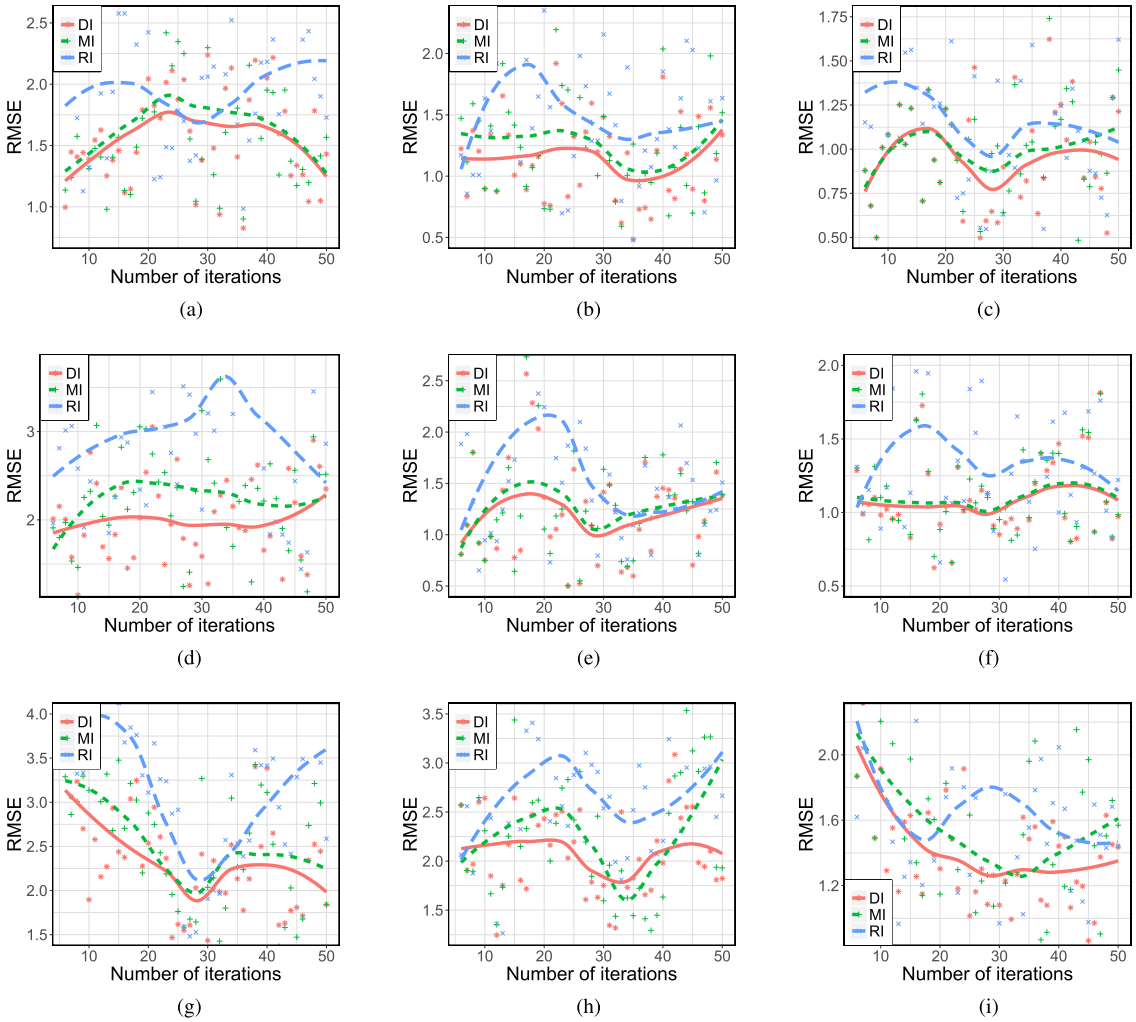


Fig. 4. Comparison of rooted mean squared errors under DI, MI, and RI methods. (a) RMSE, $J = 10$, $I = 4$. (b) RMSE, $J = 10$, $I = 6$. (c) RMSE, $J = 10$, $I = 8$. (d) RMSE, $J = 12$, $I = 4$. (e) RMSE, $J = 12$, $I = 6$. (f) RMSE, $J = 12$, $I = 8$. (g) RMSE, $J = 14$, $I = 4$. (h) RMSE, $J = 14$, $I = 6$. (i) RMSE, $J = 14$, $I = 8$.

TABLE I
PERFORMANCE OF INSPECTION METHODS WITH TWO INSPECTIONS ON SIMULATED DATA SERIES

	$I = 4$				$I = 6$				$I = 8$			
	DI	MI	RI	SSM	DI	MI	RI	SSM	DI	MI	RI	SSM
$J = 10$												
Average RMSE	1.519	1.607	1.919	11.221	1.207	1.280	1.544	4.013	1.065	1.081	1.239	6.792
SD of RMSE	0.173	0.245	0.232	2.664	0.112	0.148	0.205	1.118	0.167	0.188	0.258	0.626
$J = 12$												
Average RMSE	2.053	2.260	2.954	14.046	1.264	1.311	1.530	10.106	1.072	1.092	1.260	7.766
SD of RMSE	0.193	0.240	0.430	1.524	0.145	0.179	0.404	2.322	0.101	0.105	0.274	1.779
$J = 14$												
Average RMSE	2.341	2.510	3.172	31.451	2.088	2.237	2.673	20.326	1.427	1.550	1.598	21.028
SD of RMSE	0.411	0.520	0.706	1.887	0.206	0.446	0.322	1.527	0.174	0.206	0.185	1.219

V. REAL-WORLD APPLICATIONS

A. Type 1 Diabetes

SSMs have a wide range of applications in healthcare. Here, we illustrate how the inspection methods can be applied to the type 1 diabetes (T1D). In management of type 1 diabetes (T1D) patients, it is important for patients to monitor their immunological and metabolic variables on schedule, so that doctors can identify the change of patients' state of health in time and modify the treatment plan. However, as

TABLE II
SELECTED EXPENSIVE (LATENT) AND CHEAP (MEASUREMENT) VARIABLES

Type	Variables
Expensive	GAD, MIAA, ICA, IAA, HbA1c, HOMA-R, FPIR
Cheap	Fasting Glucose from OGTT, 60-min C-Peptide from OGTT, BMI

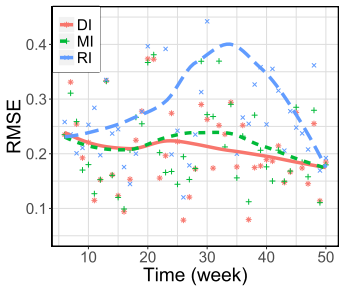


Fig. 5. RMSE of inspection methods on type 1 diabetes problem.

TABLE III
PERFORMANCE OF INSPECTION METHODS ON SIMULATED T1D SERIES

	DI	MI	RI	SSM
Average RMSE	0.205	0.213	0.295	1.663
SD of RMSE	0.072	0.074	0.128	0.164

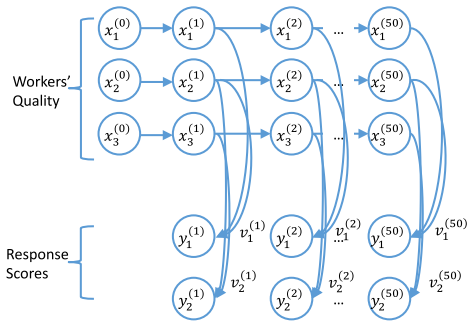


Fig. 6. Framework of crowdsourcing problem.

these variables are not easily acquired outside of clinical settings, it is more realistic for patients to monitor other variables such as glucose levels and BMI which may correlate with the inaccessible immunological and metabolic variables. This generates a typical state-space system, while the inaccessible immunological and metabolic variables are the latent variables and the accessible variables are the observable variables. Inspection methods developed in this paper provide a remedy to enhance patient monitoring that can provide valuable data-driven insight for doctors to optimize visit schedules for patients to measure their critical variables in clinics.

To demonstrate this, we leverage our previous studies on type 1 diabetes (T1D), particularly, the diabetes prevention trial—type 1 (DPT-1) [40], to build an SSM for T1D and implement the inspections methods. In this model, there are seven "expensive" variables that are critical to assessing the T1D patients' health conditions but are inaccessible to patients (thus, our latent variables) and three "cheap" variables that can be measured by patients (thus, our observable variables), as listed in Table II. The "cheap" variables can be measured through simple oral glucose tolerance testing (OGTT) by patients themselves but is less accurate than the "expensive" variables from intravenous glucose tolerance test and other tests for assessment of insulin resistance because of their invasiveness, duration, and cost of reagents [41]. Thus, rather than just monitoring the cheap but inaccurate variables, with inspections, we can enable monitoring of the patients with both the "cheap" variables and selected "expensive" variables in each week (here, each monitoring epoch is a week). Without loss of generality, we assume that $M = 2$ and $c_i = 1$ for $i = 1, \ldots, J$. The parameters of the SSM can be estimated from the scaled DPT-1 data set [40], based on which we randomly simulate a data series and use it to test our inspection methods.

The results are presented in Table III and Fig. 5. Overall, it can be observed that inspection methods outperform the SSM method, indicating their significant contribution in tracking the patients' underlying health conditions with a combination of passive monitoring and proactive inspection. Also, the DI method is more accurate and robust than the other inspection methods. These observations are consistent with the observations on the simulated SSMs in Section IV.

B. Crowdsourcing

Another example of the state-space systems is the crowdsourcing problem. Crowdsourcing is an emerging paradigm [42]–[45] for companies to dynamically recruit specialized skills for solving challenges that consist of massive online tasks. It can be used to solve a variety of tasks, such as image recognition, data entry, and innovation and product development [46]–[48]. An illustration of this crowdsourcing process is shown in Fig. 6, which includes three workers assigned to two tasks. The first and second workers are assigned to task type 1, and the second and third workers are assigned to task type 2. As shown in Fig. 6, at each time epoch, the two tasks will be completed by the three workers. The quality of the completion of the tasks can be measured, i.e., the response score for task 1 at a certain time epoch is the average of the quality of workers 1 and 2 at that certain time epoch, corrupted with a measurement error (i.e., $v_1^{(t)}$).

Obviously, estimating the quality of the workers is very crucial for monitoring and optimization of the crowdsourcing platforms. This can be cast into an SSM with the latent variables characterizing the quality of the workers, and the response scores of the tasks being the observable variables. Previous works in crowdsourcing have shown the validity of the assumption of normality on workers' quality and responses [49], [50]. Jung [49] showed that the quality of workers (latent variables) could be modeled with normal distributions using a data set from the Amazon's Mechanical Turk, which is a major crowdsourcing platform. Zhao *et al.* [50] also defined "response score" as the measurement of workers'

TABLE IV
PERFORMANCE OF INSPECTION METHODS ON SIMULATED CROWDSOURCING DATA SERIES

	DI	MI	RI	SSM
# of testing tasks = 3				
Average RMSE	2.511	2.578	6.085	73.256
SD of RMSE	0.457	0.459	8.074	1.352
# of testing tasks = 4				
Average RMSE	2.131	2.199	3.296	73.256
SD of RMSE	0.449	0.452	3.098	1.352

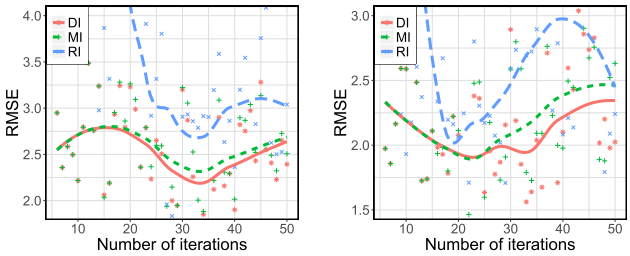


Fig. 7. RMSE of inspection methods on crowdsourcing problem with different number of testing tasks. (a) # of Testing tasks = 3. (b) # of Testing tasks = 4.

responses on the tasks, and showed the validity of normal assumptions on the scores. However, existing works only focus on statistical estimation using the measurements from the task completion. In practice, we can inspect the workers with additional testing tasks that we have known answers, to evaluate their quality. Neither the state-space formulation has been proposed nor the inspection strategy has been developed to combine measurement, monitoring, and inference. Thus, in this experiment, we articulate this state-space system and illustrate how our inspection methods can be applied to better estimate the quality of workers.

The experiment is conducted with 18 workers and 4 types of tasks on a simulated data series. Each worker is assigned to 2 of 4 types of tasks throughout the experiment, and therefore, we receive nine responses for each type of tasks at each time epoch. In our simulation setting, the quality of a worker changes over time according to a Brownian motion and is independent of other workers' quality. The response score of a task is assumed to be the average of workers' quality who are assigned to the task, plus an measurement error $\mathbf{v}_i^{(t)}$ that is randomly generated from the normal distribution $N(0, \mathbf{R}_{ii})$ where $\mathbf{R}_{ii}$ is the variance of the measurement error and is stationary. We also assume the score errors are independent of each other, i.e., $\mathbf{R}_{ij} = 0$ for $i \neq j$. The matrix $\mathbf{B}$ is set to be the identity matrix, and $\mathbf{Z}$ is specified to be a 0–1 matrix based on the allocation of tasks. The parameter $\mathbf{Q}$ is also a diagonal matrix where each element is sampled between 2 and 4. Therefore, in this experiment, the relation of latent variables is from the allocation of tasks only. Although the matrix $\mathbf{B}$ is identical in this experiment, the workers are still conditionally dependent given the response scores. At each time epoch, we can inspect M workers (here, we set M to be 3 and 4 and $c_i = 1$ for $i = 1, \ldots, J$).

The results are given in Table IV and Fig. 7. We can observe that the inspection methods can largely improve the estimation of the quality of the workers, while the DI method outperforms MI and RI in terms of both accuracy and robustness. The estimation accuracy and robustness of DI and MI improve when increasing the cost budget at each time epoch. We also note that a small number of inspections ($M \ll J$) is sufficient to reach a good estimation of the latent quality of workers most of the time. The standard deviations of RMSE of RI are much larger than those of DI, MI, and the original SSM because RI takes approximate 20 iterations to converge on this data set, and the RMSE is away from the mean before convergence, as shown in Fig. 7.

VI. CONCLUSION

In this paper, we propose a novel DI method to improve the state estimation in SSMs, by inspecting a small portion of latent variables at each time epoch. The selection of latent variables relies on the conditional estimation variances and the association of latent variables and is obtained by solving an s-t min-cut optimization problem. Simulation experiments show the DI method can achieve better performance than the original state estimation by the Kalman filter and other inspection methods. We apply the proposed methods to two real-world problems in type 1 diabetes and crowdsourcing. In the type 1 diabetes problem, we fill in the need of patients to monitor their levels of antibodies with our DI method. Compared with MI, RI, and classic SSM, the proposed DI model provides the most accurate and robust estimation of the levels of patient's antibodies. In the crowdsourcing problem, our method provides a DI strategy by assigning testing tasks to workers based on their responses and estimating the quality of workers. The proposed DI method also outperforms other methods with the most accurate and robust estimations of workers' quality.

The current method can be further improved. For example, the current DI method becomes slow when the length of data series grows, and it needs to be reestimated every time a new observation is added. It can be more efficient if the parameters and DI can be updated simultaneously. Previous works provide potential approaches to this problem but none of them solves it completely. Shwe and Yamamoto [51] showed that some parameters could be updated with given hidden states and observations. Another work in [52] proposed a method for point estimation of static parameters which did not suffer from "degeneracy problem" that many sequential Monte Carlo (SMC)-based algorithms suffered. Improved SMC method, SMC^2, can be used to update the parameters but suffers from bias or exponential cost in the dimension of the parameter space [53]. It is also interesting to explore the use of our method in the large hidden state-space scenario, e.g., a crowdsourcing project involves tens of thousands of workers that are assigned to hundreds of tasks. Therefore, it is necessary to develop an efficient inspection method which can make use of the sparsity of relation between tasks and between workers and tasks, and provide an accurate estimation of hidden states with a limited number of inspections. In the current DI formulation,

we assume a resource constraint exists at each time point. It is worth exploring how to formulate the optimization problem if we have a resource constraint for the entire process.

APPENDIX A
PROOF OF THEOREM 1

In this proof, we will introduce how to solve for the expected values and variances of the latent variables with the Kalman filter. First, the $I \times 1$ vector that represents the realization of the observable variables at time t, $\mathbf{y}^{(t)}$, is expanded to an $(I+J)\times 1$ vector $\tilde{\mathbf{y}}^{(t)}$ where $\tilde{\mathbf{y}}^{(t)}_{1:I} = \mathbf{y}^{(t)}$ and $\tilde{\mathbf{y}}^{(t)}_{I+1:I+J}$ is initialized as a missing vector. Then, the $I \times J$ matrix $\mathbf{Z}$ is expended to an $(I+J)\times J$ matrix $\tilde{\mathbf{Z}}$ where $\tilde{\mathbf{Z}}_{1:I,1:J} = \mathbf{Z}$, $\tilde{\mathbf{Z}}_{I+1:I+J,1:J} = \mathbf{I}_J$, $\mathbf{I}_J$ is the identical matrix of size J, and the rest elements are zeros. The $I \times I$ matrix $\mathbf{R}$ is extended to $\tilde{\mathbf{R}}$ where $\tilde{\mathbf{R}}_{1:I,1:I} = \mathbf{R}$ and the rest elements are zeros. The extended observation error $\tilde{\mathbf{v}}^{(t)}$is from $N(0, \tilde{\mathbf{R}})$. If variable i is inspected at time t, we replace the missing value $\tilde{\mathbf{y}}^{(t)}_{I+i}$ with the observed value. Therefore, in this problem, we actually solve (11) for the estimation of parameters and conditional variance

$$\begin{aligned}\mathbf{x}^{(t+1)} &= \mathbf{B}\mathbf{x}^{(t)} + \boldsymbol{\epsilon}^{(t)}, \boldsymbol{\epsilon}^{(t)} \sim N(0, \mathbf{Q})\\ \tilde{\mathbf{y}}^{(t+1)} &= \tilde{\mathbf{Z}}\mathbf{x}^{(t+1)} + \tilde{\mathbf{v}}^{(t)}, \tilde{\mathbf{v}}^{(t)} \sim N(0, \tilde{\mathbf{R}}).\end{aligned} \tag{11}$$

This construction benefits from the feature of the Kalman filter that it can incorporate missing values in the observations which we will discuss later. Note that based on the construction

$$P(\boldsymbol{X}^{(t)}|\tilde{\boldsymbol{Y}}^{(1:t)}) = P(\boldsymbol{X}^{(t)}|\boldsymbol{Y}^{(1:t)}, \boldsymbol{X}_I^{(1:t-1)}).$$

This equivalence is crucial for us to derive the needed statistics for implementing the DI methods. We will use $\tilde{\boldsymbol{Y}}^{(1:t)}$ in the following proof.

Then, we introduce the Kalman filter used in this problem. This section largely follows the deduction in [54, Sec. 6] with a slight abuse of notations. Let us define

$$\begin{aligned}\hat{\mathbf{x}}_s^{(t)} &= E[\boldsymbol{X}^{(t)}|\tilde{\boldsymbol{Y}}^{(1:s)}]\\ \mathbf{P}_s^{(t_1,t_2)} &= E\big[\big(\boldsymbol{X}^{(t_1)} - \hat{\mathbf{x}}_s^{(t_1)}\big)\big(\boldsymbol{X}^{(t_2)} - \hat{\mathbf{x}}_s^{(t_2)}\big)'|\tilde{\boldsymbol{Y}}^{(1:s)}\big].\end{aligned}$$

We write $\mathbf{P}_s^{(t)}$ for convenience if $t_1 = t_2 = t$, and $\mathbf{P}_t^{(t)}$ is the conditional covariance we want to obtain from the Kalman filter.

For the SSM specified in (11), first we define $(I+J)\times(I+J)$ matrix $\mathbf{I}_*^{(t)}$, where

$$\begin{aligned}\mathbf{I}^{(t)}_{*,(i,j)} &= 0, \quad \text{for } i \neq j\\ \mathbf{I}^{(t)}_{*,(i,i)} &= \begin{cases}0, & \text{if } \tilde{\mathbf{y}}_i^{(t)} \text{ is missing at } t\\ 1, & \text{otherwise}\end{cases}\end{aligned}$$

and

$$\tilde{\mathbf{y}}_*^{(t)} = \mathbf{I}_*^{(t)}\tilde{\mathbf{y}}^{(t)}, \tilde{\mathbf{Z}}_*^{(t)} = \mathbf{I}_*^{(t)}\tilde{\mathbf{Z}}.$$

Then with the initial conditions $\hat{\mathbf{x}}_0^{(0)} = \boldsymbol{\mu}_0$ and $\mathbf{P}_0^{(0)} = \Sigma_0$, we can compute $\mathbf{P}_t^{(t)}$ in the following way:

for $t = 1, \ldots, n$

$$\begin{aligned}\hat{\mathbf{x}}_{t-1}^{(t)} &= \mathbf{B}\hat{\mathbf{x}}_{t-1}^{(t-1)}\\ \mathbf{P}_{t-1}^{(t)} &= \mathbf{B}\mathbf{P}_{t-1}^{(t-1)}\mathbf{B}' + \mathbf{Q}\\ \hat{\mathbf{x}}_t^{(t)} &= \hat{\mathbf{x}}_{t-1}^{(t)} + \mathbf{K}^{(t)}\big(\tilde{\mathbf{y}}_*^{(t)} - \tilde{\mathbf{Z}}_*^{(t)}\hat{\mathbf{x}}_{t-1}^{(t)}\big)\\ \mathbf{P}_t^{(t)} &= \big(\mathbf{I} - \mathbf{K}^{(t)}\tilde{\mathbf{Z}}_*^{(t)}\big)\mathbf{P}_{t-1}^{(t)}\end{aligned}$$

where $\mathbf{K}^{(t)} = \mathbf{P}_{t-1}^{(t)}\tilde{\mathbf{Z}}_*'(\tilde{\mathbf{Z}}_*^{(t)}\mathbf{P}_{t-1}^{(t)}(\tilde{\mathbf{Z}}_*^{(t)})' + \tilde{\mathbf{R}})^{-1}$ is called the Kalman gain. The proof of this process can be found in [54]. Therefore, we can solve for the expectations and variances of the latent variables with the Kalman filter combining the inspections and observations.

APPENDIX B
PROOF OF LEMMA 2

Based on the construction in the proof of Theorem 1, if all the latent variables are inspected, we have $(\tilde{\mathbf{Z}}_*^{(t)})' = [\mathbf{Z}' \;\; \mathbf{I}_J]$. Then, we can write the multiplication of the Kalman gain and $\tilde{\mathbf{Z}}_*^{(t)}$ as

$$\begin{aligned}\mathbf{K}^{(t)}\tilde{\mathbf{Z}}_*^{(t)} &= \mathbf{P}_{t-1}^{(t)}\big(\tilde{\mathbf{Z}}_*^{(t)}\big)'\big(\tilde{\mathbf{Z}}_*^{(t)}\mathbf{P}_{t-1}^{(t)}\big(\tilde{\mathbf{Z}}_*^{(t)}\big)' + \tilde{\mathbf{R}}\big)^{-1}\tilde{\mathbf{Z}}_*^{(t)}\\ &= \mathbf{P}_{t-1}^{(t)}\big(\tilde{\mathbf{Z}}_*^{(t)}\big)'\left(\begin{bmatrix}\mathbf{Z}\\ \mathbf{I}_J\end{bmatrix}\mathbf{P}_{t-1}^{(t)}[\mathbf{Z}' \;\; \mathbf{I}_J] + \begin{bmatrix}\mathbf{R} & 0\\ 0 & 0\end{bmatrix}\right)^{-1}\tilde{\mathbf{Z}}_*^{(t)}\\ &= \mathbf{P}_{t-1}^{(t)}\big(\tilde{\mathbf{Z}}_*^{(t)}\big)'\begin{bmatrix}\mathbf{R}^{-1} & -\mathbf{R}^{-1}\mathbf{Z}\\ -\mathbf{Z}'\mathbf{R}^{-1} & \big(\mathbf{P}_{t-1}^{(t)}\big)^{-1} + \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}\end{bmatrix}\tilde{\mathbf{Z}}_*^{(t)}\\ &= \mathbf{P}_{t-1}^{(t)}\big(\mathbf{P}_{t-1}^{(t)}\big)^{-1} = \mathbf{I}.\end{aligned}$$

This implies $\mathbf{P}_t^{(t)} = 0$.

APPENDIX C
PROOF OF LEMMA 3

This follows from the proof of Theorem 1 and the deduction of the Kalman filter.

APPENDIX D
PROOF OF THEOREM 4

In this proof, we will show the optimization problem (9) is equivalent to the s-t min-cut problem introduced as follows.

An s-t cut $(C, V\backslash C)$ on a weighted graph (V, E) of size J is a partition of V such that the source $s \in C$ and the sink $t \in V\backslash C$. The cut-set E_C is the set of edges such that

$$E_C = \{(u, v) \in E | u \in C, v \in V\backslash C\}.$$

If $\mathbf{S}$ is the adjacency matrix of the graph, the s-t min-cut problem can be expressed as

$$\min_u \sum_{(u,v)\in E_C} \mathbf{S}_{u,v} = \sum_{u=1}^{J}\sum_{v=1}^{J}\boldsymbol{\delta}_u(1-\boldsymbol{\delta}_v)\mathbf{S}_{u,v}$$

where $\boldsymbol{\delta}_u = 1$ if $u \in C$ and 0 otherwise, $\boldsymbol{\delta}_s = 1$ and $\boldsymbol{\delta}_t = 0$.

The optimization problem (9) can be solved as an s-t min-cut problem mentioned earlier. First, let us define $\boldsymbol{\Omega}^{(t)}$ as the set of indices of the chosen latent variables at time t, and introduce artificial units $u_1 \in \boldsymbol{\Omega}^{(t)}$ and $u_2 \notin \boldsymbol{\Omega}^{(t)}$ as the source and the sink, while the related decision variables are fixed

to be $\boldsymbol{\delta}_{u_1,t} = 1$ and $\boldsymbol{\delta}_{u_2,t} = 0$. The covariances between the observed units and the artificial units are defined as

$$|\hat{\mathbf{S}}^{(t)}|_{u_1,i} = \left(\beta \sum_{j=1}^{J} |\hat{\mathbf{S}}_{ij}^{(t)}| + (1-\beta)|\hat{\mathbf{S}}_{ii}^{(t)}| - c_i\eta\right)_+$$

$$|\hat{\mathbf{S}}^{(t)}|_{i,u_2} = \left(c_i\eta - \beta \sum_{j=1}^{J} |\hat{\mathbf{S}}_{ij}^{(t)}| - (1-\beta)|\hat{\mathbf{S}}_{ii}^{(t)}|\right)_+$$

where $(x)_+ = \max\{x, 0\}$.

Then, the first part of (9) can be written as

$$\sum_{i=1}^{J} \delta_{it}\left(c_i\eta - \beta \sum_{j=1}^{J} |\hat{\mathbf{S}}_{ij}^{(t)}| - (1-\beta)|\hat{\mathbf{S}}_{ii}^{(t)}|\right)$$
$$= \sum_{i=1}^{J} |\hat{\mathbf{S}}^{(t)}|_{u_1,i}\delta_{u_1t}(1-\delta_{it}) + \sum_{i=1}^{J} |\hat{\mathbf{S}}^{(t)}|_{i,u_2}\delta_{it}(1-\delta_{u_2t}) - \sum_{i=1}^{J} |\hat{\mathbf{S}}^{(t)}|_{u_1,i}$$

where $\sum_{i=1}^{J} |\hat{\mathbf{S}}^{(t)}|_{u_1,i}$ is a constant.

Thus, the problem can be transformed into an s-t min-cut problem on a graph

$$\min_{\boldsymbol{\delta}_t \in \{0,1\}^J} \sum_{i=1}^{J} |\hat{\mathbf{S}}^{(t)}|_{u_1,i}\delta_{u_1t}(1-\delta_{it}) + \sum_{i=1}^{J} |\hat{\mathbf{S}}^{(t)}|_{i,u_2}\delta_{it}(1-\delta_{u_2t}) + \beta \sum_{i,j=1}^{J} |\hat{\mathbf{S}}_{ij}^{(t)}|\delta_{it}(1-\delta_{jt})$$

which can be solved by applying the maximal flow algorithm.

REFERENCES

[1] Y. Zou, S. E. Li, B. Shao, and B. Wang, "State-space model with non-integer order derivatives for lithium-ion battery," *Appl. Energy*, vol. 161, pp. 330–336, Jan. 2016.

[2] R. Moghaddass and C. Rudin, "The latent state hazard model, with application to wind turbine reliability," *Ann. Appl. Statist.*, vol. 9, no. 4, pp. 1823–1863, 2015.

[3] A. C. Harvey, *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge, U.K.: Cambridge Univ. Press, 1990.

[4] J. Jin and J. Shi, "State space modeling of sheet metal assembly for dimensional control," *J. Manuf. Sci. Eng.*, vol. 121, no. 4, pp. 756–762, 1999.

[5] J. Wang, H. Liang, and R. Chen, "A state space model approach for HIV infection dynamics," *J. Time Ser. Anal.*, vol. 33, no. 5, pp. 841–849, 2012.

[6] M. Quach, N. Brunel, and F. D'Alché-Buc, "Estimating parameters and hidden variables in non-linear state-space models based on odes for biological networks inference," *Bioinformatics*, vol. 23, no. 23, pp. 3209–3216, 2007.

[7] C. Rangel *et al.*, "Modeling T-cell activation using gene expression profiling and state-space models," *Bioinformatics*, vol. 20, no. 9, pp. 1361–1372, 2004.

[8] I. Shmulevich, E. R. Dougherty, S. Kim, and W. Zhang, "Probabilistic Boolean networks: A rule-based uncertainty model for gene regulatory networks," *Bioinformatics*, vol. 18, no. 2, pp. 261–274, 2001.

[9] H. Baltzakis and P. Trahanias, "A hybrid framework for mobile robot localization: Formulation using switching state-space models," *Auton. Robots*, vol. 15, no. 2, pp. 169–191, 2003.

[10] S. Mergner, *Applications of State Space Models in Finance: An Empirical Analysis of the Time-Varying Relationship Between Macroeconomics, Fundamentals and Pan-European Industry Portfolios*. Universitätsverlag Göttingen, 2009.

[11] J. C. Morley, "A state–space approach to calculating the Beveridge–Nelson decomposition," *Econ. Lett.*, vol. 75, no. 1, pp. 123–127, 2002.

[12] P. Kokic, S. Crimp, and M. Howden, "Forecasting climate variables using a mixed-effect state-space model," *Environmetrics*, vol. 22, no. 3, pp. 409–419, 2011.

[13] A. Stathopoulos and M. G. Karlaftis, "A multivariate state space approach for urban traffic flow modeling and prediction," *Transp. Res. C, Emerg. Technol.*, vol. 11, no. 2, pp. 121–135, 2003.

[14] H.-I. Suk, C.-Y. Wee, S.-W. Lee, and D. Shen, "State-space model with deep learning for functional dynamics estimation in resting-state fMRI," *NeuroImage*, vol. 129, pp. 292–307, Apr. 2016.

[15] Y. Ding *et al.*, "Modeling and diagnosis of multistage manufacturing processes: Part I: State space model," in *Proc. Japan/USA Symp. Flexible Autom.*, 2000, pp. 23–26.

[16] Y. Ding, D. Ceglarek, and J. Shi, "Fault diagnosis of multistage manufacturing processes by using state space approach," *J. Manuf. Sci. Eng.*, vol. 124, no. 2, pp. 313–322, 2002.

[17] Y. Ding, P. Kim, D. Ceglarek, and J. Jin, "Optimal sensor distribution for variation diagnosis in multistation assembly processes," *IEEE Trans. Robot. Autom.*, vol. 19, no. 4, pp. 543–556, Aug. 2003.

[18] M. Jin, Y. Li, and F. Tsung, "Chart allocation strategy for serial-parallel multistage manufacturing processes," *IIE Trans.*, vol. 42, no. 8, pp. 577–588, 2010.

[19] M. Jin and F. Tsung, "A chart allocation strategy for multistage processes," *IIE Trans.*, vol. 41, no. 9, pp. 790–803, 2009.

[20] E. Ertin, J. W. Fisher, and L. C. Potter, "Maximum mutual information principle for dynamic sensor query problems," in *Information Processing in Sensor Networks*. Berlin, Germany: Springer, 2003, pp. 405–416.

[21] H. Wang, K. Yao, G. Pottie, and D. Estrin, "Entropy-based sensor selection heuristic for target localization," in *Proc. 3rd ACM Int. Symp. Inf. Process. Sensor Netw.*, 2004, pp. 36–45.

[22] A. Krause, A. Singh, and C. Guestrin, "Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies," *J. Mach. Learn. Res.*, vol. 9, no. 2, pp. 235–284, 2008.

[23] X. Lin, A. Chowdhury, X. Wang, and G. Terejanu. (Mar. 2017). "Approximate computational approaches for Bayesian sensor placement in high dimensions." [Online]. Available: https://arxiv.org/abs/1703.00368

[24] G. C. Goodwin, S. F. Graebe, and M. E. Salgado, *Control System Design*. Upper Saddle River, NJ, USA: Prentice-Hall, 2001.

[25] J. H. Oud and R. A. Jansen, "Continuous time state space modeling of panel data by means of SEM," *Psychometrika*, vol. 65, no. 2, pp. 199–215, 2000.

[26] B. P. Carlin, N. G. Polson, and D. S. Stoffer, "A Monte Carlo approach to nonnormal and nonlinear state-space modeling," *J. Amer. Statist. Assoc.*, vol. 87, no. 418, pp. 493–500, 1992.

[27] G. Kitagawa, "Monte Carlo filter and smoother for non-Gaussian nonlinear state space models," *J. Comput. Graph. Statist.*, vol. 5, no. 1, pp. 1–25, 1996.

[28] T. B. Schön, F. Gustafsson, and P.-J. Nordlund, "Marginalized particle filters for mixed linear/nonlinear state-space models," *IEEE Trans. Signal Process.*, vol. 53, no. 7, pp. 2279–2289, Jul. 2005.

[29] Y. Yuan, G.-B. Stan, S. Warnick, and J. Goncalves, "Robust dynamical network structure reconstruction," *Automatica*, vol. 47, no. 6, pp. 1230–1235, 2011.

[30] J. Adebayo *et al.*, "Dynamical structure function identifiability conditions enabling signal structure reconstruction," in *Proc. IEEE 51st Annu. Conf. Decis. Control (CDC)*, Dec. 2012, pp. 4635–4641.

[31] M. Imani and U. M. Braga-Neto, "Particle filters for partially-observed Boolean dynamical systems," *Automatica*, vol. 87, pp. 238–250, 2018.

[32] J. Linder and M. Enqvist, "Identification and prediction in dynamic networks with unobservable nodes," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 10574–10579, 2017.

[33] J. M. Hendrickx, M. Gevers, and A. S. Bazanella. (Mar. 2018). "Identifiability of dynamical networks with partial node measurements." [Online]. Available: https://arxiv.org/abs/1803.05885

[34] G. Liang, B. Yu, and N. Taft, "Maximum entropy models: Convergence rates and applications in dynamic system monitoring," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun./Jul. 2004, p. 168.

[35] E. E. Holmes, E. J. Ward, and K. Wills, "MARSS: Multivariate autoregressive state-space models for analyzing time-series data," *R J.*, vol. 4, no. 1, pp. 11–19, 2012.

[36] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Trans. ASME, D, J. Basic Eng.*, vol. 82, pp. 35–45, 1960.

[37] E. E. Holmes. (Feb. 2013). "Derivation of an em algorithm for constrained and unconstrained multivariate autoregressive state-space (MARSS) models." [Online]. Available: https://arxiv.org/abs/1302.3919

[38] C. H. Papadimitriou and K. Steiglitz, *Combinatorial Optimization: Algorithms and Complexity*. North Chelmsford, MA, USA: Courier, 1998.
[39] A. V. Goldberg and R. E. Tarjan, "A new approach to the maximum-flow problem," *J. ACM*, vol. 35, no. 4, pp. 921–940, 1988.
[40] Diabetes Prevention Trial–Type 1 Diabetes Study Group, "Effects of insulin in relatives of patients with type 1 diabetes mellitus," *New England J. Med.*, vol. 346, no. 346, pp. 1685–1691, 2002.
[41] I. Aloulou, J.-F. Brun, and J. Mercier, "Evaluation of insulin sensitivity and glucose effectiveness during a standardized breakfast test: Comparison with the minimal model analysis of an intravenous glucose tolerance test," *Metabolism*, vol. 55, no. 5, pp. 676–690, 2006.
[42] V. Diederik, K. Dervojeda, F. Nagtegaal, J. Sjauw-Koen-Fa, L. Frideres, and L. Probst, "Smart factories: Crowdsourced manufacturing," *Bus. Innov. Observatory*, 2014.
[43] J. Howe, "The rise of crowdsourcing," *Wired Mag.*, vol. 14, no. 6, pp. 1–4, Jun. 2006.
[44] D. C. Brabham, "Crowdsourcing as a model for problem solving: An introduction and cases," *Converg.*, vol. 14, no. 1, pp. 75–90, 2008.
[45] C.-M. Chiu, T.-P. Liang, and E. Turban, "What can crowdsourcing do for decision support?" *Decision Support Syst.*, vol. 65, pp. 40–49, Sep. 2014.
[46] D. C. Brabham, K. M. Ribisl, T. R. Kirchner, and J. M. Bernhardt, "Crowdsourcing applications for public health," *Amer. J. Preventive Med.*, vol. 46, no. 2, pp. 179–187, 2014.
[47] S. Marjanovic, C. Fry, and J. Chataway, "Crowdsourcing based business models: In search of evidence for innovation 2.0," *Sci. Public Policy*, vol. 39, no. 3, pp. 318–332, 2012.
[48] A. Kittur, E. H. Chi, and B. Suh, "Crowdsourcing user studies with Mechanical Turk," in *Proc. ACM SIGCHI Conf. Hum. Factors Comput. Syst.*, 2008, pp. 453–456.
[49] H. J. Jung, "Quality assurance in crowdsourcing via matrix factorization based task routing," in *Proc. ACM 23rd Int. Conf. World Wide Web*, 2014, pp. 3–8.
[50] Z. Zhao, F. Wei, M. Zhou, W. Chen, and W. Ng, "Crowd-selection query processing in crowdsourcing databases: A task-driven approach," in *Proc. EDBT*, 2015, pp. 397–408.
[51] P. E. E. Shwe and S. Yamamoto, "Real-time simultaneously updating a linearized state-space model and pole placement gain," in *Proc. IEEE 55th Annu. Conf. Soc. Instrum. Control Eng. Jpn. (SICE)*, Sep. 2016, pp. 196–201.
[52] C. Andrieu, A. Doucet, and V. B. Tadic, "On-line parameter estimation in general state-space models," in *Proc. 44th IEEE Conf. Decis. Control, Eur. Control Conf. (CDC-ECC)*, Dec. 2005, pp. 332–337.
[53] Y. Zhou and A. Jasra. (Mar. 2015). "Biased online parameter inference for state-space models." [Online]. Available: https://arxiv.org/abs/1503.00266
[54] R. H. Shumway and D. S. Stoffer, *Time Series Analysis and Its Applications: With R Examples*. New York, NY, USA: Springer Science Business Media, 2006.

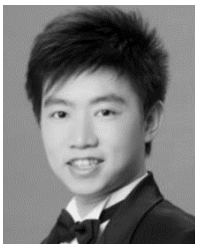
Tianshu Feng received the B.S. degree in statistics from the University of Science and Technology of China, Hefei, China, in 2015. He is currently pursuing the Ph.D. degree in industrial and systems engineering with the University of Washington, Seattle, WA, USA.

His research interests include statistical modeling, data mining, and quality engineering.

Xiaoning Qian (S'01–M'07–SM'17) received the Ph.D. degree in electrical engineering from Yale University, New Haven, CT, USA.

He is currently an Associate Professor with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, USA. He is also with the Center for Bioinformatics and Genomic Systems Engineering and the Center for Translational Environmental Health Research, Texas A&M University. His research interests include Bayesian learning, optimization, and their applications in computational network biology, genomic signal processing, biomedical signal and image analysis, as well as experimental design for novel materials discovery.

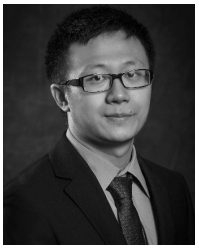
Dr. Qian was a recipient of the National Science Foundation CAREER Award, the Texas A&M Engineering Experiment Station Faculty Fellow and the Montague-Center for Teaching Excellence Scholar at Texas A&M University. His recent work on computational network biology has received the Best Paper Award at the 11th Asia Pacific Bioinformatics Conference in 2013 and the Best Paper Award in the International Conference on Intelligent Biology and Medicine in 2016.



Kaibo Liu (M'14) received the B.S. degree in industrial engineering from the Hong Kong University of Science and Technology, Hong Kong, in 2009 and the M.S. degree in statistics and the Ph.D. degree in industrial engineering from the Georgia Institute of Technology, Atlanta, GA, USA, in 2011 and 2013, respectively.

He is currently an Assistant Professor with the Department of Industrial and Systems Engineering, University of Wisconsin-Madison, Madison, WI, USA. His research interests include data fusion for process modeling, monitoring, diagnosis, prognostics, and decision making.

Dr. Liu is a member of ASQ, INFORMS, and IIE.



Shuai Huang (M'12) received the B.S. degree in statistics from the University of Science and Technology of China, Hefei, China, in 2007 and the Ph.D. degree in industrial engineering from Arizona State University, Tempe, AZ, USA, in 2012.

He is currently an Assistant Professor with the Department of Industrial and Systems Engineering, University of Washington, Seattle, WA, USA. His research interests include statistical learning and data mining with applications in healthcare and manufacturing.

Dr. Huang is a member of INFORMS, IIE, and ASQ.