A Bayesian Approach to Robust Process Identification with ARX Models

Shima Khatibisepehr and Biao Huang
Dept. of Chemical and Materials Engineering, University of Alberta, Edmonton, AB T6G 2G6, Canada

DOI 10.1002/aic.13887
Published online August 1, 2012 in Wiley Online Library (wileyonlinelibrary.com).

In the context of process industries, outlying observations mostly represent a large random error resulting from irregular process disturbances, instrument failures, or transmission problems. Statistical analysis of process data contaminated with outliers may lead to biased parameter estimation and plant-model mismatch. The problem of process identification in the presence of outliers has received great attention and a wide variety of outlier identification approaches have been proposed. However, there is a great need to seek for more general solutions and a robust framework to deal with different types of outliers. The main objective of this work is to formulate and solve the robust process identification problem under a Bayesian framework. The proposed solution strategy not only yields maximum a posteriori estimates of model parameters but also provides hyperparameters that determine data quality as well as prior distribution of model parameters. Identification of a simulated continuous fermentation reactor is considered to show the effectiveness and robustness of the proposed Bayesian framework. The advantages of the method are further illustrated through an experimental case study of a pilot-scale continuous stirred tank heater. © 2012 American Institute of Chemical Engineers *AIChE J,* 59: 845–859, 2013
Keywords: process identification, Bayesian inference, outliers

Introduction

Reliable process models are key requirements for investigating the behavior of industrial processes. Such descriptive models can help to improve process productivity, achieve safety of operation, and develop tight control policies.[1]

Depending on the level of *a priori* knowledge, different strategies have been proposed in the literature to model chemical processes. Traditionally, knowledge-driven models are developed on the basis of first principles analysis, which requires complete understanding of underlying mechanisms.[2–4] Not surprisingly, development of first principle models can often be prohibitively difficult and time-consuming due to the complexity of industrial processes. Therefore, decades of research have been devoted to develope empirical process models without complete *a priori* knowledge of the internal mechanisms governing the process dynamics. The empirical models are usually constructed based on the limited process knowledge as well as the great amount of historical data acquired for monitoring purposes.[5,6] In the context of process industries, various data-driven model structures can be used to describe the behavior of unit operations. Owing to the capability of autoregressive with exogenous (ARX) input models in approximating complex linear dynamic systems, ARX model structure is commonly adopted in industrial applications.[1] Regardless of the model structure selected, the procedures applied to identify empirical process models are often highly sensitive to the quality of identification data, that is, the varying quality of process data can greatly deteriorate the performance of data-driven identification methods.

Outlying measurements, also called ''outliers'', are one of the common factors that may affect the quality of operational and laboratory data.[7–9] Outliers are observations which appear to deviate markedly from the typical ranges of other observations.[10] The outliers in operational data mostly represent a random error caused by such issues as process disturbances, instrument degradation, and transmission problems.[11,12] Moreover, the outlying laboratory measurements may be generated due to potential human errors that may occur in collecting samples, conducting experiments, and recording results. Statistical analysis of process data contaminated with outliers may lead to biased parameter estimation and plant-model mismatch. Therefore, the problem of process model identification in the presence of outliers has received great attention during the last two decades and a wide variety of so-called outlier identification approaches have been proposed.[13,14] As pointed out by Kadlec et al.,[6] however, this issue is currently solved in a rather *ad hoc* manner, which leads to unnecessarily high costs of the process model identification. Therefore, there is a great need to seek for more advanced and more general solutions.

In principle, Bayesian formulation of empirical models suggests a general solution for many types of systems including linear and nonlinear systems, in the presence of Gaussian or non-Gaussian disturbances, with or without constraints, and in dealing with regular or irregular data samples. The main contribution of this work is to formulate and solve the ARX model identification problem in the presence of outliers under a robust Bayesian framework consisting of

Correspondence concerning this article should be addressed to B. Huang at biao.huang@ualberta.ca.

© 2012 American Institute of Chemical Engineers

consecutive levels of optimization. First, we adopt a contaminated distribution to describe the observed data and introduce a set of indicator variables to denote the quality of each data point. Next, we propose a unified objective function for model identification in the presence of outliers. The resulting optimization problem is hierarchically decomposed and a layered optimization strategy is implemented. To obtain explicit solutions, we adopt an iterative hierarchical Bayesian approach through which the solutions obtained in subsequent layers of optimization are coordinated. The proposed optimization strategy not only yields maximum *a posteriori* (MAP) estimates of model parameters but also provides an automated mechanism for determining the hyperparameters and for investigating the quality of each observation. Moreover, the developed framework allows us to incorporate the prior knowledge of the noise distribution and to include the relevant information contained in identification data. As a result, restrictive assumptions made in traditional robust methods about contaminating distributions (e.g., symmetric noise distribution) can be relaxed. Also, note that the identification procedures used in classical statistical estimation techniques often result in a set of single-valued parameter estimates. In contrast, the full Bayesian model identification results in posterior distributions over parameters to reveal how uncertain the estimated values would be.[9] More importantly, the Bayesian inference at a particular level takes into account the uncertainty in the estimates of the previous level. This is a great feature that allows us to link different levels of Bayesian inference together and, consequently, interconnect the solutions obtained in subsequent layers of optimization.

The remainder of this article is organized as follows. A brief overview of the existing outlier identification techniques is presented. The problem of ARX model identification in the presence of outliers is discussed. Our proposed objective function resulting in the consecutive layers of optimization is described. The idea of hierarchical Bayesian inference approach adopted to solve the layered optimization problem is explained. The most common outlier models are introduced. The problem of ARX model parameter estimation is formulated in a unified Bayesian framework and the details of the identification procedure are presented. The application of the developed framework is demonstrated on numerical simulation and experimental examples. These case studies will show robustness of the proposed parameter estimation method in the presence of outliers, which is an attractive feature for applying the proposed method to real world problems. Finally, this article is summarized with the concluding remarks.

An Overview of the Existing Outlier Identification Methods

Outlier identification constitutes an essential prerequisite for identification of process models and thus several outlier handling approaches have been developed in the past few decades. As the focus of this article is on the Bayesian methods, we limit our literature review to the most common "statistical" approaches. A comprehensive review of the outlier detection problem and several outlier detection algorithms is given by Hodge and Austin,[13] Kadlec et al.,[6] and Chandola et al.[14]

Statistical analysis of residuals described in Fortuna et al.[1] is one of the common outlier detection approaches. This is based on the use of a regression model between dependent and independent variables. First, the least-square method is applied to obtain an estimation of model parameters for normal operating condition. Outliers can then be detected if the model residuals of new data (r_i) lie outside a specified confidence interval. As outliers can significantly deteriorate least-squares solutions, robust regression can be applied to handle them while fitting regression models. In general, robust regression methods are designed to iteratively downweight the influence of outliers. The most common robust regression analysis is performed with M (Maximum likelihood) estimators, introduced by Huber.[15] The general M-estimator minimizes the objective function $\sum_i^n \rho(r_i)$, where the function ρ gives the contribution of each residual to the objective function (e.g., for least-squares estimation $\rho(r_i) = r_i^2$). Differentiating the objective function with respect to the parameters and setting the partial derivatives to 0, the estimating equations may be written as $\sum_i^n w_i r_i X_i = 0$, where the robustness weight assigned to the ith observation, $w_i = w(r_i)$, is obtained from the weight function defined as $w(r) = \rho'/r$. For instance, the robustness weights in the Huber robust regression technique are determined using the Huber weighting function, that is

$$w_i = \left[\max\left(1, \left| \frac{r_i}{c \times s} \right| \right) \right]^{-1} \tag{1}$$

where r_i is the residual calculated from the previous iteration, c is the tuning constant, and $s = \frac{MAD}{0.6745}$ is an estimate of the standard deviation of the error term.

Several solutions have been proposed for solving the outlier detection problem by estimating a probability density of the normal data. For instance, in Bishop[16] the density distribution of the input space is first estimated by a standard Parzen window approach with Gaussian kernel functions. Next, a suitable threshold is specified based on the identification dataset which is known to be representative of normal data. The new observation is then flagged as an outlier, if the value of the density function is above the threshold. Yu[17] proposes a Bayesian approach to first estimate the posterior probabilities of all samples within the model input space and specify appropriate confidence levels. A calibration procedure is then followed to correct the observations identified as outliers. An alternative approach for probability density estimation is to model normal instances as a mixture of parametric distributions. Bishop[16] and Agarwal[18] use Gaussian mixture models for such techniques. In Ritter and Gallegos,[19] both normal instances and outliers are modeled as separate parametric distributions. First, the ellipsoidal multivariate trimming[20] technique is used to detect outliers and to estimate distribution parameters of both outliers and regular observations. Next, a Bayesian classifier is designed to compare certain linear combinations of posterior densities of each data vector with respect to the estimated distributions. Several variations of Bayesian classification technique have further been proposed by Varbanov,[21] Ghosh-Dastidar and Schafer,[22] Das and Schneider,[23] and many others.

In this research, we take a hierarchical Bayesian approach to address the problem of model identification in the presence of outliers. We develop a robust Bayesian inference framework consisting of three consecutive steps: (1) Given an identification dataset, the posterior probability of each observation acting as an outlier is evaluated; a set of indicator

variables is specified to denote the identity of each data point. (2) The hyperparameters are then estimated by solving an optimization problem that maximizes the posterior probability distribution of hyperparameters conditional on the indicator variables. (3) Given current estimates of hyperparameters and indicator variables, the posterior probability distribution of model parameters is maximized to obtain MAP estimates. These three steps will be repeated until the estimates change within a given tolerance.

Problem Statement

In the identification of an empirical model, the overall objective is to find a model that best fits the identification dataset, $\mathcal{D} = \{(r_i,y_i)\}_{i=1}^{N}$. Bayesian models are a compact way to represent probabilistic relationships between a set of random variables in a system. Before going into details of how to learn Bayesian models, we need a more detailed definition of what the model includes. A model is defined by its functional form, f, and a set of parameters, Θ. Let us consider a general form of a nonlinear model

$$y_t = f(r_t, \Theta) + e_t \tag{2}$$

where $y_t \in \mathbb{R}$ is the output, $r_t \in \mathbb{R}^p$ is the regressor constructed from past inputs and outputs, and e_t is the noise/error term.

Suppose e_t is modeled as zero-mean Gaussian noise with constant standard deviation σ_e. Given the model structure, $\mathcal{M}$, and the model parameters, Θ, the likelihood of the data can be expressed as

$$P(\mathcal{D}|\Theta,\zeta,\mathcal{M}) = \left(\frac{\zeta}{2\pi}\right)^{N/2} \exp\left(-\zeta E_{\mathrm{D}}(\mathcal{D}|\Theta,\zeta,\mathcal{M})\right) \tag{3}$$

where ζ defines a noise level with $\sigma_e^2 = \zeta^{-1}$, and E_{D} is the error term defined as

$$E_{\mathrm{D}} = \frac{1}{2}\sum_{i=1}^{N} e_i^2 = \frac{1}{2}\sum_{i=1}^{N} \left(y_i - f(r_i,\Theta)\right)^2 \tag{4}$$

It is well-known that finding the maximum likelihood estimates of the parameters, Θ_{ML}, may be an ill-posed problem. As the Θ that minimizes E_{D} may depend sensitively on the details of the noise in the data, the maximum likelihood estimates would oscillate widely so as to fit the noise.[24] Bayesian methods solve this type of ill-posed problem by combining information contained in the observed data with available information concerning the distribution of the parameters. Introducing a regularizing constant, α, such a prior can be expressed on the parameters; $P(\Theta|\alpha,\mathcal{M})$ represents the current state of knowledge about the plausible values of model parameters. Therefore, the prior distribution of model parameters is parameterized by a set of variables called hyperparameters.* Both ζ and α are considered as hyperparameters, because they describe the overall characteristics of the priors. If a hyperparameter is not known *a priori*, its probability distributions can be estimated in an intermediate step of the model identification process.

To develop a Bayesian formulation of inferential models that is robust to inconsistent data, we need to be able to efficiently perform different levels of Bayesian inference even if the dataset is contaminated with outlying observations. Given an identification dataset $\mathcal{D}$, we can consider a set of hyperparameters $\{\zeta_1,\ldots,\zeta_N\}$. Thus, the hyperparameter ζ_i defines a noise level $\sigma_{ei}^2 = \zeta_i^{-1}$ on the ith sample in the given training dataset. When having nonconstant values of ζ_i, the outliers will be automatically handled by assigning less weights to the observations with relatively larger σ_{ei}^2. However, the underlying formulation involves a heavy nonlinear optimization problem in dealing with large datasets.

To obtain a computationally feasible formulation, we adopt a contaminated distribution to describe the observed data and then solve the problem under a unified Bayesian framework. The error distribution function is thus expressed as $F(e) = \delta\, G(e) + (1-\delta)H(e)$, where δ is the unknown prior probability of appearance of an outlier, $H(e) = N(0, \sigma_e^2)$ is a normal distribution, and $G(e)$ is a contaminating distribution. This model arises for instance if the observations are assumed to be normal with variance σ_y^2, but a fraction δ of them is affected by gross errors.[15] Moreover, a set of indicator variables $\{q_1, \ldots, q_N\}$ is introduced to denote the identity of each data point; the indicator variable associated with each data point determines whether that observation comes from the regular or contaminating distribution. However, the indicator variables are usually not known *a priori* and should be estimated in an intermediate step of the model identification process.

Outlier Models

In general, we need to tackle two types of outliers, namely scale outliers and location outliers. As the names suggested, scale and location outliers are generated by a shift in the scale (variability) or in the location (mean) of measurement noise. The process measurements that violate the physical limitations of the involved unit operations can be modeled as scale outliers, whereas the ones that violate the technological limitations of the measuring devices can often be considered as symmetric location outliers. Moreover, the outlying measurements made by a jammed instrument may be modeled as asymmetric location outliers.

In this section, we present our proposed scale and location outlier models which later will be needed to develop a robust Bayesian framework.

Scale outlier model

The error distribution affected by scale outliers is a mixture of two multivariate normal distributions centered at the same mean but with different covariance matrices, one being proportionately larger than the other. Therefore, it is assumed that the noise term, e_i, is distributed as

$$e_i \sim \delta N(0, \rho^{-1}\sigma_e^2) + (1-\delta)N(0,\sigma_e^2) \tag{5}$$

where $0 < \rho < 1$ is the variance inflation factor that indicates the magnitude of the errors leading to an outlying observation. Note that the proposed Bayesian framework does not require any knowledge of the noise distribution parameters (e.g., δ, σ_e, and ρ); these parameters are iteratively estimated in the identification process using the observations identified as outliers.

Introduce a set of indicator variables, $q_{1:N} = \{q_1, \ldots, q_N\}$, to denote identity of each data point; that is, $q_i = \rho$ if e_i is distributed as $N(0,\rho^{-1}\sigma_e^2)$ and $q_i = 1$ if e_i is distributed as $N_n(0,\sigma_e^2)$. Therefore, q_i is Bernoulli distributed with parameter δ, that is, $P(q_i;\delta) = \delta^{\left(1-\frac{q_i-\rho}{1-q_i\rho}\right)}(1-\delta)^{\left(\frac{q_i-\rho}{1-q_i\rho}\right)}$.

*The term is used to distinguish them from model parameters.

Location outlier model

Now, suppose the contaminating distribution consists of two multivariate normals such that $G(e) = N(-\Delta,\sigma_e^2) + N(\Delta,\sigma_e^2)$. To capture the presence of location outliers, it is thus assumed that the noise term, e_i, is distributed as

$$e_i \sim \delta[N(\Delta,\sigma_e^2) + N(-\Delta,\sigma_e^2)] + (1-\delta)N(0,\sigma_e^2) \quad (6)$$

where Δ indicates the location shift in the outlying observations. As mentioned previously, the proposed Bayesian framework does not require any knowledge of the noise distribution parameters (e.g., δ, σ_e, and Δ); these parameters are iteratively estimated in the identification process using the observations identified as outliers.

Introduce a set of indicator variables, $q_{1:N} = \{q_1, \ldots, q_N\}$, to denote identity of each data point; that is $q_i = +\Delta$ if e_i is generated from $N(+\Delta, \sigma_e^2)$, $q_i = -\Delta$ if e_i is generated from $N(-\Delta,\sigma_e^2)$, and $q_i = 0$ if e_i is distributed as $N_n(0,\sigma_e^2)$. Therefore, q_i has a categorical distribution such that $P(q_i;\delta) = (0.5\delta)^{\left(\frac{|q_i|-q_i}{2\Delta}\right)}(0.5\delta)^{\left(\frac{|q_i|+q_i}{2\Delta}\right)}(1-\delta)^{\left(1-\frac{|q_i|}{\Delta}\right)}$, or equivalently, $|q_i|$ has a Bernoulli distribution such that $P(|q_i|;\delta) = \delta^{\frac{|q_i|}{\Delta}}(1-\delta)^{\left(1-\frac{|q_i|}{\Delta}\right)}$.

Hierarchical Optimization Framework

In general, the identification problem is to estimate the model parameters, Θ, the hyperparameters of the prior distribution of model parameters, Φ, and the indicator variables, Q, using the process dataset, $\mathcal{D}$. To obtain MAP estimates simultaneously, the joint probability density function (JPDF), $P(\Theta,\Phi,Q|\mathcal{D})$ should be optimized. However, evaluating such posterior density functions requires a complex nonlinear optimization problem to be solved. To circumvent the difficulties associated with the direct maximization of $P(\Theta,\Phi,Q|\mathcal{D})$, the identification problem is formulated under a layered optimization framework, as we will show in the following.

First, the chain rule of probability theory is used to factorize the JPDF as

$$P(\Theta,\Phi,Q|\mathcal{D}) = P(\Theta|\Phi,Q,\mathcal{D})P(\Phi|Q,\mathcal{D})P(Q|\mathcal{D}) \quad (7)$$

Then, the optimization problem is decomposed hierarchically into three layers

$$\begin{aligned}\max_{\Theta,\Phi,Q} & P(\Theta|\Phi,Q,\mathcal{D})P(\Phi|Q,\mathcal{D})P(Q|\mathcal{D}) \\ &= \max_{\Phi,Q}\left\{P(Q|\mathcal{D})P(\Phi|Q,\mathcal{D})\max_{\Theta}\left\{P(\Theta|\Phi,Q,\mathcal{D})\right\}\right\} \\ &= \max_{Q}\left\{P(Q|\mathcal{D})\max_{\Phi}\left\{P(\Phi|Q,\mathcal{D})\max_{\Theta}\left\{P(\Theta|\Phi,Q,\mathcal{D})\right\}\right\}\right\}\end{aligned} \quad (8)$$

The three-layer optimization problem is formulated as follows:

1. Inference of model parameters Θ by maximizing the following posterior density function

$$P(\Theta|\mathcal{D},\Phi,Q) = \frac{P(\mathcal{D}|\Theta,\Phi,Q)P(\Theta|\Phi,Q)}{P(\mathcal{D}|\Phi,Q)} \quad (9)$$

2. Inference of hyperparameters Φ by maximizing the following posterior density function

$$P(\Phi|\mathcal{D},Q) = \frac{P(\mathcal{D}|\Phi,Q)P(\Phi|Q)}{P(\mathcal{D}|Q)} \quad (10)$$

3. Inference of outlier indicator variables Q by maximizing the following posterior density function

$$P(Q|\mathcal{D}) = \frac{P(\mathcal{D}|Q)P(Q)}{P(\mathcal{D})} \quad (11)$$

In this Bayesian formulation, the likelihood function at a particular level corresponds to the evidence function at the previous level. For example, the likelihood at Level 2, $P(\mathcal{D}|\Phi,Q)$, is equal to the evidence at Level 1. Through this pattern, the optimization variables are gradually integrated out at different levels of Bayesian inference. Consequently, the optimal solutions obtained in subsequent layers of optimization are coordinated. However, direct optimization of all these three layers is still not a tractable problem and further simplification is required.

To obtain a tractable explicit solution to the above layered optimization problem, we adopt a hierarchical Bayesian approach through which the posterior probability density functions are sequentially approximated in each layer and the procedure is iterated. The hierarchical Bayesian approach has been applied to a great variety of problems. For instance, MacKay[24] is the first author who proposed the heuristic Bayesian evidence framework and later on applied it to neural network modeling.[25] Molina et al.[26] and Galatsanos et al.[27] used the hierarchical Bayesian paradigm to address the image modeling and restoration problem. Kwok[28] and Suykens et al.[29] derived a probabilistic formulation of the least squares support vector machine within a hierarchical Bayesian evidence framework.

Formulation of Inferential Modeling Problem in a Unified Bayesian Framework

To derive analytical expressions for all levels of inference, here we use the popular ARX model to illustrate the design of a robust unified Bayesian framework. The application of the ideas presented in this section is not limited to ARX models. The derivations can be directly extended to other classes of dynamic models, although numerical optimization may be required.

For fixed model orders *na* and *nb*, an ARX model is defined by introducing the regression vector $r_t \in \mathbb{R}^p$

$$r_t = [y_{t-1}, \ldots, y_{t-na}, u_{t-1}^T, \ldots, u_{t-nb}^T]^T \quad (12)$$

where $u \in \mathbb{R}^m$ is the input and $p = na + m \cdot nb$.

The output can then be expressed as a linear function of r_t such that

$$y_t = \Theta^T \begin{bmatrix} r_t \\ 1 \end{bmatrix} + e_t \quad (13)$$

where y_t is the output, e_t is a zero-mean Gaussian noise with nonconstant variance and $\Theta = [\theta_1,\ldots,\theta_p,\theta_{p+1}]^T \in \mathbb{R}^{p+1}$ denotes the parameter vector including a subset of model parameters, $\Theta_{1:p} = [\theta_1,\ldots,\theta_p]^T$, and a bias term, θ_{p+1}. The reason for keeping $\Theta_{1:p}$ and θ_{p+1} distinct will become clear in deriving analytical expressions for the location outlier model.

Given the identification dataset that is contaminated by the presence of outliers, the objective is to identify model parameters Θ. The proposed hierarchical Bayesian optimization framework allows us to obtain MAP estimates of model parameters with an automated mechanism for determining

the hyperparameters and investigating the quality of each data point.

Inference of Model Parameters Θ

Given the identification dataset $\mathcal{D} = \{(r_i,y_i)\}_{i=1}^N = \{Z\}_{i=1}^N$ along with a set of indicator variables $q_{1:N} = \{q_1, \ldots, q_N\}$ and the hyperparameters $\alpha_{1:p+1} = \{\alpha_1, \ldots, \alpha_{p+1}\} = \{\sigma_{\theta 1}^{-2}, \ldots, \sigma_{\theta p+1}^{-2}\}$ and $\zeta = \sigma_e^{-2}$, the MAP estimates of model parameters are obtained by maximizing the posterior $P(\Theta|\mathcal{D},\alpha_{1:p+1},\zeta,q_{1:N})$. Thus, the formulation of Bayes' Theorem in the first level of optimization becomes

$$P(\Theta|\mathcal{D},\alpha_{1:p+1},\zeta,q_{1:N}) = \frac{P(\mathcal{D}|\Theta,\alpha_{1:p+1},\zeta,q_{1:N})P(\Theta|\alpha_{1:p+1},\zeta,q_{1:N})}{P(\mathcal{D}|\alpha_{1:p+1},\zeta,q_{1:N})} \tag{14}$$

It is reasonable to assume that the prior distribution of each parameter $\theta_i \in \Theta$ is independent of hyperparameter ζ and indicator variables $q_{1:N}$, that is, $P(\theta_i|\alpha_i,\zeta,q_{1:N}) = P(\theta_i|\alpha_i)$. In the absence of other prior information, the prior distribution of Θ is taken as independent Gaussian with zero-mean and variance of $\sigma_{\theta j}^2 = \alpha_j^{-1}$

$$\begin{aligned} P(\Theta|\alpha_{1:p+1}) &= \prod_{j=1}^{p+1} P(\theta_j|\alpha_j) \\ &= \prod_{j=1}^{p+1} \sqrt{\frac{\alpha_j}{2\pi}} \exp\left(-\frac{1}{2}\alpha_j\theta_j^2\right) \end{aligned} \tag{15}$$

It is noteworthy that a set of independent hyperparameters $\{\alpha_1, \ldots, \alpha_{p+1}\}$ is specified to obtain sparsity. Considering that the bias could be any value, an uniform prior is chosen for θ_{p+1}; that is, $\alpha_{p+1} \to 0$ to approximate a uniform distribution, which can also be considered as a Gaussian distribution in the limit. Plugging in our assumptions, the prior is then expressed as follows

$$P(\Theta|\alpha_{1:p},\zeta,q_{1:N}) \propto \prod_{j=1}^{p} \sqrt{\frac{\alpha_j}{2\pi}} \exp\left(-\frac{1}{2}\alpha_j\theta_j^2\right) \tag{16}$$

The chain rule of probability theory allows us to factorize joint probabilities as

$$\begin{aligned} P(\mathcal{D}) &= P(Z_1, Z_2, \ldots, Z_N) \\ &= \prod_{i=1}^{N} P(Z_i|Z_{1:i-1}) \end{aligned} \tag{17}$$

Given Θ, the sampled data $\mathcal{D}$ would be independent of hyperparameters $\alpha_{1:p}$ (inverse of the variance of the prior distribution of model parameters), that is, $P(\mathcal{D}|\Theta,\alpha_{1:p},\zeta,q_{1:N}) = P(\mathcal{D}|\Theta,\zeta,q_{1:N})$. Applying the chain rule, therefore, the likelihood can be further expressed as

$$\begin{aligned} P(\mathcal{D}|\Theta,\zeta,q_{1:N}) &= \prod_{i=1}^{N} P(Z_i|Z_{1:i-1},\Theta,\zeta,q_i) \\ &\propto \prod_{i=1}^{N} P(e_i|\Theta,\zeta,q_i) \end{aligned} \tag{18}$$

where

$$P(e_i|\Theta,\zeta,q_i) = \sqrt{\frac{\zeta q_i}{2\pi}} \exp\left(-\zeta q_i \frac{1}{2} e_i^2\right) \tag{19}$$

if the identification dataset is contaminated with scale outliers and

$$P(e_i|\Theta,\zeta,q_i) = \sqrt{\frac{\zeta}{2\pi}} \exp\left(-\zeta \frac{1}{2}(e_i - q_i)^2\right) \tag{20}$$

if the identification dataset is contaminated with location outliers.

To be able to carry forward the derivations, we need to take the underlying outlier model into account.

Scale Outlier Model Combining Eqs. 16 and 18 (along with Eq. 19), the posterior probability of the model parameters is then

$$\begin{aligned} P(\Theta|\mathcal{D},\alpha_{1:p},\zeta,q_{1:N}) &\propto \exp\left(-\frac{1}{2}\sum_{j=1}^{p}\alpha_j\theta_j^2 - \frac{1}{2}\sum_{i=1}^{N}\zeta q_i e_i^2\right) \\ &= \exp\left(-\sum_{j=1}^{p}\alpha_j E_{\theta_j} - \zeta\sum_{i=1}^{N} q_i E_{e_i}\right) \\ &= \exp\left(-\mathcal{J}_1(\Theta)\right) \end{aligned} \tag{21}$$

where $E_{\theta_j} = \theta_j^2/2$ and $E_{e_i} = e_i^2/2$. All constants are neglected in Eq. 21, because the optimal solution will not be affected by constant terms in the objective function.

One then proceeds to estimate the most probable values of the model parameters, Θ^{MP}, by maximizing the posterior probability, or equivalently, by minimizing the negative logarithm of Eq. 21. The gradient of the cost function $\mathcal{J}_1(\Theta)$ is

$$\frac{\partial \mathcal{J}_1}{\partial \Theta_{1:p}} = D_\alpha\Theta_{1:p} - \zeta R D_q Y + \zeta R D_q R^T \Theta_{1:p} + \zeta R D_q \vec{\mathbf{1}}_N \theta_{p+1} \tag{22}$$

$$\frac{\partial \mathcal{J}_1}{\partial \theta_{p+1}} = \zeta \vec{\mathbf{1}}_N^T D_q Y - \zeta \vec{\mathbf{1}}_N^T D_q R^T \Theta_{1:p} - \zeta s_q \theta_{p+1} \tag{23}$$

where $\vec{\mathbf{1}}_N = [1,\ldots,1]^T \in \mathbb{R}^N$, $Y = [y_1,\ldots,y_N]^T \in \mathbb{R}^N$, $R = [r_1,\ldots,r_N] \in \mathbb{R}^{p\times N}$, $D_\alpha = \mathrm{diag}(\alpha_1, \ldots, \alpha_p) \in \mathbb{R}^{p\times p}$, $D_q = \mathrm{diag}(q_1,\ldots,q_N) \in \mathbb{R}^{N\times N}$ and $s_q = \sum_{i=1}^N q_i$. Note that D_q may be viewed as a weighing matrix constructed to downplay the effect of scale outliers on the parameter estimates.

Making the partial derivatives expressed in Eqs. 22 and 23 equal to zero, the analytical expressions for $\Theta_{1:p}^{\mathrm{MP}}$ and $\theta_{p+1}^{\mathrm{MP}}$ can be derived

$$\Theta_{1:p}^{\mathrm{MP}} = \left(\mathrm{RCR}^T + \frac{1}{\zeta}D_\alpha\right)^{-1}\mathrm{RCY} \tag{24}$$

$$\theta_{p+1}^{\mathrm{MP}} = \frac{1}{s_q}\left(\vec{\mathbf{1}}_N^T D_q Y - \vec{\mathbf{1}}_N^T D_q R^T \Theta_{1:p}^{MP}\right) \tag{25}$$

where $C = D_q - {s_q}^{-1} D_q \vec{\mathbf{1}}_N \vec{\mathbf{1}}_N^T D_q$.

The posterior given by Eq. 21 is complex in general and cannot be directly used for the three-layer optimization of Eq. 8. The key to the hierarchical Bayesian approach is to obtain an approximation of the posterior. This approach of MacKay[30] is adopted here to obtain an approximated solution first and then the optimization problem is solved through iteration. Approximating the logarithm of the posterior distribution by its second-order Taylor expansion around $\Theta_{1:p+1}^{\mathrm{MP}}$, we obtain

$$\log P(\Theta|\mathcal{D},\alpha_{1:p},\zeta,q_{1:N}) \approx \log\ P(\Theta|\mathcal{D},\alpha_{1:p},\zeta,q_{1:N})\big|_{\Theta^{\mathrm{MP}}} + \nabla \log P(\Theta|\mathcal{D},\alpha_{1:p},\zeta,q_{1:N})\big|_{\Theta^{\mathrm{MP}}} m + \frac{1}{2} m^T \nabla\nabla \log P(\Theta|\mathcal{D},\alpha_{1:p},\zeta,q_{1:N})\big|_{\Theta^{\mathrm{MP}}} m \tag{26}$$

where $m = \left[\Theta - \Theta^{MP}\right]$.

As Θ^{MP} corresponds to a maximum of the logarithm of the posterior, the second term on the right hand side of Eq. 26 evaluates to zero. A Gaussian approximation of the posterior distribution can therefore be obtained as

$$P(\Theta|\mathcal{D},\alpha_{1:p},\zeta,q_{1:N}) \approx P(\Theta^{\mathrm{MP}}|\mathcal{D},\alpha_{1:p},\zeta,q_{1:N}) \exp\left(-\frac{1}{2} m^T H m\right) = \frac{1}{\sqrt{(2\pi)^{(p+1)} \det H^{-1}}} \exp\left(-\frac{1}{2} m^T H m\right) \tag{27}$$

where H is the Hessian of the cost function $\mathcal{J}_1(\Theta)$ evaluated at Θ^{MP}. The Hessian of the cost function $\mathcal{J}_1(\Theta)$ is defined as

$$H = \begin{bmatrix} H_{11} & H_{12} \\ H_{12}^T & H_{22} \end{bmatrix} = \begin{bmatrix} \frac{\partial^2 \mathcal{J}_1}{\partial \Theta_{1:p}^2} & \frac{\partial^2 \mathcal{J}_1}{\partial \Theta_{1:p} \partial \theta_{p+1}} \\ \frac{\partial^2 \mathcal{J}_1}{\partial \theta_{p+1} \partial \Theta_{1:n}} & \frac{\partial^2 \mathcal{J}_1}{\partial \theta_{p+1}^2} \end{bmatrix} \tag{28}$$

where

$$H_{11} = D_\alpha + \zeta R D_q R^T \tag{29}$$

$$H_{12} = \zeta R D_q \vec{\mathbf{1}}_N \tag{30}$$

$$H_{22} = \zeta s_q \tag{31}$$

Using the Schur complement of the Hessian matrix, we obtain,[29]

$$H = \begin{bmatrix} I_n & H_{12}H_{22}^{-1} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} H_{11} - H_{12}H_{22}^{-1}H_{12}^T & 0 \\ 0 & H_{22} \end{bmatrix} \begin{bmatrix} I_n & 0 \\ H_{22}^{-1}H_{12}^T & 1 \end{bmatrix} \tag{32}$$

Hence,

$$\begin{aligned} \det H &= \det \begin{bmatrix} H_{11} - H_{12}H_{22}^{-1}H_{12}^T & 0 \\ 0 & H_{22} \end{bmatrix} \\ &= H_{22} \det(H_{11} - H_{12}H_{22}^{-1}H_{12}^T) \\ &= \zeta s_q \det(D_\alpha + \zeta G) \\ &= \zeta s_q \prod_{j=1}^{p} (\alpha_j + \zeta \lambda_{G,j}) \end{aligned} \tag{33}$$

where $\lambda_{G,j}$ are the eigenvalues of the symmetric matrix $G = RCR^T$; the eigenvalue problem is:

$$R\left(D_q - \frac{1}{s_q} D_q \vec{\mathbf{1}}_N \vec{\mathbf{1}}_N^T D_q\right) R^T v_{G,j} = \lambda_{G,j} v_{G,j} \tag{34}$$

Location Outlier Model. Combining Eqs. 16 and 18 (along with Eq. 20) and neglecting all constants, the posterior probability of the model parameters is then

$$\begin{aligned} P(\Theta|\mathcal{D},\alpha_{1:p},\zeta,q_{1:N}) &\propto \exp\left(-\frac{1}{2}\sum_{j=1}^{p} \alpha_j \theta_j^2 - \frac{1}{2}\sum_{i=1}^{N} \zeta (e_i - q_i)^2\right) \\ &= \exp\left(-\sum_{j=1}^{p} \alpha_j E_{\theta_j} - \zeta \sum_{i=1}^{N} E_{e_i'}\right) \\ &= \exp\left(-\mathcal{J}_2(\Theta)\right) \end{aligned} \tag{35}$$

where $E_{e_i'} = E_{e_i} + q_i^2/2 - e_i q_i$.

The MAP estimates of the model parameters, Θ^{MP}, are obtained by maximizing the posterior probability, or equivalently, by minimizing the negative logarithm of Eq. 35. The gradient of the cost function $\mathcal{J}_2(\Theta)$ is

$$\frac{\partial \mathcal{J}_2}{\partial \Theta_{1:p}} = D_\alpha \Theta_{1:p} - \zeta R Y + \zeta R R^T \Theta_{1:p} + \zeta R \vec{\mathbf{1}}_N \theta_{p+1} + \zeta R D_q \vec{\mathbf{1}}_N \tag{36}$$

$$\frac{\partial \mathcal{J}_2}{\partial \theta_{p+1}} = \zeta \vec{\mathbf{1}}_N^T Y - \zeta \vec{\mathbf{1}}_N^T R^T \Theta_{1:p} - \zeta N \theta_{p+1} - \zeta \vec{\mathbf{1}}_N^T D_q \vec{\mathbf{1}}_N \tag{37}$$

Note that D_q may be viewed as a correction matrix constructed to reduce the effect of location outliers on the parameter estimates.

Making the partial derivatives expressed in Eqs. 36 and 37 equal to zero, the analytical expression for $\Theta_{1:p}^{\mathrm{MP}}$ and $\theta_{p+1}^{\mathrm{MP}}$ can be derived

$$\Theta_{1:p}^{\mathrm{MP}} = \left(RC'R^T + \frac{1}{\zeta} D_\alpha\right)^{-1} RC'\left(Y - D_q \vec{\mathbf{1}}_N\right) \tag{38}$$

$$\theta_{p+1}^{\mathrm{MP}} = N^{-1} \vec{\mathbf{1}}_N^T \left(Y - D_q \vec{\mathbf{1}}_N - R^T \Theta_{1:p}^{\mathrm{MP}}\right) \tag{39}$$

where $C' = I_N - N^{-1} \vec{\mathbf{1}}_N \vec{\mathbf{1}}_N^T$.

As explained previously, a Gaussian approximation of the posterior distribution is given as

$$P(\Theta|\mathcal{D},\alpha_{1:p},\zeta,q_{1:N}) \approx \frac{1}{\sqrt{(2\pi)^{(p+1)} \det H^{-1}}} \exp\left(-\frac{1}{2} m^T H m\right) \tag{40}$$

where H is the Hessian of the cost function $\mathcal{J}_2(\Theta)$ evaluated at Θ^{MP}.

It can be shown that the elements of the Hessian are

$$H_{11} = D_\alpha + \zeta R R^T \tag{41}$$

$$H_{12} = \zeta R \vec{\mathbf{1}}_N \tag{42}$$

$$H_{22} = \zeta N \tag{43}$$

The Cholesky factorization of the Hessian would be similar to Eq. 32. To obtain an expression for $\det H$, thus, one needs to solve the following eigenvalue problem

$$R\left(I_N - \frac{1}{N} \vec{\mathbf{1}}_N \vec{\mathbf{1}}_N^T\right) R^T v_{G,j} = \lambda'_{G,j} v_{G,j} \tag{44}$$

where $\lambda'_{G,j}$ are the eigenvalues of the symmetric matrix $G' = RC'R^T$.

Finally, we obtain

$$\det H = \zeta N \det(D_\alpha + \zeta G') = \zeta N \prod_{j=1}^{p} (\alpha_j + \zeta \lambda'_{G,j}) \tag{45}$$

Inference of hyperparameters $\alpha_{1:p}$ and ζ

Hyperparameters $\alpha_{1:p}$ and ζ are inferred from the identification data $\mathcal{D}$ by applying Bayes' rule in the second layer of optimization. First, the posterior distribution of the hyperparameters is written as

$$P(\alpha_{1:p},\zeta|\mathcal{D},q_{1:N}) = \frac{P(\mathcal{D}|\alpha_{1:p},\zeta,q_{1:N})P(\alpha_{1:p},\zeta|q_{1:N})}{P(\mathcal{D}|q_{1:N})} \tag{46}$$

As priors, it is assumed that the hyperparameters are statistically independent, that is, $P(\alpha_{1:p},\zeta|q_{1:N}) = P(\zeta|q_{1:N})\prod_{j=1}^{p}P(\alpha_j|q_{1:N})$. If there is no explicit information available for the hyperparameters, a uniform distribution can then be used to describe appropriate noninformative priors on $\log\ \alpha_j$ and $\log\ \zeta$. To incorporate precise prior knowledge, however, *conjugate priors*[31] are commonly assigned for which the resulting posterior distribution can be conveniently evaluated. To assure generality, we consider the following gamma distributions as hyperpriors:

$$P(\alpha_j|q_{1:N}) = \frac{s_j^{k_j}\alpha_j^{k_j-1}}{\Gamma(k_j)}\exp(-s_j\alpha_j) \propto \alpha_j^{k_j-1}\exp(-s_j\alpha_j) \tag{47}$$

$$P(\zeta|q_{1:N}) = \frac{s_0^{k_0}\zeta^{k_0-1}}{\Gamma(k_0)}\exp(-s_0\zeta) \propto \zeta^{k_0-1}\exp(-s_0\zeta) \tag{48}$$

where k_j is the shape parameter and s_j is the inverse of the scale parameter. Therefore, gamma distribution is a simple peaked distribution for which mean and variance are defined by k_j/s_j and k_j/s_j^2, respectively. The fact that the gamma distribution is the conjugate prior to many likelihood distributions justifies the choice of gamma hyperpriors.

Under the stated assumptions, the prior distribution over hyperparameters is expressed as

$$P(\alpha_{1:p},\zeta|q_{1:N}) \propto \zeta^{k_0-1}\exp(-s_0\zeta)\prod_{j=1}^{p}\alpha_j^{k_j-1}\exp(-s_j\alpha_j) \tag{49}$$

Hereinafter, the underlying outlier model will be taken into account to lay out a computational procedure for the inference of hyperparameters.

Scale Outlier Model. The likelihood $P(\mathcal{D}|\alpha,\zeta,q_{1:N})$ is equal to the normalizing constant in Eq. 14 for the first level of inference. Substituting Eqs. 16, 18 (along with 19) and 27 in Eq. 14, we can derive the following expression for the likelihood

$$P(\mathcal{D}|\alpha_{1:p},\zeta,q_{1:N}) \propto \frac{\prod_{j=1}^{p}\sqrt{\alpha_j}\prod_{i=1}^{N}\sqrt{\zeta q_i}}{\sqrt{\det H}}\exp\Big(-\mathcal{J}_1(\Theta)+\frac{1}{2}m^THm\Big)\Big|_{\Theta^{MP}} \tag{50}$$

Substituting Eqs. 49 and 50 into Eq. 46, the posterior probability of the hyperparameters becomes

$$P(\alpha_{1:p},\zeta|\mathcal{D},q_{1:N}) \propto \sqrt{\frac{\zeta^N\prod_{j=1}^{p}\alpha_j\prod_{i=1}^{N}q_i}{\zeta s_q\prod_{j=1}^{p}(\alpha_j+\zeta\lambda_{G,j})}}\exp\Big(-\mathcal{J}_1\big(\Theta^{\mathrm{MP}}\big)\Big) \times \zeta^{k_0-1}\exp(-s_0\zeta)\prod_{j=1}^{p}\alpha_j^{k_j-1}\exp(-s_j\alpha_j) \tag{51}$$

Minimizing the negative logarithm of Eq. 51 leads to the following optimization problem

$$\min_{\alpha_{1:p},\zeta}\mathcal{J}_1(\alpha_{1:p},\zeta) = \sum_{j=1}^{p}\alpha_j\Big[s_j+E_{\theta_j}(\theta_j^{\mathrm{MP}})\Big]+\zeta\Big[s_0+\sum_{i=1}^{N}q_iE_{e_i}(\Theta^{\mathrm{MP}})\Big] - \frac{N+2k_0-3}{2}\log\zeta-\frac{1}{2}\sum_{j=1}^{p}(2k_j-1)\log\alpha_j+\frac{1}{2}\sum_{j=1}^{p}\log(\alpha_j+\zeta\lambda_{G,j}) \tag{52}$$

The gradient of the cost function $\mathcal{J}_1(\alpha_{1:p},\zeta)$ is

$$\frac{\partial\mathcal{J}_1}{\partial\alpha_{1:p}} = \Big(D_s+E_\Theta(\Theta^{\mathrm{MP}})+\frac{1}{2}(D_\alpha+\zeta D_\lambda)^{-1}-\frac{1}{2}D_\alpha^{-1}(2D_k-I_p)\Big)\vec{\mathbf{1}}_p \tag{53}$$

$$\frac{\partial\mathcal{J}_1}{\partial\zeta} = s_0+\vec{\mathbf{1}}_N^TD_qE_e(\Theta^{\mathrm{MP}})\vec{\mathbf{1}}_N+\frac{1}{2}\vec{\mathbf{1}}_p^TD_\lambda(D_\alpha+\zeta D_\lambda)^{-1}\vec{\mathbf{1}}_p - \frac{N+2k_0-3}{2\zeta} \tag{54}$$

where $D_s = \mathrm{diag}(s_1,\ldots,s_p) \in \mathbb{R}^{p\times p}$, $E_\Theta = \mathrm{diag}(E_{\theta_1},\ldots,E_{\theta_p}) \in \mathbb{R}^{p\times p}$, $E_e = \mathrm{diag}(E_{e_1},\ldots,E_{e_N}) \in \mathbb{R}^{N\times N}$, $D_\lambda = \mathrm{diag}(\lambda_{G,1},\ldots,\lambda_{G,p}) \in \mathbb{R}^{p\times p}$, $D_k = \mathrm{diag}(k_1,\ldots,k_p) \in \mathbb{R}^p$, and $\vec{\mathbf{1}}_p = [1,\ldots,1]^T \in \mathbb{R}^p$.

Setting the partial derivatives equal to zero and carrying out a few algebraic manipulations, the following expressions are obtained in the optimum of $\mathcal{J}_1(\alpha_{1:p},\zeta)$

$$D_\alpha^{\mathrm{MP}} = \Big(D_s+E_\Theta(\Theta^{\mathrm{MP}})\Big)^{-1}\Big(\frac{1}{2}D_\gamma+D_k-I_p\Big) \tag{55}$$

$$\zeta^{\mathrm{MP}} = \frac{1}{2}\Big(s_0+\vec{\mathbf{1}}_N^TD_qE_e(\Theta^{\mathrm{MP}})\vec{\mathbf{1}}_N\Big)^{-1}\Big(N+2k_0-3-\vec{\mathbf{1}}_p^TD_\gamma\vec{\mathbf{1}}_p\Big) \tag{56}$$

where $D_\gamma = \mathrm{diag}(\gamma_1,\ldots,\gamma_p)$. The jth diagonal element of D_γ is defined as

$$\gamma_j = \frac{\zeta^{\mathrm{MP}}\lambda_{G,j}}{\alpha_j^{\mathrm{MP}}+\zeta^{\mathrm{MP}}\lambda_{G,j}} \tag{57}$$

where $\lambda_{G,j}$ is obtained by solving the eigenvalue problem of Eq. 34. Thus, $\gamma_j \in [0,1]$ is a measure of the strength of the likelihood in relation to the prior in determining θ_j. For instance, $\gamma_j \to 0$ (i.e., $\lambda_j \ll \alpha_j$) indicates that θ_j is poorly measured from the identification data. Consequently

$$\gamma_{\mathrm{eff}} = 1+\sum_{j=1}^{p}\frac{\zeta^{\mathrm{MP}}\lambda_{G,j}}{\alpha_j^{\mathrm{MP}}+\zeta^{\mathrm{MP}}\lambda_{G,j}} = 1+\vec{\mathbf{1}}_p^TD_\gamma\vec{\mathbf{1}}_p \tag{58}$$

is the number of well-determined parameters.[25]

Since $\alpha_{1:p}$ and ζ are positive scale variables, we can consider a separable Gaussian distribution for $P(\log\ \alpha_{1:p},\ \log\ \zeta|\mathcal{D},q_{1:N})$ such that[†]

$$P(\log\alpha_{1:p},\log\zeta|\mathcal{D},q_{1:N}) \approx \frac{1}{2\pi\sqrt{\det A^{-1}}}\exp\Big(-\frac{1}{2}f^TAf\Big) \tag{59}$$

where $f = \Big[\log\alpha_{1:p}-\log\alpha_{1:p}^{MP}\quad \log\zeta-\log\zeta^{MP}\Big]^T$ and A is the Hessian of the cost function $\mathcal{J}_1(\alpha_{1:p},\zeta)$ evaluated at $\alpha_{1:p}^{MP},\zeta^{MP}$.

[†]It is natural to represent the uncertainty associated with positive scale variables on a log scale.

It is pointed out by MacKay[32] that the Gaussian approximation over $\log \alpha_j^{\mathrm{MP}}$ and $\log \zeta$ holds good if the model parameters are all well-determined in relation to their prior range by the identification data.

Having obtained MAP estimates of hyperparameters, the elements of the A are calculated as follows

$$\begin{aligned} A_{11} &= \frac{\partial^2 \mathcal{J}_1(\alpha_{1:p},\zeta)}{\partial(\log \alpha_{1:p})^2}\Big|_{\alpha_{1:p}^{\mathrm{MP}},\zeta^{\mathrm{MP}}} \\ &= \Big(D_s + E_\Theta(\Theta^{\mathrm{MP}})\Big)D_\alpha^{\mathrm{MP}} + \frac{1}{2}\zeta^{\mathrm{MP}} D_\alpha^{\mathrm{MP}} D_\lambda \Big(D_\alpha^{\mathrm{MP}} + \zeta^{\mathrm{MP}} D_\lambda\Big)^{-2} \\ &\approx \Big(D_s + E_\Theta(\Theta^{\mathrm{MP}})\Big)D_\alpha^{\mathrm{MP}} \\ &= \frac{1}{2}D_\gamma + D_k - I_p \end{aligned} \tag{60}$$

$$\begin{aligned} A_{22} &= \frac{\partial^2 \mathcal{J}_1(\alpha_{1:p},\zeta)}{\partial(\log \zeta)^2}\Big|_{\alpha_{1:p}^{\mathrm{MP}},\zeta^{\mathrm{MP}}} \\ &= \zeta^{\mathrm{MP}}\Big(s_0 + \vec{\mathbf{1}}_N^T D_q E_e(\Theta^{\mathrm{MP}})\vec{\mathbf{1}}_N\Big) + \frac{1}{2}\zeta^{\mathrm{MP}}\vec{\mathbf{1}}_p^T D_\alpha^{\mathrm{MP}} D_\lambda \Big(D_\alpha^{\mathrm{MP}} + \zeta^{\mathrm{MP}} D_\lambda\Big)^{-2}\vec{\mathbf{1}}_p \\ &\approx \zeta^{\mathrm{MP}}\Big(s_0 + \vec{\mathbf{1}}_N^T D_q E_e(\Theta^{\mathrm{MP}})\vec{\mathbf{1}}_N\Big) \\ &= \frac{1}{2}\big(N + 2k_0 - 2 - \gamma_{eff}\big) \end{aligned} \tag{61}$$

These approximations are valid if $\gamma_j + 2k_j - 2 \gg 1$ and $N + 2k_0 - 2 - \gamma_{\mathrm{eff}} \gg 1$.[32]

From Eqs. 60 and 61, it is straightforward to show that

$$\det A = \frac{1}{2}(N + 2k_0 - 2 - \gamma_{\mathrm{eff}})\prod_{j=1}^{p}\Big(\frac{\gamma_j}{2} + k_j - 1\Big) \tag{62}$$

Location Outlier Model. When the identification dataset is contaminated with location outliers, Eqs. 16, 18 (along with 20), and 40 are substituted in Eq. 14 to obtain an expression for the likelihood

$$P(\mathcal{D}|\alpha_{1:p},\zeta,q_{1:N}) \propto \frac{\sqrt{\zeta^N}\prod_{j=1}^{p}\sqrt{\alpha_j}}{\sqrt{\det H}}\exp\Big(-\mathcal{J}_2(\Theta) + \frac{1}{2}m^T H m\Big)\Big|_{\Theta^{\mathrm{MP}}} \tag{63}$$

Substituting Eqs. 49 and 63 into Eq. 46, the posterior distribution of the hyperparameters $\alpha_{1:p}$ and ζ becomes:

$$\begin{aligned} P(\alpha_{1:p},\zeta|\mathcal{D},q_{1:N}) \propto & \sqrt{\frac{\zeta^N\prod_{j=1}^{p}\alpha_j}{\zeta N\prod_{j=1}^{p}(\alpha_j + \zeta\lambda'_{G,j})}}\exp\Big(-\mathcal{J}_2\big(\Theta^{\mathrm{MP}}\big)\Big) \\ & \times \zeta^{k_0-1}\exp(-s_0\zeta)\prod_{j=1}^{p}\alpha_j^{k_j-1}\exp(-s_j\alpha_j) \end{aligned} \tag{64}$$

One can then proceed to infer the hyperparameters in a similar way as for the scale outlier model. The condition for optimality is thus expressed as

$$D_\alpha^{\mathrm{MP}} = \Big(D_s + E_\Theta(\Theta^{\mathrm{MP}})\Big)^{-1}\Big(\frac{1}{2}D_{\gamma'} + D_k - I_p\Big) \tag{65}$$

$$\zeta^{\mathrm{MP}} = \frac{1}{2}\Big(s_0 + \vec{\mathbf{1}}_N^T E_{e'}(\Theta^{\mathrm{MP}})\vec{\mathbf{1}}_N\Big)^{-1}\Big(N + 2k_0 - 3 - \vec{\mathbf{1}}_p^T D_{\gamma'}\vec{\mathbf{1}}_p\Big) \tag{66}$$

where $D_{\gamma'} = \mathrm{diag}(\gamma'_1,\ldots,\gamma'_p)$. The jth diagonal element of D'_γ is defined as

$$\gamma_j = \frac{\zeta^{\mathrm{MP}}\lambda'_{G,j}}{\alpha_j^{\mathrm{MP}} + \zeta^{\mathrm{MP}}\lambda'_{G,j}} \tag{67}$$

where $\lambda'_{G,j}$ is obtained by solving the eigenvalue problem of Eq. 44.

As explained previously, a separable Gaussian approximation of $P(\log \alpha_{1:p}, \log \zeta|\mathcal{D},q_{1:N})$ can be obtained as

$$P(\log\alpha_{1:p},\log\zeta|\mathcal{D},q_{1:N}) \approx \frac{1}{2\pi\sqrt{\det A^{-1}}}\exp\Big(-\frac{1}{2}f^T A f\Big) \tag{68}$$

where A is Hessian of the cost function $\mathcal{J}_2(\alpha_{1:p},\zeta)$ evaluated at $\alpha_{1:p}^{\mathrm{MP}}$, ζ^{MP}.

Finally, it can be shown that

$$\det A = \frac{1}{2}(N + 2k_0 - 2 - \gamma'_{\mathrm{eff}})\prod_{j=1}^{p}\Big(\frac{\gamma'_j}{2} + k_j - 1\Big) \tag{69}$$

Inference of outlier indicator variables $\mathbf{q}_{1:N}$

So far, in our derivations, we have assumed that the indicator variables $q_{1:N}$ are known. As $q_{1:N}$ are unobserved variables, they still need to be estimated from the identification dataset. Applying Bayes' rule in the third level of optimization, we obtain the following posterior distribution

$$P(q_{1:N}|\mathcal{D}) = \frac{P(\mathcal{D}|q_{1:N})P(q_{1:N})}{P(\mathcal{D})} \tag{70}$$

where the prior distribution of $q_{1:N}$ is expressed as

$$P(q_{1:N}) = \prod_{i=1}^{N}P(q_i) = \prod_{i=1}^{N}\delta^{(1-\frac{q_i-\rho}{1-q_i\rho})}(1-\delta)^{(\frac{q_i-\rho}{1-q_i\rho})} \tag{71}$$

and as

$$\begin{aligned} P(q_{1:N}) = \prod_{i=1}^{N}P(q_i) &= \prod_{i=1}^{N}(0.5\delta)^{(\frac{|q_i|-q_i}{2\Delta})}(0.5\delta)^{(\frac{|q_i|+q_i}{2\Delta})}(1-\delta)^{(1-\frac{|q_i|}{\Delta})} \\ &= \prod_{i=1}^{N}(0.5\delta)^{\frac{|q_i|}{\Delta}}(1-\delta)^{(1-\frac{|q_i|}{\Delta})} \end{aligned} \tag{72}$$

for the scale and location outliers, respectively. In deriving Eqs. 71 and 72, we assumed that the occurrence of the outliers is completely random.

The likelihood $P(\mathcal{D}|q_{1:N})$ can be obtained by integrating over $\alpha_{1:p}$ and ζ, and then an approximate solution is obtained[25]

$$\begin{aligned} P(\mathcal{D}|q_{1:N}) &= \int P(\mathcal{D}|q_{1:N},\alpha_{1:p},\zeta)P(\alpha_{1:p},\zeta|q_{1:N})d\alpha_{1:p}d\zeta \\ &\approx P(\mathcal{D}|q_{1:N},\alpha_{1:p}^{\mathrm{MP}},\zeta^{\mathrm{MP}})P(\alpha_{1:p}^{\mathrm{MP}},\zeta^{\mathrm{MP}}|q_{1:N})2\pi\sqrt{\det A^{-1}} \end{aligned} \tag{73}$$

At this stage, the type of outliers should be determined to obtain explicit expressions for evaluating the posterior probability of indicator variables.

Scale Outlier Model. Combining Eqs. 49, 50, and 62 and neglecting all constants, the likelihood of the third level of Bayesian inference is expressed as

$$P(\mathcal{D}|q_{1:N}) \propto \prod_{j=1}^{p}(\alpha_j^{\mathrm{MP}})^{k_j-1}\exp(-s_j\alpha_j^{\mathrm{MP}})\sqrt{\frac{\alpha_j^{\mathrm{MP}}}{(0.5\gamma_j+k_j-1)(\alpha_j^{\mathrm{MP}}+\zeta^{\mathrm{MP}}\lambda_{G,j})}}\sqrt{\frac{(\zeta^{\mathrm{MP}})^{N+2k_0-3}\prod_{i=1}^{N}q_i}{s_q(N+2k_0-2-\gamma_{eff})}}\exp\Big(-s_0\zeta^{\mathrm{MP}}-\mathcal{J}_1\big(\Theta^{\mathrm{MP}}\big)\Big) \tag{74}$$

The posterior probability of the indicator variables is obtained by substituting Eqs. 71 and 74 into Eq. 70:

$$P(q_{1:N}|\mathcal{D}) \propto \prod_{j=1}^{p}(\alpha_j^{\mathrm{MP}})^{k_j-1}\exp(-s_j\alpha_j^{\mathrm{MP}})\sqrt{\frac{\alpha_j^{\mathrm{MP}}}{(0.5\gamma_j+k_j-1)(\alpha_j^{\mathrm{MP}}+\zeta^{\mathrm{MP}}\lambda_{G,j})}}\sqrt{\frac{(\zeta^{\mathrm{MP}})^{N+2k_0-3}}{s_q(N+2k_0-2-\gamma_{\mathrm{eff}})}}\exp\Big(-s_0\zeta^{\mathrm{MP}}-\mathcal{J}_1\big(\Theta^{\mathrm{MP}}\big)\Big)$$
$$\times\prod_{i=1}^{N}\delta^{\left(1-\frac{q_i-\rho}{1-q_i\rho}\right)}(1-\delta)^{\left(\frac{q_i-\rho}{1-q_i\rho}\right)}\sqrt{q_i} \tag{75}$$

To assess the quality of each data pair, $Z_i = (r_i,y_i)$, the posterior probability $P(q_i|\mathcal{D})$ is first evaluated for $q_i \in \{1,\rho\}$. The normalized probabilities are then used to estimate the expected value of q_i as follows:

$$E[q_i|\mathcal{D}] = P(q_i = 1|\mathcal{D}) + \rho P(q_i = \rho|\mathcal{D}) \tag{76}$$

Location Outlier Model. Combining Eqs. 49, 63, and 69, the likelihood of the third level of Bayesian inference is expressed as

$$P(\mathcal{D}|q_{1:N}) \propto \prod_{j=1}^{p}(\alpha_j^{\mathrm{MP}})^{k_j-1}\exp(-s_j\alpha_j^{\mathrm{MP}})\sqrt{\frac{\alpha_j^{\mathrm{MP}}}{(0.5\gamma'_j+k_j-1)(\alpha_j^{\mathrm{MP}}+\zeta^{\mathrm{MP}}\lambda'_{G,j})}}\sqrt{\frac{(\zeta^{\mathrm{MP}})^{N+2k_0-3}}{N(N+2k_0-2-\gamma'_{\mathrm{eff}})}}\exp\Big(-s_0\zeta^{\mathrm{MP}}-\mathcal{J}_2\big(\Theta^{\mathrm{MP}}\big)\Big) \tag{77}$$

Substituting Eqs. 72 and 77 into Eq. 70, the posterior probability of the indicator variables becomes

$$P(q_{1:N}|\mathcal{D}) \propto \prod_{j=1}^{p}(\alpha_j^{\mathrm{MP}})^{k_j-1}\exp(-s_j\alpha_j^{\mathrm{MP}})\sqrt{\frac{\alpha_j^{\mathrm{MP}}}{(0.5\gamma'_j+k_j-1)(\alpha_j^{\mathrm{MP}}+\zeta^{\mathrm{MP}}\lambda'_{G,j})}}\sqrt{\frac{(\zeta^{\mathrm{MP}})^{N+2k_0-3}}{N(N+2k_0-2-\gamma'_{\mathrm{eff}})}}\exp\Big(-s_0\zeta^{\mathrm{MP}}-\mathcal{J}_2\big(\Theta^{\mathrm{MP}}\big)\Big)$$
$$\times\prod_{i=1}^{N}(0.5\delta)^{\frac{|q_i|}{\Delta}}(1-\delta)^{\left(1-\frac{|q_i|}{\Delta}\right)} \tag{78}$$

For the data pair, $Z_i = (r_i,y_i)$, the posterior probability $P(q_i|\mathcal{D})$ is first evaluated over the set of possible values $q_i \in \{0, -\Delta, +\Delta\}$. The expected value of q_i is then estimated from the normalized probabilities

$$E[q_i|\mathcal{D}] = \Delta P(|q_i| = \Delta|\mathcal{D})\mathrm{sign}\big[P(q_i = +\Delta|\mathcal{D}) - P(q_i = -\Delta|\mathcal{D})\big] \tag{79}$$

Robust model identification procedure

To summarize our discussion, the implementation procedure of the proposed robust identification approach is outlined as follows. First, a few preparatory steps are completed to incorporate the relevant prior knowledge. Given a contaminated identification dataset,

1. Specify a set of indicator variables, $q_{1:N} = \{q_1, \ldots, q_N\}$, to denote the quality of each data point.

2. Select an appropriate outlier model to describe the contaminating distribution (Eqs. 5 and 6).

3. Include the noise distribution information to describe the prior distribution of $P(q_{1:N})$ (Eqs. 71 and 72). In the absence of relevant prior information, the 3σ edit rule is used to detect potential outliers and hence to initialize the estimation of noise distribution parameters, that is, $\delta^{(0)}$, ${\sigma_e}^{(0)}$, and $\Delta^{(0)}$ or $\rho^{(0)}$.

4. Characterize the prior distribution of hyperparameters $P(\alpha_{1:p},\zeta|q_{1:N})$ based on the explicit prior knowledge. The prior information over hyperparameters can be generally well-represented by gamma distributions (Eqs. 47 and 48). If there is no explicit information available for the hyperparameters, a uniform distribution can then be used to describe appropriate noninformative priors on $\log \alpha_j$ and $\log \zeta$.

5. Determine the prior distribution of model parameters $P(\Theta|\alpha_{1:p},\zeta)$ based on the available background information. In the absence of other prior information, the prior probability of Θ can be approximated by independent Gaussian distributions (Eq. 15). Depending on the model structure, it might be reasonable to assume that $\alpha_1 = \alpha_2 = \ldots = \alpha_p$.

6. Choose a set of initial values for indicator variables, $q_{1:N}^{(0)}$, and hyperparameters, $\alpha_{1:p}^{(0)}$ and $\zeta^{(0)}$.

Next, the following steps will be repeated iteratively until no further improvements are gained:

1. Maximize $P(\Theta^{(l)}|\mathcal{D},\alpha_{1:p}^{(l-1)},\zeta^{(l-1)},q_{1:N}^{(l-1)})$ to update the MAP estimates of model parameters, $\Theta^{(l)}$ (Eqs. 24 and 25 or 38 and 39).

2. Maximize $P(\alpha_{1:p}^{(l)}, \zeta^{(l)}|\mathcal{D},q_{1:N}^{(l-1)})$ to update the MAP estimates of hyperparameters, $\alpha_{1:p}^{(l)}$ and $\zeta^{(l)}$ (Eqs. 55 and 56 or 65 and 66).

3. Evaluate the posterior probability of each observation acting as an outlier to update the MAP estimates of indicator

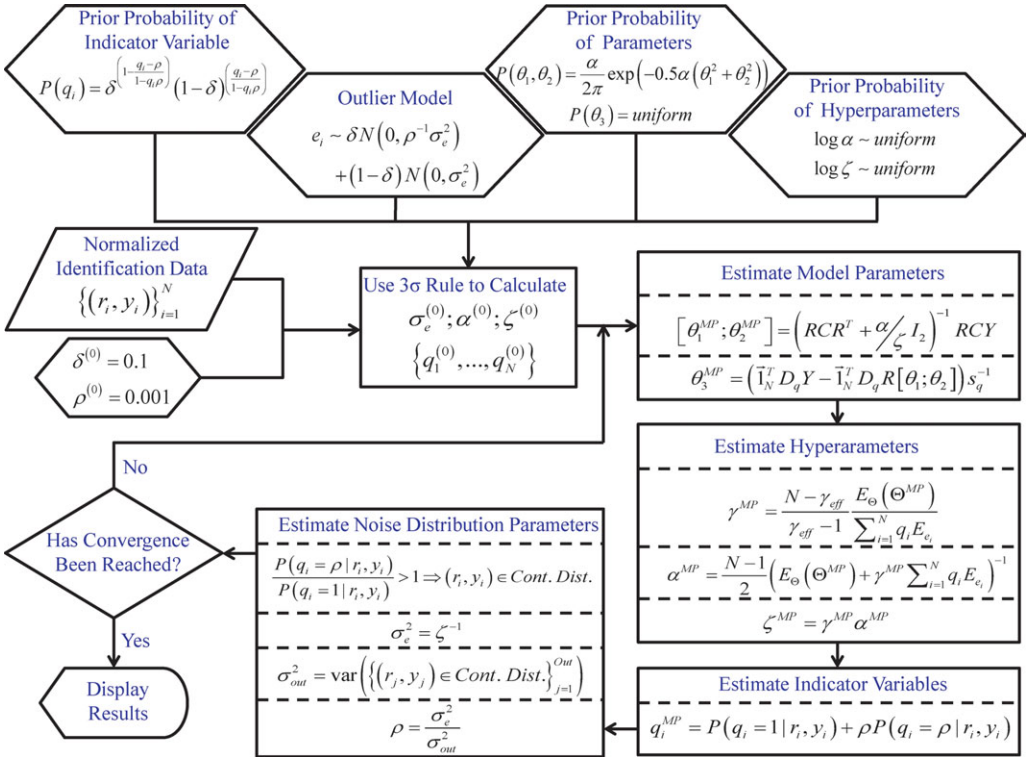


Figure 1. The flowchart of the Bayesian procedure followed for robust identification of the second-order FIR model in the presence of scale outliers.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

variables, $q_{1:N}^{(l)}$ (Eqs. 75 and 76 and 78 and 79); the updated estimates are used in the next iteration.

4. Update the estimated values of the noise distribution parameters, $\delta^{(l)}$, $\sigma_e{}^{(l)}$, and $\Delta^{(l)}$ or $\rho^{(l)}$, using the observations identified as outliers.

Although Gaussian approximations to posterior density functions may not be always adequate, the application of the robust identification procedure proposed in this article is not limited to ARX models. For robust identification of nonlinear models with non-Gaussian noise distributions, it is often required to adopt more sophisticated approximation methods such as variational Bayes methods or Monte Carlo methods with various Bayesian sampling schemes. The derivations can thus be directly extended to other classes of dynamic models, although numerical optimization may be required.

Simulation and Experimental Study

In this section, we demonstrate the effectiveness of the proposed identification approach through the simulated and experimental datasets. The purpose is to verify the performance of the Bayesian-based outlier detection algorithm and to evaluate the overall robust behavior of the proposed framework. The robustness of the Bayesian framework is compared with that of the Huber estimator, which is one of the most widely used methods of robust regression.

It is noteworthy that the M-estimation with various weighing functions were performed. In general, the results were similar to those of the Huber robust regression.

Second-order finite impulse response model

Consider the following linear second-order finite impulse response (FIR) model

$$y_t = [\,6.5 \quad -2 \quad -1\,]\begin{bmatrix} x_t \\ 1 \end{bmatrix} + e_t \tag{80}$$

with $x_t = [\,u_1(t) \quad u_2(t)\,]^T$. Three different scenarios will be considered to simulate noise distribution:

Case I. $e_t \sim 0.85N(0,4) + 0.15N(0,40)$

Case II. $e_t \sim 0.85N(0,2.25) + 0.15[N(-5,2.25),N(5,2.25)]$

Case III. $e_t \sim 0.85N(0,2.25) + 0.15N(5,2.25)$

Total number of data points is set to $N = 200$ in which around 15% have been generated from the contaminating distribution. The comparison is performed between the following methods using standard implementations:

1. Ordinary least-square (OLS) regression: The most straightforward method for identification of ARX models is the OLS method, relying on minimization of the sum of squared errors between measurements and model predictions.

2. Regular Bayesian: The identification dataset used in the proposed Bayesian method is considered to be healthy, that is, $\delta = 0$.

3. Robust regression: The M-estimation with the Huber weight function is performed. The tuning parameters of this algorithm are adjusted based on the recommended settings in MATLAB.

4. Robust Bayesian: The robust Bayesian framework does not require any knowledge of the noise distribution parameters as these parameters are actually iteratively estimated in the identification process. To illustrate how the proposed Bayesian identification framework is implemented, Figure 1 shows a detailed flowchart demonstrating the sequence of steps performed for Case I. Basic MATLAB commands can be used to execute each step.

Table 1 shows the mean relative estimation errors (MRE) and mean squared errors (MSE) of prediction averaged more

Table 1. Comparison of Estimated Parameters of the Second-Order FIR Model

	OLS Regression	Robust Regression	Regular Bayesian	Robust Bayesian
Case I: Scale outliers in the identification data				
MRE of θ_1 (%)	2.48 ± 1.84	2.02 ± 1.46	2.46 ± 1.84	1.86 ± 1.24
MRE of θ_2 (%)	3.96 ± 3.42	3.51 ± 2.42	3.95 ± 3.42	3.01 ± 2.15
MRE of θ_3 (%)	23.18 ± 15.63	16.30 ± 13.32	23.18 ± 15.63	12.27 ± 11.72
MSE of Prediction	0.161 ± 0.135	0.098 ± 0.093	0.160 ± 0.135	0.069 ± 0.074
Case II: Symmetric location outliers in the identification data				
MRE of θ_1 (%)	2.95 ± 2.04	2.27 ± 1.57	2.95 ± 2.03	1.08 ± 0.88
MRE of θ_2 (%)	4.17 ± 3.57	3.52 ± 2.65	4.16 ± 3.55	2.18 ± 1.98
MRE of θ_3 (%)	19.26 ± 12.79	14.48 ± 11.25	19.26 ± 12.78	7.44 ± 5.44
MSE of Prediction	0.149 ± 0.100	0.092 ± 0.069	0.148 ± 0.100	0.029 ± 0.030
Case III: Asymmetric location outliers in the identification data				
MRE of θ_1 (%)	1.80 ± 1.33	1.72 ± 1.22	1.76 ± 1.32	1.46 ± 0.97
MRE of θ_2 (%)	3.34 ± 2.70	2.85 ± 2.56	3.34 ± 2.70	2.59 ± 1.99
MRE of θ_3 (%)	65.30 ± 10.21	40.40 ± 11.77	65.3 ± 10.21	7.60 ± 7.04
MSE of Prediction	0.481 ± 0.137	0.215 ± 0.095	0.480 ± 0.136	0.042 ± 0.036

than 50 trials with different noise sequences. Not surprisingly, the robust methods improve the parameter estimation performance by detecting and accommodating outlying observations of the identification dataset. The smaller values of MRE indicate that the robust Bayesian framework outperforms the Huber robust regression in terms of accuracy of the parameter estimates. As a result, the models identified using robust Bayesian framework show better predictive performance, with smaller values of MSE. Specifically, the Huber robust regression can suffer from the effect of outliers when the contaminating distribution is asymmetric. In general, traditional robust regression methods assume a symmetric Gaussian distribution for the contaminating distribution and assign robustness weights accordingly. Therefore, in the case of asymmetric contaminating distribution (e.g., the noise term, e_i, is distributed as $e_i \sim \delta N(\Delta,\sigma_e^2) + (1 - \delta)N(0,\sigma_e^2)$), downweighing the outliers causes a strong bias to the estimates.

Figures 2a, 3a, and 4a show the number of iterations required for the convergence of model parameter estimates, whereas Figures 2b, 3b, and 4b present the percentage of the outliers detected in the individual runs of Monte Carlo simulation. Although the iterations needed for the robust Bayesian and Huber methods are comparable, the former is capable of successfully detecting a higher percentage of outliers.

Continuous fermentation reactor simulation

To illustrate potential applications of the proposed method in process industries, identification of a simulated continuous fermentation reactor (CFR) is considered in this section. The nonlinear dynamic behavior of a CFR is described as follows[33]

$$\dot{X} = -DX + \mu X \tag{81}$$

$$\dot{S} = D(S_f - S) - \frac{1}{Y_{X/S}}\mu X \tag{82}$$

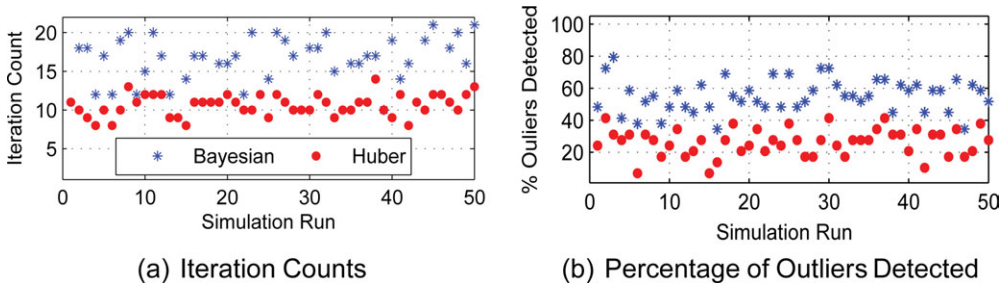


Figure 2. Scale outlier.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

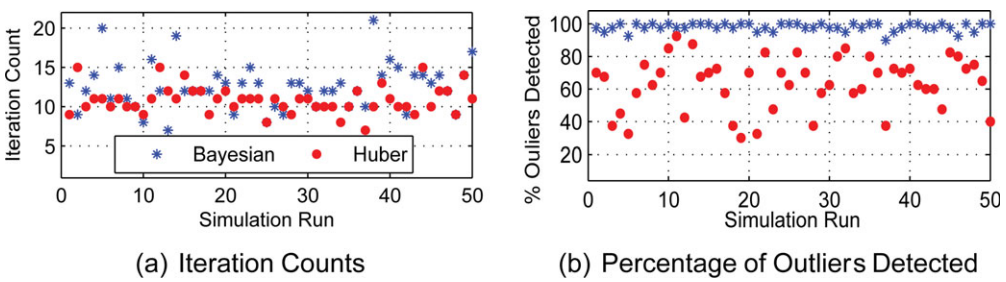
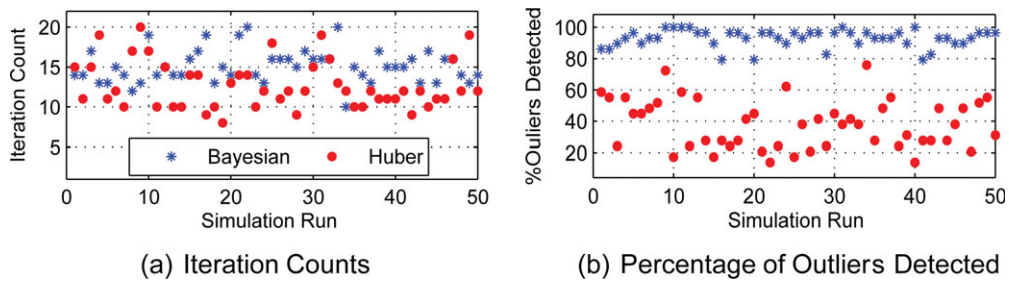


Figure 3. Symmetric location outlier.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]



(a) Iteration Counts (b) Percentage of Outliers Detected

Figure 4. Asymmetric location outlier.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

$$\dot{P} = -DP + (\alpha\mu + \beta)X \tag{83}$$

where specific growth rate (μ) is defined as

$$\mu = \frac{\mu_m\left(1 - \frac{P}{P_m}\right)S}{K_m + S + \frac{S^2}{K_i}} \tag{84}$$

X, S, and P are the state variables of the system representing the biomass concentration, substrate concentration and product concentration, respectively. Dilution rate (D) and feed substrate concentration (S_f) are normally treated as the system inputs. The cell-mass yield ($Y_{X/S}$), the yield parameters (α, β), the maximum specific growth rate (μ_m), the product saturation constant (P_m), the substrate saturation constant (K_m), and the substrate inhibition constant (K_i) are the model parameters. In this study, the case where the CFR has a single stable steady-state is considered for which the parameter settings and the operating conditions are given by Henson and Seborg.[33] The objective is to identify a MISO model relating the two input variables, dilution rate (u_1), and feed substrate concentration (u_2), with product concentration (y_1); the identification dataset is contaminated with the scale or location outliers. Dilution rate is assumed to vary between 0.13 and 0.17 hr^{-1}, whereas feed substrate concentration is assumed to vary between 18 and 22 kg/m^3.

For both steady-state and dynamic modeling exercises presented below, Gaussian noise with a relative variance of 10% was added to the outputs. To test the robustness of the proposed Bayesian framework, several outliers are randomly added to the simulated identification dataset. To fairly investigate the performance of different identification procedures in the presence of outliers, Monte–Carlo simulation is performed. The percentage of the observations generated from the contaminating distribution is fixed at 15%. Three different scenarios will be considered to simulate the contaminating distribution

Case I. Identification dataset is contaminated with scale outliers.

Case II. Identification dataset is contaminated with symmetric location outliers.

Case III. Identification dataset is contaminated with asymmetric location outliers.

The steady-state model to be identified is chosen as the form

$$y_1(k) = \theta_1 u_1(k) + \theta_2 u_2(k) + \theta_3 \tag{85}$$

We also consider the dynamic ARX-based identification of the fermentation problem in the neighbor of the nominal operating point to approximately capture the dynamic relationship between the input and output variables. The model to be identified is of the form

$$y_1(k) = \theta_1 u_1(k) + \theta_2 u_2(k) + \theta_3 y_1(k-1) + \theta_4 \tag{86}$$

OLS regression, regular Bayesian, Huber robust regression, and robust Bayesian are applied for identification of the steady-state and dynamic models. To evaluate the robustness of these methods, the prediction performance of the identified models is compared in Tables 2 and 3 for validation datasets; the results are summarized from 50 simulation runs. MSE, mean absolute error (MAE), and standard deviation of error (STDE) are the performance metrics evaluated. It can be observed that the models identified using robust Bayesian framework are both more accurate (with smaller MAE) and more reliable (with smaller STDE). Moreover, the relatively smaller values of MSE imply the overall better prediction performance in terms of both accuracy and reliability. The advantage of the proposed robust framework over the traditional robust regression techniques is highlighted specially when the identification dataset is contaminated with the location outliers.

Table 2. Comparison of the Prediction Performance of the Identified Steady-State Models on the Validation Data

	OLS Regression	Robust Regression	Regular Bayesian	Robust Bayesian
Case I: Scale outliers in the identification data				
MSE of Prediction	0.313 ± 0.067	0.286 ± 0.033	0.313 ± 0.045	0.250 ± 0.039
STDE of Prediction	0.230 ± 0.042	0.220 ± 0.031	0.230 ± 0.041	0.215 ± 0.024
MAE of Prediction	0.254 ± 0.059	0.230 ± 0.045	0.253 ± 0.045	0.198 ± 0.033
Case II: Symmetric location outliers in the identification data				
RMSE of Prediction	0.308 ± 0.076	0.249 ± 0.036	0.307 ± 0.075	0.208 ± 0.020
STDE of Prediction	0.265 ± 0.057	0.224 ± 0.030	0.265 ± 0.056	0.201 ± 0.018
MAE of Prediction	0.251 ± 0.067	0.202 ± 0.030	0.250 ± 0.066	0.165 ± 0.016
Case III: Asymmetric location outliers in the identification data				
RMSE of Prediction	0.487 ± 0.086	0.383 ± 0.059	0.486 ± 0.084	0.229 ± 0.039
STDE of Prediction	0.229 ± 0.040	0.227 ± 0.036	0.228 ± 0.038	0.212 ± 0.025
MAE of Prediction	0.432 ± 0.084	0.321 ± 0.054	0.432 ± 0.083	0.182 ± 0.029

Table 3. Comparison of the Prediction Performance of the Identified Dynamic Models on the Validation Data

	OLS Regression	Robust Regression	Regular Bayesian	Robust Bayesian
Case I: Scale outliers in the identification data				
RMSE of prediction	0.462 ± 0.124	0.445 ± 0.120	0.460 ± 0.142	0.424 ± 0.119
STDE of prediction	0.400 ± 0.120	0.394 ± 0.112	0.396 ± 0.129	0.387 ± 0.113
MAE of prediction	0.363 ± 0.080	0.346 ± 0.074	0.363 ± 0.107	0.324 ± 0.078
Case II: Symmetric location outliers in the identification data				
RMSE of prediction	0.478 ± 0.108	0.448 ± 0.081	0.475 ± 0.109	0.409 ± 0.072
STDE of prediction	0.435 ± 0.088	0.415 ± 0.072	0.433 ± 0.088	0.393 ± 0.066
MAE of prediction	0.376 ± 0.090	0.343 ± 0.0.62	0.374 ± 0.089	0.312 ± 0.054
Case III: Asymmetric location outliers in the identification data				
RMSE of prediction	0.745 ± 0.113	0.633 ± 0.097	0.739 ± 0.111	0.438 ± 0.097
STDE of prediction	0.390 ± 0.087	0.390 ± 0.086	0.390 ± 0.087	0.395 ± 0.089
MAE of prediction	0.660 ± 0.116	0.541 ± 0.094	0.657 ± 0.112	0.341 ± 0.080

The averaged noise variance (σ_e^2) estimates along with the standard deviation of the estimated values obtained from each of the robust methods are presented in Table 4. The reported results show that the robust Bayesian outperforms the robust regression in terms of the accuracy of the noise variance estimates. Therefore, another advantage of the developed Bayesian framework is that it provides much more accurate estimates of hyperparameters such as the measurement noise variance.

Continuous stirred tank heater experiment

To further demonstrate the capability of the proposed Bayesian method, identification of an ARX model using the experimental data obtained from a pilot-scale continuous stirred tank heater (CSTH) is considered. The CSTH pilot plant is located in the Computer Process Control Laboratory in the Department of Chemical and Materials Engineering at the University of Alberta. As illustrated in Figure 5, the feed stream of the cold water flows into a well-stirred heated tank. The cold water is heated using saturated steam through a heating coil and drained from the tank through a long pipe.[34] Given a fixed volume of water in the tank, it is desired to heat the inlet stream to a higher setpoint temperature. To achieve this control objective, the outflow temperature is measured and the steam flow rate is adjusted accordingly.

We consider the problem of identifying a dynamic model relating the steam flow rate (input u) to the outlet water temperature (output y); the experimental data was taken from Jin.[35] A random binary sequence based variation in the steam flow rate was used to sufficiently excite the process for collecting identification data; the input was varied between 10 and 15 kg/hr. It is noteworthy that the level of water in the tank is controlled at 25 cm to isolate the significant effect of level variations on the process dynamics. The input–output data collected from the CSTH pilot plant is shown in Figure 6. The identification dataset is used to identify empirical models of the form

$$y(k) = \theta_1 u(k) + \theta_2 y(k-1) + \theta_3 \tag{87}$$

OLS regression, regular Bayesian, Huber robust regression, and robust Bayesian are applied to estimate the model parameters. The prediction performance of the identified models is then tested on the validation dataset.

As the main focus of this study is to investigate the robustness of different identification procedures, an identification dataset contaminated with outliers is of interest. Therefore, several outliers are randomly added to the identification dataset. It is expected that the presence of outliers will decrease the performance of various identification procedures. Thus, the parameter estimates obtained from the original identification dataset are considered as reference values.

Table 5 compares the parameter estimates obtained from the original dataset with the ones identified from the contaminated datasets. In the absence of contamination, parameter estimation results from the investigated identification methods are comparable. Regardless of the form of contamination, however, presence of outliers in the identification data generally destroys the performance of nonrobust estimators. Also, it can be clearly observed that the OLS regression and the regular Bayesian methods fail similarly. In contrast, the robust methods provide reasonably accurate parameter estimates, even when the identification dataset is contaminated by either scale or location outlying observations. Having included the contaminating model in the identification procedure, it is evident that the proposed Bayesian approach outperforms the Huber estimator in robustness especially in the presence of location outliers.

To evaluate the performance of the identified models, infinite horizon predictions (simulation) are performed on the

Table 4. Comparison of the Noise Variance Estimates Obtained Using Different Robust Methods

	Simulated Value	Robust Regression	Robust Bayesian
Case I: Scale outliers in the identification data			
Steady-state model	0.490 ± 0.048	1.092 ± 0.124	0.408 ± 0.098
Dynamic model	0.490 ± 0.040	1.387 ± 0.109	0.645 ± 0.126
Case II: Symmetric location outliers in the identification data			
Steady-state model	0.494 ± 0.053	1.554 ± 0.204	0.451 ± 0.044
Dynamic model	0.507 ± 0.046	2.201 ± 0.278	0.636 ± 0.073
Case III: Asymmetric location outliers in the identification data			
Steady-state model	0.495 ± 0.046	1.219 ± 0.186	0.352 ± 0.066
Dynamic model	0.512 ± 0.044	2.238 ± 0.293	0.462 ± 0.074

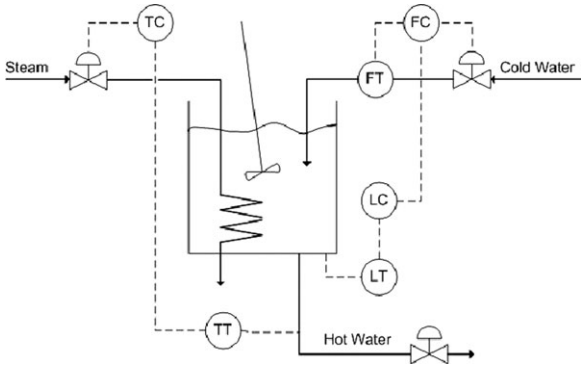


Figure 5. A simplified configuration of the CSTH.

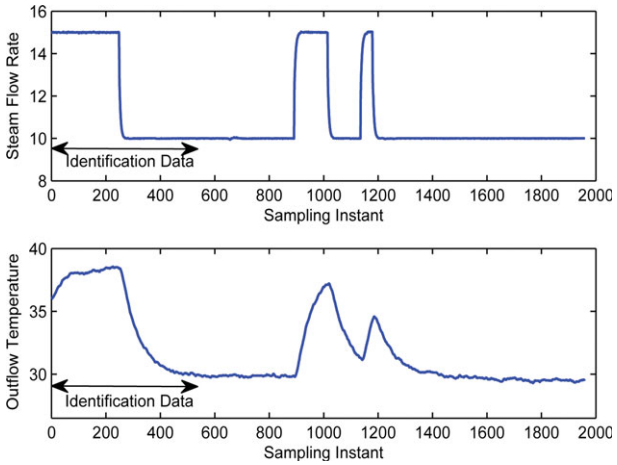


Figure 6. Input–output experimental data from a pilot-scale CSTH.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

validation dataset. The results are compared in Figures 7, 8, and 9. In the case of the contaminated identification data, the models identified through the use of nonrobust methods exhibit poor prediction performance. However, it can be observed that the robustness of the proposed Bayesian approach and Huber estimator significantly improve the predictive accuracy of the identified models. The prediction performance of the models identified using the robust Bayesian framework is better than that of the ones identified using Huber robust regression.

Table 5. Comparison of Estimated Parameters of the CSTH Model

	OLS Regression	Robust Regression	Regular Bayesian	Robust Bayesian
Case I: No outlier in the identification data				
θ_1	0.0243	0.0248	0.0243	0.0254
θ_2	0.9866	0.9862	0.9866	0.9859
θ_3	0.1547	0.1599	0.1547	0.1639
Case II: Scale outliers in the identification data				
θ_1	0.0537	0.0252	0.0539	0.0241
θ_2	0.9626	0.9856	0.9624	0.9865
θ_3	0.6236	0.1789	0.6271	0.1608
Case III: Symmetric location outliers in the identification data				
θ_1	0.0613	0.0285	0.0615	0.0250
θ_2	0.9566	0.9831	0.9564	0.9865
θ_3	0.7393	0.2212	0.7432	0.1513
Case IV: Asymmetric location outliers in the identification data				
θ_1	0.0642	0.0294	0.0644	0.0249
θ_2	0.9545	0.9823	0.9542	0.9860
θ_3	0.7834	0.2404	0.7838	0.1684

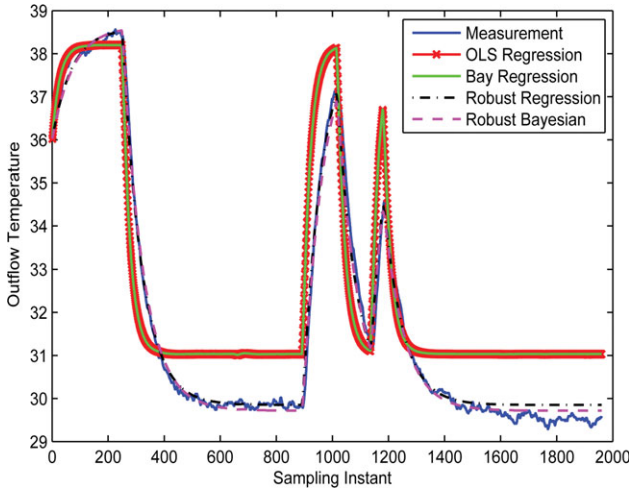


Figure 7. Prediction performance of the identified CSTH models; identification dataset is contaminated with the scale outliers.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

To summarize, this experimental study shows that the proposed robust Bayesian framework performs well under a wide variety of circumstances: with or without contamination, with scale or location outliers, and with symmetric or asymmetric contaminating distributions.

Concluding Remarks

Identification of ARX models in the presence of outliers was considered in this article. To obtain a computationally feasible formulation, a set of indicator variables was introduced to denote the identity of each data point. Also, a contaminated Gaussian distribution was adopted to describe the observed data. The ARX identification problem was then

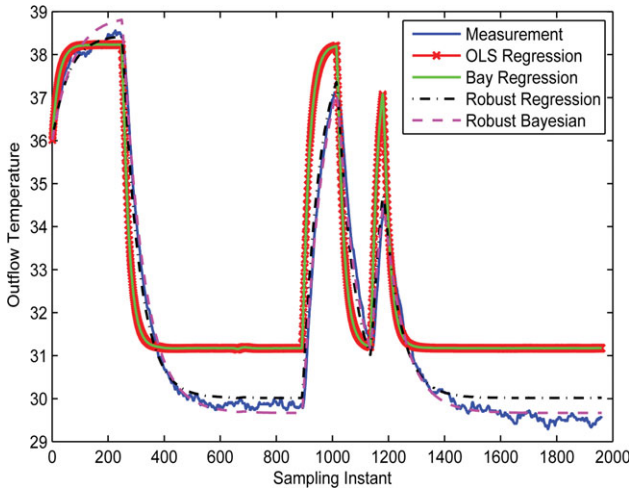


Figure 8. Prediction performance of the identified CSTH models; identification dataset is contaminated with the symmetric location outliers.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

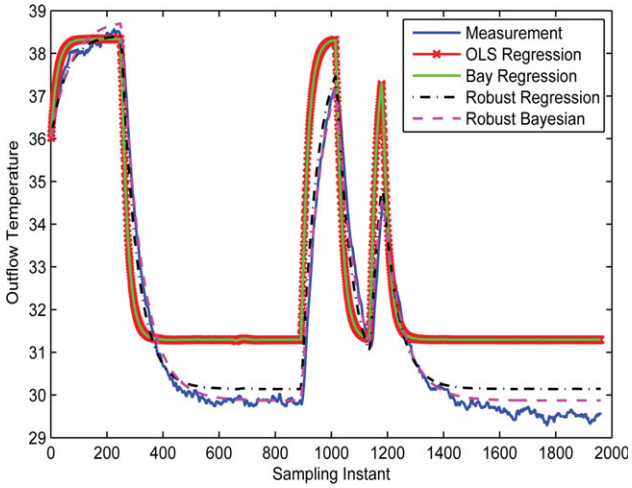


Figure 9. Prediction performance of the identified CSTH models; identification dataset is contaminated with the asymmetric location outliers.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

formulated and solved under an iterative hierarchical Bayesian optimization framework. The layered optimization scheme allows us to obtain MAP estimates of model parameters with an automated mechanism for determining the hyperparameters and investigating the identity of each data point. The effectiveness of the developed framework for robust identification was demonstrated on the simulated and experimental datasets. The layered optimization solution builds a unified framework that ensures that the model identification process is not significantly affected by outliers, which makes this method more applicable to the real world problems.

Acknowledgments

The authors gratefully acknowledge the financial support from the Natural Sciences and Engineering Research Council of Canada (NSERC).

Literature Cited

1. Fortuna L, Graziani S, Rizzo A, Xibilia MG. *Soft Sensors for Monitoring and Control of Industrial Processes*. London: Springer Verlag, 2007.
2. Prasad V, Schley M, Russo LP, Bequette BW. Product property and production rate control of styrene polymerization. *J Process Control*. 2002;12:353–372.
3. Muller CJ, Craig IK, Ricker NL. Modelling, validation, and control of an industrial fuel gas blending system. *J Process Control*. 2011;21:852–860.
4. Sabbe MK, Geem KMV, Reyniers MF, Marin GB. First principle-based simulation of ethane steam cracking. *AIChE J*. 2011;57:482–496.
5. Kano M, Nakagawa Y. Data-based process monitoring, process control, and quality improvement: Recent developments and applications in steel industry. *Comput Chem Eng*. 2008;32:12–24.
6. Kadlec P, Gabrys B, Strandt S. Data-driven soft sensors in the process industry. *Comput Chem Eng*. 2009;33:795–814.
7. Chiang LH, Pell RJ, Seasholtz MB. Exploring process data with the use of robust outlier detection algorithms. *J Process Control*. 2003;13:437–449.
8. Liu H, Shah SL, Jiang W. Online outlier detection and data cleaning. *Comput Chem Eng*. 2004;28:1635–1647.
9. Khatibisepehr S, Huang B. Dealing with irregular data in soft sensors: Bayesian method and comparative study. *Ind Eng Chem Res*. 2008;47:8713–8723.
10. Grubbs FE. Procedures for detecting outlying observations in samples. *Technometrics*. 1969;11:1–21.
11. Zeng J, Gao C. Improvement of identification of blast furnace ironmaking process by outlier detection and missing value imputation. *J Process Control*. 2009;19:1519–1528.
12. Lee J, Kang B, Kang S. Integrating independent component analysis and local outlier factor for plant-wide process monitoring. *J Process Control*. 2011;21:1519–1528.
13. Hodge VJ, Austin J. A survey of outlier detection methodologies. *Artifi Intelli Rev*. 2004;22:85–126.
14. Chandola V, Banerjee A, Kumar V. Anomaly detection : A survey. *ACM Comput Surv*. 2009;41:124–129.
15. Huber PJ. *Robust Statistics*. New York: Wiley, 1981.
16. Bishop C. Novelty detection and neural network validation. *IEE Proc Visi Image Signal Process Special Iss Appl Neural Netw*. 1994;141:217–222.
17. Yu J. A bayesian inference based two-stage support vector regression framework for soft sensor development in batch bioprocesses. *Comput Chem Eng*. 2012;41:134–144.
18. Agarwal D. Detecting anomalies in cross-classified streams: A Bayesian approach. *Knowl Inform Syst*. 2006;11:29–44.
19. Ritter G, Gallegos MT. Outliers in statistical pattern recognition and an application to automatic chromosome classification. *Pattern Recogn Lett*. 1997;18:525–539.
20. Rousseeuw P, Leroy A. *Robust Regression and Outlier Detection*. 2nd ed. New York: Wiley, 1996.
21. Varbanov A. Bayesian approach to outlier detection in multivariate normal samples and linear models. *Communi Stat Theory Methods*. 1998;27:547–557.
22. Ghosh-Dastidar B, Schafer JL. Outlier detection and editing procedures for continuous multivariate data. *J Offic Stat*. 2006;22:487–506.
23. Das K, Schneider J. *Detecting anomalous records in categorical datasets*. In: *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Jose, CA: ACM Press, 2007.
24. MacKay DJC. Bayesian interpolation. *Neural Comput*. 1992;4:415–447.
25. MacKay DJC. Probable networks and plausible predictions — a review of practical bayesian methods for supervised neural networks. *Netw Comput Neural Syst*. 1995;6:469–505.
26. Molina R, Vega M, Mateos J, Katsaggelos AK. Variational posterior distribution approximation in Bayesian super resolution reconstruction of multispectral images. *Appl Comput Harmonic Anal*. 2008;24:251–267.
27. Galatsanos NP, Mesarović VZ, Molina R, Katsaggelos AK. Hierarchical bayesian image restoration from partially known blurs. *IEEE Trans Image Process*. 2000;9:1784–1797.
28. Kwok JT. The evidence framework applied to support vector machines. *IEEE Trans Neural Network*. 2000;11:1162–1173.
29. Suykens JAK, Gestel TV, Brabanter JD, Moor BD, Vandewalle J. *Least Squares Support Vector Machines*. River Edge, NJ: World Scientific, 2002.
30. Mackay DJC. *Information Theory, Inference, and Learning Algorithm*. UK: Cambridge University Press, 2003.
31. Raiffa H, Schlaifer R. *Applied Statistical Decision Theory*. Boston: Harvard University, 1961.
32. MacKay DJC. Comparison of approximate methods for handling hyperparameters. *Neural Comput*. 1999;11:1035–1068.
33. Henson MA, Seborg DE. *Nonlinear Process Control*. USA: Prentice-Hall Inc., 1997.
34. Thornhill NF, Patwardhan SC, Shah SL. A continuous stirred tank heater simulation model with applications. *J Process Control*. 2008;18:347–360.
35. Jin X. Multiple ARX model based identification for switching/nonlinear systems with EM algorithm. Master's thesis, University of Alberta; Edmonton, Canada, 2010.

Manuscript received Apr. 6, 2012, and revision received Jun. 23, 2012.