Maximum-Likelihood Maximum-Entropy Constrained Probability Density Function Estimation for Prediction of Rare Events

Taha Mohseni Ahooyi and Masoud Soroush
Dept. of Chemical and Biological Engineering, Drexel University, Philadelphia, PA 19104

Jeffrey E. Arbogast
Air Liquide, Delaware Research & Technology Center, Newark, DE 19702

Warren D. Seider
Dept. of Chemical and Biomolecular Engineering, University of Pennsylvania, Philadelphia, PA 19104

Ulku G. Oktem
Risk Management and Decision Processes Center, Wharton School, Philadelphia, PA 19104

DOI 10.1002/aic.14330
Published online February 2, 2014 in Wiley Online Library (wileyonlinelibrary.com)

This work addresses the problem of estimating complete probability density functions (PDFs) from historical process data that are incomplete (lack information on rare events), in the framework of Bayesian networks. In particular, this article presents a method of estimating the probabilities of events for which historical process data have no record. The rare-event prediction problem becomes more difficult and interesting, when an accurate first-principles model of the process is not available. To address this problem, a novel method of estimating complete multivariate PDFs is proposed. This method uses the maximum entropy and maximum likelihood principles. It is tested on mathematical and process examples, and the application and satisfactory performance of the method in risk assessment and fault detection are shown. Also, the proposed method is compared with a few copula methods and a nonparametric kernel method, in terms of performance, flexibility, interpretability, and rate of convergence. © 2014 American Institute of Chemical Engineers
AIChE J, 60: 1013–1026, 2014
Keywords: rare events, Bayesian network modeling, probability density function estimation, risk assessment, fault detection

Introduction

Fault detection and risk assessment are of great importance in the process industries. These analyses allow one to detect and quantify risk-prone spots within a processing plant and then mitigate or eliminate risks to the plant.[1] Tools such as support vector machines,[2] causal dependency,[3] fuzzy logic,[4] event trees,[5] filter-based methods,[6] improved kernel component analysis,[7] and Bayesian networks (BNs)[8] have been successfully applied to conduct probabilistic inference, sensitivity analysis, and detection and isolation of most probable causes of abnormal events. Methods have also been developed for fault detection and isolation under nonlinear closed-loop process conditions. In these methods, various statistical tests along with control system reconfiguration have been utilized to identify deviations and take proper control actions to mitigate the risk of such abnormalities.[9,10]

Calculating risk (probability of an abnormal event times the severity of the consequences of the event) in a processing plant whose database has no historical information on the abnormal event is a major challenge in risk prediction. This incompleteness of plant information can be due to the plant data having been collected during time intervals when no abnormal event occurred, or the plant having been controlled so tightly that its variables never entered into "unsafe" ranges. The severity of the problem of addressing this data incompleteness increases significantly when no first-principles model of the plant is available. The problem of estimating the probability of an abnormal event whose occurrence has never been recorded is often referred to as"rare-event" probability estimation.[11]

There are two major rare-event probability estimation problems. The more common and easier one deals with the estimation of marginal distributions of independent variables. Many approaches have been suggested to address this problem.[12–14] Conversely, the estimation of conditional probabilities of dependent variables is more complicated. To address this, one needs to calculate joint (multivariate) probability densities as well as marginal densities. Joint probability densities describe the dependence of effect variables (child nodes) on cause variables (parent nodes) probabilistically.

Most rare-event probability estimation methods are based on sampling.[15,16] These methods estimate rare-event

Correspondence concerning this article should be addressed to M. Soroush at soroushm@drexel.edu.

probabilities by drawing large numbers of samples from appropriate models describing target systems. There are many variants of such methods for different types of underlying models. Monte-Carlo (MC) sampling is the core of many of these methods.[17,18] To address the slow convergence rate of traditional MC methods, modified versions of random samplings have been proposed. Importance sampling uses a change of measure, takes samples from an alternative distribution, and maps the outcome to the original space.[19–21] Splitting methods divide the range of each random variable into intervals and use random walk to generate rare-event missing data.[22,23] Finally, Markov-chain MC methods are those utilizing Markov chains to produce a random walk.[24,25]

Although, the sampling techniques have shown good performance in many applications, they have drawbacks that have prevented their widespread use. One drawback is that simulation of infinitesimal probabilities using these methods takes very long times in practice[26,27]; calculation of a probability as small as 10^{-8} on a computer generating one sample every millisecond can take more than 30 years using standard MC simulations. Another drawback is that they can be used only when a model exists. In other words, every sample is the outcome of a computational process that needs a model. In the absence of a reliable model, when only data are available, probability density function (PDF) estimation methods are useful to model the behavior of a stochastic system.[13] PDF estimation has its own variants, divided into parametric[28] and nonparametric types.[29] As shown in this article, despite many appealing features of existing PDF estimation methods, these methods are not general enough to address all rare-event probability estimation problems. Existing multivariate PDF estimation methods are unable to provide acceptable estimates in all regions where no data have been observed, especially when the relations among the field variables are nonmonotonic.

In this work, a method of estimating multivariate PDFs that have maximum entropy (ME) and maximum likelihood (ML) is presented. As shown herein, although this method provides continuous probability distributions for continuous random variables, it can be extended easily to discrete random variables. To derive such a PDF, PDFs that maximize entropy[30] and likelihood[31] simultaneously are sought. Therefore, herein, this method is referred to as a MLME method of PDF estimation. The method uses information available in historical datasets to estimate a global probability rule applicable to all regions of each random variable domain. Another advantage over existing parametric and nonparametric methods is that this method allows for effectively considering higher moments of each random variable PDF (e.g., skewness and kurtosis).

The rest of the article is organized as follows. The problem of estimating the probability of rare events within the framework of BNs is stated, and its significance is shown using a simple example in the next section. Some preliminaries are then presented, followed by the MLME PDF estimation method. The method is then applied to two examples, and its performance is discussed and compared with those of several widely used PDF estimation techniques. Finally, conclusions are drawn.

Problem Statement

In this section, a very simple example is considered to describe the rare-event probability estimation problem and show the importance of the problem solution in BN inference. The example involves two variables, Y and Z, where Z depends on Y. Throughout this article, each random variable is denoted by a capital letter and its numerical value denoted by a lowercase letter. Random variables are assumed to have five states: Low–Low (LL : $[\mu-4\delta, \mu-3\delta)$), Low (L : $[\mu-3\delta, \mu-2\delta)$), Normal (N : $[\mu-2\delta, \mu+2\delta]$), High (H : $[\mu+2\delta, \mu+3\delta]$), High–High (HH : $[\mu+3\delta, \mu+4\delta]$), where μ and δ are real numbers, which can be the sample mean and standard deviation, respectively.

BNs are directed acyclic graphs, which have been used extensively for probabilistic modeling, especially after Spiegelhalter[32] proposed algorithms that made probabilistic inference computationally tractable. BNs can account for the intrinsic uncertainties hidden in historical data without viewing uncertainties as noise. They are very flexible in terms of training information; they can be trained using many types of data such as historical data, data from simulated first principles, empirical and/or probabilistic process models, expert knowledge, discrete data, categorical data, continuous data, and incomplete/censored data, or a combination of these.[33] BNs require training information in every state of each variable; in the case that historical data is the only information from a process, the historical data should include data in every state of each variable.

BNs rely on training information to construct prior and conditional probability distributions.[34,35] These probability distributions are building blocks of the network and are necessary for performing inference.[36] If the distributions are estimated solely based on the ML principle, then the frequentists approach[37] should be used. In this case, the probability of the variable Y being in a state s_k is defined as the relative recurrence of the random variable Y visiting the state s_k:

$$P(y \in s_k) = \frac{n(y \in s_k)}{\sum_{i=1}^{m} n(y \in s_i)} \tag{1}$$

and the conditional probability of the variable Z being in a state r_i given the variable Y in a state s_k is defined as:

$$P(z \in r_i | y \in s_k) = \frac{n(z \in r_i, y \in s_k)}{n(y \in s_k)} \tag{2}$$

where n denotes the number (frequency) of observed samples within a specified state. Assume for the example under consideration frequencies of observed samples are those given in Table 1. Note that in some states no data have been observed. According to Eqs. 1 and 2, the probabilities of Y and Z being in these "null" states are zero. However, in most cases this situation occurs due to small sample sizes and near-zero (but not necessarily zero) probabilities. For this reason, these events are called "rare events." According to the law of large numbers, the relative frequency of the observations of a random event converges to the actual probability of the event when the number of random experiments/observations approaches infinity.

Now suppose that despite zero empirical possibility of having Y in HH, Y has been observed in this state. Since Z is a function of Y, it is affected by the state HH of Y. To calculate this impact (conduct probabilistic inference), we use Bayes' rule:

$$P(z | y \in \mathrm{HH}) = \frac{P(y \in \mathrm{HH} | z) P(z)}{P(y \in \mathrm{HH})} = \frac{0 \times P(z)}{0} = \frac{0}{0} \tag{3}$$

indicating that such an inference is impossible. Because Bayesian inference is highly dependent on the availability of

Table 1. Frequency (Number) of Y and Z Sample Data in Each State

		State of Z				
State of Y	No. of Y	LL	L	N	H	HH
LL	0	0	0	0	0	0
L	38	5	31	2	0	0
N	1093	0	11	1058	23	1
H	16	0	0	1	13	2
HH	0	0	0	0	0	0

the conditional and prior probabilities, the probabilistic inference does not yield a reasonable result for cases for which no data are available. Knowledge of the probability of such "rare" states/events is of great importance, as in many cases a random variable taking an extreme value is indicative of an unsafe (highly risky) condition. This is the main motivation for this research that is aimed at: (i) solving the problem of rare-event probability estimation from historical data and (ii) using the estimates in probabilistic inference in the framework of BNs.

Preliminaries

Moments of a probability distribution function

Moments of a random variable (vector) X with a PDF $f(\mathrm{x})$ are defined as expected values of arbitrary functions of the random variable (vector). The most common moments are the first-order moment ($E(\mathrm{X})$ or mean) and the second-order moment ($E(\mathrm{X}^2)$).[38,39] Ordinarily, there are no limitations on the form of moment functions selected, but polynomial functions are often preferred, because their analytical integral is more likely to have a closed form.

Let $g_i(\vec{x}) : R^d \rightarrow R$ be a moment function of a d-dimensional random vector $\vec{x}=[x^1 \ldots x^d]^T \in \Omega \subseteq R^d$ with a PDF $f(\vec{x}) : R^d \rightarrow R^+$, where Ω is the domain of $\vec{x}$. The moment of the random vector $\vec{x}$ with respect to the moment function $g_i(\vec{x})$ is defined as:

$$\mu_i = E(g_i(\vec{x})) = \int_\Omega g_i(\vec{x}) f(\vec{x}) d\vec{x}, \; i=0,1,2,\ldots \tag{4}$$

For a sample population, the moment of the population with respect to the moment function $g_i(\vec{x})$ is calculated using sample moments:

$$\overline{\mu}_i = n^{-1} \sum_{j=1}^{n} g_i(\vec{x}_j), \quad i=0,1,2, \ \ldots \tag{5}$$

where n is the number of samples of $\vec{x}_j$.

Entropy of a random variable

In information theory, the entropy of a random variable is a measure of the uncertainty in the random variable.[40] In this context, the term usually refers to the Shannon entropy,[41] which quantifies the expected value of the information contained in a message. Shannon entropy of a random variable is a measure of unpredictability or information content of the variable. In the case of a coin with one tail and one head having equal probabilities, the entropy of the coin toss is highest. This is because it is not possible to predict the outcome of the coin toss before tossing the coin. However, a coin toss with a coin that has no tails and two heads has zero entropy because the coin toss outcome is always known and can be predicted perfectly. Most real-world data fall between these two extremes. So, as the entropy of a random variable increases, its unpredictability (uncertainty) increases, and vice versa.

For a continuous random vector $\vec{X}$ with a PDF $f(\vec{x})$ on a domain Ω the information entropy is defined as[39]:

$$S(\vec{X}) = -\int_\Omega f(\vec{x}) \ln f(\vec{x}) d\vec{x} \tag{6}$$

with $0 \times \ln 0 = 0$. This notion of entropy is similar to the notion of entropy in thermodynamics. Physically, systems tend to evolve into states with higher entropy. In the probabilistic context, $S(\vec{X})$ is viewed as a measure of the information carried by $\vec{X}$, and as data are communicated/transmitted more, they are corrupted with more noise (entropy increases) and therefore they carry less information.

Method

Given a dataset, to estimate a PDF of a random vector, a PDF with the following two properties is sought: (a) a selected set of the moments of the PDF should be the same as the moments of the available data on the variables and (b) the PDF should have the highest level of uncertainty amongst all possible PDFs satisfying the first property. In other words, a PDF $f(\vec{x})$ is sought that is the solution to the constrained optimization problem:

$$\max_f \left\{ -\int_\Omega f(\vec{x}) \ln f(\vec{x}) d\vec{x} \right\} \tag{7}$$

subject to the equality constraints

$$\int_\Omega g_i(\vec{X}) f(\vec{X}) d\vec{X} = \overline{\mu}_i, \quad i=0,\ldots,m \tag{8}$$

where $\overline{\mu}_i$ is the ith moment of the sample data. The integer m is the number of moments of the PDF that the user chooses to match with the moments of the data sample, in addition to the zeroth moment, μ_0, which corresponds to the zeroth-moment function, $g_0(\vec{x})$. One should always set $g_0(\vec{x}) = 1$ and make sure to include this moment function in the search for the optimal PDF. The zeroth-moment equality constraint simply ensures that the calculated PDF always satisfies $\int_\Omega f(\vec{x}) d\vec{x} = \overline{\mu}_0 = 1$. This PDF estimation formulation is a multivariate version of the univariate formulation introduced by Zellner and coworkers.[42,43] This method determines the PDF that represents the data and accounts for the maximum uncertainty that exists in the data. As it does not impose any prior assumptions on the underlying distribution to be estimated, the method allows for the estimation of PDFs with minimum bias. The constrained optimization of Eqs. 7 and 8 is a classical optimization problem, whose solution minimizes the Lagrange function:

$$\begin{aligned} \tilde{L}(f, \ \lambda_0, \ldots, \lambda_m) &= \int_\Omega f(\vec{x}) \ln f(\vec{x}) d\vec{x} \\ &+ \sum_{i=0}^{m} \lambda_i \left(\int_\Omega g_i(\vec{x}) f(\vec{x}) d\vec{x} - \overline{\mu}_i \right) \end{aligned} \tag{9}$$

where $\lambda_0, \ldots, \lambda_m$ are the Lagrange multipliers. The solution to the optimization problem satisfies the following necessary conditions of optimality:

$$\frac{\partial \tilde{L}}{\partial \hat{f}} = 0, \quad \int_{\Omega} g_i(\vec{x})\hat{f}(\vec{x})d\vec{x} = \overline{\mu}_i, \quad i=0,\ldots,m \tag{10}$$

where $\hat{f}(\vec{x})$ is the estimated PDF. The first algebraic equation from the left in Eq. 10 yields:

$$\frac{\partial \tilde{L}}{\partial \hat{f}} = \frac{\partial}{\partial \hat{f}}\left\{\int_{\Omega}\left[\hat{f}(\vec{x})\ln\hat{f}(\vec{x}) + \sum_{i=0}^{m}\lambda_i g_i(\vec{x})\hat{f}(\vec{x})\right]d\vec{x} - \sum_{i=0}^{m}\lambda_i\overline{\mu}_i\right\} = 0,$$

Using the Leibniz integral rule, the preceding equation simplifies to:

$$\int_{\Omega}\left[\ln\hat{f}(\vec{x}) + 1 + \sum_{i=0}^{m}\lambda_i g_i(\vec{x})\right]d\vec{x} = 0$$

Therefore,

$$\ln\hat{f}(\vec{x}) + 1 + \sum_{i=0}^{m}\lambda_i g_i(\vec{x}) = 0$$

leading to the closed-form analytical solution:

$$\hat{f}(\vec{x}) = \exp\left(-1-\lambda_0 - \sum_{i=1}^{m}\lambda_i g_i(\vec{x})\right)$$

which can be written in the form:

$$\hat{f}(\vec{x}) = \frac{1}{e^{1+\lambda_0}}\exp\left(-\sum_{i=1}^{m}\lambda_i g_i(\vec{x})\right) \tag{11}$$

Requiring $\hat{f}(\vec{x})$ to satisfy the zeroth-moment equality constraint:

$$\int_{\Omega}\hat{f}(\vec{x})d\vec{x} = \int_{\Omega}\frac{1}{e^{1+\lambda_0}}\exp\left(-\sum_{i=1}^{m}\lambda_i g_i(\vec{x})\right)d\vec{x} = 1,$$

implies that

$$e^{1+\lambda_0} = \int_{\Omega}\exp\left(-\sum_{i=1}^{m}\lambda_i g_i(\vec{x})\right)d\vec{x}$$

There are different ways to calculate the rest of the Lagrange multipliers. For example, the Lagrange multipliers can be found by requesting that the theoretical moments described by the estimated PDF be equal to the empirical moments evaluated by taking the average over the sampled data. This procedure is usually referred to as the method of moments (MM).[44] Different versions of MM along with the generalized MM[45,46] have been proposed. Requesting equal data and model moments seems reasonable by the law of large numbers—which results from the ML estimation (MLE) method when the distribution belongs to the exponential family. The MLE is a probabilistic approach for minimum-variance estimation of PDF parameters.[47] As shown later, the use of the MLE to estimate the Lagrange multipliers (model parameters) requires that all moment constraints are satisfied.

Given sample points that are independent and identically distributed, using the MLE method, the unknown parameters (Lagrange multipliers) of the PDF are obtained from:

$$\begin{aligned}\vec{\lambda}_{\mathrm{MLE}} &= \arg\max_{\vec{\lambda}} L(\vec{\lambda}|D) = \arg\max_{\vec{\lambda}} f(D|\vec{\lambda}) \\ &= \arg\max_{\vec{\lambda}} \prod_{j=1}^{n}\frac{\exp\left(-\sum_{i=1}^{m}\lambda_i g_i(\vec{x}_j)\right)}{\theta}\end{aligned} \tag{12}$$

where L is called the likelihood function, $\vec{\lambda}$ is the vector of the Lagrange multipliers, n is the number of samples, D denotes the data samples forming an $(n \times d)$ matrix, and

$$\theta = \mathrm{e}^{1+\lambda_0} = \int_{\Omega}\exp\left(-\sum_{i=1}^{m}\lambda_i g_i(\vec{x})\right)d\vec{x}$$

which is often called the partition function. The MLE method requires the Hessian matrix of the likelihood function to be absolutely negative definite at $\vec{\lambda}_{\mathrm{MLE}}$. Since ln is a monotonically increasing function, the model parameters can also be calculated by maximizing ln of the likelihood function:

$$\begin{aligned}\vec{\lambda}_{\mathrm{MLE}} &= \arg\max_{\vec{\lambda}} \ln L(\vec{\lambda}|D) \\ &= \arg\max_{\vec{\lambda}}\left\{-n\ln[\theta] - \sum_{j=1}^{n}\sum_{i=1}^{m}\lambda_i g_i(\vec{x}_j)\right\}\end{aligned} \tag{13}$$

This is usually known as the log-likelihood of the parameters given the data.

Existence and uniqueness of the MLE solution

In this section, the existence and uniqueness of the MLE solution is investigated. The MLE optimization problem of Eq. 13 may have multiple local optima in addition to the global one. A unique solution (global optimum) exists when the likelihood function is strictly convex. The number of solutions usually increases, as the degree of nonlinearity of the PDF model increases, the number of parameters of the model increases, or the size of the data sample decreases. The number of solutions also depends on the family of distributions to which the PDF belongs. Since the ME moment-constrained estimator is from the class of exponential distributions,[48] the MLE problem of Eq. 13 is expected to have a unique optimum (global maximum).[49]

First it is proven that the MLE problem described by Eq. 13 has a solution. This can be achieved simply by showing that the system of partial derivatives of the log-likelihood function with respect to the Lagrange multipliers set to zero has a solution:

$$\frac{\partial}{\partial\lambda_i}\ln L\left(\vec{\lambda}|D\right) = -n\frac{\partial\ln\theta}{\partial\lambda_i} - \sum_{j=1}^{n} g_i\left(\vec{x}_j\right) = 0, \quad i=1,\ldots,m \tag{14}$$

leading to:

$$\frac{\partial\ln\theta}{\partial\lambda_i} = -n^{-1}\sum_{j=1}^{n} g_i\left(\vec{x}_j\right) = -\overline{\mu}_i, \quad i=1,\ldots,m \tag{15}$$

Hence, the model parameters should be estimated by satisfying the m nonlinear algebraic equations in Eq. 15. The right-hand sides of Eq. 15 are simply the empirical moments, indicating that larger sample sizes do not add any additional computational burden to the calculation of the model parameters, because only moments of the sample data are needed. The system of nonlinear equations in Eq. 15 that the Lagrange multipliers should satisfy can be solved using a root-finding method, such as the Newton–Raphson method.[42]

Furthermore, according to the definition of the partition function, θ:

$$\frac{\partial \ln\theta}{\partial\lambda_i}=\frac{1}{\theta}\frac{\partial}{\partial\lambda_i}\int_\Omega \exp\left(-\sum_{i=1}^{m}\lambda_i g_i(\vec{x})\right)d\vec{x}$$
$$=-\int_\Omega g_i(\vec{x})\frac{1}{\theta}\exp\left(-\sum_{i=1}^{m}\lambda_i g_i(\vec{x})\right)d\vec{x} \qquad (16)$$
$$=-\int_\Omega g_i(\vec{x})\hat{f}(\vec{x})d\vec{x}=-\hat{\mu}_i$$

Therefore, Eq. 14 simply requires that:

$$\hat{\mu}_i=\overline{\mu}_i,\quad i=1,\ldots,m$$

which implies that if the empirical moments, $\overline{\mu}_i, i=1,\ldots,\ m$, are finite, then the likelihood function has a critical point.

Now, this critical point that exists is shown to be a unique maximum. The entries of the Hessian matrix of the log-likelihood function are given by:

$$\frac{\partial}{\partial\lambda_k}\left(\frac{\partial}{\partial\lambda_i}\ln\left(L(\vec{\lambda}|D)\right)\right)=\frac{\partial}{\partial\lambda_k}\left(-n\frac{\partial\ln\theta}{\partial\lambda_i}-\sum_{j=1}^{n}g_i(\vec{x}_j)\right)$$
$$=-n\frac{\partial}{\partial\lambda_k}\frac{\partial\ln\theta}{\partial\lambda_i}=n\frac{\partial}{\partial\lambda_k}\left(\frac{1}{\theta}\int_\Omega g_i(\vec{x})\exp\left(-\sum_{i=1}^{m}\lambda_i g_i(\vec{x})\right)d\vec{x}\right)$$
$$=-n\frac{1}{\theta}\int_\Omega g_i(\vec{x})g_k(\vec{x})\exp\left(-\sum_{i=1}^{m}\lambda_i g_i(\vec{x})\right)d\vec{x}$$
$$+n\left(\frac{1}{\theta^2}\int_\Omega g_i(\vec{x})\exp\left(-\sum_{i=1}^{m}\lambda_i g_i(\vec{x})\right)d\vec{x}\right)$$
$$\times\int_\Omega g_k(\vec{x})\exp\left(-\sum_{i=1}^{m}\lambda_i g_i(\vec{x})\right)d\vec{x}$$
$$=-n\left[\int_\Omega g_i(\vec{x})g_k(\vec{x})\hat{f}(\vec{x})d\vec{x}\ -\int_\Omega g_i(\vec{x})\hat{f}(\vec{x})d\vec{x}\ \ \int_\Omega g_k(\vec{x})\hat{f}(\vec{x})d\vec{x}\right]$$
$$=-n[E(g_i(\vec{x})g_k(\vec{x}))-E(g_k(\vec{x}))E(g_k(\vec{x}))]$$
$$=-n\mathrm{CoV}\,(g_i(\vec{x}),g_k(\vec{x})),\quad k,i=1,\ldots,m \qquad (17)$$

where $\mathrm{CoV}\,(a,b)$ is the covariance of two random numbers a and b. Equation 17 indicates that the Hessian matrix is symmetric and strictly negative definite for every value of the vector of the Lagrange multipliers, implying that the critical point is a unique maximum. Equation 17 also indicates that the use of larger-size samples (larger n) gives the likelihood function a sharper peak, allowing one to calculate the maximum with less number iterations. In summary, the MLE solution of the moment-constrained ME problem is exactly the same widely used MM where $\hat{\mu}_i=\overline{\mu}_i, i=1,\ldots,m$.

Selection of moment functions

The type of the moment functions not only affects the estimated density functions, but it can affect significantly the computational complexity in the parameter estimation and the calculation of probabilities using the resulting PDF models. For a systematic selection of the moment functions, a criterion-based algorithm is suggested. In the MLME PDF of Eq. 11, if each $g_i(\vec{x})$ is replaced with a truncated Taylor series expansion of $g_i(\vec{x})$ around the expectation of $\vec{x}$, then the problem of looking for proper $g_i(\vec{x})$ moment functions is converted to that of finding an optimal order of the truncation for each of the expansions:

$$\vec{\lambda}_{\mathrm{MLE}}=\arg\max_{\vec{\lambda}}\ L\left(\vec{\lambda}|D\right)$$
$$=\arg\max_{\vec{\lambda}}\prod_{j=1}^{n}\frac{1}{\theta}\exp\left(-\sum_{i=1}^{m}\lambda_i\left[a_i+b_i\vec{x}_j+\vec{x}_j^{T}c_i\vec{x}_j+\ldots\right]\right)$$
$$=\arg\max_{\vec{\lambda}}\prod_{j=1}^{n}\frac{1}{\theta}\exp\left(\beta_0+\beta_1\vec{x}_j+\vec{x}_j^{T}\beta_2\vec{x}_j+\ldots\right)$$

where $\beta_0,\beta_1,\beta_2\ldots$ are constants to be estimated. For simplicity, one can seek equal truncation orders, denoted by O, for all of the moment functions. With this simplification, the search for the moment functions is converted to a search for an optimal truncation order, O_{opt}, that yields the best fit of $f(\vec{x})$ to the data. Measures like mean square error (MSE) and ML are often used to find an optimal level of the model complexity (O_{opt}). It is known that ML estimates tend to over-fit data, if model complexity exceeds a certain limit.[50–52] Such a limit exists here as well. However, since this optimum usually occurs at a high level of complexity at which the MSE and ML measures are insensitive to the complexity, a method is proposed herein to find an optimal value of the truncation order (O_{opt}) that provides adequate complexity/nonlinearity at a reasonable computational cost. A plot of the natural logarithm of the likelihood function at $\vec{\lambda}_{\mathrm{MLE}}$ vs. the order of the truncated Taylor series usually shows that the natural logarithm approaches a limit as the order of the truncation increases. This implies that an optimal truncation order (O_{opt}) can be calculated, for example, by using:

$$O_{\mathrm{opt}}=\arg\min_{O}\left(\frac{\partial\ln L(\vec{\lambda}'_{\mathrm{MLE}}(O)|D)}{\partial O}-\alpha\right)^2 \qquad (18)$$

where $L(\vec{\lambda}'_{\mathrm{MLE}}(O)|D)$ is the maximum of the likelihood function using an Oth -order truncated Taylor series expansion of every $g_i(\vec{x}),\ i=1,\ldots,m$. α is a positive scalar design parameter; a higher value of α leads to a lower value of O_{opt} and lower computational complexity and time needed to estimate PDF parameters and use the estimated PDFs. Therefore, the MLME PDF estimation provides a goodness-of-fit measure that can be used to systematically evaluate the advantages and disadvantages of selecting each moment function.

Application to Two Examples

In this section, two examples are considered to show the application and performance of the MLME PDF estimation method.

Example 1: A bivariate BN

Consider two random variables Y and Z described by:

$$Y\sim N(0,0.25) \qquad (19)$$
$$Z=\cos(Y)+\varepsilon(0,\ 0.01) \qquad (20)$$

where $N(0,0.25)$ represents a normal distribution with a mean of 0 and a variance of 0.25. $\varepsilon(0,\ 0.01)$ is white noise with a variance of 0.01 (a normal distribution with a mean of 0 and a variance of 0.01). The BN of this example is shown in Figure 1. The MLME method of PDF estimation is applied, and the resulting MLME-estimated PDF is compared with PDFs estimated from the same dataset using Student's t and Gumbel copulas and the method of kernel.[53,54] Student's t and Gumbel copulas were chosen, as they represent two distinct classes of elliptical and Archimedean copulas, respectively, and the kernel method is a widely used

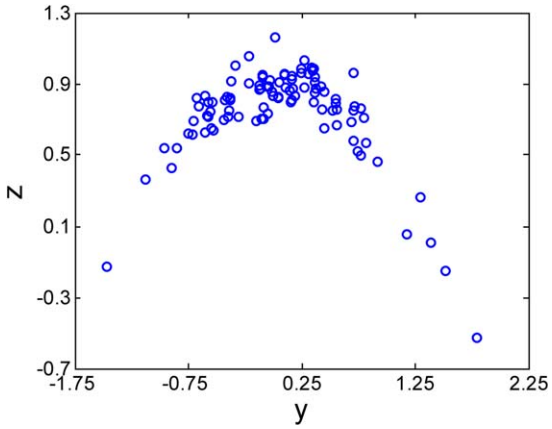

Figure 1. Scatter plot of the 100 (*Y, Z*) samples.

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

nonparametric approach to probability estimation. All of these powerful methods have been extensively used to estimate the behavior of uncertain variables.[53,54]

First, 100 samples of Y are generated followed by 100 samples of Z using Eqs. 19 and 20. Figure 1 shows a scatter plot of the 100 (Y, Z) samples. When the random numbers are discretized into the five intervals (states), Low–Low (LL), Low (L), Normal (N), High (H), and High–High (HH), according to the rule described in the second section, the marginal probabilities given in Table 2 are obtained.

As can be seen in Table 2, none of the samples are within an LL or a HH state. Therefore, when there is an evidence that lies within one of these states, no inference can be made. However, as shown in Figure 2, when there is an evidence that lies within a state other than the LL and HH states, partial inference is possible. Note that this network was constructed using Netica,[55] which does not show states that have zero probability. To be able to conduct complete inference, the MLME PDF estimation method with $O_{\text{opt}}=7$ is used herein to estimate complete PDFs of the Y and Z from the 100 samples. A few low-order moments of the random variable Z, and the combinatorial random variable $\vec{X}=(Y,Z)$ are given in Table 3. Figure 3 shows that as the order of the truncations, O, increases, the maximized logarithm of the likelihood function converges to a higher limit. Figure 3a compares the true $f_Z(z)$ described by Eq. 20 and the $f_Z(z)$ estimated using $O=2$, 4, and 6. Probabilities of the random variables Y and Z being inside the selected states/intervals are calculated using:

Table 2. Marginal Probabilities (Relative Frequencies of the Samples) of *Y* and *Z* Being in the LL, L, N, H, and HH States

	States				
Variable	LL	L	N	H	HH
Y	0.000	0.039	0.932	0.029	0.000
Z	0.000	0.063	0.915	0.022	0.000

$$\hat{P}(y\in s_i)=\int_{s_i}\hat{f}_Y(y)dy \tag{21}$$

$$\hat{P}\left(z\in r_j|y\in s_i\right)=\frac{\int_{r_j}\int_{s_i}\hat{f}_Y(y)\hat{f}_{\vec{X}}(z|y)dydz}{\hat{P}(Y\in s_i)} \tag{22}$$

where s_i and r_j denote the ith state of Z and the jth state of Y, respectively.

Comparison with conventional copulas

Copulas are a class of multivariate probability distribution functions primarily defined for continuous random variables and used to estimate multivariate PDFs.[56,57] They are particularly useful due to the fact that they use a predetermined dependence structure between the random variables, indicating the extent to which random variables are dependent on each other. This dependence structure is reflected in the form of the copula cumulative distribution function (CDF),

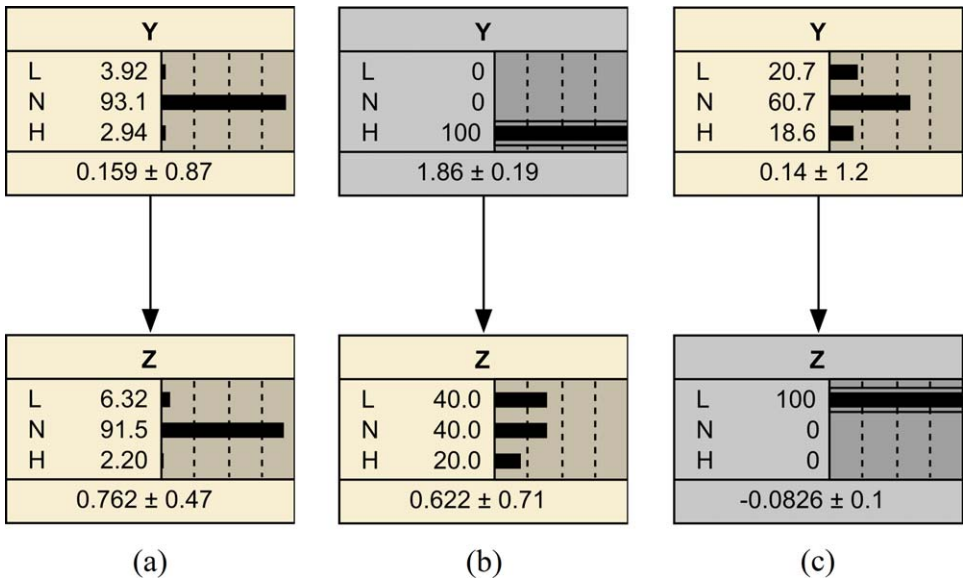

Figure 2. Bivariate BN for *Y* and *Z* trained with 100 samples in Netica.

(a) Normal operation network, (b) Predictive inference (evidence is for *Y*), and (c) Diagnostic inference (evidence is for *Z*). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Table 3. Moments of Z and (Y, Z) in an Increasing Order of Moments, Calculated Using the Data Samples Given in Figure 1

Moment Function	z^0	z	z^2	z^3	z^4	z^5	z^6	z^7	z^8	z^9
Moment value	1.000	0.745	0.621	0.520	0.452	0.397	0.357	0.325	0.301	0.283
Moment Function	y^0z^0	y	z	y^2	yz	z^2	y^3	y^2z	yz^2	z^3
Moment value	1.000	0.062	0.745	0.338	0.010	0.621	0.113	0.126	0.027	0.520

denoted by C, or its equivalent PDF, denoted by c, which are related to each other according to:

$$c(u_1,\ldots,u_d)=\frac{\partial^d C(u_1,\ldots,u_d)}{\partial u_1\ldots\partial u_d} \tag{23}$$

C is actually the probability integral transform of a multivariate PDF; that is, it develops a multivariate CDF over the marginal CDF of individual random variables of interest.

After choosing an appropriate copula, its parameter(s) are adjusted with respect to the available data. This copula is then utilized to estimate a multivariate PDF using:

$$f(x_1,\ldots,x_d)=c(u_1,\ldots,u_d)\prod_{i=1}^{d} f_{X_i}(x_i) \tag{24}$$

where f and f_{X_i} are multivariate and univariate marginal PDFs, respectively, and

$$u_i=\int_{D_{x_i}} f_{X_i}(x_i)dx_i,\qquad i=1,\ldots,d \tag{25}$$

with D_{x_i} being the domain of x_i. There are several families of copulas. The elliptical copulas that are based on well-known multivariate distributions (e.g., Gaussian copula) and Archimedean copulas (e.g., Frank and Gumbel copulas) have been used widely to estimate multivariate probability functions.[58] Despite their many advantages such as low computational complexity and the ability to capture nonlinearity, conventional copulas are only applicable to random variables whose relationships can be described by monotonic functions. This weakness is a result of function parameter(s) of copulas, which are supposed to describe the degree of correlation between random variables based on the covariance of data and its derivatives.[59] Since the covariance between two random variables can only capture monotonic dependence (as in linear or logarithm functions), it cannot describe the true dependence in cases where nonmonotonic dependence exists.

Herein, the joint PDF $f(y,z)$ is estimated from the same dataset using Student's t and Gumbel copulas for the bivariate case:

$$C_{t,v}(u_1,u_2)=\int_{-\infty}^{t_v^{-1}(u_1)}\int_{-\infty}^{t_v^{-1}(u_2)}\frac{1}{2\pi(1-\rho^2)^{\frac{1}{2}}}\left\{1+\frac{y^2-\rho yz+z^2}{v(1-\rho^2)}\right\}^{-(v+2)/2}dydz \tag{26}$$

$$C_{Gu}(u_1,u_2)=\exp\left\{-[(-\ln(u_1))^{\rho}+(-\ln(u_2))^{\rho}]^{\frac{1}{\rho}}\right\} \tag{27}$$

where u_1 and u_2 are defined according to Eq. 25, v is the degree of freedom of the univariate Student's t distribution (t), and ρ is Spearman's rank correlation:

$$\rho=\frac{\mathrm{CoV}(u_1,u_2)}{\sqrt{\mathrm{Var}(u_1)\mathrm{Var}(u_2)}}$$

with $\mathrm{Var}(X)$ denoting the variance of random variable X. The multivariate PDF of the random vector (Y, Z) is then calculated using Eqs. 23 and 24, where the marginal PDFs f_Y and f_Z are obtained using a nonparametric kernel method,[60] also described in the next section. Figures 4a, b, and c compare three PDFs (multivariate MLME estimated PDF, and Student's t and Gumbel copulas estimated PDFs) estimated from the same data, with the data. The circles represent the actual data, and the solid line contours represent the estimated bivariate joint PDF of random variables Y and Z. Figure 4a shows the estimated PDF by the MLME method using a seventh order of truncation. Compared to the actual PDF shown in Figure 4e, the MLME PDF can capture the nonmonotonic behavior of the sine function around $Y=0$.

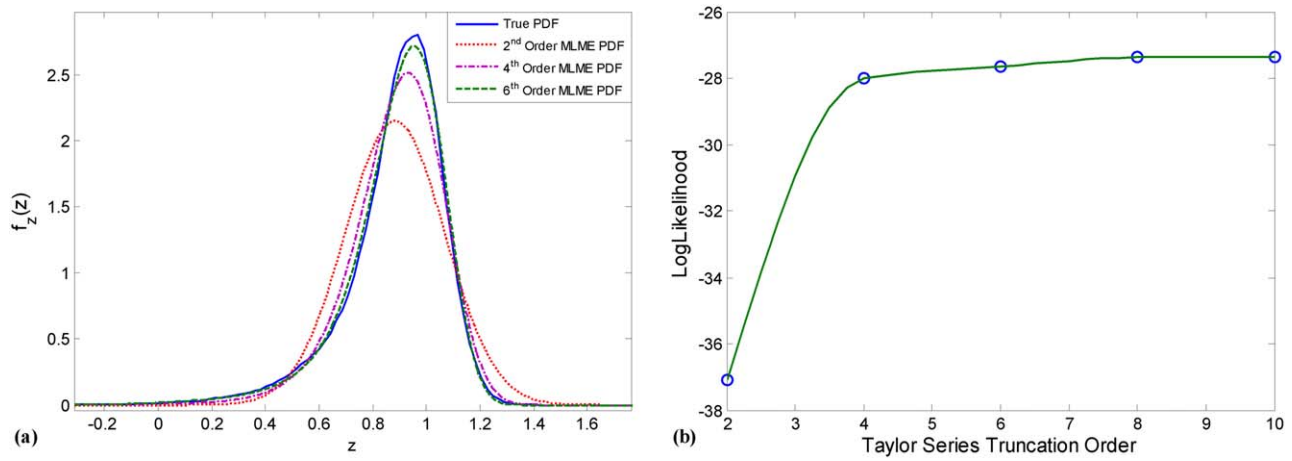

Figure 3. (a) Univariate MLME PDF estimated using different truncation orders and (b) Log-likelihood of the PDF of Z, MLME estimated with different moment orders.

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

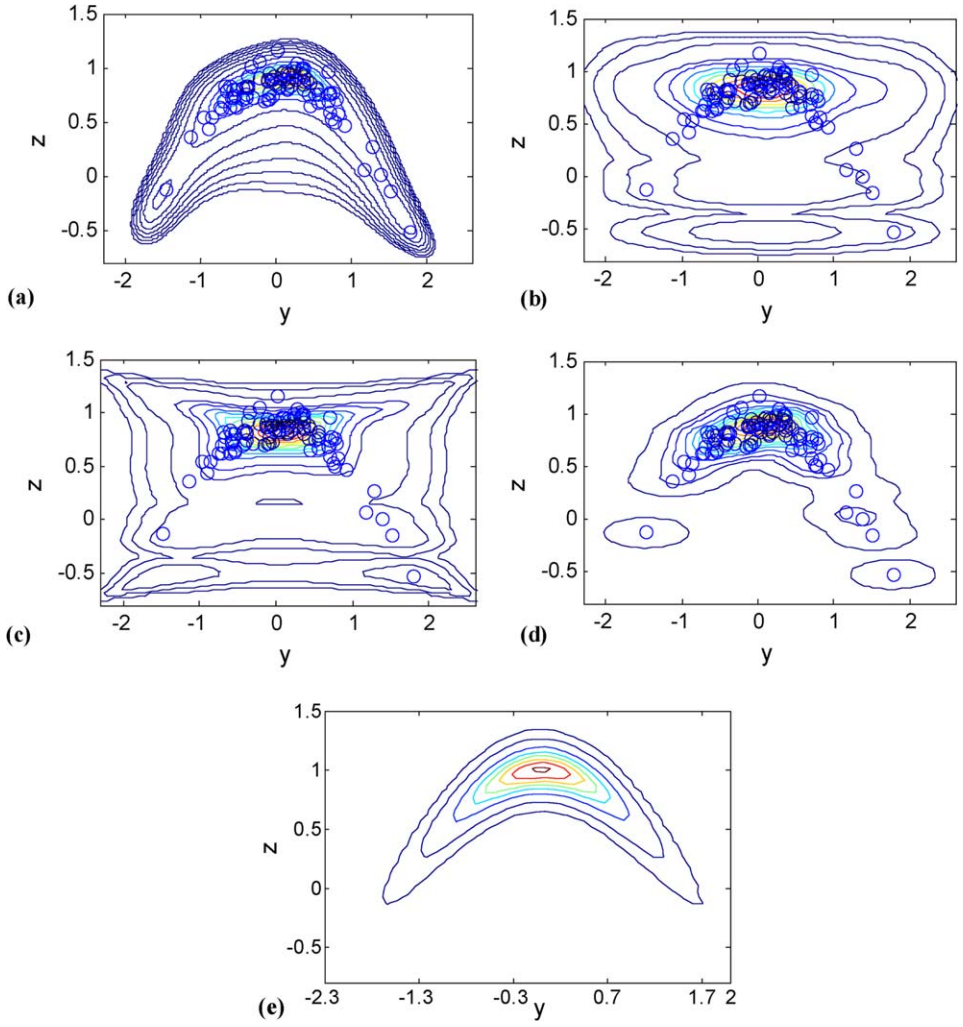

Figure 4. Contour plots of estimated joint PDFs of (*Y, Z*) and samples shown by the small circles.

(a) MLME PDF estimated using a seventh-order truncated Taylor series, (b) PDF estimated using Gumbel copula, (c) PDF estimated using Student's t copula, (d) PDF estimated using nonparametric method of kernel, and (e) True PDF. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

As the copulas use the covariance matrix of Y and Z to capture the correlation between these variables, as can be seen in Figures 4b and c, Student's t and Gumbel copulas fail to predict the actual behavior of the data inside and outside the range of the data. In summary, the copula functions are incapable of providing estimates that agree with the PDF of the actual data.

Comparison with nonparametric Kernel method

Kernel density estimation methods are a subclass of nonparametric density estimation techniques in which a simplified probability distribution called kernel is considered for each sample point. A weighted sum of these kernel functions over the entire sample set is then the kernel density estimator[60]:

$$\hat{f}(\vec{x}|H)=\frac{\sum_{j=1}^{n}K_H\left(\vec{x}-\vec{x}_j\right)}{n}=\frac{\sum_{j=1}^{n}K\left((H)^{-1}\left(\vec{x}-\vec{x}_j\right)\right)}{n.\det\left(H\right)} \quad (28)$$

where $\hat{f}$ is the PDF estimated using a kernel method with a scaled kernel function $K_H(.)$, $K(.)$ is the kernel PDF, and H is called the bandwidth matrix of the kernel function $K_H(.)$. The bandwidth matrix is estimated by minimizing a measure of the error between the sample and estimated PDFs. Examples of such measures are the mean integrated square error or the mean integrated absolute error. Kernel estimators are applicable to both univariate and multivariate problems. In the univariate case, H is a scalar, generally known as a smoothing parameter. Kernel density estimation methods are not considered model-based in the sense that no closed-form model is used to describe the underlying PDF. However, they require kernel models. As when expressing a function in terms of eigen functions, a PDF is expressed in terms of kernels; that is, as a weighted sum of PDFs (kernels), where each sample point is observed in R^d.[61]

In practice, the kernel estimators have shown satisfactory performance and stability for random vectors with low dimensions only. For higher dimensions, however, estimating the optimal bandwidth becomes increasingly complicated. Another shortcoming of the kernel methods is that their rate of convergence with respect to sample size n is lower than that of their counterpart parametric methods ($n^{-\gamma}$ compared to n^{-1} where $0 \leq \gamma < 1$).[62] This means that with small sample sizes it is not possible to remove nonsmoothness caused by individual data points. A large increase in the smoothing parameter may eventually lead to oversmoothness and

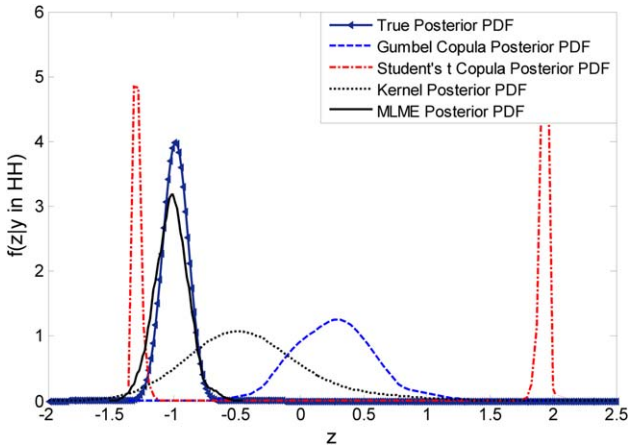

Figure 5. Comparison of posterior conditional PDFs of the random variable *Z* given its parent (*Y*) in its High–High (HH) state, when no data in the state HH provided by the historical dataset.

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

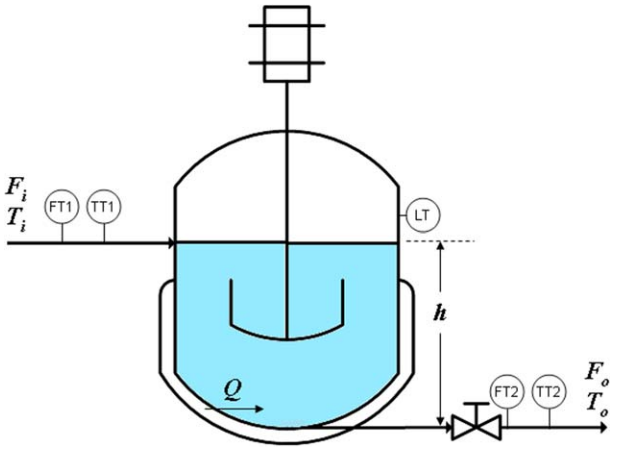

Figure 6. Schematic of the heating tank example (Example 2).

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

valuable information loss about the underlying PDF such as multimodality. As a result, to obtain smaller estimation errors, larger sample sizes should be used, which can lead to a very large analytical expression without a closed form. However, in the case of the MLME estimation method single sample points are not taken into account individually, but their cumulative properties are compacted and exploited in the collective form of moments. Conversely, as described in Eq. 17, the use of larger-size samples (larger n), not only does not decelerate the probability estimation, but also gives the likelihood function a sharper peak, allowing the maximum to be calculated with fewer iterations. However, it should be noted that increasing the degree of connectivity of nodes (not necessarily the network size) affects the parameter estimation step by increasing the number of parameters needed as coefficients of the multivariate polynomial moment functions defined in the previous section.

Kernel density estimators also have the same disadvantage that copula methods have; their constant parameter matrix for the multivariate PDF estimation cannot capture nonmonotone behavior in historical data, resulting in the estimation of PDFs that describe uncorrelated random variables. Furthermore, in kernel methods, even though the bandwidths are calculated to obtain the PDF with minimum error inside the region where samples are taken, the predictions made by kernel methods outside the observed zone are unreliable, unless the variables are monotonically related. Therefore, unlike the MLME method, the kernel methods do not introduce a general solution to the rare-event probability estimation problem.

To estimate the bivariate PDF of Y and Z, bivariate Gaussian kernels with a smoothing parameter equal to the square root of the data-based covariance matrix of the random numbers are used herein. As can be seen in Figures 4d, the nonparametric kernel method also fails to predict the actual behavior of the data outside the range of the data. However, since the kernel method uses an averaging algorithm to estimate probability values, its predictions are reliable locally within the range of the data.

Figure 5 compares the posterior conditional PDF of variable Z given Y observed in its HH state, estimated by the MLME method, the Student's t and Gumbel copula methods, and the kernel method. As can be seen, the only reliable estimation is that of the MLME method. As mentioned earlier, due to the nonmonotonic dependence of Z on Y, covariance-based approaches are unable to capture the actual relation hidden in the data. This inability increases in regions distant from the mean of the sampled population.

Example 2: A process example

Consider the stirred heating tank shown in Figure 6. A steady-state first-principles mathematical model of the process is:

$$\rho(F_i - F_o) = 0 \tag{29}$$

$$\rho C_P(F_i(T_i - T_r) - F_o(T_o - T_r)) + Q + \varepsilon_1 = 0 \tag{30}$$

$$F_{\text{out}} = \frac{h^{1/2}}{R} + \varepsilon_2 \tag{31}$$

PDFs of the root nodes (independent variables) of this process and the two noise signals are given in Table 4. This first-principles model is used to extract the causal relations among the variables to construct a BN, to generate a normal operation dataset, which plays the role of historical dataset in this example, and finally to describe the actual behavior of the process to be compared with the behavior predicted by the estimated MLME PDFs.

PDFs of the independent variables and white noise signals are chosen such that the random samples fall entirely in their normal operation states. The reason behind this selection is to replicate the situation where the information available in the historical data includes no faulty operation records. This allows determination of whether the MLME PDF yields correct predictions when no abnormal-condition data is present.

Table 4. Probability Distributions of Root Nodes (Variables) and Noise Signals in the Heating Tank Example

Variable or Noise	Distribution
F_i	Normal (0.01, 10^{-6})
T_i	Normal (25, 1)
Q	Normal (10^6, 10^5)
ε_1	Normal (0, 4×10^{-8})
ε_2	Normal (0, 0.25)

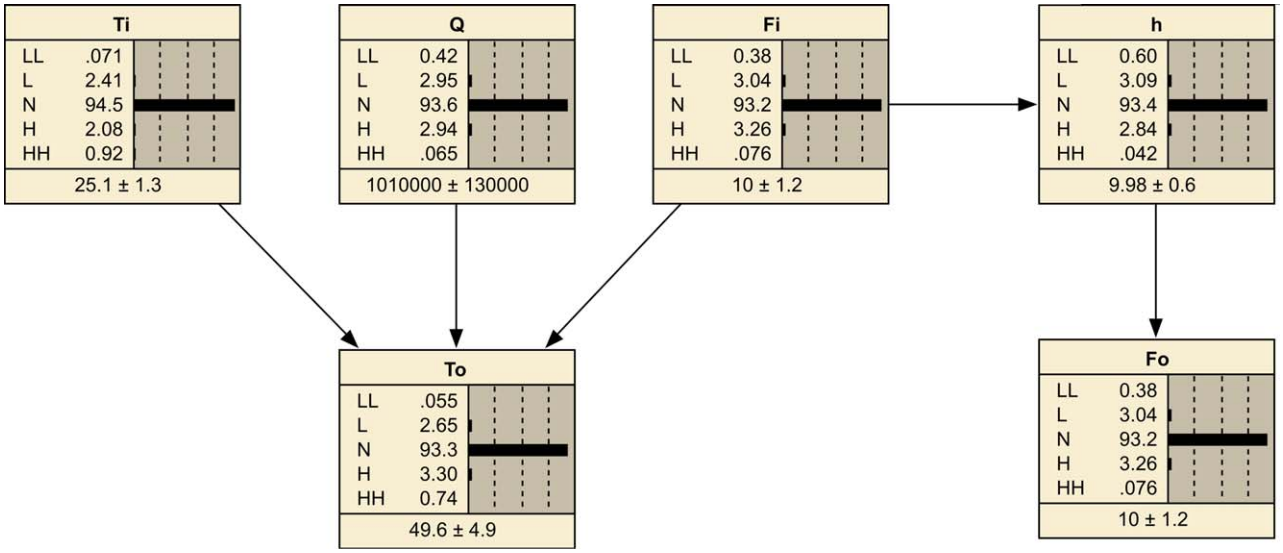

Figure 7. BN of Example 2 trained using complete PDFs estimated using the MLME method to cover the extreme states, LL and HH, as well.

The shown probabilities are normal operation probabilities. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Figure 7 shows the BN representing the system's normal operation data. As in the bivariate Example 1, each observed region is split into three state; Low (L), Normal (N), and High (H). Using the MLME method, the states for each variable can be extended to a level satisfying our design needs by adding the Low–Low (LL) and High–High (HH) states. All states are defined according to the rule stated in Section 2. Inference is conducted using the Netica software of Norsys Corp.[55]

After estimating complete joint and conditional PDFs using the MLME method, inference (from evidence) can be conducted using the BN. Once the network is provided with evidence; that is, probability distribution(s) of evidence node(s) are set according to the evidence, the probabilities of all other nodes are updated. These updated probabilities are indeed posterior probabilities. Two types of studies can then be conducted. If one is interested in how the evidence has altered the probability distributions of the nodes/variables that are affected by the evidence node(s)/variable(s) in the BN, the inference is called a "predictive" inference. Conversely, if one is interested in how the evidence has altered the probability distributions of the nodes/variables that affect the evidence node(s)/variable(s) in the BN, the inference is called a "diagnostic" inference. The diagnostic inference can be used for fault detection.

To quantify the difference between the posterior and prior probabilities of each variable, a useful measure is the relative Kullback–Liebler divergence[63] that is applicable to both continuous and discrete random variables and to individual probability values as well. For a node X_j, the RKLD is defined as:

$$\mathrm{RKLD}_{X_j} = \frac{\mathrm{KLD}_{X_j}}{\sum_{j=1}^{w} \mathrm{KLD}_{X_j}} \tag{32}$$

where

$$\mathrm{KLD}_{X_j} = \sum_{i=1}^{r} P_{X_{j,i}} \log\left(\frac{P_{X_{j,i}}}{Q_{X_{j,i}}}\right) \tag{33}$$

where $P_{X_{j,i}}$ and $Q_{X_{j,i}}$ are the prior and posterior probabilities of the ith state of node X_j with r states, respectively. w is the number of nodes of the network under consideration. KLD can be viewed as the expected value of $\log\left(\frac{P_{X_j}}{Q_{X_j}}\right)$ with respect to the prior probability P_{X_j}. If the prior probability of a state is zero, its corresponding term in the KLD expression is zero, since $0 \times \log(0) = 0$.

Forward inference (Prediction)

In the context of predictive inference, the variable with the highest RKLD value is the variable mostly affected by the applied change (evidence). Hence, this index can be used to perform risk assessment. Outcome of such an analysis together with the costs of the associated abnormal events can be used to quantify risks. Such an analysis can be implemented offline and online. Offline predictive Bayesian inference is a powerful tool for risk assessment and risk scenario development, as it provides valuable information about most probable consequences of changes applied to the system and can be utilized to detect or remove risky features from processing plants. Online (real-time) predictive Bayesian inference can provide important information about the consequences of an observed evidence. This information can be used immediately to take a series of preventing actions leading to loss reduction.

Figure 8a shows the BN of Example 2 with updated (posterior) probabilities when the inlet flow is at its HH state, and Figure 8b shows the corresponding RKLD values of the nodes. The RKLD values indicate that when the inlet flow moves to its HH state, its most severe effect is on the water level, h, with a probability of more than 50% being in the HH state.

Backward (diagnostics) inference: Fault detection

The RKLD values can also be used to identify: (a) the most-likely cause root variable/node whose change has led to the observed evidence fed to the network, and (b) the most likely state of the most-likely cause root node. After identifying the most-likely cause root node for the evidence (root node with the highest RKLD value), for each state i of

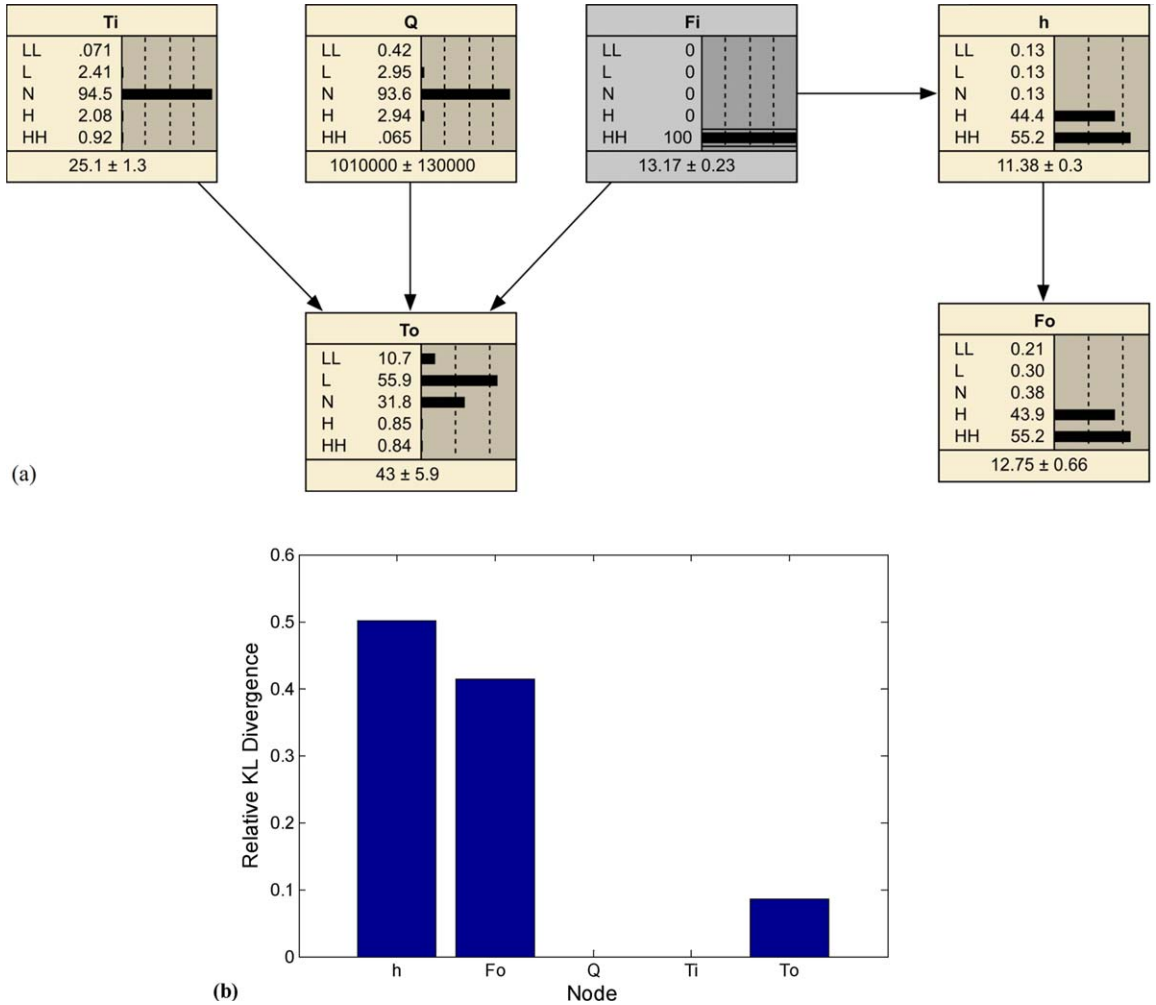

Figure 8. (a) BN of Example 2 showing updated (posterior) probabilities when evidence F_i in HH was given to the network and (b) RKLD values of the five nodes.

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

the most-likely cause root node X_{mlc} the difference between the posterior and prior probability of the state i is calculated:

$$d_{X_{\text{mlc},i}} = Q_{X_{\text{mlc},i}} - P_{X_{\text{mlc},i}}, \quad i = 1, \ldots, m \tag{34}$$

where $d_{X_{\text{mlc},i}}$, $Q_{X_{\text{mlc},i}}$, and $P_{X_{\text{mlc},i}}$ denote the deviation index, and the posterior and prior probabilities of state i of the most likely cause node for the observed evidence. As implied by the definition, a positive value of the deviation index indicates an increase in the probability of state i; larger values indicate greater contributions of the abnormal event to the state.

Figure 9a depicts the BN of Example 2. The probabilities given in this figure are updated (posterior) probabilities corresponding to the evidence that T_o is in the state L. The corresponding calculated RKLD values shown in Figure 9b indicate that the most-likely cause root node is Q. Figure 9c showing the differences between posterior and prior probabilities of the states of Q points to the state of N of the root node Q having the largest prior-to-posterior probability change. An interesting implication of constructing a BN model from the historical data can be seen in this example. Although the inlet temperature, the rate of heat transfer to the tank, Q, and the inlet flow rate, F_i, all affect the outlet temperature, T_o, but do not have equal contributions to the changes observed in the outlet temperature. As Figure 9a shows, given the evidence of T_o being in the LL state, change in the Q's probability distribution is higher than the changes in the two other parents of T_o. Therefore, backward Bayesian inference identified a change in Q as the most probable cause of T_o being in the LL state. Similar arguments can be made to find the most deviated state from Figure 9c. Figure 10 compares the posterior probabilities of the parents of the node T_o given T_o in its LL state and calculated using two different historical datasets for calculating the parameters of the network. The blue bars represents posterior probabilities calculated by a network trained by one million samples drawn out of the system's governing equations, while the green bars represents posterior probabilities calculated by a network trained using the MLME completed conditional probabilities. This figure clearly reveals the high reliability of the MLME PDF estimation method for use in probabilistic inference.

Conclusions

The problem of rare-event probability estimation was studied. A moment-constrained, ML, maximum-entropy method of multivariate PDF estimation was proposed. This method

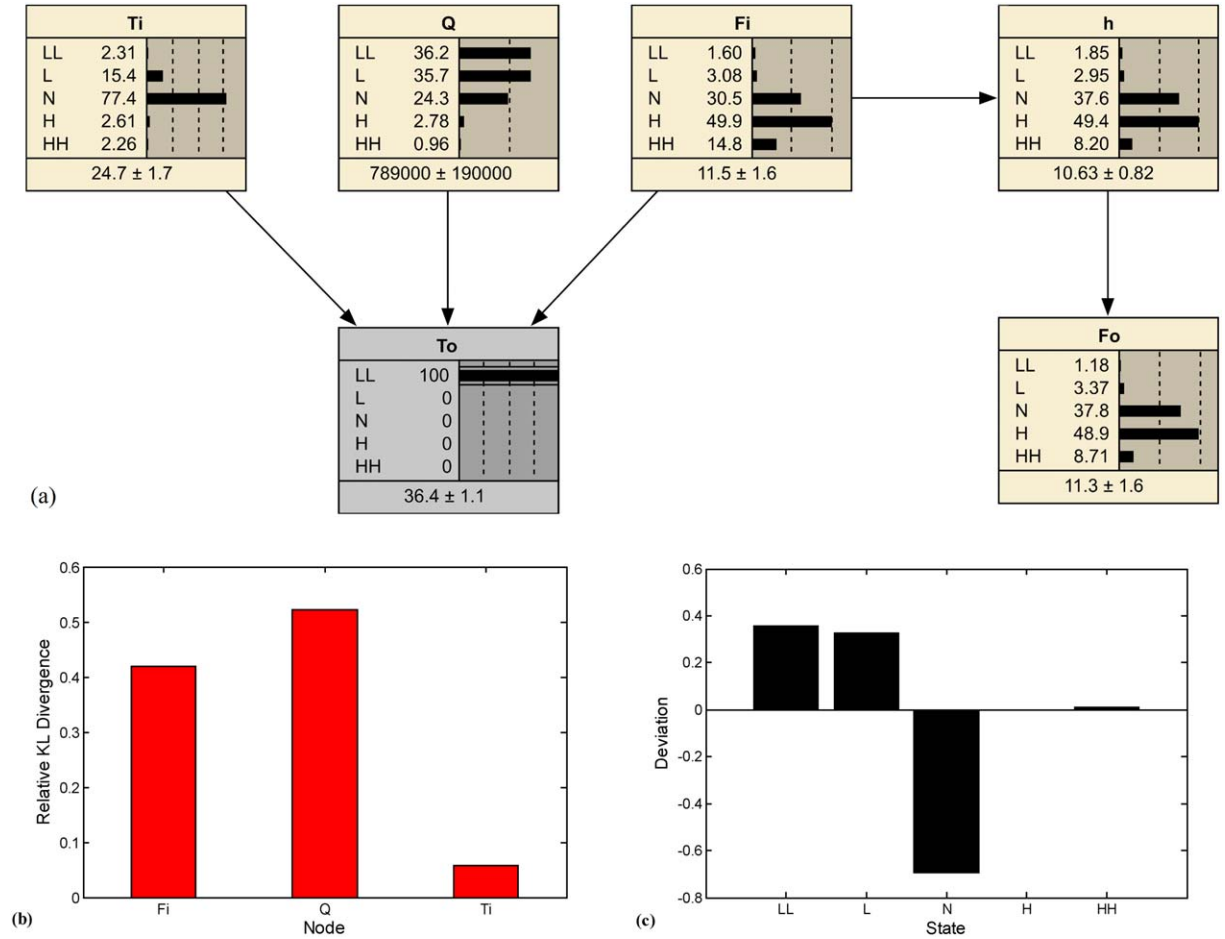

Figure 9. (a) BN of Example 2 showing updated (posterior) probabilities when evidence T_o in LL was given to the network, (b) RKLD values of the three root nodes, and (c) Differences between posterior and prior probabilities of the most-likely cause root node, Q.

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

is superior to other widely used approaches such as copula densities and nonparametric kernel methods because it applies when relations among the variables are nonmonotonic. Copula and kernel estimators, despite their power in capturing highly nonlinear behavior, predict poorly in regions where no data have been observed. Another advantage of the MLME method is its capability in replicating the complex behavior of probability densities in a natural way

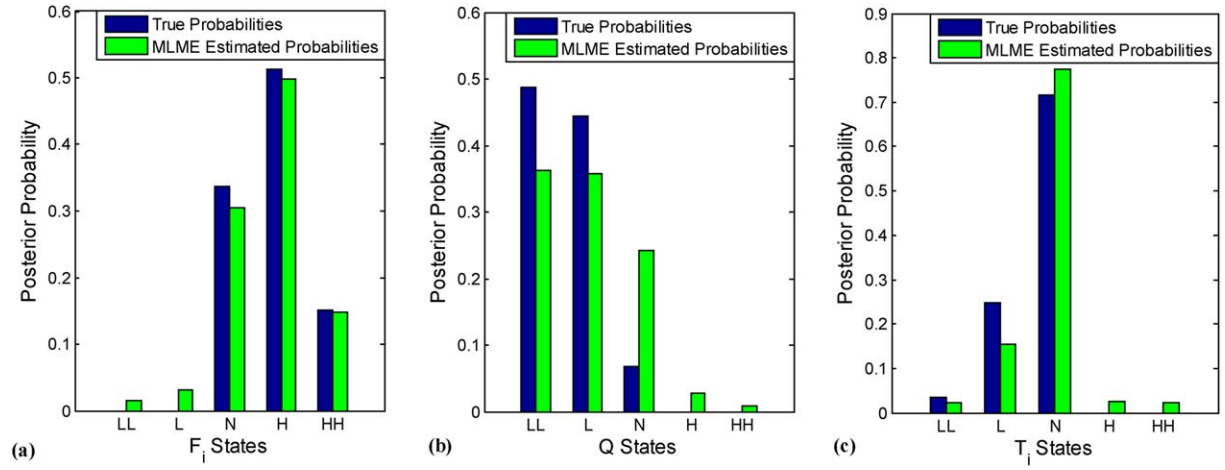

Figure 10. Diagnostic Bayesian inference with 1,000,000 samples and with the MLME estimated network.

(a) Posterior probability distribution of F_i, (b) posterior probability distribution of Q, and (c) posterior probability distribution of T_i. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

using moments introduced by the sampled population. The MLME PDFs are highly interpretable in terms of their closed-form formulas using the statistical properties of the data itself (skewness, peakness, etc.). Moreover, since PDFs estimated by the MLME method belong to the class of parametric PDFs, the convergence rate of the method is higher than other nonparametric PDF estimation methods.[62] To take advantage of the likelihood function as a goodness-of-fit measure, a method of selecting the moment functions was presented. Finally, unlike nonparametric methods, the computational load of the parameter estimation step of the MLME method is not affected negatively by the number of samples being processed—primarily because MLME PDFs use cumulative characteristics of data in moment values rather than individual data points. Larger sample sizes yield steeper peaks for the likelihood function, which lead to computationally faster optimizations.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grants No. CBET-1066461 and CBET-1066475. T.M. was supported by National Science Foundation Grant 1066461. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The authors would like to thank Amogh V. Prabhu, Darrin Feather, Benjamin J. Jurcik, and Brian M. Besancon for their invaluable comments on this work.

Literature Cited

1. Qin SJ. Statistical process monitoring: basics and beyond. *J Chemom*. 2003;17:480–502.
2. Mahadevan S, Shah SL. Fault detection and diagnosis in process data using one-class support vector machines. *J Process Control*. 2009;19:1627–1639.
3. Chiang LH, Braatz RD. Process monitoring using causal map and multivariate statistics: fault detection and identification. *Chemom Intell Lab Syst*. 2003;65:159–178.
4. Castillo I, Edgar TF, Dunia R. Nonlinear model-based fault detection with fuzzy set fault isolation. In; IECON 2010–36th Annual Conference on IEEE Industrial Electronics Society. Glendale, AZ, 2010: 174–179.
5. Mhaskar P, McFall C, Gani A, Christofides PD, Davis JF. Isolation and handling of actuator faults in nonlinear systems. *Automatica*. 2008;44:53–62.
6. Pariyani A, Seider WD, Oktem UG, Soroush M. Improving process safety and product quality using large databases. In: Pierucci S, Buzzi Ferraris G, editors. Computer Aided Chemical Engineering. Naples, Italy: Elsevier, 2010;28:175–180.
7. Zhang Y, Qin SJ. Improved nonlinear fault detection technique and statistical analysis. *AIChE J*. 2008;54(12):3207–3220.
8. Mehranbod N, Soroush M, Panjapornpon C. A method of sensor fault detection and identification. *J Process Control*. 2005;15:321–339.
9. Mhaskar P, Gani A, McFall C, Christofides PD, Davis JF. Fault-tolerant control of nonlinear process systems subject to sensor faults. *AIChE J*. 2007;53:654–668.
10. Ohran B, Muñoz de la Peña D, Davis JF, Christofides PD. Enhancing data-based fault isolation through nonlinear control. *AIChE J*. 2008;54:223–241.
11. Taleb NN. *The Black Swan: The Impact of the Highly Improbable, 2nd ed*. New York: Random House Trade Paperbacks, 2010.
12. Härdle W, Müller M, Sperlich S, Werwatz A. *Nonparametric and Semiparametric Models*. New York: Springer, 2004.
13. Good IJ. *The Estimation of Probabilities: An Essay on Modern Bayesian Methods (Research Monograph)*. Cambridge: The MIT Press, 2003.
14. McLachlan GJ, Peel D. *Finite Mixture Models*. New York: Wiley-Interscience, Inc., 2000.
15. Juneja S, Shahabuddin P. Rare-Event simulation techniques: an introduction and recent advances. In: Henderson SG, Nelson BL, editors. *Handbooks in Operations Research and Management Science*. The Netherlands: Elsevier, 2006;13:291–350.
16. Berryman JT, Schilling T. Sampling rare events in nonequilibrium and nonstationary systems. *J Chem Phys*. 2010;133:244101.
17. Hiemstra C, Kelejian HH. A rare events model: Monte Carlo results on sample design and large sample guidance. *Econ Lett*. 1991;37: 255–263.
18. Doucet A, de Freitas N, Gordon N. *Sequential Monte Carlo Methods in Practice*. New York: Springer, 2001.
19. Asmussen S, Kroese DP, Rubinstein RY. Heavy tails, importance sampling and cross entropy. *Stochastic Models*. 2005;21:57–76.
20. Uneja S. Efficient rare-event simulation using importance sampling: an introduction. In Misra JC, editor. *Computational Mathematics, Modeling and Algorithms*. New Delhi: Narosa Publishing House, 2003:357–396.
21. Ahamed TPI, Borkar VS, Juneja S. Adaptive importance sampling technique for Markov chains using stochastic approximation. *Oper Res*. 2006;54:489–504.
22. Dean T, Dupuis P. Splitting for rare event simulation: a large deviation approach to design and analysis. *Stochastic Processes Appl*. 2009;119:562–587.
23. Shortle JF. Efficient simulation of blackout probabilities using splitting. *Int J Electr Power Energy Syst*. 2013;44:743–751.
24. Campillo F, Rakotozafy R, Rossi V. Parallel and interacting Markov chain Monte Carlo algorithm. *Math Comput Simul*. 2009;79:3424–3433.
25. Kaynar B, Ridder A. The cross-entropy method with patching for rare-event simulation of large Markov chains. *Eur J Oper Res*. 2010; 207:1380–1397.
26. Cerou F, LeGland F, Del Moral P, Lezaud P. Limit theorems for the multilevel splitting algorithm in the simulation of rare events. In: Proceedings of the 2005 Winter Simulation Conference. Orlando, FL, 2005:682–691.
27. Qiu Y, Zhou H, Wu Y. An importance sampling method with applications to rare event probability. In: Proceedings of 2007 IEEE International Conference on Grey Systems and Intelligent Services. Nanjing, China, 2007:1381–1385.
28. Devroye L, Gábor L. *Combinatorial Methods in Density Estimation*. New York: Springer, 2001.
29. Tsybakov AB. *Introduction to Nonparametric Estimation*. New York: Springer, 2009.
30. Heylighen F, Joslyn C. Cybernetics and second order cybernetics. In: Meyers RA, editor. *Encyclopedia of Physical Science and Technology, 3rd ed*. New York: Academic Press, 2001:155–170.
31. Aldrich JRA. Fisher and the making of maximum likelihood 1912–1922. *Stat Sci*. 1997;12:162–176.
32. Lauritzen L, Spiegelhalter J. Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *J R Stat Soc*. 1988;50:157–224.
33. Koski T, Noble JM. *Bayesian Networks: An Introduction*. Hoboken: Wiley, Inc., 2009.
34. Korb KB, Nicholson AE. *Bayesian Artificial Intelligence*, 2nd ed. Boca Raton: CRC Press, 2010.
35. Daly R, Qiang S, Stuart A. Learning Bayesian networks: approaches and issues. *Knowl Eng Rev*. 2011;26:99–157.
36. Casella G, Berger RL. *Statistical Inference*, 2nd ed. Australia: Cengage Learning, 2002.
37. Gillies D. *Philosophical Theories of Probability*. New York: Routledge, 2000.
38. Beesack PR. Inequalities for absolute moments of a distribution: from Laplace to von Mises. *J Math Anal Appl*. 1984;98:435–457.
39. Koralov L, Sinai YG. *Theory of Probability and Random Processes*, 2nd ed. Berlin: Springer, 2012.
40. Shannon CE. Prediction and entropy of printed English. *Bell Syst Tech J*. 1951;30:50–64.
41. Shannon CE. A mathematical theory of communication. *Bell Syst Tech J*. 1984;27:379–423.
42. Zellner A, Highfield AR. Calculation of maximum entropy distribution and approximation of marginal posterior distributions. *J Econometrics*. 1988;37:195–209.
43. Golan A, Judge G, Miller D. *Maximum Entropy Econometrics Robust Estimation with Limited Data*. New York: Wiley, Inc., 1996.

44. Lindsay BG. Moment matrices: applications in mixtures. *Ann Stat*. 1989;17:722–740.
45. Hall AR. *Generalized Method of Moments*. Oxford: Oxford University Press, 2005.
46. Carrasco M, Florens JP. Generalization of GMM to a Continuum of Moment Conditions. *Econometric Theory*. 2000;20:797–834.
47. Zacks S, Even M. Minimum variance unbiased and maximum likelihood estimators of reliability functions for systems in series and in parallel. *J Am Stat Assoc*. 1966;61:1052–1062.
48. Schmidt DF, Makalic E. Universal models for the exponential distribution. *IEEE Trans Inf Theory*. 2009;55:3087–3090.
49. Chris DO, Paul AR. On the uniqueness of the maximum likelihood estimator. *Econ Lett*. 2002;75:209–217.
50. Moons KGM, Rogier A, Donders T, Steyerberg EW, Harrell FE. Penalized maximum likelihood estimation to directly adjust diagnostic and prognostic prediction models for overoptimism: a clinical example. *J Clin Epidemiol*. 2004;57:1262–1270.
51. Benedikt M, Pötscher HL. On the distribution of penalized maximum likelihood estimators: the LASSO, SCAD, and thresholding. *J Multivariate Anal*. 2009;100:2065–2082.
52. Eliason SR. *Maximum Likelihood Estimation: Logic and Practice*. Newbury Park: SAGE Publications, Inc., 1993.
53. Piotr J, Durante F, Härdle WK, Rychlik T. Copula theory and its applications. In: Proceedings of the Workshop Held in Warsaw 2009. Heidelberg: Springer, 2010.
54. Li Q, Racine JS. Advances In Econometrics: Nonparametric Econometric Methods, Volume 25. Bingley, UK: Emerald, 2009.
55. Netica v. 4.16 [Software], Norsys Software Corporation. Vancouver, BC. 2010. Available at http://www.norsys.com.
56. Nelsen RB. An Introduction to Copulas. Vol. 139. *Lecture Notes in Statistics*. Berlin Heidelberg New York:Springer-Verlag, 1999.
57. Embrechts P, Lindskog F, McNeil A. Modelling dependence with copulas and applications to risk management. In: Rachev S, editor. *Handbook of Heavy Tailed Distributions in Finance*. New York: Elsevier, 2003: 331–385.
58. McNeil AJ, Nešlehová J. From Archimedean to Liouville copulas. *J Multivariate Anal*. 2010;101:1772–1790.
59. Kolev N, Paiva D. Copula-based regression models: a survey. *J Stat Plann Inference*. 2009;139:3847–3856.
60. Silverman BW. *Density Estimation for Statistics and Data Analysis*. London: Chapman & Hall/CRC, 1998.
61. Epanechnikov VA. Non-parametric estimation of a multivariate probability density. *Theory Probability Appl*. 1969;14:153–158.
62. Duong T, Hazelton ML. Convergence rates for unconstrained bandwidth matrix selectors in multivariate kernel density estimation. *J Multivariate Anal*. 2005;93:417–433.
63. Kullback S. Letter to the Editor: The Kullback–Leibler distance. *Am Stat*. 1987;41:340–341.

Manuscript received Aug. 14, 2013, and revision received Dec. 9, 2013.