Probabilistic Modeling of Process Systems with Application to Risk Assessment and Fault Detection

A Thesis

Submitted to the Faculty

of

Drexel University

by

Taha Mohseni Ahooyi

in Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

December 2015

Dedications

To my beloved Farzaneh, my loving parents and family for their endless support, love and patience

And in memory of my grandparents

Acknowledgments

First, I would like to specifically extend my deepest appreciations to my PhD advisor Prof. Masoud Soroush for his invaluable contribution, guidance and support during the course of my doctoral research. Doubtlessly, accomplishment of this research would not have been possible without his supervision and patience. Prof. Soroush not only provided me with the resources and scientific inspiration and support essential to conduct my PhD research, but also I learned important life lessons from him.

I would also like to sincerely thank my committee members, Prof. Warren Seider, Prof. Giuseppe Palmese, Prof. Kenneth Lau and Dr. Jeffrey Arbogast for their willingness to be on my PhD committee and for their valuable encouragement, support and advice throughout my PhD studies.

I am also deeply grateful to all our great collaborates, Dr. Ulku Oktem, Dr. Ankur Pariyani and Ian Moskowitz as well as my wonderful friends and lab fellows Mona Bavarian, Siamak Nejati, Nazanin Maghaddam, Yuriy Smolin and many others at Drexel University and University of Pennsylvania.

I would also like to express my sincerest gratitude to my loving wife, parents and family for their unconditional love, support, encouragement and patience during my PhD research, for whom I would not have made it this far without.

Finally I would like to give special thanks to the funding resources which allowed me complete this research. This research was funded by the National Science Foundation (NSF grants CBET-1066461 and CBET-1066475).

Table of Contents

List of Tables

List of Figures

Abstract

Three new methods of joint probability estimation (modeling), a maximum-likelihood maximum-entropy method, a constrained maximum-entropy method, and a copula-based method called the rolling pin (RP) method, were developed. Compared to many existing probabilistic modeling methods such as Bayesian networks and copulas, the developed methods yield models that have better performance in terms of flexibility, interpretability and computational tractability. These methods can be used readily to model process systems and perform risk analysis and fault detection at steady state conditions, and can be coupled with appropriate mathematical tools to develop dynamic probabilistic models. Also, a method of performing probabilistic inference using RP-estimated joint probability distributions was introduced; this method is superior to Bayesian networks in several aspects. The RP method was also applied successfully to identify regression models that have high level of flexibility and are appealing in terms of computational costs.

Chapter 1: Background

1.1. Introduction

Successful management of industrial processes needs adequate information and wise judgment. Today, the necessity of evaluating frequencies and consequences of hazardous accidents is becoming one the most attractive fields in the process engineering, along which legislators are placing an increasing stress on the control of the strength of risky events. Various process risk analysis techniques have been developed over the last decade to equip decision makers with tools to estimate the impacts of undesired events on the personnel, economic matters, society and the environment. The availability of technical background and information resources to perform the analysis is the primary constraint on the completeness of risk assessment. Managers must consider the value of risk analysis results in their decision making to reduce the intensity of probable accidents.

1.2. Stochastic Modeling of Operational Risks

Operational risks are those events imposing a loss to an operating system mainly due to failures in internal process or anomalies applied by external environment. These failures are usually a result of gradual depraving processes, finally leading to an intolerance point, beyond which the system cannot continue its routine function. Although in most cases these risks give rise to small to medium scale losses, but there is always a potential danger of a single faulty operation, through certain chain or cascade interactions, undergoes a “snowball” effect eventually resulting in an irrecoverable catastrophic event or calamity.[1]

Since real-world systems, especially those with many variables or components, bear a large degree of uncertainty in them, as far as the observer's "epistemic" knowledge about the system could reach, traditional deterministic models fail to depict the system and the manner it really behaves properly. Deterministic models, for many reasons, are only good approximations sufficient to describe systems under certain simplified conditions with point estimates, i.e. without providing information on how uncertain the result is. The first of such reasons is that deterministic models which are constructed on physical laws are supposedly reflecting the same variables which are considered noteworthy from the viewpoint of the scientist or engineer. That gives rise to a model lacking many sources of information disregarded unintentionally or deliberately for the sake of feasibility of computations.[2] Furthermore, in addition to variables omitted from the model as mentioned above, there are variables that can barely be taken into account in a deterministic model and are actually almost uncontrollable. These variables, introducing an "aleatory" uncertainty to the model, are also known as noises or disturbances. Finally, we have to rely on sensors which are intermediates between us and measurable quantities, providing us with the only immediately discernible information from the reality. Sensors, on the other side, carry uncertainties in terms of their bias and variance, making the corresponding deterministic model parameters and outcomes less trustworthy and dignified than they have displayed so far.[3-5]

All these facts suggest that in order to model such a complex entity as risk represented by a multivariate system, there should be introduced a comprehensive stochastic framework to allow for different sources of uncertainty being incorporated and employed to generate more reality-consistent results. The rest of this document is dedicated to proposing such a model and introducing its features.

1.3. Rare Event Probability Estimation

Risk assessment is usually referred to as a set of actions implemented to evaluate risk distribution over the components of a system.[6] Resulting analysis then will be used as an input to risk management strategies utilized to mitigate or remove risks to which a complex system is exposed. This process is directly connected to estimate the likelihood of different possible risky situation scenarios that could happen for the system and their associated costs. To this end, risk assessment must overcome multiple obstacles simultaneously, many of them have received much more attention within the past decades.[7]

Most of the efforts done to estimate risks have focused on those abnormal situations with higher probabilities and moderate costs, whereas the major part of catastrophic and large scale incidents imposing highly destructive consequences to a system are caused by some triggering events whose probabilities have been considered infinitesimal when performing the risk assessment procedure. This class of abnormal events are usually referred to as “rare events” and categorized into two major groups: those which are such rare and far-fetched that their probabilities may be considered to be practically zero[8]; e.g. industrial plant destruction due to a meteor colliding exactly with the plant site, and those that are actually predictable, but showing a minor recurrence frequency compared to the system’s expected lifetime; e.g. control system failure. Throughout this text we use the term “rare events” for the latter type.[9]

Although the second type of events mentioned above is predictable and its consequence can be well avoided, even modern day industrial establishments are still suffering from the resulting catastrophes.[10-12] Two major reasons underlie such disastrous consequences. First and apparently main cause is that the probability of those so called rare events is usually underestimated intentionally or inadvertently. That misunderstanding eventually leads to a common thought that corresponding risk values are negligible as well. Therefore, according to such a perception the associated risks simply taken not very seriously and specified with minimum degree of precautionary schemes. Another major reason, on the other hand, is the fact that estimating the probabilities of events that have seldom or never happened, observed or recorded in the course of system's operation carries a great deal of uncertainty in its outcome, mainly because a general framework integrating between the rarity of sample realizations and their generalizability to the future has not been introduced yet and as a result point estimates presented by the mentioned methods are hardly applicable to an actual operating system. Hence the problem is to estimate probability values that are unknown, infinitesimally small and hard to predict, which in most cases tend to be ignored.

Timely performing a thorough risk assessment procedure is literally crucial to prevent any future large-impact incident, making it not only lifesaving but only a profitable task.

As a key element of risk assessment, rare event probability estimation can be applied in two different phases of a complex system undergoing development. Sometimes, particularly when we are dealing with small to average scale settings not including intricate interactions, it is more convenient to evaluate risks in the design or pre-production stages. However, this is not the preferred solution in cases where the system is supposed to involve a large number of interconnected field variables with probably hidden or unknown ones and subjected to stochasticity highly incorporated in forming the system's behavioral nature. So, oftentimes risk assessment is implemented over the systems that are already existing and working.[13,14] This is mainly because the signicance of risk assessment has not been known to many until recently, or the mathematical tools required as an infrastructure to the modeling step were not available. More importantly, computational power and software backbone needed for huge numerical calculations of complicated mathematical models have become widely accessible only in the past couple of decades. Another reason, from which our proposed research receives a considerable incentive, is the fact that the best way of characterizing a system's future behavior is to construct the predicting models based upon information from the recorded historical past.

After a historical database reflecting the previous trends of the system becomes available, some further important steps should be taken toward the complete risk determination.[15]

To reconstruct a full profile of the likelihood of any specific variable taking a predetermined value or occupying some certain state, there should exist a model to get trained by the available dataset. This model can either stem from the fundamental physicochemical rules governing the system's behavior and derived by differential conservation laws, or get developed based upon techniques suggested by statistical analysis. It is claimed here that for some reasons, in the context of risk assessment statistical models are superior to any other class of models. Firstly, they can be established in a much shorter time than a regular systematic formulating of astronomically large number of conservation equations could take, probably without requiring much knowledge about the underlying actual mechanisms. Moreover, they are capable of incorporating uncertainties appearing in the data with minimum level of artificial assumptions. Finally, inference over different fault scenarios is performed instantly compared to that of large systems of coupled equations describing a system using differential conservation laws.

Many traditional statistical learning approaches, on the other hand, are susceptible to the quality of the data provided: large historical datasets may contain information differing from the current state of the system. This situation happens since large scale and complex systems are continuously subject to changes. For example, replacing a process component with an upgraded one can render the failure profile of the older piece partly useless. On the other side small historical datasets may not contain adequate information required to accurately estimate probabilities.[16]

According to the above facts, statistical risk models are specifically attractive in addressing rare probability estimation when a reliable first-principles model is not available, or creating such a model is not feasible at least with limited budget or time. Even though a reliable first-principle model is attainable, picking up an adequately large sample size where all possible instantiation of the variables, including those of rare events, are reflected takes much longer time than needed for performing similar calculations in statistical models to derive the corresponding probabilities. Indeed, many of the widely used approaches to rare event probability estimation follow sampling formalisms, for which the essential factor is the presence of a more or less reliable mathematical model containing epistemic and aleatory uncertainties existing in a different level of interactions between the field variables.[17-21]

1.4. Bayesian Networks

Since introducing the tractable probabilistic inference rules[22,23], Bayesian networks have monopolized attentions in the context of stochastic modeling of highly complex systems. A Bayesian network is constituted of two basic parts: 1) directed acyclic graph representing the cause and effect interrelationship between variables and 2) probability values, quantifying the causal links, by relating the probability distribution of each variable (node) to its parents' probability distribution laws. Discretized Bayesian networks can effectively handle different types of variables (continuous or discrete) and present updated probabilities almost at once when new evidences are given. This task is called "inference" in the context of Bayesian statistics.[24-27]

Central to Bayesian networks inference engine is the Bayes' rule, allowing one to update beliefs about correlated random variables once getting informed about the state of some of them

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1.1}$$

where $P(B|A)$ and $P(A|B)$ denote the likelihood and conditional probabilities of event A given event B. In Bayesian terminology cause is called a parent and effect is its child. Different children can share a common cause, and a cause may have multiple children.

To work properly, Bayesian network topological structure (graph) and probability values must be well established. Developing the Bayesian network's structure is mostly done using the previous knowledge coming from information about the real-world system being modeled. However, systematically learning the structure from the data is a rapidly growing area.[28,29] Estimating prior and conditional probabilities, on the other hand, is to great extent dependent on the data. This stage is usually referred to as *parameter learning*.

As we remarked earlier in previous sections, many actual systems tend to take on certain states more frequently. This behavior, which particularly holds for systems under control, reveals a quality in which the system represents probability distributions with only a few modes. Whenever the variance of the said distribution becomes smaller, a narrower distribution is yielded. In such cases, the probability of observing a sample far from the mean of the distribution dramatically shrinks by the Chebyshev's inequality.[30] If parameter learning of Bayesian networks is intended, with limited number of samples, inadequate information is often obtained for the states or events with probabilities lower than some thresholds dictated by the number of samples. This phenomenon is the same “rare event” situation revisited here for the Bayesian networks. Consequently, more samples should be taken from the system to make sure the historical dataset includes information on the states with small probabilities; otherwise proper learning and inference over the unobserved regions become impossible. The situation gets even more severe when one is looking for data to estimate conditional probabilities of extreme values of a child variable given extreme values of its parents (compound risk situation).

Despite of the fact that in general finding a solution to the rare probability estimation could be problematic, fortunately there was a hope to find an acceptably accurate result for industrial systems with which Bayesian networks are concerned in the current research. Insofar as the target variables to be modeled are numerical and continuous, their associated rare event equivalently implies the extreme values. That is; since in most cases, continuous variables of interest (temperature, flow rate, etc.) are to be controlled at some specific design values, the extreme values far from these set points are considered to be unwanted. As a result, the probability of such an extreme states converges to zero, where the rarity usually emanates. Therefore, within our framework of interest where rare events of continuous entities can be interpreted as extreme values, rare events are no longer unknown. Such rare events are metaphorically called “grey swans”. The difference of grey swan with the so called “black swan” event is that the former is a predictable event, but with unknown probability. Knowing the facts above, we will be able to propose a methodology consolidating our decentralized knowledge.

In the current work we propose a rigorous mathematical modeling technique based on established fundamental laws of probability, statistics and information theory to estimate probability distributions of continuous multivariate random variables from an optimal probability density. This density unifies information coming from every individual sample points and provides a framework for maximum use of information encoded in the data. Unlike traditional approaches to Bayesian network parameter estimation using the local relative frequency technique to estimate probabilities; our improved method incorporates all information presented by finite datasets to set up a unique multivariate probability density function extendable to unobserved regions. Using such a density, not only calculating unknown and near-zero conditional probabilities becomes possible, but also it will be carried out much faster than sampling techniques. Thereafter, our enhanced Bayesian networks would be capable of performing inference over the regions which are recognized by the data itself.

1.5. Bayesian Networks for Risk Assessment and Fault Detection

Causal models in general and our equipped Bayesian networks in particular can be effectively exploited to construct probabilistic models to determine operational risk within industrial systems. By introducing a framework to estimate the system's overall status given its input, such a model is capable of calculating the corresponding deviations and unfavorable events likelihoods, guiding to detect system's weak points and vulnerabilities. This valuable information will further be utilized to calculate risks, in combination with related loss severity and costs. This model also enables us to assess the existing risk controllability, e.g. controllers' robustness. Prior probabilities of the root variables and their behavior, on the other hand, have also much to say about which input parameters, whether internal or external, are more probable to impose risks to the system. In this regard, Bayesian networks can be comparable to traditional risk assessment procedures, e.g. performance indicators, score cards, etc.[31] In addition to training from the historical data, Bayesian networks take the advantage of ability to bind different sources of information, such as expert knowledge. The outcome of this type of analyses will further conduce to large mitigation in risks via risk management formalities.

Another significant application of Bayesian modeling of industrial processes is fault detection.[32] Fault detection can be performed either real-time (online analysis) or as a tool to figure out the most probable reasons of an already happened accident in order to diagnose and prevent similar future events. Although by adding any evidence to the network the whole network gets updated, but two kind of different studies can be done over the updated network simultaneously. If one is interested in how the effect nodes of the given evidence differ from their prior probabilities, the inference called "predictive".

On the other hand, if the ways by which the cause nodes of the given evidence deviate from their prior state are matter of interest, the inference called “diagnostic”.

Finally, the Bayesian methodology furnishes risk evaluation, as well as additionally empowering incorporation of different types of information, where quantified data and subjective knowledge meet each other. This can result some outstandingly influential achievements not conceivable by alternative methods. This extent of profits, along with the unequivocal assessment of stochasticity and capability to convey the outcomes effortlessly and graphically to users, renders Bayesian networks a unique solution for risk determination under uncertainty.[33]

1.6. Bayesian Network Structure Learning from Data

In many industrial applications, such as risk and failure modeling, there is an urgent need for discovering cause and effect relationships among the domain variables.[34] Number of variables under study and the state of the knowledge of the model builder strongly affect the quality of the model. When the internal mechanisms are not fully understood, traditional methods of finding this causal structure by the expert knowledge may lead to poor or even misleading outcomes. Because of the importance of relationships in characterizing complex systems, automatic data-driven approaches have received more attention in recent years.[35,36]

When comes to BNs, the above issue is translated into learning BN topological structure, or DAG, which encodes the conditional dependencies amongst nodes. Besides plenty of hybrid and heuristic methods[37- 39], there are two major classes of data-driven BN structure learning strategies: (1) score and search, which searches for the structure maximizing an objective function;[40,41] and (2) conditional independence (CI) tests like χ^2, which involves accepting or rejecting pairwise independence hypotheses.[42,43]

Despite advances made in this area, some major problems still persist. First, search through a large possible structure space, despite huge computational improvement in the past decade, still takes considerable computational time, which exponentially increases with number of nodes of the model. Second, conditional independence tests often perform poorly in capturing complex dependencies or at their best give an undirected version of the underlying network's graph.[44] In addition, the available methods are susceptible to scarce data sets.

In view of these, there is an increasing demand for a practical method that can automatically produce BN causal structures from data using simplifying assumptions or new analytical approaches. Such a method not only provides the BN calculation with a strong explanatory backbone, but also provides the researchers with a better understanding of complex and large-scale phenomena encountered in real world.

Therefore, finding the causal model which describes the observed data becomes an optimization problem, with exponentially increasing in the number of candidates with respect to size of the variable set. In view of this, and since conditional independence tests cannot theoretically determine directed causal relationship, an efficient search-based algorithm must be developed such that it will be able to explore the search space in minimum possible time. Such capability sounds more critical if we consider the fact that the major complexity of finding the optimal Bayesian network structure via search methods arises due to estimating the goodness of fit measure (search score) for every candidate being studied by the search algorithm. This process which plays an important role in the final result of the optimization process, renders increasingly time consuming with the number of nodes included. That means, if an appropriate strategy is not selected to pick up samples from the search space, most of the computational power will be wasted on redundant cases.

To address this problem, it seems necessary to develop methods to constrain the search space to a hyperspace polytope of best featured candidates. This task would be more feasible if a convenient way is introduced to map the space of DAG to an equivalent space of encoded entities with respect to the nature of the systems being studied (trees, polytrees, densely connected, etc.). A possible solution to this problem can be using CI test to establish such constraints by specifying equivalent classes of undirected graphs representing the conditional dependence among the domain variables. However, such a technique should be carefully developed for the cases where overly connected, dynamic Bayesian networks (DBNs) or recursive causality has to be dealt with.

Another important improvement to the available search-based methods is to develop new scores for the optimization. As mentioned before, since a significant amount of computational capacity is committed to considering unlikely candidate graphs, the need for a metric that can unwind this burden is highlighted. In other words, the selected metric’s ease of calculation expedites the entire arithmetic operations. On the other hand, an efficient score function is required to recognized between more probable nominated samples with less suitable ones more profoundly; that is, implausible samples should be scored such that the similar configurations instantly get penalized by the algorithm by leaving the associated neighborhood.

Finally, oftentimes when optimizing traditional score functions over the assumed sample space multiple optima may be observed, where all of which, except the global optimum one, are local optima and not indicating the true structure of the observed multivariate data. A solution to this problem can be achieved by adopting an appropriate global optimization technique, such as evolutionary optimization (genetic algorithm) or swarm intelligence (particle swarm optimization). Again defining a heuristic score function or transforming the search space into equivalent sets based upon the type of the space being explored may be useful to address the problem of existence of multiple local optima.

1.7. Dynamic Bayesian Networks

Cyclic causality happens when one variable simultaneously affects and is affected by another node in a causal network, or a set of subsequent arrows starting from one node finally ends to the same node. This situation usually observed in process industries, when control loops are present or thermo-sensitive reactions are present. Such a causal relationship is not supported by traditional Bayesian network update rules.[45] Although there have recently appeared works on developing cyclic causal network, but their applications are still limited to certain cases. Furthermore, industrial processes usually undergo transient behavior; they move towards or away from a steady state or operate around a steady state. In modeling such processes, time has to be taken into account.

To resolve these problems we propose applying Dynamic Bayesian networks (DBNs) as they allows one to eliminate cyclic causality as well as modeling time dependent phenomena. In BNs this feature is added by considering different time-step nodes for each individual time-varying node.[45] In addition to regular conditional probabilities that quantify the uncertain interactions of two different variables, this model relates current state of transient nodes to their immediate past or even more steps deep into their historical behavior.[46,47] This extension will allow us to capture the most probable causes for observed evidences in transient operation mode, which is an essential step in real-time process monitoring and fault detection. This interpretation of the cyclic causality is in special compliance with understating of the nature's laws. In most real-world systems, when two objects are in cyclic interaction (where the first objects affects the other and vice versa), this effect is not immediate, since no message can be transmitted faster that the speed of light. Therefore when the second object receives this action sent at moment (t_i) it is at another time instant, say (t_{i+1}), and the reaction reaches the first object at moment (t_{i+2}) and so forth. Hence a cyclic behavior is actually consecutive messages being sent to one object from its partner in a previous time step and cycles can be decomposed using this fact (Figure 1.1).

An important barrier which is revisited here is to infer the DBN causal structure from observational data. Like static BNs, the amount of information encoded in observational data takes an important role in success of the structure learning scheme. However in many actual systems, this information which is carried by time series data don't includes all possible abnormal events. A probable solution for lab to pilot plant scale systems is to use active learning procedure, in which system inputs are manipulated by being taken to certain abnormal states and then corresponding systemic response recorded and employed to give a better view to the system's behavior. Definitely active learning is not applicable for large scale industrial systems; as taking the variables close to extreme events is not allowed by different criteria dictated by system's design, operational and safety guidelines. Therefore this research is going to focus on capturing the DBN causal structure from bounded time series data. To this end, we are going to use some intermediate modeling environments such as Markov chains, and in a long run, sufficient number of samples are provided to feed into the structure learning algorithms.

As mentioned above, unlike static BNs, in DBNs some nodes are specified to link the current or future state of the variables to their past. The manner by which the current system is being affected by its past and the extent to which these messages are effective (long range time dependence) is also a critical issue in designing DBNs, particularly for the industrial systems containing innumerable complicated and in many cases, unknown interactions. To discern such phenomena some indices must be developed (e.g. by generalizing available measures like Hurst's exponent). All the above efforts have as well to be generalized in the proposed research to DBNs which aim at modeling recursive causality, as it imposes extra complexity due to ambiguous demonstration of such relationships in the observational sample data, otherwise a circular causality may mistakenly be recognized as a one way regular effect and vice versa.[48]

Finally for cases of tightly controlled systems (as can be usually found in industry), we will consider the possibility of developing DBNs from regular BNs and its impact on the quality and performance of the achieved models.

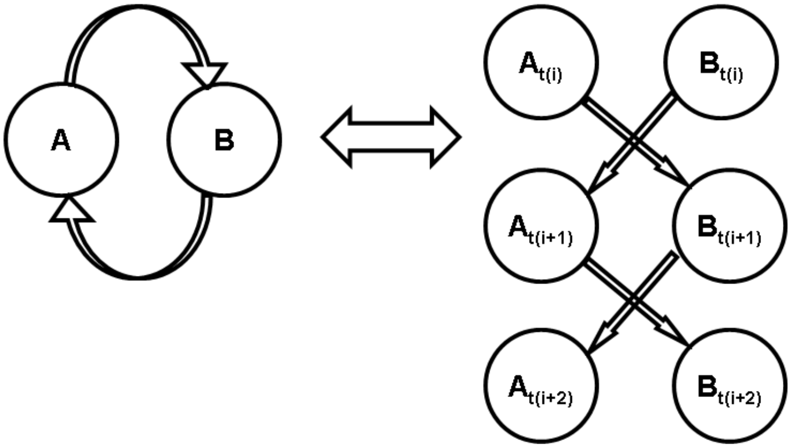

Figure 1. 1: Cyclic causality decomposition using DBN with two time increments.

1.8. Large-Scale BN Inference

Real world systems have hundreds or thousands of variables; because BNs of these systems have hundreds or thousands of nodes, inference using such large-scale networks is computationally infeasible at the present time, especially when real-time network updating and inference are desired.

In BN modeling when the number of variables (nodes) grows, the first problem appears when automatic data-driven structure learning is intended. Learning BN structure has been proven to be NP-hard[49,50], in the sense that number of possible directed acyclic graphs (DAGs) available in the place of candidates for the true underlying network grows astronomically as the number of variables increases.[51] Furthermore, the presence of more variables in the BN model leads to more intricate and hard-to capture interactions whose discovery is a difficult task both manually and automatically. A similar complexity exists when structure discovery is carried out by independence tests[52,53], where much more pairwise-conditional-independence assessments must be made for larger BNs. This problem becomes even more difficult when the database is scarce.

Once the network topology is established, available data is used to estimate conditional probabilities. Larger number of nodes and dense networks can significantly slow down the parameter estimation step, due in large part to higher number of calculations needed to classify the observed data over the assigned states, particularly when a streaming input of data must be processed in real-time to update the network's parameters. This problem further magnifies when data is scarce, as the data required to describe probabilities of combinatorial extreme states of a cluster of parent nodes and their child become less available, and leads to incomplete conditional probability tables. The most severe condition imposed by large BNs is associated with Bayesian inference. Even though the network consists of discrete nodes rather than continuous variables, Bayesian inference is still considered as an NP-hard problem.[54] Despite great deal of work in the literature[55] on the Bayesian inference problem, there is no general method to update probabilities given new evidences in polynomial time. More complexity is faced when the network moves away from sparse configurations, in which case number of states increases or conditional probabilities becomes incomplete.

All of the above challenges seem more severe when viewed from the real-time inference stand point, which is essential for fault detection and process monitoring roles of the BNs. To address this problem, the existence of an integrated framework is critical in modern day application. More attention must be focused on “anytime” algorithms[56-59], which incrementally and iteratively present more accurate solutions for the updated network. As an alternative resolution, since no single algorithm can handle all kinds of possible BNs, as stated by No Free Lunch (NFL) theory[60,61], a library consisting of the most efficient available methods can be developed as a toolbox to work with any sort of

BNs according to their type. Finally, approximate solutions can be employed to simplify large BNs by doing local inference, variable elimination, stochastic samplings, node reduction, state merging and so on, based on the type of the network being dealt with, to get around the problems caused by large scale Bayesian networks. The performance of these techniques can be improved significantly if meta-level reasoning is utilized to explore the space of the candidate methods and characterize the best possible solution[62,63].

To address multiple inference challenges arisen by large BNs, a unifying framework consisting of solutions to different possible network complexity levels is proposed. As mentioned earlier, depending on the size of the causal network being analyzed, its density, number of states of each node, the purpose of inference (real-time computations or offline studies), etc., an appropriate inference technique is present in the proposed framework. If enough random samples are already produced before an abnormal situation is met, approximate inference methods are superior over the exact algorithm. On the other side, anytime algorithms start with approximate solutions to satisfy some urgent real-time inference needs and gradually give more accurate results as time goes by. These stepwise approximations apply to the network structure simplifications, reduced number of samples used to perform approximate inference, minimized number of states and so on. On the other hand, heuristic solutions are to be developed to perform Bayesian inference locally; that is, given a set of evidences, the computational power is intelligently spent on updating those parts of the network which are most likely to have resulted in the observed evidences without needing to redundantly update the entire network. Finally, an optimization scheme can be exploited to discretize the variables within the network considering the traditional objective functions (maximum entropy principle, maximum consistency, etc.) together with considering the cost of Bayesian inference computations when constructing the network to achieve minimal possible inference time. Developing new local message passing algorithms, developing novel measures for structure decomposition to equivalent set of polytrees and performing stepwise inference algorithm on them, and developing new measure for selecting optimal subset of nodes required to speed up the Bayesian inference based upon the location where the evidence is introduced to the network, demand received from the user about the nodes of interest and the variables experiencing maximum deviation from their normal values are under active research.

References

1. Kaplan K., Garrick S., “On the Quantitative Definition of Risk,” *Risk Analysis,* **1,** 11-37 (1981).

2. Loisel, S., Milhaud, X., “From deterministic to stochastic surrender risk models: impact of correlation crises on economic capital,” *European Journal of Operational Research* **214**(2), 348-357 (2011).

3. Baker, S., Ponniah, D., Smith, S., “Techniques for the analysis of risks in major projects,” *Journal of the Operational Research Society*, **49**(6), 567-572 (1998).

4. Haimes, Y. Y., Risk Modeling, Assessment, and Management. A John Wiley & Sons Inc. publication, 2009.

5. Dehling, H. G., Gottschalk, T., Hoffman, A. C., Stochastic Modeling in Process Technology, Elsevier, 2007.

6. Harms-Ringdahl, L., Safety Analysis, Principles and Practice in Occupational Safety, CRC Press, 2001.

7. Hallenbeck, W.H., Quantitative risk assessment for environmental and occupational health, Lewis Publishers, 1986.

8. Cohen, F., “Risk Management: There Are No Black Swans,” *Fred Cohen & Associates: Analyst Report and Newsletter*, **4** (2009).

9. Taleb, N. N, The Black Swan: The Impact of the Highly Improbable, Random House, 2007.

10. “Washburn 'A' Mill Explosion”, Library: History Topics. Minnesota Historical Society, 2010.

11. “Fire Investigation Summary: Grain Elevator Explosion – Haysville, Kansas, June 8, 1998”, *National Fire Protection Association (NFPA), Fire Investigations Department*, 1999.

12. “Savar collapse death toll reaches 1,126” , *The Daily Ittefaq*, 2013.

13. Vernon H. Guthrie and David A. Walker, Enterprise Risk Management, Proceedings of the 17th International System Safety Conference, Orlando, FL, (1999).

14. Kontogiannis, T., Leopoulos, V., Marmaras, N. , “A comparison of accident analysis techniques for safety-critical man-machine systems,” *International Journal of Industrial Ergonomics*, **25**, 327-347 (2000).

15. Park, I. S., Park. J. W., “Determination of a risk management primer at petroleum-contaminated sites: Developing new human health risk assessment strategy,” *Journal of Hazardous Materials*, **185**(2–3), 1374–1380 (2011).

16. Jodrá, P., “A closed-form expression for the quantile function of the Gompertz–Makeham distribution,” *Mathematics and Computers in Simulation* **79** (10), 3069–3075 (2009).

17. Asmussen, S., Kroese, D. P.,”Improved algorithms for rare event simulation with heavy tails,” *Advances in Applied Probability*, **38**, 545-558 (2006).

18. Dupuis, P., Wang, H., “Dynamic Importance Sampling for Uniformly Recurrent Markov Chains,” *Annals Appl. Probabilit.*, **15**, 1–38 (2005).

19. Glasserman, P., Heidelberger, P., Shahabuddin, P., Zajic, T., “A large deviations perspective on the efficiency of multilevel splitting,” *IEEE Trans. Automat. Control,* **43**, 1666–1679 (1998).

20. Shahabuddin, P., “Rare event simulation of stochastic systems,” In *Proceedings of the 1995 Winter Simulation Conference,* Alexopoulos, C., Kang, K., Lilegdon, W. R., Goldsman, D. (eds), 178–185: IEEE Press (1995)

21. Rubinstein, R. Y., “The stochastic minimum cross-entropy method for combinatorial optimization and rare-event estimation,” *Methodology and Computing in Applied Probability,* **7**, 5–50 (2005).

22. Lauritzen, L., Spiegelhalter, J., “Local computations with probabilities on graphical structures and their application to expert systems (with discussion),” *Journal of the Royal Statistical Society*, **50(**2), 157-224 (1988).

23. Shenoy P., Shafer G., “Axioms for probability and belief-function propagation,” *Readings in Uncertain Reasoning*, 575–610, (1990).

24. Duda, R. O., Hart, P. E., Nilsson, N. J., “Subjective Bayesian methods for rule-based inference systems,” in *AFIPS*, **45**, 1075–1082 (1976).

25. Pearl, J., “Reverend Bayes on inference engines: a distributed hierarchical approach,”In *Proceedings of the Second National Conference on Artificial Intelligence*, 133–136 (1982).

26. Murphy, K. P., “Dynamic Bayesian networks: Representation, inference and learning,” Ph. D. thesis, Department of Computer Science, University of California, Berkeley, (2002).

27. Zweig, G., “Bayesian network structures and inference techniques for automatic speech recognition,” *Computer Speech & Language* **17**(2-3), 173–193 (2003).

28. Heckerman, D., “A Bayesian approach to learning causal networks,” In *Advances in Decision Analysis: from Foundations to Applications,* Edwards, W. & Miles R. F. Jr (eds). Chapter 11, Cambridge University Press, 202–220 (2007).

29. Lucas, P. , “Restricted Bayesian network structure learning,” In *Proceedings of the First European Workshop on Probabilistic Graphical Models* (PGM 2002), Gamez J. A. & Salmeron A. (eds). 117–126 (2002).

30. Marshal, A. W., Olkin, I., “Multivariate Chebyshev Inequalities,” *The Annals of Mathematical Statistics,* **31**(4), 1001-1014 (1960).

31. Pariyani A., Seider, W. D., Oktem, U., and Soroush, M., “Incidents Investigation and Dynamic Analysis of Large Alarm Databases in Chemical Plants: An FCCU Case Study,” *Ind. Eng. Chem. Res.,* **49,** 8062-8079 (2010).

32. Mehranbod, N., Soroush, M., Piovoso, M., Ogunnaike, B.A., “A probabilistic model for sensor fault detection and identification,” *AIChE Journal* **49** (7) 1787-1802 (2003).

33. Cruz, M.G.(Ed.), Operational risk modeling and Analysis, London, England: Risk Books, (2004).

34. Verma, T., Pearl, J., “An algorithm for deciding if a set of observed independencies has a causal explanation,” In *Proceedings of the Eighth Conference on Uncertainty in Artificial Intelligence (UAI-92)*, Dubois, D., Wellman, M. P., D’Ambrosio, B., Smets, P. (eds.), Morgan Kaufmann, 323–330 (1992).

35. Heckerman, D., “A Tutorial on Learning with Bayesian Networks,” *Technical report MSR-TR-95-06,* Microsoft Research (1995).

36. Geiger, D., Heckerman, D., King, H., Meek, C., “Stratified exponential families: graphical models and model selection,” *The Annals of Statistics* **29**(2), 505–529 (2001).

37. Dash, D., Druzdzel, M. J., “A hybrid anytime algorithm for the construction of causal models from sparse data,” In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI-99),* Prade, H., Laskey, K. (eds.). Morgan Kaufmann, 142–149 (1999).

38. Guo, Y.-Y., Wong, M.-L. Cai, Z. H., “A novel hybrid evolutionary algorithm for learning Bayesian networks from incomplete data,” In *Proceedings of the IEEE Congress on Evolutionary Computation (CEC 2006),* 916–923 (2006).

39. Wang, M., Chen, Z., Cloutier, S., “A hybrid Bayesian network learning method for constructing gene networks,” *Computational Biology and Chemistry,* **31**(5–6), 361–372 (2007).

40. Chickering, D. M., Heckerman, D., “Efficient approximations for the marginal likelihood of Bayesian networks with hidden variables,” *Machine Learning*, **29(**2–3), 181–212 (1997).

41. Shaughnessy, P., Livingston, G., “Evaluating the causal explanatory value of Bayesian network structure learning algorithms,” *Research paper 2005-013*, Department of Computer Science, University of Massachusetts Lowell (2005).

42. Brown, L. E., Tsamardinos, I., Aliferis, C. F., “A comparison of novel and state-of-the-art polynomial Bayesian network learning algorithms,” In *Proceedings of the Twentieth National Conference On Artificial Intelligence*, Veloso, M. M., Kambhampati, S. (eds). 2, AAAI Press, 739–745 (2005).

43. Yu, K., Wang, H., Wu, X., “A parallel algorithm for learning Bayesian networks,” In *Proceedings of the Eleventh Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD 2007)*, Lecture Notes in Artificial Intelligence 4426, Springer,1055–1063 (2007).

44. Fung, R. M., Crawford, S. L., “Constructor: a system for the induction of probabilistic models,” In *Proceedings of the Eighth National Conference on Artificial Intelligence 2*, AAAI Press, 762–769 (1990).

45. Lauritzen L., Spiegelhalter J., “Local computations with probabilities on graphical structures and their application to expert systems (with discussion),” *Journal of the Royal Statistical Society*, **50(**2), 157-224 (1988).

46. Dean, T., Kanazawa, K. “A model for reasoning about persistence and causation,” Computational Intelligence **5**(2), 142–150 (1989).

47. Friedman, N., Murphy, K., Russell, S., “Learning the structure of dynamic probabilistic networks,” In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI-98)*, Cooper, G. F., Moral, S. (eds), Morgan Kaufmann, 139–148 (1998).

48. Chabanenko., D. M., “Detection of short- and long-term memory and time series prediction methods of complex Markov chains,” In *Visnyk Natsionalnogo tehnichnogo universitetu Kharkivsky politehnichny institut. Zbirnik Naukovyh pratz. Tematichny vypusk: Informatika i modelyuvannya*, **31**, NTU KHPI, Kharkov, 184-190 (2010).

49. Chickering, D. M., Heckerman, D., Meek, C., “Large-sample learning of Bayesian networks is NP-hard,” *Journal of Machine Learning Research* **5**, 1287–1330 (2004).

50. Chickering, D. M., “Learning Bayesian networks is NP-complete,” In *Learning from Data: Artificial Intelligence and Statistics V*, Fisher, D., Lenz, H. J. (eds.), Lecture Notes in Statistics **112**, 121–130. Springer (1996).

51. Heckerman, D., “A Bayesian approach to learning causal networks,” In *Advances in Decision Analysis: from Foundations to Applications*, Edwards, W., Miles R. F. Jr (eds.), Chapter 11, Cambridge University Press, 202–220 (2007).

52. Cowell, R., “Conditions under which conditional independence and scoring methods lead to identical selection of Bayesian network models,” In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence (UAI-01),* Breese, J., Koller, D. (eds.), Morgan Kaufmann, 91–97.

53. de Campos, L. M., Huete, J. F., “Approximating causal orderings for Bayesian networks using genetic algorithms and simulated annealing,” In *Proceedings of the Eight Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems,* Madrid, Spain, 333–340 (2000).

54. Cooper, G. F., “The computational complexity of probabilistic inference using Bayesian belief networks,” *Artificial Intelligence* **42**(2–3), 393–405 (1990).

55. Larrañaga, P., Karshenas, H., Bielza, C., Santana, R., “A review on evolutionary algorithms in Bayesian network learning and inference tasks,” *Information Sciences* **233**, 109-125 (2013).

56. Boddy, M., “Anytime problem solving using dynamic programming,” In *Proceedings of the Ninth National Conference on Artificial Intelligence (AAAI91),* AAAI Press, San Mateo, CA, 738-743 (1991).

57. Horvitz, E., Suermondt, H. J., Cooper, G. F., “Bounded conditioning: Flexible inference for decisions under scarce resources,” *Proc. 5th conference on Uncertainty in Artificial Intelligence,* Windsor, Ontario, 182--193, (1989).

58. Garvey, A. J., Lesser, V. R., “Design-to-time real-time scheduling,” *IEEE Transactions on Systems, Man and Cybernetics*, **23**(6), 1491-1502 (1993).

59. Zilberstein, S., “Using anytime algorithms in intelligent systems,” *AI Magazine*, **17**(3):73-83 (1996).

60. Droste, S., Jansen, T., Wegener, I., “Optimization with randomized search heuristics: the NFL theorem, realistic scenarios, and difficult functions,” *Theoretical Computer Science*, **287**(1), 131–144 (2002).

61. English, T., “No More Lunch: Analysis of Sequential Search,” In *Proceedings of the 2004 IEEE Congress on Evolutionary Computation,* 227–234 (2004).

62. Santos, E. J., Shimony, S. E., Solomon, E., Williams, E., “On a distributed anytime architecture for probabilistic reasoning,” AFIT/EN/TR95-02, Department of Electrical and Computer Engineering, Air Force Institute of Technology, (1995).

63. Dagum, P., Luby, M., “Approximating probabilistic inference in Bayesian belief networks is NP hard,” *Artificial Intelligence*, **60**, 141-153 (1993).

Chapter 2: Maximum-Likelihood Maximum-Entropy Constrained Probability Density Function Estimation for Prediction of Rare Events

2.1. 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[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.[12,13,14] On the other hand, 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 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.[19,20,21] 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 Monte-Carlo 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 Monte-Carlo 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 non-parametric types.[29] As shown in this chapter, 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 non-monotonic.

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 maximum-likelihood, maximum-entropy (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 non-parametric methods is that this method allows for effectively considering higher moments of each random variable (e.g., skewness and kurtosis).

The rest of the chapter is organized as follows. The problem of estimating the probability of rare events within the framework of Bayesian networks 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.

2.2. 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 Bayesian network inference. The example involves two variables, Y and Z, where Z depends on Y. Throughout this chapter, each random variable is denoted by a capital letter and its numerical value denoted by a lower-case letter. Random variables are assumed to have 5 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.

Bayesian networks (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.

Bayesian networks 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 maximum likelihood principle, then the frequentists approach[37] should be employed. 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:

$$P(y \in s_k) = \frac{n(y \in s_k)}{\sum_{i=1}^{m} n(y \in s_i)} \tag{2.1}$$

and the conditional probability of the variable Z being in a state $\mathrm{r_i}$ given the variable Y in a state s_k is defined as:

$$P(z \in r_i | y \in s_k) = \frac{n(z \in r_i, y \in s_k)}{n(y \in s_k)} \tag{2.2}$$

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 2.1. Note that in some states no data have been observed. According to Eqs. (2.1) and (2.2), 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”.

Table 2. 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

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:

$$P(z|y \in HH) = \frac{P(y \in HH|z)P(z)}{P(y \in HH)} = \frac{0 \times P(z)}{0} = \frac{0}{0} \qquad (2.3)$$

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

2.3. Preliminaries

2.3.1. Moments of a Probability Distribution Function

Moments of a random variable (vector) X with a probability density function $f(x)$ are defined as expected values of arbitrary functions of the random variable (vector). The most common moments are the first-order moment ($E(X)$ or mean) and the second-order moment ($E((X - E(X))^2)$).[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 $g_i(\mathbf{x}): \mathbb{R}^d \rightarrow \mathbb{R}$ be a moment function of a d-dimensional random vector $\mathbf{X} = [X_1, \dots, X_d]^T \in \Omega \subseteq \mathbb{R}^\mathrm{d}$ with a PDF $f(\mathbf{x}): \mathbb{R}^d \rightarrow \mathbb{R}^+ \cap \{0\}$, where Ω is the domain of **X**. The moment of the random vector **X** with respect to the moment function $g_i(\mathbf{x})$ is defined as:

$$\mu_i = E\big(g_i(\mathbf{x})\big) = \int_\Omega g_i(\mathbf{x}) f(\mathbf{x}) d\mathbf{x}, \qquad i = 0, 1, 2, \dots \tag{2.4}$$

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

$$\bar{\mu}_i = n^{-1} \sum_{j=1}^{n} g_i\left(\boldsymbol{\chi}_j\right), \qquad i = 0,1,2, \dots \tag{2.5}$$

where n is the number of samples of $\boldsymbol{\chi}_j$.

2.3.2. Entropy of a Random Variable

In information theory, the entropy of a random variable is a measure of the uncertainty of 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 $\mathbf{X}$ with a PDF $f(\mathbf{x})$ on a domain Ω the information entropy is defined as[39]:

$$S(\mathbf{X}) = -\int_{\Omega} f(\mathbf{x}) \ln f(\mathbf{x}) d\mathbf{x} \tag{2.6}$$

with 0×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, $S(\mathbf{X})$ is viewed as a measure of the information carried by $\mathbf{X}$, and as data are communicated/transmitted more, they are corrupted with more noise (entropy increases) and therefore they carry less information.

2.4. Method

Given a data set, 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 $f(\mathbf{x})$ is sought that is the solution to the constrained optimization problem:

$$\max_f \left\{ -\int_{\Omega} f(\mathbf{x}) \ln f(\mathbf{x}) d\mathbf{x} \right\} \tag{2.7}$$

subject to the equality constraints:

$$\int_{\Omega} g_i(\mathbf{x})f(\mathbf{x})d\mathbf{x} = \bar{\mu}_i, \quad i = 0, \dots, m \tag{2.8}$$

where $\bar{\mu}_i$ is the i-th 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, $g_0(\mathbf{x})$. One should always set $g_0(\mathbf{x}) = 1$ 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 $\int_{\Omega} f(\mathbf{x})d\mathbf{x} = \bar{\mu}_0 = 1$. This PDF estimation formulation is a multivariate version of the univariate formulation introduced by Zellner et al.[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 many 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. (2.7) and (2.8) is a classical optimization problem, whose solution minimizes the Lagrange function:

$$\tilde{L}(f, \lambda_1, \dots, \lambda_m) = \int_{\Omega} f(\mathbf{x}) \ln f(\mathbf{x})d\mathbf{x} + \sum_{i=0}^{m} \lambda_i \left(\int_{\Omega} g_i(\mathbf{x})f(\mathbf{x})d\mathbf{x} - \bar{\mu}_i \right) \tag{2.9}$$

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

$$\frac{\partial \tilde{L}}{\partial \hat{f}} = 0, \quad \int_{\Omega} g_i(\mathbf{x})\hat{f}(\mathbf{x})d\mathbf{x} = \bar{\mu}_i, \quad i = 0, \dots, m \tag{2.10}$$

where $\hat{f}(\mathbf{x})$ is the estimated PDF. The first algebraic equation from the left in Eq. (2.10) yields:

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

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

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

Therefore,

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

leading to the closed-form analytical solution:

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

which can be written in the form:

$$\hat{f}(\mathbf{x}) = \frac{1}{e^{1+\lambda_0}} \exp(-\textstyle\sum_{i=1}^{m} \lambda_i g_i(\mathbf{x})) \tag{2.11}$$

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

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

implies that

$$e^{1+\lambda_0} = \int_{\Omega} \exp(-\textstyle\sum_{i=1}^{m} \lambda_i g_i(\mathbf{x}))\, d\mathbf{x}$$

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 method of moments[45,46] have been proposed. Requesting equal data and model moments seems reasonable by the law of large numbers – which results from the maximum likelihood 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:

$$\vec{\lambda}_{MLE} = \arg\max_{\vec{\lambda}} L(\vec{\lambda}|\mathbf{D}) = \arg\max_{\vec{\lambda}} f(\mathbf{D}|\vec{\lambda}) = \arg\max_{\vec{\lambda}} \prod_{j=1}^{n} \frac{\exp(-\sum_{i=1}^{m} \lambda_i g_i(\boldsymbol{\chi}_j))}{\theta} \quad (2.12)$$

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

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

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

$$\vec{\lambda}_{MLE} = \arg\max_{\vec{\lambda}} \ln L(\vec{\lambda}|\mathbf{D}) = \arg\max_{\vec{\lambda}} \left\{-n\ln[\theta] - \sum_{j=1}^{n} \sum_{i=1}^{m} \lambda_i g_i(\boldsymbol{\chi}_j)\right\} \quad (2.13)$$

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

2.4.1. 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. (2.13) 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. (2.13) is expected to have a unique optimum (global maximum).[49]

First it is proven that the MLE problem described by Eq. (2.13) 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:

$$\frac{\partial}{\partial\lambda_i}\ln L\left(\vec{\lambda}\middle|\mathbf{D}\right) = -\,n\frac{\partial\theta}{\partial\lambda_i} - \sum_{j=1}^{n} g_i\left(\boldsymbol{\chi}_j\right) = 0, \qquad i = 1, \dots, m \tag{2.14}$$

leading to:

$$\frac{\partial\theta}{\partial\lambda_i} = -n^{-1}\sum_{j=1}^{n} g_i\left(\boldsymbol{\chi}_j\right) = -\bar{\mu}_i, \qquad i = 1, \dots, m \tag{2.15}$$

Hence, the model parameters should be estimated by satisfying the m nonlinear algebraic equations in Eq. (2.15). The right-hand sides of Eq. (2.15) 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. (2.15) 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:

$$\frac{\partial \ln\theta}{\partial\lambda_i} = \frac{1}{\theta}\frac{\partial}{\partial\lambda_i}\int_\Omega \exp\left(-\sum_{i=1}^m \lambda_i g_i(\mathbf{x})\right) d\mathbf{x} = -\int_\Omega g_i(\mathbf{x})\frac{1}{\theta}\exp\left(-\sum_{i=1}^m \lambda_i g_i(\mathbf{x})\right) d\mathbf{x} =$$

$$-\int_\Omega g_i(\mathbf{x})\hat{f}(\mathbf{x})d\mathbf{x} = -\hat{\mu}_i \qquad (2.16)$$

Therefore, Eq. (2.14) simply requires that:

$$\hat{\mu}_i = \bar{\mu}_i, \qquad i = 1, \ldots, m$$

which implies that if the empirical moments, $\bar{\mu}_i, i = 1, \ldots, m,$ 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:

$$\frac{\partial}{\partial\lambda_k}\left(\frac{\partial}{\partial\lambda_i}\ln\left(L(\vec{\lambda}|\mathbf{D})\right)\right) = \frac{\partial}{\partial\lambda_k}\left(-n\frac{\partial\ln\theta}{\partial\lambda_i} - \sum_{j=1}^n g_i\left(\boldsymbol{\chi}_j\right)\right)$$

$$= -n\frac{\partial}{\partial\lambda_k}\frac{\partial\, ln\,\theta}{\partial\lambda_i} = n\frac{\partial}{\partial\lambda_k}\left(\frac{1}{\theta}\int_\Omega g_i(\mathbf{x})\exp\left(-\sum_{i=1}^m \lambda_i g_i(\mathbf{x})\right)d\mathbf{x}\right)$$

$$= -n\frac{1}{\theta}\int_\Omega g_i(\mathbf{x})g_k(\mathbf{x})\exp\left(-\sum_{i=1}^m \lambda_i g_i(\mathbf{x})\right)d\mathbf{x}$$

$$+ n\left(\frac{1}{\theta^2}\int_\Omega g_i(\mathbf{x})\exp\left(-\sum_{i=1}^m \lambda_i g_i(\mathbf{x})\right)d\mathbf{x}\right)\int_\Omega g_k(\mathbf{x})\exp\left(-\sum_{i=1}^m \lambda_i g_i(\mathbf{x})\right)d\mathbf{x}$$

$$= -n\left[\int_\Omega g_i(\mathbf{x})g_k(\mathbf{x})\hat{f}(\mathbf{x})d\mathbf{x} - \int_\Omega g_i(\mathbf{x})\hat{f}(\mathbf{x})d\mathbf{x}\int_\Omega g_k(\mathbf{x})\hat{f}(\mathbf{x})d\mathbf{x}\right]$$

$$= -n\left[E\left(g_i(\mathbf{x})g_k(\mathbf{x})\right) - E\left(g_k(\mathbf{x})\right)E\left(g_k(\mathbf{x})\right)\right] = -n.\operatorname{cov}(g_i(\mathbf{x}), g_k(\mathbf{x})), \quad i = 1, \ldots, m$$

(2.17)

where $\mathrm{cov}(\mathrm{a},\mathrm{b})$ is the covariance of two random numbers a and b. Eq. (2.17) 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. Eq. (2.17) 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 method of moments where $\hat{\mu}_i = \bar{\mu}_i\,,\ i = 1, \dots, m$.

2.4.2. Selection of Moment Function

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. (2.11), if each $g_i(\mathbf{x})$ is replaced with a truncated Taylor series expansion of $g_i(\mathbf{x})$ around the expectation of $\mathbf{x}$, then the problem of looking for proper $g_i(\mathbf{x})$ moment functions is converted to that of finding an optimal order of the truncation for each of the expansions:

$$\vec{\lambda}_{MLE} = \arg\max_{\vec{\lambda}} L(\vec{\lambda}|\mathbf{D}) = \arg\max_{\vec{\lambda}} \prod_{j=1}^{n} \frac{1}{\theta} \exp\left(-\sum_{i=1}^{m} \lambda_i \left[a_i + b_i \boldsymbol{\chi}_j + \boldsymbol{\chi}_j^T c_i \boldsymbol{\chi}_j + \cdots\right]\right)$$

$$=\arg\max_{\vec{\lambda}} \prod_{j=1}^{n} \frac{1}{\theta} \exp\left(\beta_0 + \beta_1 \boldsymbol{\chi}_j + \boldsymbol{\chi}_j^T \beta_2 \boldsymbol{\chi}_j + \cdots\right)$$

where $\beta_0, \beta_1, \beta_2$ ... 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 $f(\mathbf{x})$ to the data. Measures like mean square error

(MSE) and maximum likelihood 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.[50,51,52] 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 $\vec{\lambda}_{MLE}$ versus 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:

$$O_{opt} = \arg\min{}_{O}\left(\frac{\partial \ln\, L\left(\vec{\lambda}'_{MLE}(O)\middle|\mathbf{D}\right)}{\partial O} - \alpha\right)^2 \tag{2.18}$$

where $L\left(\vec{\lambda}'_{MLE}(O)\middle|\mathbf{D}\right)$ is the maximum of the likelihood function using an O^{th}-order truncated Taylor series expansion of every $g_i(\mathbf{x}), i = 1, \dots, m.$ 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.

2.5. Application to Two Examples

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

2.5.1. Example 1: A Bivariate Bayesian Network

Consider two random variables Y and Z described by:

$$Y \sim N(0{,}0.25) \tag{2.19}$$

$$Z = \cos(Y) + \epsilon(0, 0.1) \tag{2.20}$$

where $N(0{,}0.25)$ represents a normal distribution with a mean of 0 and a standard deviation of 0.25. $\epsilon(0, 0.01)$ is white noise standard deviation of 0.1 (a normal distribution with a mean of 0 and a variance of 0.01). The Bayesian network of this example is shown in Figure 2.1. 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 non-parametric 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. (2.19) and (2.20). Figure 2.1 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),

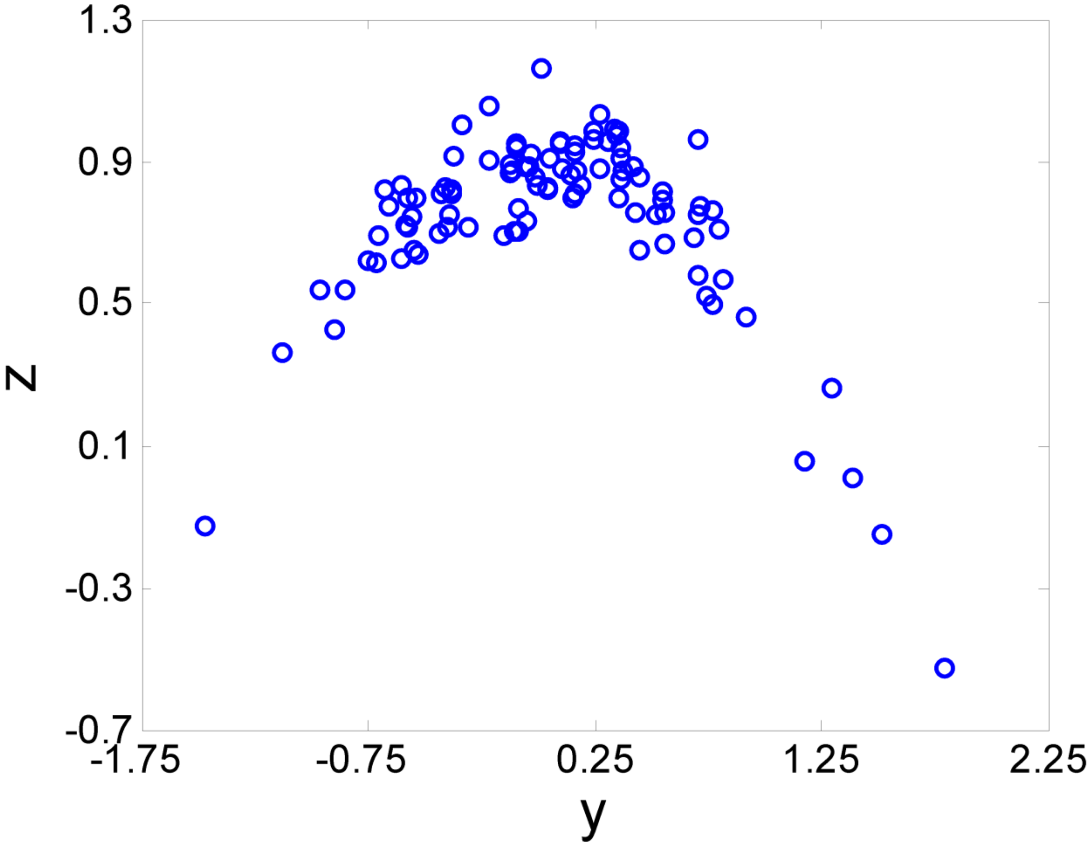

Figure 2. 1: Scatter plot of the 100 (Y, Z) samples.

Normal (N), High (H) and High-High (HH), according to the rule described in the second section, the marginal probabilities given in Table 2.2 are obtained.

As can be seen in Table 2.2, 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 2.2, 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 $O_{opt} = 7$ is used herein to estimate complete PDFs of the Y and Z from the 100 samples. A few low-order moments of the random variables Y and Z, and the combinatorial random variable ($\mathbf{X} = (Y, Z)^T$) are given in Table 2.3. Figure 2.3 shows that as the order of the truncations, O, increases, the maximized logarithm of the likelihood function converges to a higher limit. Figure 2.3a compares the true $f_Z(z)$ described by Eq. (2.20) 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:

$$\hat{P}(y \in s_i) = \int_{s_i} \hat{f}_Y(y) dy \tag{2.21}$$

$$\hat{P}\left(z \in r_j \middle| y \in s_i\right) = \frac{\int_{r_j} \int_{s_i} \hat{f}_Y(y) \hat{f}_{\vec{\lambda}}(z|y) dy dz}{P(Y \in s_i)} \tag{2.22}$$

where s_i and r_j denote the i-th state of Z and the j-th state of Y, respectively.

Table 2. 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

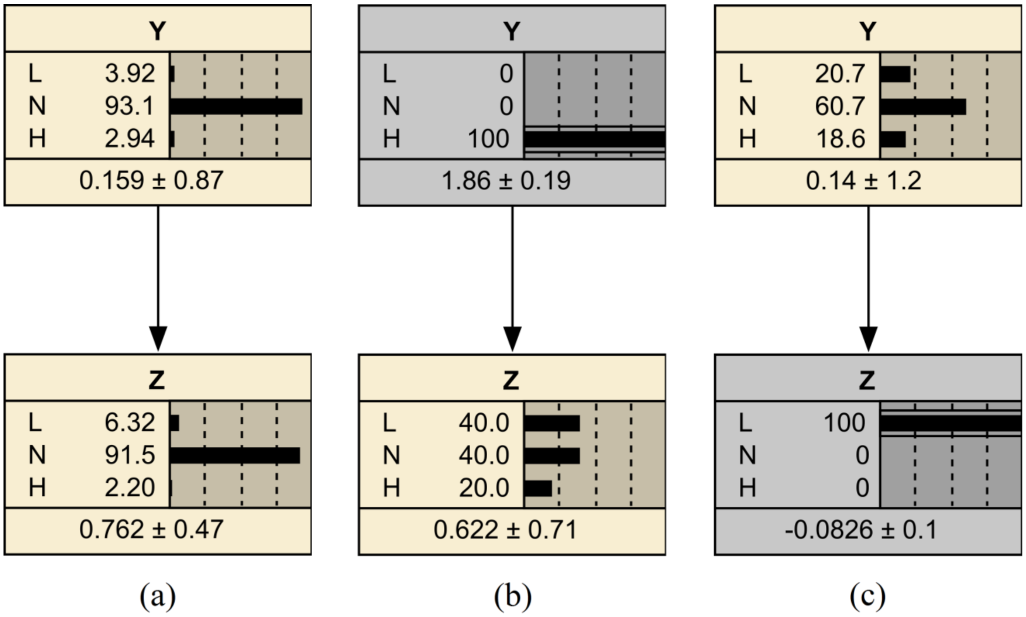

Figure 2. 2: Bivariate Bayesian network for *Y* and *Z* trained with 100 samples in Netica. (a) Normal operation network. (b) Predictive inference (evidence is for *Y*). (c) Diagnostic inference (evidence is for *Z*).

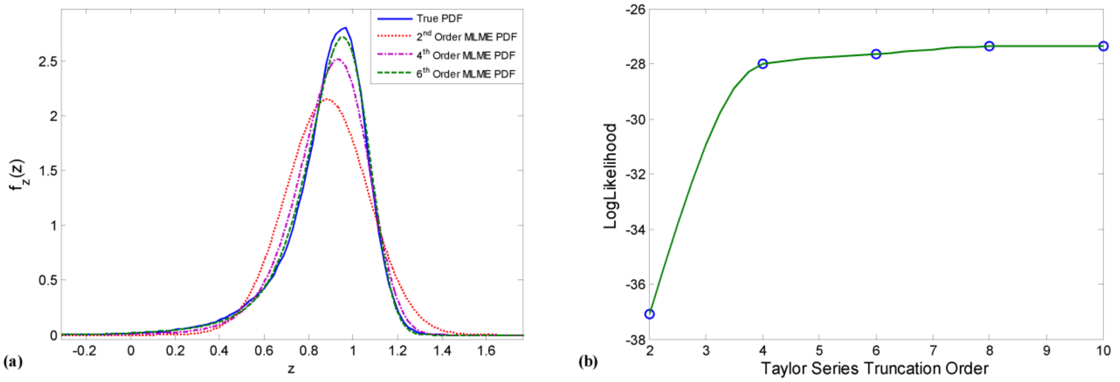

Figure 2. 3: (a) Univariate MLME PDF estimated using different truncation orders. (b) Log-likelihood of the PDF of *Z*, MLME estimated with different moment orders.

Table 2. 3: Moments of *Z* and (Y, Z) in an increasing order of moments, calculated using the data samples given in Figure 2.1.

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

2.5.1.1. 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 copula cumulative distribution function (CDF), denoted by C, or its equivalent PDF, denoted by c, which are related to each other according to:

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

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:

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

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

$$u_i = \int_{\Omega_{X_i}} f_{X_i}(x_i) dx_{i,}\ i = 1, \dots, d \tag{25}$$

with Ω_{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 non-monotonic 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:

$$C_{t,\nu}(u_1,u_2) = \int_{-\infty}^{t_\nu{}^{-1}(u_1)} \int_{-\infty}^{t_\nu{}^{-1}(u_2)} \frac{1}{2\pi(1-\rho^2)^{\frac{1}{2}}} \{1 + \frac{y^2-\rho yz+z^2}{\nu(1-\rho^2)}\}^{-(\nu+2)/2} dydz \qquad (2.26)$$

$$C_{Gu}(u_1,u_2) = \exp\ \{-[(-\ln\,(u_1))^{\varphi} + (-\ln\,(u_2))^{\varphi}]^{1/\varphi}\} \qquad (2.27)$$

where u_1 and u_2 are defined according to Eq. (2.25), ν is the degree of freedom of the univariate Student's t distribution (t), $\varphi \in [1, +\infty)$ 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. (2.23) and (2.24), where the marginal PDFs f_Y and f_Z are obtained using a non-parametric kernel method,[60] also described in the next section. Figures 2.4a, 2.4b and 2.4c 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 little circles represent the actual data, and the solid line contours represent the estimated bivariate joint PDF of random variables Y and Z. Figure 2.4a shows the estimated PDF by the MLME method using a 7-th order of truncation.

Compared to actual PDF shown in Figure 2.4e, the MLME PDF can capture the non-monotonic behavior of the sine function around $Y = 0$. As the copulas use the covariance matrix of Y and Z to capture the correlation between these variables, as can be seen in Figures 2.4b and 2.4c, 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.

2.5.1.2. Comparison with Non-parametric Kernel Method

Kernel density estimation methods are a sub-class of non-parametric 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]:

$$\hat{f}(\mathbf{x}|H) = \frac{\sum_{j=1}^{n} K_H(\mathbf{x}-\chi_j)}{n} = \frac{\sum_{j=1}^{n} K\left((H)^{-1}(\mathbf{x}-\chi_j)\right)}{n.\det(H)} \qquad (2.28)$$

where $\hat{f}$ is the PDF estimated using a kernel method with a scaled kernel function $K_H(.)$, $K(.)$ is the kernel probability density function, 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 ($n^{-\alpha}$ compared to n^{-1} where $0 \leq \alpha < 1$).[62] This means that with small sample sizes it is not possible to remove non-smoothness caused by individual data points. A large increase in the smoothing parameter may eventually lead to over-smoothness and 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. On the other hand, as described in Eq. (2.17), the use of larger-size samples (larger n), not only doesn't decelerate the probability estimation, but also gives the likelihood function a sharper peak, allowing the maximum to be calculated with less iteration. 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 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 non-monotone 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 kernel 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 2.4d, the non-parametric 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 2.5 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 non-monotonic 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.

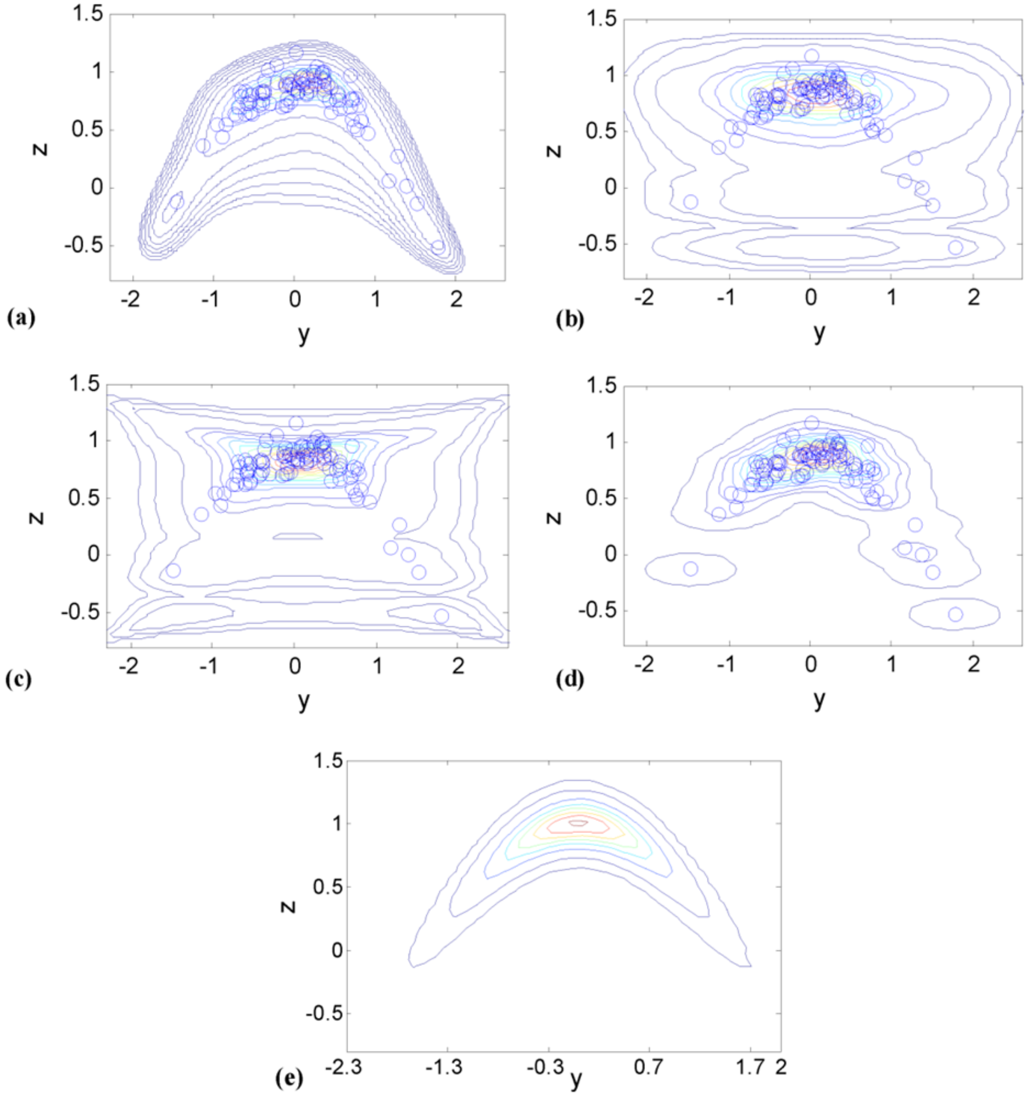

Figure 2. 4: Contour plots of estimated joint PDFs of (Y, Z) and samples shown by the small circles. a) MLME PDF estimated using a 7th-order truncated Taylor series. (b) PDF estimated using Gumbel copula. (c) PDF estimated using Student’s t copula. (d) PDF estimated using Gaussian kernel. (e) True PDF.

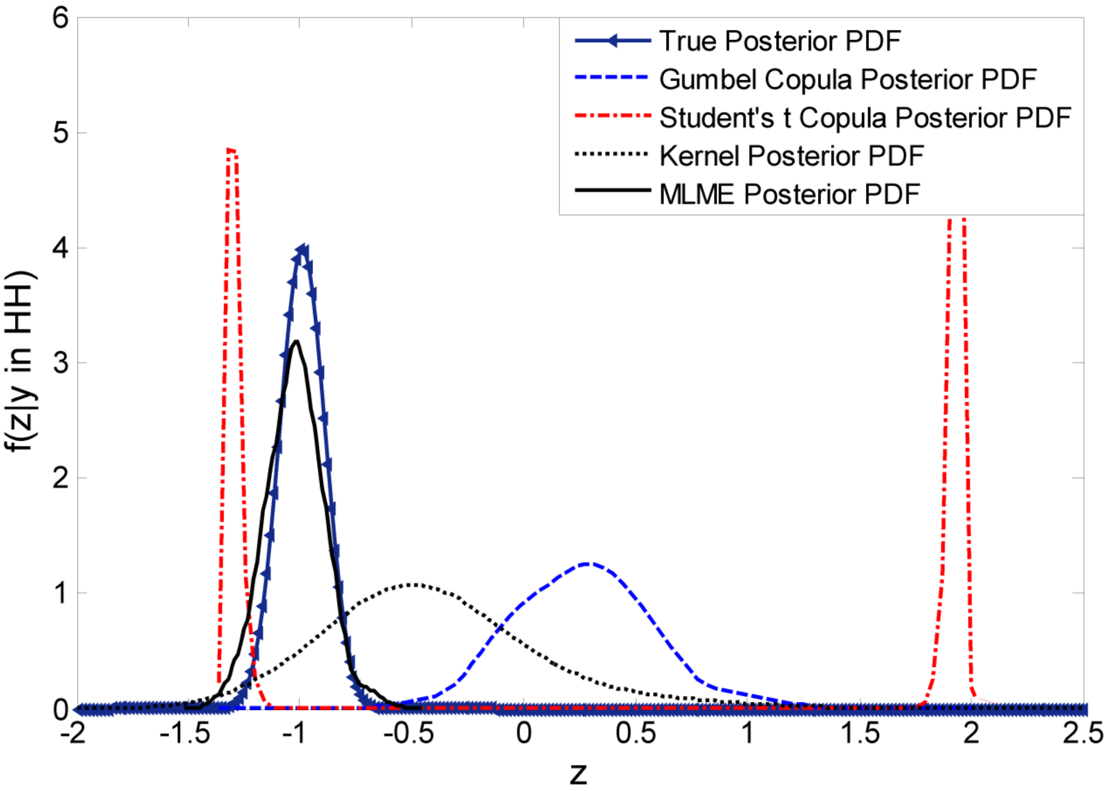

Figure 2. 5: Comparison of posterior conditional PDFs of the random variable *Z* given its parent (*Y*) in its High-High (*HH*) state, when no data in the state *HH* provided by the historical dataset.

2.5.2. Example 2: A Process Example

Consider the stirred heating tank shown in Figure 2.6. A steady-state first-principles mathematical model of the process is:

$$\rho(F_i - F_o) = 0 \tag{2.29}$$

$$\rho C_P\big(F_i(T_i - T_r) - F_o(T_o - T_r)\big) + Q + \epsilon_1 = 0 \tag{2.30}$$

$$F_{out} = \frac{h^{1/2}}{R} + \epsilon_2 \tag{2.31}$$

PDFs of the root nodes (independent variables) of this process and the two noise signals are given in Table 2.4. This first-principles model is used to extract the causal relations among the variables to construct a Bayesian network, 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. Figure 2.7 shows the Bayesian network 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)

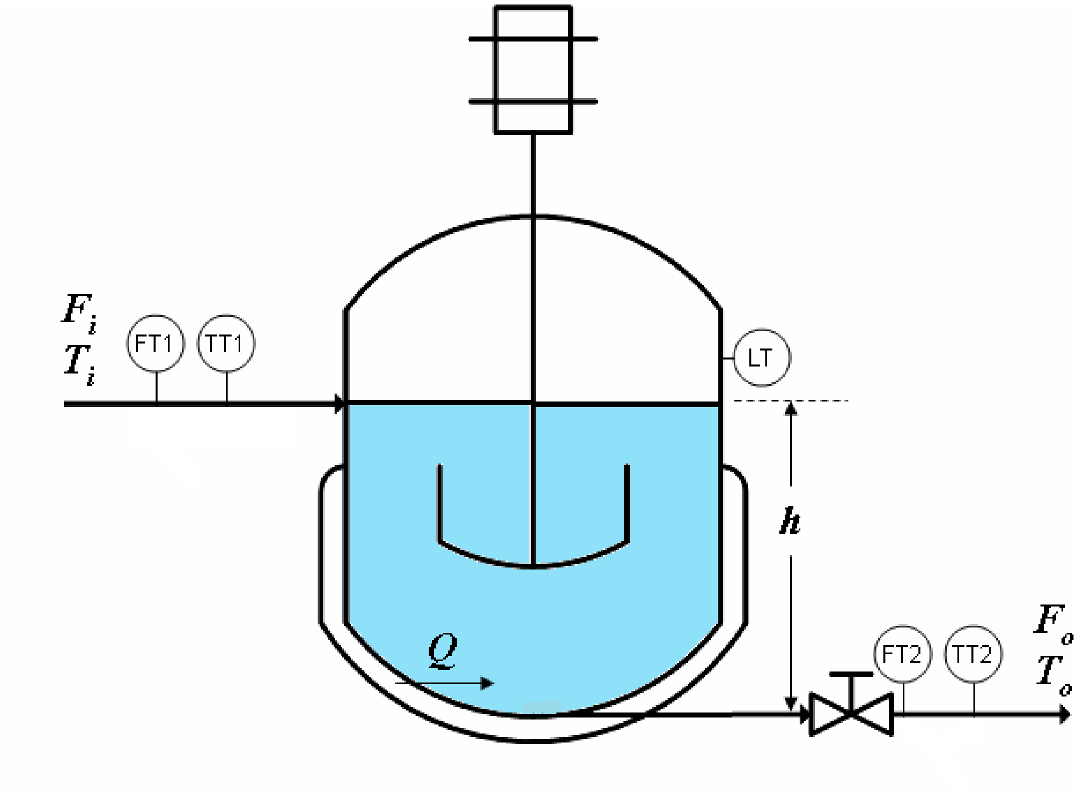

Figure 2. 6: Schematic of the heating tank example.

Table 2. 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	$Normal(0,4 \times 10^{-8})$
ϵ_2	$Normal(0,0.25)$

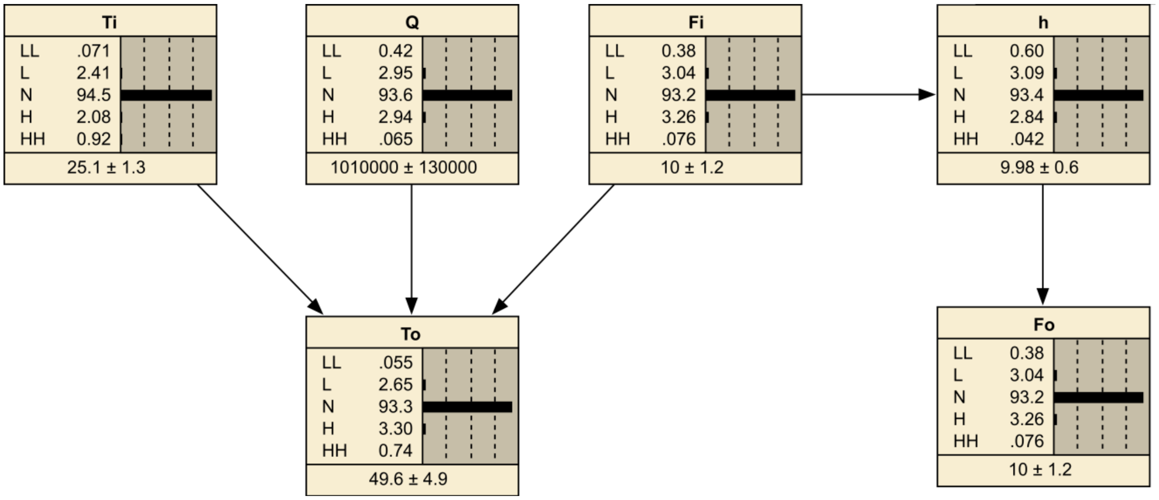

Figure 2. 7: Bayesian network of Example 2 trained using complete PDFs estimated using the MLME method to cover the extreme states, *LL* and *HH*, as well. The shown probabilities are normal operation probabilities.

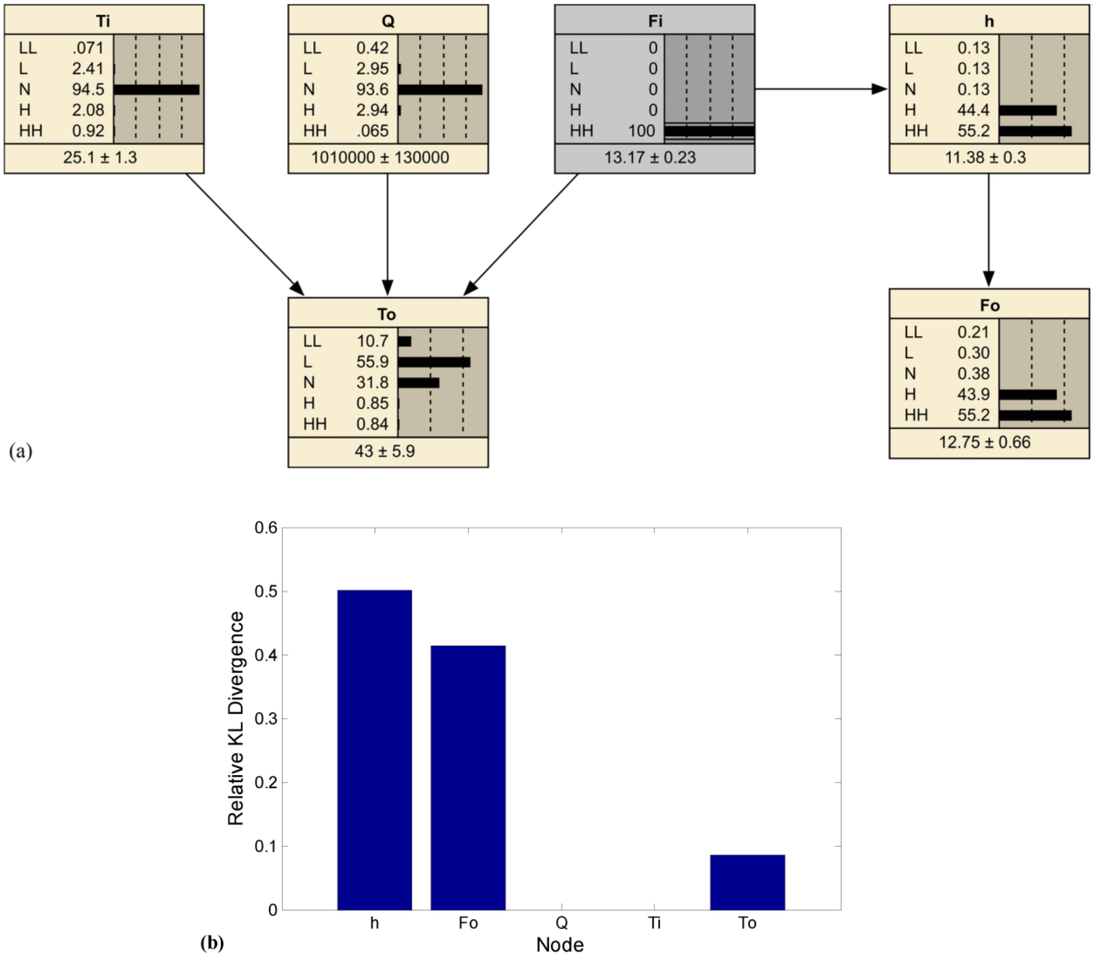

Figure 2. 8: (a) Bayesian network of Example 2 showing updated (posterior) probabilities when evidence F_i in *HH* was given to the network. (b) RKLD values of the five nodes.

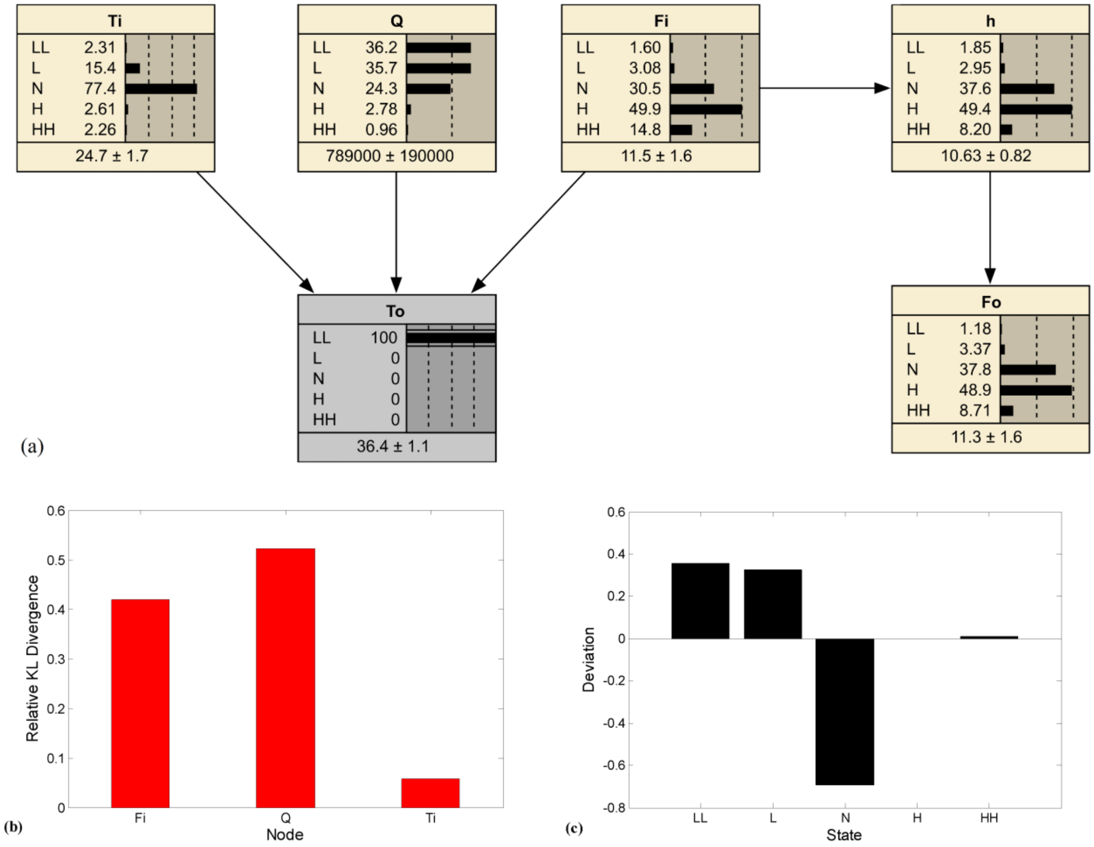

Figure 2. 9: (a) Bayesian network of Example 2 showing updated (posterior) probabilities when evidence T_o in *LL* was given to the network. (b) RKLD values of the three root nodes. (c) Differences between posterior and prior probabilities of the most-likely-cause root node, *Q*.

states. All states are defined according to the rule stated in Section 2.2. 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 Bayesian network. 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 Bayesian network*, the inference is called a “predictive” inference. On the other hand, 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 Bayesian network*, 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 (RKLD)[63] that is applicable to both continuous and discrete random variables and to individual probability values as well. For a node $\mathrm{X_j}$, the RKLD is defined as:

$$\mathrm{RKLD}_{X_j} = \frac{KLD_{X_j}}{\sum_{j=1}^{w} KLD_{X_j}} \tag{2.32}$$

where

$$\mathrm{KLD}_{X_j} = \sum_{i=1}^{r} P_{X_{j,i}} \log\left(\frac{P_{X_{j,i}}}{Q_{X_{j,i}}}\right) \tag{2.33}$$

where $P_{X_{j,i}}$and $Q_{X_{j,i}}$are the prior and posterior probabilities of the i-th 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 0, its corresponding term in KLD expression is 0, since $0 \times \log(0) = 0$.

2.5.2.1. 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 off-line and on-line. 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 observed evidences. This information can be used immediately to take a series of preventing actions leading to loss reduction.

Figure 2.8a shows the Bayesian network of Example 2 with updated (posterior) probabilities when the inlet flow is at its HH state, and Figure 2.8b 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.

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

$$d_{X_{mlc,i}} = Q_{X_{mlc,i}} - P_{X_{mlc,i}}, \qquad i = 1, \dots, m \tag{2.34}$$

where $d_{X_{mlc,i}}$, $Q_{X_{mlc,i}}$ and $P_{X_{mlc,i}}$ 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 2.9a depicts Bayesian network 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 2.9b indicate that the most-likely-cause root node is Q. Figure 2.9c 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

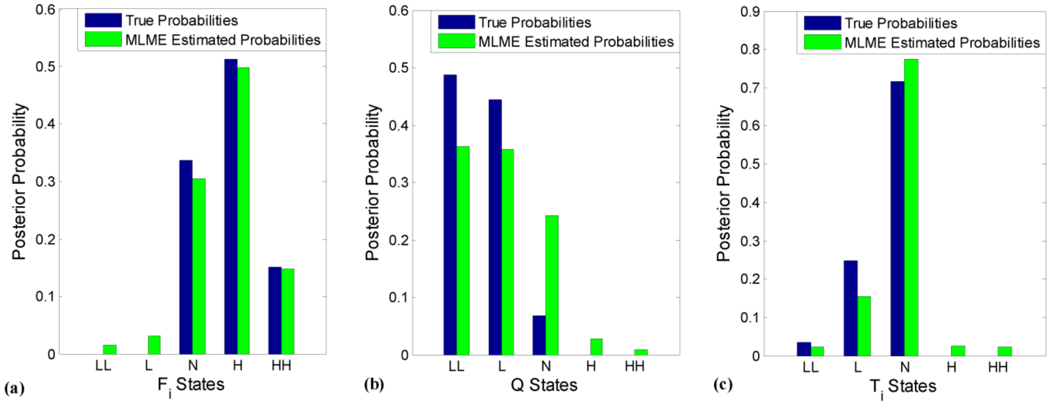

Figure 2. 10: Diagnostic Bayesian inference with 1,000,000 samples and with the MLME estimated network. (a) Posterior probability distribution of $\boldsymbol{F_i}$, (b) posterior probability distribution of $\boldsymbol{Q}$, and (c) posterior probability distribution of $\boldsymbol{T_i}$.

affect the outlet temperature, T_o, but they do not have equal contributions to the changes observed in the outlet temperature. As Figure 2.9a 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 2.9c. Figure 2.10 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 data sets 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.

2.6. Conclusions

The problem of rare-event probability estimation was studied. A moment-constrained, maximum-likelihood, maximum-entropy method of multivariate PDF estimation was proposed. This method is superior to other widely used approaches such as copula densities and non-parametric kernel methods because it applies when relations among the variables are non-monotonic. 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 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 non-parametric 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 non-parametric 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.

References

1. Qin, SJ. Statistical process monitoring: basics and beyond. Journal of Chemometrics. 2003;17:480-502.

2. Mahadevan S, Shah SL. Fault detection and diagnosis in process data using one-class support vector machines. Journal of Process Control. 2009;19:1627-1639.

3. Chiang LH, Braatz RD. Process monitoring using causal map and multivariate statistics: fault detection and identification. Chemometrics and Intelligent Laboratory Systems. 2003;65: 159-178,

4. Castillo I, Edgar TF, Dunia R. Nonlinear model-based fault detection with fuzzy set fault isolation. IECON 2010 - 36th Annual Conference on IEEE Industrial Electronics Society. 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. Computer Aided Chemical Engineering. Elsevier, 2010;28:175-180.

7. Zhang Y, Qin SJ. Improved nonlinear fault detection technique and statistical analysis. AIChE Journal. 2008;54 (12):3207-3220.

8. Mehranbod N, Soroush M, Panjapornpon C. A method of sensor fault detection and identification. Journal of 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 Journal. 2007;5:3 654-668.

10. Ohran B, Muñoz de la Peña D, Davis JF, Christofides PD. Enhancing_Data-Based Fault Isolation through Nonlinear Control. AIChE Journal. 2008;54:223-241.

11. Taleb NN. The Black Swan: the Impact of the Highly Improbable (2nd edition). 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: Shane G, Henderson, Barry LN. Handbooks in operations research and management science. Elsevier, 2006;13:291-350.

16. Berryman JT, Schilling T. Sampling rare events in nonequilibrium and nonstationary systems. The Journal of Chemical Physics. 2010;133:244101.

17. Hiemstra C, Kelejian HH. A rare events model: Monte Carlo results on sample design and large sample guidance. Economics Letters. 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. 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. Operations Research. 2006;54:489-504.

22. Dean T, Dupuis P. Splitting for rare event simulation: A large deviation approach to design and analysis. Stochastic Processes and their Applications. 2009;119:562-587.

23. Shortle JF. Efficient simulation of blackout probabilities using splitting. International Journal of Electrical Power & Energy Systems. 2013;44:743-751.

24. Campillo F, Rakotozafy R, Rossi V. Parallel and interacting Markov chain Monte Carlo algorithm. Mathematics and Computers in Simulation. 2009;79:3424-3433.

25. Kaynar B, Ridder A. The cross-entropy method with patching for rare-event simulation of large Markov chains. European Journal of Operational Research. 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. Proceedings of the 2005 Winter Simulation Conference. 2005:682–691.

27. QIU Y, ZHOU H, WU Y. An importance sampling method with applications to rare event probability. 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. Encyclopedia of Physical Science & Technology (3rd edition). New York: Academic Press, 2001:155-170.

31. Aldrich J. R. A. Fisher and the making of maximum likelihood 1912–1922. Statistical Science. 1997;12:162–176.

32. Lauritzen L, Spiegelhalter J. Local computations with probabilities on graphical structures and their application to expert systems (with discussion). Journal of the Royal Statistical Society. 1988;50: 157-224.

33. Koski T, Noble JM. Bayesian networks: an introduction. Hoboken: Wiley, Inc., 2009.

34. Korb K B, Nicholson AE. Bayesian artificial intelligence (2nd edition). Boca Raton: CRC Press, 2010.

35. Daly R, Qiang S, Stuart A. Learning Bayesian networks: approaches and issues. The Knowledge Engineering Review. 2011;26:99-157

36. Casella G, Berger RL. Statistical Inference (2nd edition). 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. Journal of Mathematical Analysis and Applications. 1984;98:435–457.

39. Koralov L, Sinai YG. Theory of probability and random processes (2nd edition). Berlin: Springer, 2012.

40. Shannon CE. Prediction and entropy of printed English. The Bell System Technical Journal. 1951;30:50–64.

41. Shannon CE. A mathematical theory of communication. The Bell System Technical Journal. 1984;27:379–423.

42. Zellner A, Highfield AR. Calculation of maximum entropy distribution and approximation of marginal posterior distributions. Journal of 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. Annals of Statistics. 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. Journal of the American Statistical Association. 1966;61:1052-1062.

48. Schmidt DF, Makalic E. Universal Models for the Exponential Distribution. IEEE Transactions on Information Theory. 2009;55:3087–3090.

49. Chris DO, Paul AR. On the uniqueness of the maximum likelihood estimator. Economics Letters. 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. Journal of Clinical Epidemiology. 2004;57:1262-1270

51. Benedikt M, Pötscher HL. On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD, and thresholding. Journal of Multivariate Analysis. 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: Proceedings of the Workshop Held in Warsaw 2009. Heidelberg: Springer, 2010.

54. Li Q, Racine JS. Nonparametric Econometric Methods (advances in Econometrics). Emerald Group Publishing Limited, 2009.

55. Netica (v. 4.16) [Software], Norsys Sofware Corp., 2010.

56. Nelsen RB. An introduction to copulas. Volume 139 of 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. Handbook of heavy tailed distributions in finance. Elsevier, 2003:331-385.

58. McNeil AJ, Nešlehová J. From Archimedean to Liouville copulas. Journal of Multivariate Analysis. 2010;101:1772-1790.

59. Kolev N., Paiva D. Copula-based regression models: A survey. Journal of Statistical Planning and 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 of Probability and its Applications. 1969;14:153–158.

62. Duong T, Hazelton ML. Convergence rates for unconstrained bandwidth matrix selectors in multivariate kernel density estimation. Journal of Multivariate Analysis . 2005;93:417–433.

63. Kullback S. Letter to the Editor: The Kullback–Leibler distance. The American Statistician. 1987;41:340–341.

Chapter 3: Estimation of Complete Discrete Multivariate Probability Distributions from Scarce Data with Application to Risk Assessment and Fault Detection

3.1. Introduction

Risk assessment usually refers to a set of analyses that identify potential hazards and evaluate possible consequences of the hazards if they occur.[1] It involves estimating (a) the likelihoods of different possible risky situation scenarios and (b) the costs associated with the risks. More specifically, risk assessment includes simultaneous failure cost estimation, development of realistic fault scenarios, and quantification of risk probabilities.

Most of the current risk assessment and fault detection schemes have focused on abnormal situations with high probabilities and moderate costs,[2] whereas a major fraction of catastrophic and large scale incidents with highly destructive consequences are caused by some triggering events whose probabilities had been found infinitesimal by risk assessment. This class of abnormal events are usually referred to as “rare events”,[3] which are of two major types: (a) those that are so rare and far-fetched that their probabilities may be considered to be practically zero and (b) those that are actually predictable but show a minor recurrence frequency compared to the plant’s expected lifetime. An example of the first type is industrial plant destruction due to a meteor hitting the plant, and an example of a rare event of the second type is a control system failure. Throughout this chapter we simply use the term “rare events” for those of the second type.

Although the probabilities of the rare events may be predictable and thus the negative impacts of their consequences can be reduced, modern day industrial establishments are still suffering from the resulting catastrophes for two major reasons. First, the probabilities of rare events are usually underestimated intentionally or inadvertently. This underestimation eventually leads to a false assurance that the associated risks are negligible as well. Because of such a perception, the associated risks are not taken very seriously, and rudimentary precautionary schemes are used to mitigate the risks. Second, the estimation of the probabilities of events that have seldom or never happened, observed or recorded in the course of plant operation carries a great deal of uncertainty in its outcome. This estimation uncertainty is mainly due to the lack of a general method of integrating the rarity and “extrapolation into the future” of sample realizations. Consequently, the estimates are hardly useful in practice. Hence, an open problem is reliable estimation of probabilities that are unknown, infinitesimally small, and hard to predict. This problem becomes even more challenging and at the same time more interesting when it comes to studying the complex failure scenarios, where the rare event simulation is not simply to determine the failure probability of an individual component, e.g. a pump, but calculating the probability of a series of subsequent failures, when due to complicated interactions between components fault can propagate and finally lead to a catastrophe.

Estimating probability of rare events has been under active research in the past decade. In cases where an accurate plant model is available, most research has focused on sampling from the model. Methods such as Markov Chains Monte Carlo,[4] importance sampling,[5] and splitting[6] have been employed extensively to calculate probability distributions to identify abnormal situations. However, in cases where a reliable model is not available, especially one that accounts for uncertainties in the system, sampling methods fail to provide a thorough representation of the system's behavior. In such cases, probability estimation methods have been developed to reconstruct probabilities of possible random events from the data. The most primary method of this type is the histogram method,[7] which itself belongs to the non-parametric probability estimation[8] group. Parametric methods[9], on the other hand, have also been employed widely to estimate probabilities by considering a parametric family behind the observed data. In this context, sophisticated probabilistic structures such as copula densities[10] and moment based probabilities[11] have been proposed to add maximum flexibility to the estimated models. Nevertheless, despite all advantages of probability density estimation techniques such as their purely data-based framework, they still suffer from high computational cost (as of the non-parametric method of kernel[12]) or lack of extendibility to the general dependence structure observed in the data (as of conventional copula methods).[13]

Once complete probability distributions were estimated, then one can conduct (a) prediction to assess risk and (b) inference to perform fault detection and identification using Bayesian networks.[14] Alternative fault detection and identification methods use Kalman filtering,[15] principle component analysis,[16] fault and event trees,[17] artificial neural networks,[18] fuzzy logic based modeling,[19] or other concepts.[20]

In this chapter we propose a method of estimating discrete multivariate probability distributions from scarce historical data with a special attention to the states with no observations (rare states). The method is based on a constrained maximization of the information entropy function. It considers information coming from every individual sample points in the form of sample moments, which provides a framework for maximum use of information encoded in the data. Such a model will further be applied to estimating parameters of Bayesian networks and eventually provide a stochastic modeling framework for risk analysis and fault detection under rare event regime. Unlike traditional approaches to Bayesian network parameter estimation using the local relative frequency technique to estimate probabilities, the method incorporates all information presented by finite datasets to set up discrete multivariate probability distributions extendable to unobserved regions. With such probability distributions, the calculation of unknown and near-zero probabilities becomes possible and much faster than sampling from the first principles models. Furthermore, the method is able to model nonlinear and non-monotonic relations with an optimal level of model's complexity. Moreover, combination of the proposed method with Bayesian networks provides an important tool in modeling and calculating the probability of multilevel risk scenarios, where due to the causal interrelationships between the process components failure can propagate through the system. This work is an extension the method presented in Chapter 2[21] on estimating probability density functions of continuous random variables. We also present two approaches of finding the optimal complexity level of the estimated probability distributions without over-fitting or losing of flexibility. These objectives are met through controlling the likelihood (as a goodness-of-fit measure) with respect to the model's level of complexity.

The rest of the chapter is organized as follows. Section 3.2 describes the constrained maximum-entropy probability estimation method for discrete random variables. Section 3.3 begins with an example on how rare events are connected to small samples sizes. The probability distribution estimation method is then applied to an example Bayesian network, and the estimated and true probability distributions are compared. Finally, conclusions are made in Section 3.4.

3.2. Method

3.2.1. Entropy Maximization

To estimate the probabilities of unobserved events from the probabilities of observed ones, we combine two concepts widely used in statistical learning[22] to estimate complete probability distributions. The first concept is the information entropy introduced by Claude Shannon[23] as an informatics equivalent of the thermodynamics entropy, which represents disorder. The maximum entropy principle[24] states that every system loses its information content gradually, whether intrinsically (similar to what seen in the nature) or observationally, where the degree of uncertainty (lack of predictability) about the system grows with time from the last available observation. The information entropy of a discrete random variable X, denoted by $S(X)$, is defined as:[25]

$$S(X) = -\sum_{i=1}^{N} P(X = x_i) \ln P(X = x_i) \qquad (3.1)$$

where $P(X = x_i)$ is the probability of $X = x_i$, and N is the number of the states of X (discrete values that X can take).

In information theory, maximization of the entropy function is frequently used to ensure that minimum prior artificial assumptions are included in knowledge-based systems.[26] This procedure leads to minimum bias models. If one tries to estimate a mass probability distribution (PMF) by maximizing the entropy function without imposing any constraints on the shape of the distribution, then the result will trivially be a uniform distribution in which all states have equal probabilities of occurrence. In such a system, the outcome of the random process is absolutely uncertain; that is, there is no outcome that is more likely than the others. However, for every process there is usually some information, no matter how uncertain, that can be used to impose some constraints on the entropy maximization so that a more specific and informative probability distribution can be obtained.

3.2.2. Moments of a Probability Distribution

Given the probability distribution of a discrete random variable X, the theoretical moment of this distribution with respect to a moment function $g_k(X)$ is given by:

$$\mu_k(X) = \sum_{i=1}^{N} g_k(X = x_i) P(X = x_i)\,, \qquad k = 0, \dots, q \tag{3.2}$$

For example, when $\mathrm{g_k(X) = X}$,

$$\mu_k(X) = \sum_{i=1}^{N} x_i P(X = x_i) = E(X) \tag{3.3}$$

where $E(X)$ is the expectation or mean of X, and when $g_k(X) = (X - E(X))^2$, then $\mu_k = E((X - E(X))^2)$ represents the degree of diffuseness of the PMF or variance, and when $g_k(X) = (\frac{X-\mu}{\sigma})^3$, $E\left((\frac{X-\mu}{\sigma})^3\right)$ provides information on the skewness or asymmetry of the distribution, where μ and σ refer to the mean and standard deviation of the distribution respectively.

The definition of moments for the univariate PMFs can be extended to the multivariate ones. Such multivariate PMFs are of particular interest in the current work because of their capability of modeling joint probabilities. Estimation of joint PMFs allows the user to obtain complete conditional probabilities, required to train the

Bayesian networks. In the multivariate case, the moments of a d-dimensional vector of random variables, $\mathbf{X} = [X_1 \dots X_d]^T \in \Lambda$, are given by:

$$\mu_k(\mathbf{X}) = \sum_{i=1}^{N} g_k(\mathbf{X} = \mathbf{x}_i) P(\mathbf{X} = \mathbf{x}_i) \qquad k = 0, \dots, q \tag{3.4}$$

Where Λ denotes the d-dimensional discrete state-space, $g_k(\mathbf{X}): \Lambda \rightarrow \mathbb{R}$ is a moment function, and $P(\mathbf{X}): \Lambda \rightarrow [0,1]$ is a joint PMF. In this multivariate case, the information entropy is given by:

$$S(\mathbf{X}) = -\sum_{i=1}^{N} P(\mathbf{X} = \mathbf{x}^i) \ln P(\mathbf{X} = \mathbf{x}^i)$$

Given historical data on a discrete random vector $\mathbf{X}$, using Eq.(3.4) sample moments can be calculated. The empirical (sample) moment $\bar{\mu}_{\mathrm{k}}$ of the corresponding sampled population is given by:

$$\bar{\mu}_k = \frac{1}{n} \sum_{j=1}^{n} g_k(\boldsymbol{\chi}_j), \qquad k = 0, \dots, q \tag{3.5}$$

where $\boldsymbol{\chi}_j \in \mathbb{R}^d$ represents the $j -$ th sample, and n is the number of the samples. The sample moments are forms of encoded information about the structure of the data. This information is used to estimate probability distributions that govern the samples. If the probability distribution is sought solely based on the sample moments, then the resulting PMF will be the maximum-likelihood estimated PMF.[27] As more moments are included in the form of constraints, more information from the samples is included in the estimated probability distribution.

3.2.3. Constrained Entropy Maximization (ME Method)

To constrain the maximization of the entropy function, a PMF whose moments are the same sample moments, is sought. In other words, given a set of moment functions, we seek optimal model probabilities, $\hat{P}_1, \dots, \hat{P}_N$, which are the solution to the following constrained optimization problem:

$$\max_{\hat{P}_1,\dots,\hat{P}_N} S(\mathbf{X}) = \min_{\hat{P}_1,\dots,\hat{P}_N} \sum_{i=1}^{N} \hat{P}_i \ln \hat{P}_i \qquad (3.6)$$

subject to:

$$\sum_{i=1}^{N} g_k\big(\mathbf{X} = \mathbf{x}^i\big)\hat{P}_i = \bar{\mu}_k, \qquad k = 0, \dots, q \qquad (3.7)$$

where $\hat{P}_i = \hat{P}\big(\mathbf{X} = \mathbf{x}^i\big) \in [0,1]$ and $g_0(\mathbf{X}) = 1$. Therefore, given the moment functions, the constrained optimization problem of Eqs.(3.6) and (3.7) is a conventional nonlinear program, which is easy to solve numerically, preferably using a global optimization technique to avoid possible local optima.[28]

3.2.3. Selection of the Moments

To estimate PMFs reliably, it is essential to select appropriate moment functions. These moment functions provide the PMFs with sufficient cumulative information extracted from data. The moment function selection not only can significantly improve the accuracy of the estimation, but it also provides a means to control the computational complexity of the optimization step with minimum information loss due to coarse discretization of distribution functions. As an example, estimation of the Gaussian distribution requires only the first moment (mean) and the second central moment (variance). Additional moments do not provide considerable additional knowledge from the data to arrive at a substantially different estimated distribution. Table 3.1 lists the minimal set of moment functions that are needed to characterize a number of well-known continuous distribution functions. However, since the true distribution that has given rise to observed data is generally unknown, a decomposition technique is used here to approximate the true moment functions underlying the observed samples.

In order to simplify the search for the appropriate moment functions, it is proposed here to search for the moment functions $g_k(\vec{x}) \in \bigcup_{i=0}^{O} \Omega_i$, where O is a positive integer to be selected by the user, and

$$\begin{gathered} \Omega_0 = \{1\}, \\ \Omega_1 = \{x_1, \dots, x_d\}, \\ \Omega_2 = \{x_i x_j |\, i, j = 1, \dots, d\}, \\ \Omega_3 = \{x_i x_j x_k |\, i, j, k = 1, \dots, d\}, \\ \Omega_4 = \{x_i x_j x_k x_l |\, i, j, k, l = 1, \dots, d\} \\ \dots \end{gathered} \tag{3.9}$$

Thus, Eq.(3.7) becomes

$$\sum_{i=1}^{N} g_k(\mathbf{X} = \mathbf{x}^i) \hat{P}(\mathbf{X} = \mathbf{x}^i) = \frac{1}{n} \sum_{j=1}^{n} g_k(\vec{\chi}_j), \qquad \forall g_k(\mathbf{x}) \in \bigcup_{i=0}^{O} \Omega_i \tag{3.10}$$

then with an appropriate O, the appropriate parameter vector $\vec{a}$ of the Taylor series power expansion and Eq.(3.10), the equality between the expectation of the true moment function G and its associated moment M will be achieved. This decomposition indicates that the user just needs to choose an appropriate value for the positive integer O, instead of choosing appropriate moment functions needed in the original formulation of Eq.(3.7). The use of a higher O may seem to provide higher accuracy at the first look. However, in general, because (a) the use of a higher O imposes higher computational costs and (b) more complex moment functions often fail to predict the actual probability behavior

Table 3.1: Some exponential probability distribution functions and their characteristic moments.

Distribution	Moments	Density Function
Exponential	$\int xf(x) = \mu$	$f(x) = \frac{1}{\mu}\exp\,(\frac{-x}{\mu})$
Gaussian	$\int xf(x) = \mu\,, \int (x-\mu)^2 f(x) = \sigma^2$	$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\,(\frac{-(x-\mu)^2}{2\sigma^2})$
Beta	$\int \ln x\, f(x) = \frac{\Gamma'(a)}{\Gamma(a)} - \frac{\Gamma'(a+b)}{\Gamma(a+b)},$ $\int \ln(1-x)\, f(x) = \frac{\Gamma'(b)}{\Gamma(b)} - \frac{\Gamma'(a+b)}{\Gamma(a+b)}$	$f(x) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}$
Gamma	$\int xf(x) = a, \int \ln \mathrm{x}\, f(x) = \frac{\Gamma'(a)}{\Gamma(a)}$	$f(x) = \frac{x^{a-1}\exp\,(-x)}{\Gamma(a)}$
Weibull	$\int x^a f(x) = 1,\ \int \ln x\, f(x) = -b/a$	$f(x) = \ ax^{a-1}\exp\,(-x^a)$

outside the range of the data (due to over-fitting),[29] one should use an adequately large O, as suggested by Occam's razor principle.[30] Hence, there is a tradeoff between informationloss in lower- order moment functions and high variance of higher-order ones, particularly when prediction of probabilities outside the observed region is intended. In view of these, the lowest level of complexity, O, which satisfies an error tolerance threshold, should be chosen. To find an optimal O systematically, we propose two methods that consider the tradeoff between bias and variance of the estimator.

Maximum Likelihood Estimation of the Truncation Orders

Likelihood of a parameter O given a data base **D** is simply defined as conditional probability of **D** given O or

$$L(O|\mathbf{D}) = P(\mathbf{D}|O) \tag{3.11}$$

where $L(.|.)$ and $P(.|.)$ are the likelihood function and conditional probability, respectively. Therefore, to calculate the likelihood function, conditional probabilities must be available. Such a definition is the basis of the maximum likelihood estimation.[27] The likelihood function indicates how well the observed data samples are described by the parameters, O.

$$L(O|\mathbf{D}) = P(\mathbf{D}|O) = \prod_{i=1}^{N} \hat{P}_{i,o}{}^{n_i} \tag{3.12}$$

where $\hat{P}_{i,o}$ denotes the model-prediction of the probability of state i using of the moment functions up to order O, and n_i represents the number of data points in the $i-$th state. Note that $\sum_{i=1}^{N} n_i = n$.

Similar to the behavior observed in the case of mean square error and the bias-variance tradeoff[31] as the complexity level of a model increases, the model fit the data better and the likelihood function increases. These trends continue up to a certain complexity level beyond which these trends reverse; beyond this level of complexity (here, O) the likelihood of the data (as measure of accuracy of the model) decreases but the computational cost increases. The value of O that yields the best fit is called the maximum likelihood estimate (MLE) of the parameter O:

$$O_{MLE} = \arg\max_{q} \; P(\mathbf{D}|O) \tag{3.13}$$

which agrees with Occam's razor principle. However, since this maximum occurs at high orders of O while showing no significant increase through a wide range of lower values of O, user may decide to select a lower order O which satisfies some minimal goodness-of-fit criterion while keeping the computations more tractable.

Maximum a Posteriori Estimation of the Truncation Orders

If the Bayes rule is used to relate the likelihood and a priori probability over the model's complexity, one can setup a framework to incrementally update our belief about the complexity level. Unlike the MLE, which defines a point-wise estimation, the Bayesian model selection provides a distribution for the complexity level; i.e., we can derive confidence intervals for our parameter, in addition to other statistical characteristics. Using the Bayesian model averaging[32] we obtain

$$P(X = x_i|\mathbf{D}) = \sum_{O=1}^{O_{max}} P(X = x_i|\mathbf{D}, O)\, P(O|\mathbf{D}) \tag{3.14}$$

in which O_{max} stands for the maximum truncation order when equal orders for all truncations are used. Eq.(3.14) allows us to average over different complexity levels to derive a distribution for $P(X|\mathbf{D})$. However, it is oftentimes not possible to calculate this sum. As a general solution, we approximate $P(X|\mathbf{D})$ by

$$P(X|\mathbf{D}) \simeq P(X|\mathbf{D}, O_{MAP}) \quad (3.15)$$

where

$$O_{MAP} = \arg\max_{O} \; P(O|\mathbf{D}) = \arg\max_{O} \; \frac{P(\mathbf{D}|O)P(O)}{P(\mathbf{D})} = \arg\max_{O} \; P(\mathbf{D}|O)P(O) \quad (3.16)$$

The right hand side equation is based on Bayesian belief updating. The parameter O is also known as complexity controlling parameter. $P(O)$ denotes the prior probability of $O.$ Since the likelihood function, $P(\mathbf{D}|O)$, as stated by Eq.(3.12), does not have a closed form in general, setting up a conjugate prior for the likelihood function is not possible. However, we can still assign an informative prior, for example a normal distribution with a zero mean and some positive number as the variance. As more information is incorporated into this function through the likelihood term, the updated belief about O approaches its true value. If the mode of this posterior is used as our point estimate of O, this estimation is called maximum a posteriori (MAP) estimation. Eq.(3.16) implies if a uniform distribution is used as the prior, MLE and MAP estimates indicate the same result for O. In the next section we apply these concepts and algorithms to an example Bayesian network.

3.3. Application to an Example

To demonstrate the performance of the PMF estimation method, we apply the method to a plant with five variables governed by:

$$X \sim \text{PMF1} \quad (\text{Figure} \quad 3.1\text{a}) \quad (3.17)$$

$$Y \sim \text{PMF2} \quad (\text{Figure} \quad 3.1\text{b}) \quad (3.18)$$

$$Z = X^{1/2} + \epsilon(0,0.1) \quad (3.19)$$

$$U = \log(Z) + \epsilon(0{,}0.03) \tag{3.20}$$

$$W = \frac{YZ}{Z+1} + \epsilon(0{,}0.2) \tag{3.21}$$

where $\epsilon(0, b)$ is a white noise with a variance of b. PMF1 and PMF2 are shown in Figures 3.1a and 3.1b, respectively. The white noise represents internal process uncertainty. The nonlinearities are added to increase the problem's complexity and make the estimation problem more challenging. Figure 3.2 shows a Bayesian network representation of the plant example. This structure includes three types of causal structures: common cause, common effect, and chain causation relationships.

Figure 3.1 shows the true underlying distributions of the nodes. To simulate scarce information condition, we take only 100 random samples from the root nodes (X and Y) probability mass function, as shown in Figures 3.1a and 3.1b. The 100 samples are taken simply by using a pseudo-random number generator. First, a number π is picked up randomly from a uniform distribution defined on the support of (0,1), $\mathrm{U}(0{,}1)$. This number plays the role of a cumulative probability value. π is then transformed to the original random variable (X or Y) space by utilizing its inverse cumulative probability function, F^{-1}. For example for random variable X,

$$\chi_i = F_X^{-1}(\pi_i) \tag{22}$$

where χ_i and $\pi_i$are the i-th random numbers taken from the distribution of X and $\mathrm{U}(0{,}1)$ respectively. F_X^{-1} and F_Y^{-1}can be derived from the discrete probability distributions defined in Figures 3.1a and 3.1b. Figure 3 illustrates this approach. The corresponding values for the child nodes are then calculated. This calculation can be either performed using the true conditional probability tables of the network (if available) or by generating samples from the set of governing equations and constructing the corresponding discrete probabilities. Randomly selected samples from X and Y are converted to Z, U and W using their related class values where seven preset states are used for discretization of these child nodes (Figures 3.1c, 3.1d and 3.1e). This set of 100 samples constitutes the basis for our probability distribution estimation.

The first step in training our Bayesian network to properly do inference under the rare event regime is to complete marginal PMFs for the root nodes and conditional probabilities for the rest of the network. Although these tasks share the same theoretical background to be implemented, computing the conditional probabilities, as implied by our generative statistical approach, requires to first calculating discrete joint PMFs. These multidimensional arrays of multivariate probabilities must further be converted to conditional probability tables of the child node by being divided by the marginal probabilities of the parent nodes.[36]

As clearly shown in Figure 3.1, when the number of samples is not large enough, the so-called rare states are not likely enough to appear in the historical data. This fact is particularly in accordance with the vector form of Chebyshev's inequality[33] ,

$$P(\left|\left|\vec{V}-\vec{\mu}\right|\right| \geq \delta\left|\left|\sigma\right|\right|) \leq \frac{1}{\delta^2} \tag{3.23}$$

where $\vec{V}$ is a d -dimensional random vector with mean $\vec{\mu}$ and $\sigma^2 = ({\sigma^2}_1, \dots, {\sigma^2}_d)$ is the variance vector. $|| \;||$ is the vector's norm. This relation states that the majority of data are close to the mean of the distribution. More precisely, in a general probability distribution, the probability of a random number being equal to δ standard deviations away from the mean of the number is less than or equal to $\frac{1}{\delta^2}$. The following example helps to explain this.

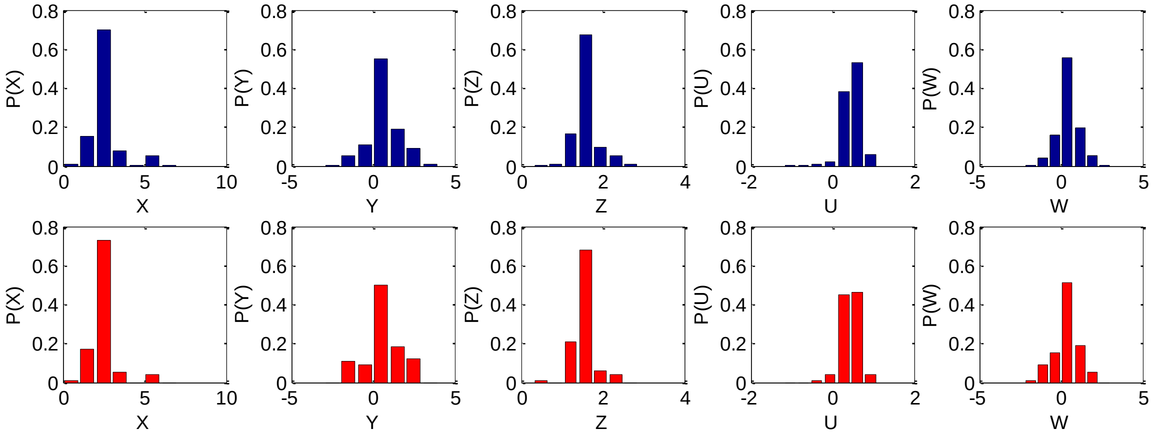

Figure 3.1: Actual probability distributions (upper row) and probability distributions based on randomly selected 100 samples (lower row).

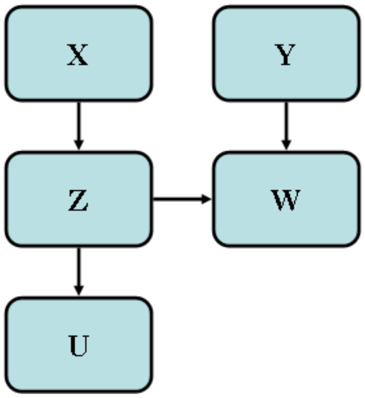

Figure 3.2: Bayesian network of the example.

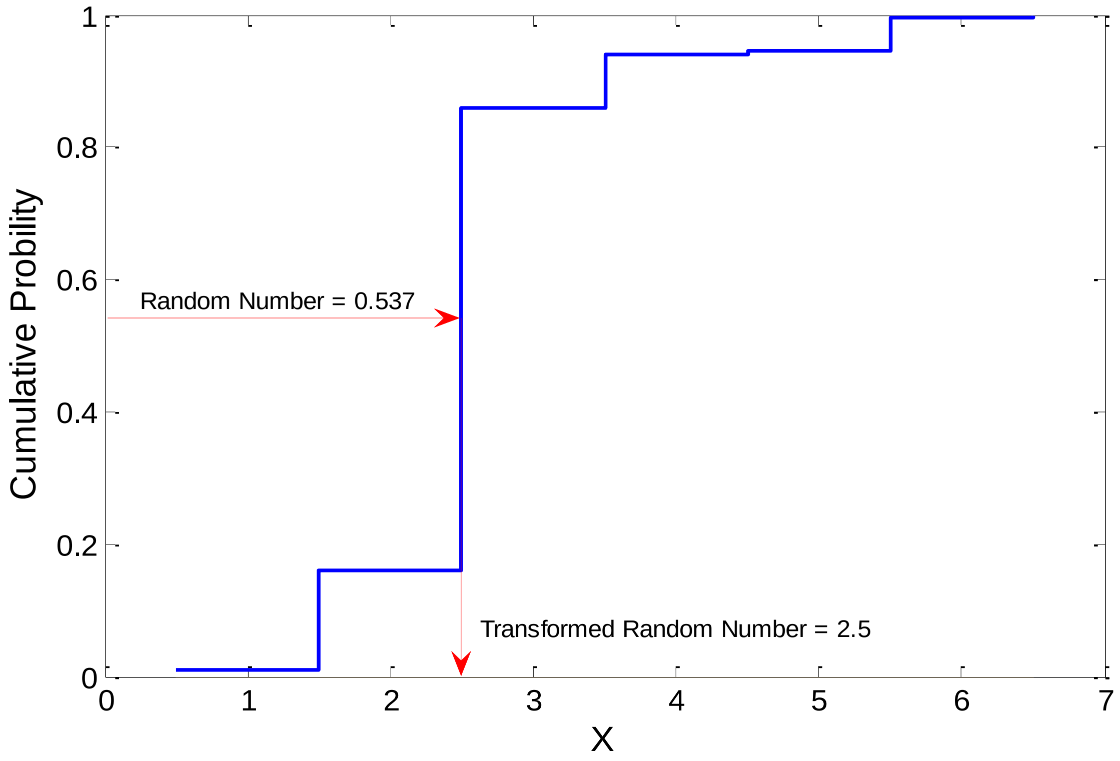

Figure 3.3: random number generation from a given cumulative distribution function.

To see the effect of the sample size on observing the data from low probability regions, consider a one-dimensional case where we are interested in estimating the probability of obtaining at least one sample beyond 13 standard deviations or $\sqrt{13}$ variance far from the mean of a general univariate distribution in 100 trial, named event E here. The probability of getting such a sample in one trial, called $e,$ is:

$$P(e) = P(|X-\mu| \geq 13\sigma) \leq \frac{1}{169}$$

If the maximum value of $\frac{1}{169}$ is taken as $P(e),\ P(E)$ is defined by a multinomial distribution:

$$P(E) = 1 - P(E^c) = 1 - \frac{100!}{100!0!}\left(1 - \frac{1}{169}\right)^{100} = 0.448$$

where E^c is the complement of event E; that is none of the samples are observed within 13 standard deviations from the mean. This result suggests that, even by taking the supremum of the inequality above, the probability of observing event e in 100 samples is less than it not having been observed. For elliptical distributions such as normal distribution this probability is even smaller. Conversely, if we use 1,000 samples $P(E)$ will reach 0.997. This example shows that with an inadequate sample size, some possible states that possess an infinitesimally non-zero probabilities are not visited in the data at all; therefore their probabilities are considered to be empirically zero by traditional statistical approaches. As suggested by Chebyshev’s inequality, for random vectors this situation is even worse, and this makes it impossible to train complete arrays of multivariate PMFs.

In fault detection applications, Bayesian inference should be feasible for all possible states, including the rare states; therefore such zero empirical probabilities are problematic. For instance, assume that we intend to perform Bayesian inference (backward or forward) when a rare state is introduced to the network. When conditional probabilities of other states given the rare state are undetermined, the inference cannot be made.

Estimating Probabilities

To address the important issue raised above, the method presented in the previous section is first applied to the univariate root nodes of the example system, i.e., X and Y. Figures 3.1a and 3.1b compares the true and data-based distributions of random variables X and Y. For an efficient estimation of rare states probabilities we initially need to apply the method with different values of O. The value, which gives rise to either maximum likelihood of the model's complexity (MLE) or maximum likelihood of the data (MAP), is used to estimate the model's parameters (probabilities).

Figures 3.4 and 3.5 compares the MLE and MAP estimates for the model's complexity of X and Y, respectively. Obviously because of applying a non-conjugate prior normal distribution $P(O)$ with mean 2, O_{MAP} is smaller than O_{MLE}. This can be thought of as the effect of our prior assumption that a simple second degree model can generally be a good fit for the many elliptical distributions, as suggested in X and Y. This prior assumption then is updated when additional information from the data is incorporated in the Bayesian parameter estimation approach. As a consequence of multimodality of the true PMF of X, higher orders of moment functions (complexity) must be employed to model X data compared to that of Y. Figures 3.6a and 3.6b compare MLE and MAP estimates of X and Y with their true PMFs.

An important fact to note here is that the likelihood functions stay constant for a wide range of O. This suggests that for the cases that computational tractability of the algorithm is the limiting factor, the lowest order that satisfies some likelihood threshold can be used to model the probability mass function.

A similar approach is utilized to estimate the joint and conditional probability tables. First, the joint PMF of each set parents and child nodes are estimated with the same procedure outlined in Eqs.(3.6) and (3.7) and using the multivariate moments and moment functions with the form as of Eq.(3.9), Then normalizing the probability array of the child node given the states of its parents will lead to the entire set of conditional probabilities required by the network. For example

$$P(U|Z = z) = \frac{P(Z = z, U)}{P(Z = z)}$$

Figure 3.7 illustrates a comparison of the true and estimated conditional probabilities of W given Z and Y with the ME estimated probabilities. It can be observed that the complete set of probabilities are reconstructed by constrained maximum entropy method, using the information collected from entire dataset, over the states where initially considered to have zero probabilities. Figure 3.8a shows the Bayesian network trained using constrained maximum entropy method. We use Netica[34] software for Bayesian analysis and network visualization.

Confidence Intervals for the Estimated Probabilities

Since the solution to the nonlinear programming of Section 3.2.3 results in point estimates of the probabilities for any given dataset, it cannot be used to find the confidence intervals of the estimated probabilities. To derive confidence intervals for the estimated probabilities, here we use a resampling method called the Jacknife or (leave-one-out procedure) described in[35] . In this method, a distribution for the estimated probability is calculated by performing an optimization procedure multiple times, each time with a different sample set derived by systematically leaving one of the sample points out. Each resampled data give rise to a different estimate for the probabilities of X's states. The mean and variance of the estimated probabilities is then found by

$$\bar{P} = \frac{1}{n}\sum_{j=1}^{n} P_j \tag{24}$$

$$\mathrm{Var}(P) = \frac{n-1}{n}\sum_{j=1}^{n}(P_j - \bar{P})^2 \tag{25}$$

where $\bar{P}$ and $\mathrm{Var}(P)$ refer to the mean and standard deviation of the estimated probability $\bar{P}$ and P_j is the estimated value of P using the dataset without the j-th observation. Characteristic statistical parameters of these distributions (mean and standard deviations) are listed in Table 3.2. To calculate these quantities, the original dataset of 100 samples are used. Each distribution indicates how reliable the estimated parameters are. Assuming the estimator is unbiased, that is $\bar{P}$ is equal to the true value of P, the narrower distributions (smaller relative standard deviation) indicate that the mean value suggested by the distribution is more likely than the actual value of the parameters under investigation. It can be shown that such narrow distributions are associated with larger sample size; however, the achievement of a narrow distribution with relatively small sample size can be a sign of the consistency of the estimation method; that is, with increasing the sample size the estimated parameter converge to its true value. It should be noted that in this research the distributions are defined on a bounded support, as the

Table 3.2: Mean and standard deviations of the estimated probabilities for the random variable X.

Parameter	Mean	Standard Deviation
P_1	0.0104	2.8053x10^{-7}
P_2	0.1676	0.0057
P_3	0.7320	0.0062
P_4	0.0460	0.0013
P_5	0.0060	1.9348x 10^{-6}
P_6	0.0376	0.0023
P_7	0.0004	2.4674x10^{-8}

parameters under estimation are probabilities by their own and can only take values in [0,1]. It should be noted that, in addition to studying the behavior of the estimated parameters for different input training data, the outlined approach can be used to avoid uncertainty due to the randomness of the small datasets, i.e. while two different small datasets might give rise to noticeable discrepancy between the results, this resampling technique can present more reliable estimation of the parameters by aggregation. Therefore for the applications where a single value variable is required (as the case is here), the mean of the distribution can be used as an alternative. For more detail the reference[35] may be helpful.

Bayesian Network Risk Analysis

Once the complete sets of marginal and conditional probabilities including those related to rare events are estimated using the ME method outlined in section 3.2, the Bayesian network enables us to probabilistically model the system's behavior using Bayes' rule[32] . Generally, there are two kinds of Bayesian inference, forward and backward, depending on how the updated network (with posterior probabilities) is treated as the evidence is introduced to the network

Predictive (forward). In this case, the flow of information is from parent nodes/variables to child nodes/variables. Probabilities of the child variables are updated given the state of their parent(s). This type of inference is especially important to develop abnormal event propagation scenarios and to find the most probable abnormal consequences encountered within the system and their failure probabilities through risk assessment procedures. Figure 3.8b shows the updated network once an example evidence indicating that X is in the state S_1, is introduced. Such evidence along with the conditional probabilities and the Bayesian network update rules[36] updates the probabilistic belief about each variable's states, including those which have not shown up in the historical data. Such analysis enables the analyst to draw inferential information about the system's tendency to deviate from its normal operation state, particularly those variables that show more potential to be at a dangerous abnormal state. In this case, once the evidence X in S_1 is given to the network, the network shows a large deviation in its most closely connected child node, Z, and a weaker deviation in Z. However, as also implied by Eq.(3.21), these deviations do not affect W strongly. We will later introduce an index to quantify and compare these deviations.

Diagnostic (backward). In this case, the flow of information is from children toward parents. A change in the states of a child variable updates its parents' states using the Bayesian network belief propagation rules, and conclusions can then be made about the most probable causes of the observed anomaly. Such an analysis is mostly used in real-time analysis for fault detection and isolation.

In both cases, making decisions about the abnormal states (which in most cases are the same as rare states) is of critical importance; that is, states of most interest, whether in risk assessment or fault detection, are exactly those states which are poorly reflected in the data due to their small probabilities (which are indeed desirable from the scope of process control and safety, indicating an efficient safety system). To tackle this issue, the proposed constrained maximum- entropy algorithm is employed to estimate these probabilities. Figure 3.8c illustrates the updated network given an evidence in the node W. As can be seen, the flow of information updates the entire network, giving a

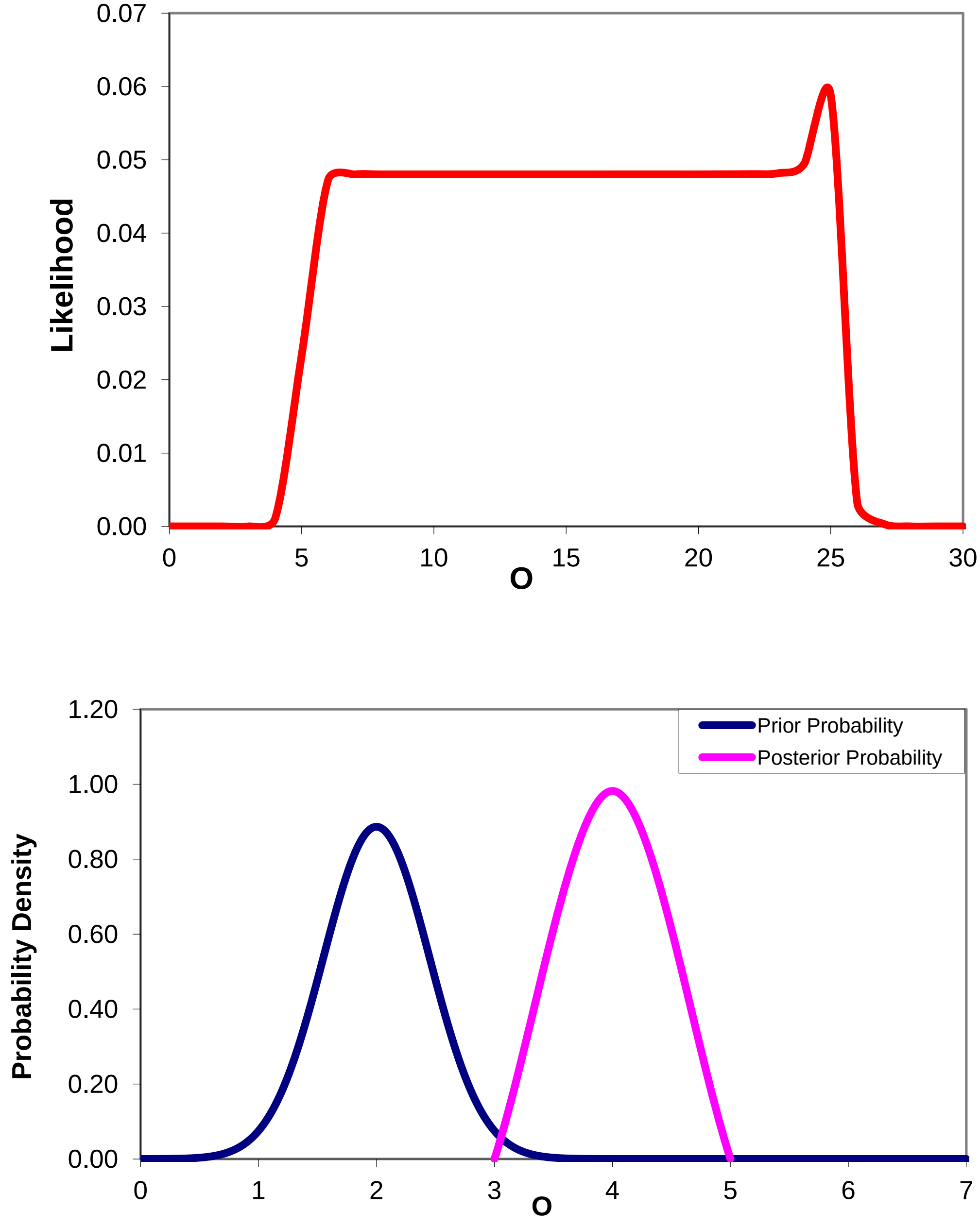

Figure 3.4: (a) Likelihood and (b) prior and posterior probabilities versus the model's complexity level, O, for node X.

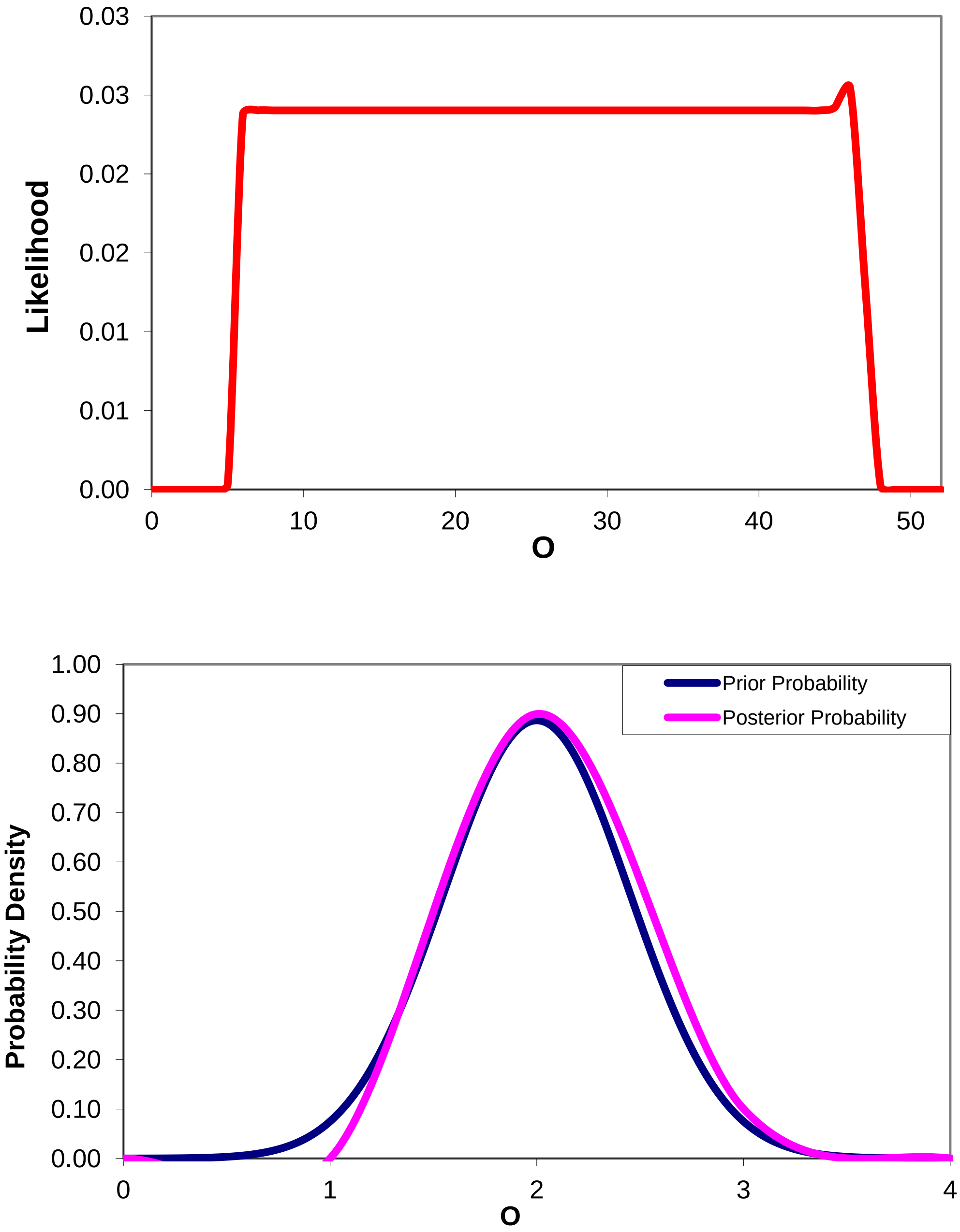

Figure 3.5: (a) Likelihood and (b) prior and posterior probabilities versus the model's complexity level, O, for X.

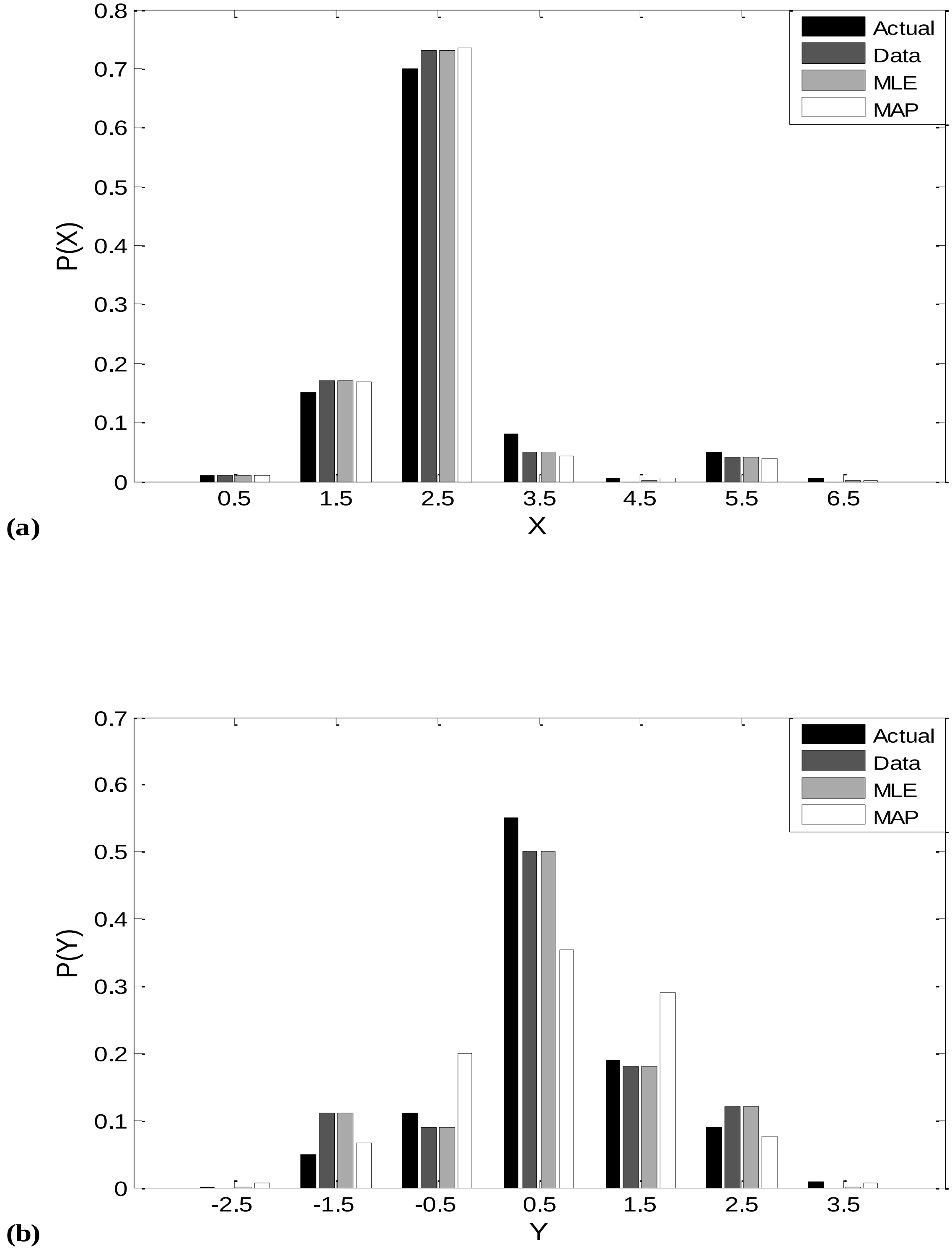

Figure 3.6: Actual, based on 100 samples, MLE-estimated, and MAP-estimated probabilities. (a) Node X. (b) Node Y.

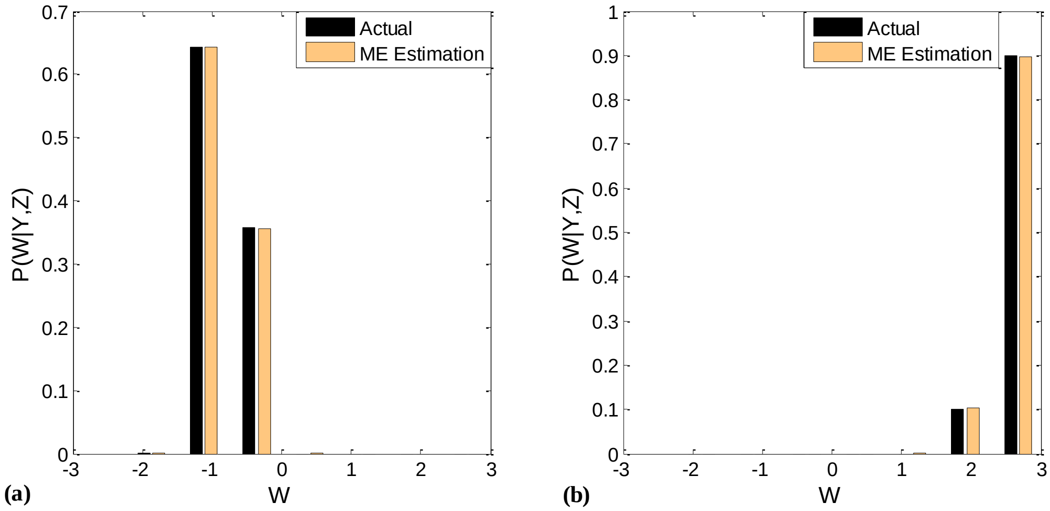

Figure 3.7: Actual and ME-estimated conditional probabilities for node W given (a) both parent nodes Y and Z in their lowest states, and (b) both parent nodes Y and Z in their highest states.

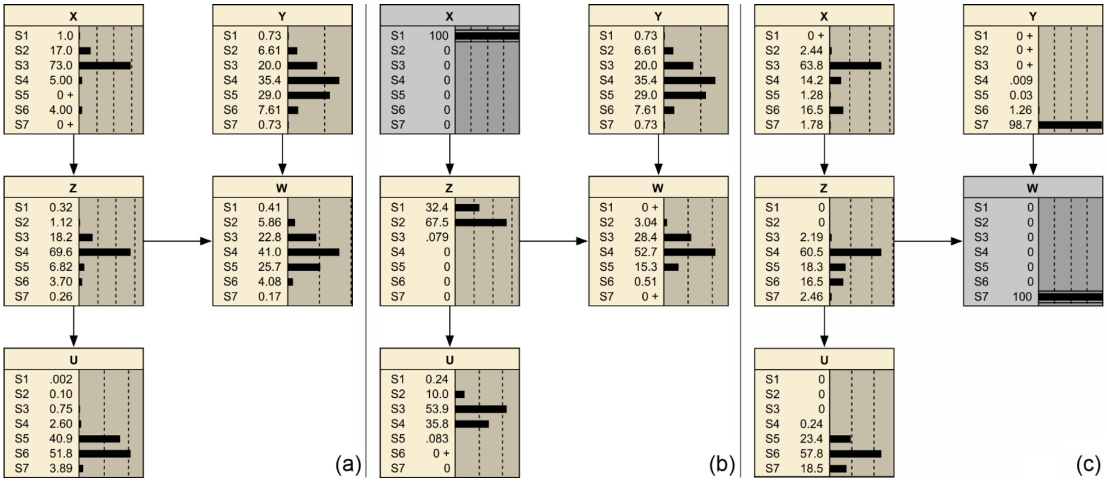

Figure 3.8: Bayesian network of the example system based on the ME-estimated PMFs. (a) Normal operation network. (b) Predictive inference (evidence is for X). (c) Diagnostic inference (evidence is for W).

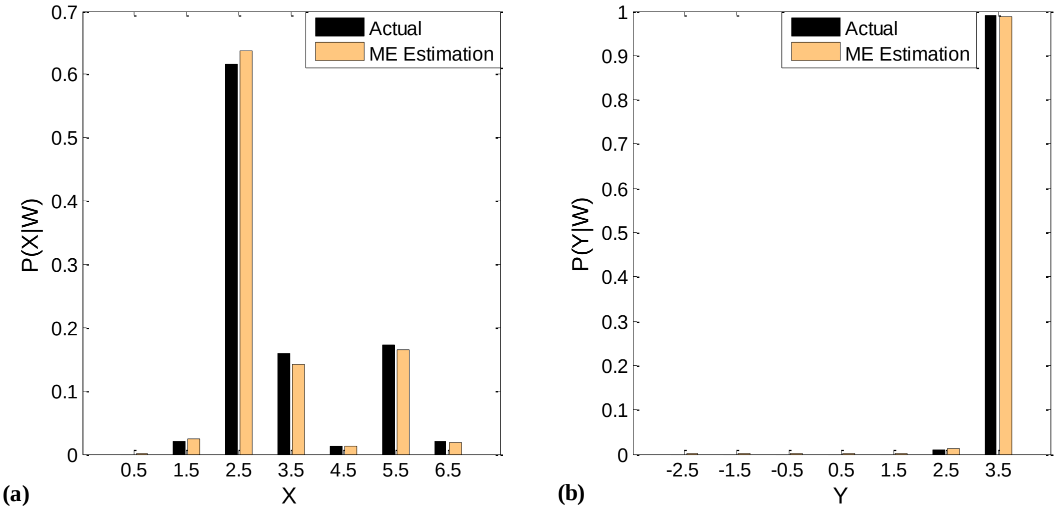

Figure 3.9: Diagnostic Bayesian inference based on 1 million samples and the ME-estimated PMFs. (a) X. (b) Y.

clue about the most probable reasons to the observed evidence. In such cases, partial predictive inference can also be conducted through updating the probabilistic belief about child nodes located at downstream of the root nodes, implying how the deviation occurred at the upstream root nodes affects the nodes which do not share a causal path with evidence node. Figures 3.9a and 3.9b compare the result of such diagnostic Bayesian inference for the nodes X and Y derived by actual model and ME estimation method. This plot indicates how accurate the estimation with only one hundred samples is compared to a sample with 10,000 time larger size.

Risk Quantification

Since the variables in the Bayesian network context are treated as random variables, one direct way of studying their behavior is to consider their (marginal) probability distributions. To quantify the changes in marginal probabilities caused by the evidences and identify the node(s) that undergo the most significant change, we use a probability distance measure called Kullback-Liebler divergence (KLD) or information gain[37] . This measure reveals information about the relative entropy of two probability distributions defined over the same set of states, and therefore indicates how different two marginal probabilities are compared to each other:

$$KLD_j = \sum_{i=1}^{N} P(X_j = x_j^i) \ln\left(\frac{P(X_j = x_j^i)}{Q(X_j = x_j^i)}\right) \tag{3.26}$$

where P and Q are the prior and posterior probabilities of node X_j with N states, respectively. To measure the most significant deviations in the network, we then use the relative Kullback-Liebler divergence (RKLD) . For a node $\mathrm{X_j}$, the RKLD is defined as:

$$\mathrm{RKLD_j} = \frac{\mathrm{KLD_j}}{\sum_{\mathrm{j=1}}^{\mathrm{w}} \mathrm{KLD_j}} \tag{3.27}$$

If a prior probability of a state is 0, then its corresponding term in KLD expression is 0, since $0 \times \log(0) = 0$. Figure 3.10 shows a comparison of the relative KL divergence values. As shown in Figure 3.10a, based on the relative KL divergence, moving X to its lowest state causes the largest changes in nodes Z and U, respectively. This prediction makes sense, since U is connected to X through Z.

The same procedure is applied to identify the most probable cause to an abnormality observed in node W by backward Bayesian inference (Figure 3.8c). Since deviation of the node Y from its normal operation PMF indicated in Figure 3.10b is more than that of the node X, the former is the variable which has most likely led to the observed anomaly in node W. It is important to note that the prior probability of each node has a strong effect on its updated probability. In other words, more diffuse PMFs (larger variance) are more likely to have caused the abnormal states, as confirmed by our example. The generalization of this procedure to larger and denser is straightforward; the same principles for reconstructing the probabilities will apply regardless of the size of the problem or its degree of complexity.

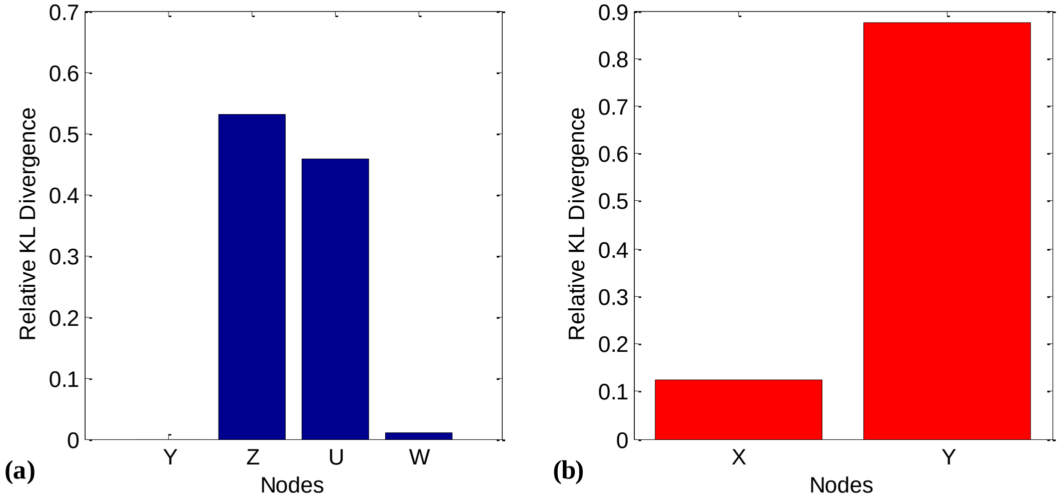

Figure 3.10: RKLD values for the nodes. (a) Predictive inference. (b) Diagnostic inference.

3.4. Conclusions

This chapter presented a method of estimating the probabilities of states with no observed data (rare states) of a multivariate probability distribution with finite or countable number of states. The method maximizes the Shannon's information entropy subject to constraints imposed by empirical moments of the sampled population. At the same time, two approaches to select the optimal constraints are also investigated. This method is especially beneficial as the model's parameters (PMF probability values) are to be utilized to train directly Bayesian networks, where discretized random variables are usually preferred. Advantages of this method over other probability estimation techniques are as follows. Firstly, since the information content of individual sample points are compacted in the form of moments, larger sample sizes do not affect the speed of convergence. Second, as no limiting assumptions are made on the dependence structure of the domain variables, it is capable of modeling highly nonlinear dependence structures. Third, the MLE and MAP criteria to determine the optimal level of model's complexity enables one to use the highest possible flexibility in modeling while avoiding over-fitting. Because of these features, the method provides reliable probability estimates for regions outside the observed region, where most of the risky events are likely to occur. Since the method is primarily developed to estimate discrete multivariate probabilities and therefore is suitable to calculate conditional probabilities, it is particularly advantageous when combined with Bayesian networks. This combination allows the user to calculate the probability of abnormal event propagation, where an abnormal situation in one component of the systems increases the chance of abnormal situation in another component. However, although the Bayesian network allows decomposing the high dimensional multivariate distributions, leading to tractability of the method, care must be taken when working with overly dense networks and/or high number of states which may significantly increase the number of variables and the resulting computational cost. Finally, the estimates can be used to perform risk assessment and fault detection effectively, e.g., through Bayesian networks.

References

1. Harms-Ringdahl, L. *Safety Analysis, Principles and Practice in Occupational Safety*; CRC Press: New York, 2001.

2. Garvey, P. R. *Analytical Methods for Risk Management: a Systems Engineering Perspective*; Chapman and Hall/CRC: Boca Raton, FL, 2009.

3. Taleb, N. N. *The Black Swan: The Impact of the Highly Improbable*; Random House: New York, 2007.

4. Kaynar, B.; Ridder, A. The cross-entropy method with patching for rare-event simulation of large Markov chains. *European Journal of Operational Research* **2010**, *207*, 1380-1397.

5. Shahabuddin, P. Rare event simulation of stochastic systems. In *Proceedings of the 1995 Winter Simulation Conference*. IEEE Press: Washington, D.C., 1995: 178–185.

6. 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.

7. Scott, D. W. *Multivariate Density Estimation: Theory, Practice, and Visualization*; John Wiley: New York, 1992.

8. Tsybakov, A. B. *Introduction to Nonparametric Estimation*; Springer: New York, 2009.

9. Devroye, L.; Gábor, L. *Combinatorial methods in Density Estimation*; Springer: New York, 2001.

10. Piotr, J.; Durante, F.; Härdle, W. K.; Rychlik, T. *Copula Theory and Its Applications: Proceedings of the Workshop Held in Warsaw 2009*; Springer: Heidelberg, 2010.

11. Zellner, A.; Highfield, A. R. Calculation of maximum entropy distribution and approximation of marginal posterior distributions. *Journal of Econometrics* **1988**, *37*, 195-209.

12. Duong, T.; Hazelton, M. L. Convergence rates for unconstrained bandwidth matrix selectors in multivariate kernel density estimation. *Journal of Multivariate Analysis* **2005**, *93*, 417–433.

13. 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*. Elsevier: New York, 2003: 331–385.

14. Mehranbod, N.; Soroush, M.; Panjapornpon, C. A method of sensor fault detection and identification. *Journal of Process Control* **2005**, *15*, 321-339.

15. Villez, K.; Srinivasan, B.; Rengaswamy, R.; Narasimhan, S.; Venkatasubramanian, V. Kalman-based strategies for Fault Detection and Identification (FDI): Extensions and critical evaluation for a buffer tank system. *Computers & Chemical Engineering* **2011**, *35* (5), 806-816.

16. Bin Shams, M. A.; Budman, H. M.; Duever, T. A. Fault detection, identification and diagnosis using CUSUM based PCA. *Chemical Engineering Science* **2011**, *66* (20), 4488-4498.

17. Pariyani, A.; Seider, W. D.; Oktem, U. G.; Soroush. M. Improving process safety and product quality using large databases. In Pierucci S, Buzzi Ferraris G., eds., *Computer Aided Chemical Engineering*. Elsevier: Naples, Italy, 2010, *28*: 175-180.

18. Castillo, I.; Edgar, T. F.; Dunia, R. Nonlinear model-based fault detection with fuzzy set fault isolation. *IECON 2010 - 36th Annual Conference on IEEE Industrial Electronics Society*. Phoenix, AZ, 2010: 174-179.

19. Özyurt, B.; Kandel, A. A hybrid hierarchical neural network-fuzzy expert system approach to chemical process fault diagnosis. *Fuzzy Sets and Systems* **1996**, *83* (1), 11-25.

20. Isermann, R. *Fault-Diagnosis Systems: An Introduction from Fault Detection to Fault Tolerance*; Springer: Berlin, 2006.

21. Mohseni Ahooyi, T.; Arbogast, J. E.; Oktem, U.; Seider, W. D.; Soroush, M. Maximum-Likelihood Maximum-Entropy Constrained Probability Density Function Estimation for Prediction of Rare Events. *AIChE Journal* **2014**, *60* (3), 1013-1026.

22. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer: New York, 2003.

23. Shannon, C. E. A mathematical theory of communication. *The Bell System Technical Journal* **1984**, *27*, 379–423.

24. Williams, D. *Weighing the Odds: a Course in Probability and Statistics*; Cambridge University Press: New York, 2001.

25. Shannon, C. E. Prediction and entropy of printed English. *The Bell System Technical Journal* **1951**, *30*, 50–64.

26. Gray, R. M. *Entropy and Information Theory*, 2nd ed.; Springer: New York, 2011.

27. Eliason, S. R. *Maximum Likelihood Estimation: Logic and Practice*; SAGE Publications, Inc.: Newbury Park, 1993.

28. Ruszczyński, A. *Nonlinear Optimization*; Princeton University Press: Princeton, NJ, 2006.

29. Subramanian, J.; Simon, R. Overfitting in prediction models – Is it a problem only in high dimensions?. *Contemporary Clinical Trials* **2013**, *36* (2), 636-641.

30. Jefferys, W. H.; Berger, J. O. Ockham's Razor and Bayesian Statistics. *American Scientist* **1991**, *80*, 64–72.

31. Briscoe, E.; Feldman, J. Conceptual complexity and the bias/variance tradeoff. *Cognition* **2011**, *118* (1), 2-16.

32. Ando, T. *Bayesian Model Selection and Statistical Modeling*; Chapman and Hall/CRC: Boca Raton, 2010.

33. Ferentinos, K. On Tchebycheff's type inequalities. *Trabajos de Estadística e Investigación Operativa* **1982**, *33* (1), 125-132.

34. Netica v. 4.16 [Software], Norsys Software Corporation. Vancouver, BC. 2010. Available at http://www.norsys.com.

35. Quenouille, M. H. "Notes on Bias in Estimation". Biometrika **1956**, *43* (3-4): 353–360.

36. Lauritzen, L.; Spiegelhalter, J. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society* **1988**, *50*, 157-224.

37. Kullback, S. Letter to the Editor: The Kullback–Leibler distance. *The American Statistician* **1987**, *41*, 340–341.

Chapter 4: Rolling Pin Method: Efficient General Method of Joint Probability Modeling

4.1. Introduction

Complex real world systems with a great deal of uncertainty often cannot be properly represented by deterministic models, as these models are conditioned approximations of reality. Deterministic models only provide point-wise estimate predictions with no information on the uncertainty of their predictions and do not provide a systematic way of accounting for noise and stochastic disturbances. For such systems, *probabilistic modeling* techniques have become more popular in recent years.[1]

Joint probability distributions are key mathematical elements in modeling uncertain knowledge and stochastic systems. They assign a probability (or probability density) to each state (or point) in the multidimensional space of the domain variables.[2] A probability distribution may be used to make predictions about the likelihood of a query state or to perform inference about the query variables as evidential knowledge becomes available. Also, joint probability distributions over parameters with uncertainty can be used to conduct parameter estimation.[3] These models are particularly useful in performing predictions under uncertainty. The uncertainty can be an intrinsic property of the systems, can be due to the lack of adequate knowledge about the systems, or a combination of the two.

The need for a reliable method of estimating joint probability distributions has motivated numerous studies on this topic in the past few decades.[4] The existing methods of joint probability estimation can be divided into three main groups: parametric methods,[5] nonparametric methods,[6] and combined parametric and nonparametric methods (semi-parametric methods).[7]

The most simplistic parametric method is to use a pre-defined multivariate probability distribution such as a multivariate elliptical distribution to describe the observed data. Generally, these parametric distributions are computationally favorable in terms of training their parameters and sampling.[8] However, the flexibility of such distributions in modeling real-world systems is quite limited.[9] Another widely-used parametric method is the parametric copula method, which decomposes a joint distribution into a dependence structure represented by the copula and univariate marginal distributions of the domain variables.[10] The standard parametric copulas are very good at modeling nonlinear relationships that are monotonic and they do this by a relatively small number of parameters. However, they are unable to describe joint probability distributions of variables with non-monotonic relationships. Furthermore, finding the true dependence structure may not always be easy and in many cases there is no conventional parametric copula corresponding the system's true dependence structure. The problem can be more severe when the pairwise dependence structures of the variables are not the same. Another parametric method is the moment-based approach, which presents a highly flexible way to model arbitrary joint distributions from the data moments.[11,12] Despite such flexibility, the method suffers from high computational cost when system's number of dimensions grows.

The non-parametric methods of probability distribution estimation assume no predefined model for the observed data; rather they construct the distribution function using the simple functions assigned to each point in the dataset. The histogram methods is the simplest method of this kind where the density value of each state is calculated using the data. In addition to inaccuracies due to discretizing the attributes, the histogram method may suffer from very high number of density values, which increases exponentially as the number of variables increases.[13] The kernel density is an example of a continuous nonparametric model.[14] It provides high flexibility, but similar to the moment-based approach it can be computationally expensive and its bandwidth matrix (of the smoothing parameters) becomes unstable for high dimensional systems.[15] For this reasons the kernel method is prescribed for low to moderate number of dimensions, with an upper limit of six. Finally, compared to the parametric methods, non-parametric methods have a slower convergence rate to the actual probability distribution, as the size of training data increases. For example, the convergence rate of the Gaussian kernel error to zero is $O(n^{-0.8})$, which is lower than that of parametric methods, $O(n^{-1})$, where n is the number of training data.

Considering the drawbacks described above, semi-parametric methods that combine the computational tractability of the parametric methods with the flexibility of non-parametric methods, have been proposed. Olkin and Spiegelman in their pioneering work[16] proposed a semi-parametric distribution as a weighted linear combination of parametric and non-parametric densities, where the weights were estimated using the maximum likelihood principle. Another semi-parametric method is a combinatorial copula method, where a parametric copula is combined with a non-parametric method of estimating the marginal densities, e.g. by the kernel method.[17] This method models nonlinear monotonic relationships satisfactorily, however because of the limitations of the available parametric families of copula, it fails to present reliable distributions for non-monotonic and complex dependence structures.

This chapter presents a novel efficient method of estimating joint probability distributions of continuous random variables with non-monotonic or monotonic relationships. As the backbone of the method is a set of monotonization transformations that 'roll out' the relationships, the method is named the *rolling pin method*. The rolling pin method allows one to estimate joint probability distributions when the actual causal structure of the attributes is unknown or extremely intricate to be accurately determined. This method aims at addressing the common drawbacks of the existing joint probability estimation methods, as well as limitations of the ordinary parametric copula method, as discussed above. The rolling pin method offers the following advantages over the existing joint probability estimation methods: 1) it doesn't require any knowledge of the causal structure among variables; 2) unlike conventional copulas, it is capable of modeling non-monotonic relationships between variables; 3) it enables the user to model joint probability distributions over multiple (more than two) random variables with the same parametric family of copula, regardless of possible differences in joint probability dependence structure of each pair of variables; 4) it may be programmed such that unknown joint probability dependence structures of the variables is modeled with a known parametric copula; 5) it is computationally efficient, as the joint probability distribution is fully specified with $O(d^2)$ parameters, where d denotes the number of random variables; 6) its estimated probability densities may be used to quantify the probability of rare events (i.e., events having no data available in the historical data), as well as compound (multivariate) risks, where a rare event in a variable may lead to a rare event in another variable. This is possible, as the derived continuous probability distributions can be defined and evaluated over the regions (rare states) which historical data lacks information on.[11]

The rest of the chapter is organized as follows. Section 4.2 provides some preliminaries. Section 4.3 describes the rolling pin method. It also compares the rolling pin method and the conventional parametric copula method. Section 4.4 presents the application of the rolling pin method to two mathematical and process examples. Section 4.5 includes concluding remarks.

4.2. Preliminaries

Let $\mathbf{W} = (W_1, \dots, W_d)^T$denote a vector of continuous random variables with an unknown dependence structure in a d-dimensional space of real numbers. Each pair $\left(W_i, W_j\right)$, $i, j \in \{1, \dots, d\}, i \neq j,$ is assumed to have an arbitrary relationship. The objective is to construct a joint probability density function of $\mathbf{W}$, $f_{\mathbf{W}}(\boldsymbol{w})\colon \mathbb{R}^d \to \mathbb{R}^+$, given the observed dataset $\mathbf{D}$. The joint density function$f_{\boldsymbol{W}}$ represents a mathematical model of a stochastic system that has d random variables. In this chapter, random variables are shown by capital letters and their numerical values by small letters.

In many real-world applications, variables describing systems often have different orders of magnitude. Since this is a potential source of inaccuracy, to address this problem, as a standard practice throughout this chapter, we obtain normalized variables corresponding to W_i's using

$$X_i = \frac{W_i - \mu(W_i)}{\sqrt{\mathrm{Var}(W_i)}} \tag{4.1}$$

where $\mu(W_i)$ is the empirical mean of W_i defined as

$$\mu(W_i) \ = \frac{1}{n}\sum_{k=1}^{n}(w_{i,k}) \tag{4.2}$$

and $\mathrm{Var}(W_i)$ denotes the empirical variance of W_i:

$$\mathrm{Var}(W_i) = \frac{1}{n}\sum_{k=1}^{n}\left(w_{i,k} - \mu(W_i)\right)^2 \tag{4.3}$$

where n is the number of samples of $w_{i,k}$ for W_i. Therefore, samples of X_i has a mean value of 0 and a variance of 1. Throughout the rest of this chapter it is assumed that variables are already normalized using Eq.(4.1).

4.2.1. Modeling Joint Distributions Using Copulas

A copula is a multivariate probability distribution of a set of random variables that have uniform univariate marginal probability densities. Copulas are employed to describe the dependence structure of random variables. According to the Sklar's Theorem,[18] every multivariate joint distribution can be written in terms of univariate marginal distributions and a copula. Indeed, the main strength of copula density estimation is that it enables one to decompose and describe a joint probability distribution into univariate marginal cumulative distribution function (CDFs) of random variables and a dependence structure (copula function) of the variables. Parametric copulas have parameters, which allow one to adjust the strength of dependence among random variables. Copulas may be utilized as a basis to model dependence structures based on Sklar’s theorem:

Theorem 1.[18] For every multivariate joint probability distribution $F(x_1, \dots, x_d) = \Pr(X_1 \leq x_1, \dots, X_d \leq x_d)$, there is a copula function C: $[0,1]^d \rightarrow [0,1]$ such that

$$\mathrm{C}(u_1, \dots, u_d) = F\left({F_{X_1}}^{-1}(u_1), \dots, {F_{X_d}}^{-1}(u_d)\right) \tag{4.4}$$

where ${F_{X_i}}^{-1}: [0, 1] \to \mathbb{R}$ is the marginal quantile function (inverse cumulative distribution function (CDF)) of random variable X_i. If the margins of F are continuous, the copula function will be unique, otherwise it will be uniquely defined on $(\text{range}(F_{X_1}), \dots, (\text{range}(F_{X_d})))$. In view of this, copula may be thought as a joint cumulative probability distribution function over the uniform random variables $F_{X_1}, \dots, F_{X_d}$ distributed as $\mathrm{U}(0,1)$, where $\mathrm{U}(a, b)$ is the uniform probability density function between a and b. Every copula has the following basic properties:[19]

- It is a *grounded* function; i.e., $\mathrm{C}(u_1, \dots u_{i-1}, 0, u_{i+1} \dots, u_d) = 0$.
- $\mathrm{C}(1, \dots 1, u_i, 1 \dots, 1) = u_i$. (4.5)
- If $b \geq a,$ then $\mathrm{C}(u_1, \dots, u_{i-1}, b, u_{i+1}, \dots, u_d) \geq \mathrm{C}(u_1, \dots, u_{i-1}, a, u_{i+1}, \dots, u_d)$. (4.6)
- If the set of random variables $T_1(X_1), \dots, T_d(X_d)$ are derived by strictly increasing transformations of $X_1, \dots, X_d$, then $\mathrm{C}_{\boldsymbol{T}} = \mathrm{C}$. In other words, the copula (dependence structure) is preserved under strictly increasing transformations.

The ability of a d-dimensional copula to decompose a joint probability distribution into a d univariate marginal CDFs of random variables and a dependence structure in terms of the copula function allows one to exactly reconstruct a joint probability distribution given its true copula and d univariate marginal CDF functions, $F_{X_i}(x_i)$, $F_{X_i}: \mathbb{R} \to [0, 1]$, $i \in \{1, \dots, d\}$:

$$F(x_1, \dots, x_d) = \mathrm{C}\left(F_{X_1}(x_1), \dots, F_{X_d}(x_d)\right) \tag{4.7}$$

As a result of this decomposition, one can model the marginal distributions

$(F_{X_1}(x_1), \dots, F_{X_d}(x_d))$ and the dependence structure (represented by the copula) separately; i.e., the parameters involved are adjusted independently to model the marginal distributions and the dependence structure. This procedure allows one to model highly nonlinear joint probability distributions by adjusting a few parameters.

There are many techniques to model univariate marginal distributions. In many practical applications, parametric probability distributions (such as the Gaussian and gamma distributions) offer a good representation of the data behavior. More complex marginal distributions can be non-parametrically defined by an empirical CDF:

$$\hat{F}_X(x) = \frac{\text{number of data points} \leq x}{n} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{1}\{\chi_k \leq x\} \tag{4.8}$$

where $\mathbf{1}\{.\}$ is the indicator function and χ_k denotes the k-th sample of X. As an alternative, the kernel distribution function can be used to obtain a smoother CDF that is extendable to the unobserved regions

$$\hat{F}_{X,h}(x) = \frac{1}{n}\int_{-\infty}^{x} \frac{1}{h}\sum_{k=1}^{n} \mathrm{K}\left(\frac{t-\chi_k}{h}\right) dx \tag{4.9}$$

where $\hat{F}_{X,h}(x)$ is the CDF estimated by the kernel method, $\mathrm{K}(.)$ is the kernel probability density function, and h is a scalar called the smoothing parameter or bandwidth. The smoothing parameter is calculated through minimizing an error measure such as the mean integrated squared error or the mean integrated absolute error.[20]

The copula density is the basis for the definition of the joint probability density of **X**

$$\mathrm{c}(u_1, \dots, u_d) = \frac{\partial^d \mathrm{C}(u_1, \dots, u_d)}{\partial u_1 \dots \partial u_d} \tag{4.10}$$

$$f_{\mathbf{X}}(x_1, \dots, x_d) = \mathrm{c}\big(F_{X_1}(x_1), \dots, F_{X_d}(x_d)\big) \prod_{i=1}^{d} f_{X_i}(x_i) \tag{4.11}$$

where $\mathrm{c}: [0,1]^d \rightarrow \mathbb{R}^+$, $f_{\mathbf{X}}: \mathbb{R}^d \rightarrow \mathbb{R}^+ \cap \{0\}$ and $f_{X_i}: \mathbb{R} \rightarrow \mathbb{R}^+ \cap \{0\}$ denote the copula density, joint density and marginal density functions, respectively. Once all marginal CDFs are estimated, they can be used in combination with a copula function to generate a joint probability distribution. It is important to note that the final joint probability distribution is as dependent on the choice of copula function as it is on the marginal CDFs; that is, different copula functions give rise to totally different joint probabilities given the same marginal CDFs. There are plenty of choices for the copula function. Well-known parametric copulas are based on random processes (e.g., Marshall-Olkin family), defined using the dependence structures of widely-used joint probabilities (e.g., elliptical family), or developed from the so-called generator functions (e.g., Archimedean family). Non-parametric empirical copulas are defined in a similar way as in Eq. (4.6) .[21]

Parametric copulas are becoming a popular and standard framework for multivariable probabilistic modeling, mainly because they are easily formulated, parameterized and sampled. However, they suffer from the following important limitations:

1. When the objective is data-driven construction of a joint probability distribution, while its actual dependence structure unknown, the availability of a strategy to systematically choose the right copula from the wide range of parametric copulas is of critical importance. In the absence of such a strategy, there is no guarantee that a chosen copula replicates the actual dependence structure accurately, particularly over the regions where no sample is observed, this is where the tail dependence behavior of copulas play its determining role.

2. There is an ever-growing number of different parametric copulas covering different types of dependence structures. Despite this progress, every possible dependence structure which underlies the data cannot be captured by the existing copulas yet.
3. While copulas can model highly nonlinear monotonic relationships, the most serious problem with the conventional parametric copulas is that they are unable to capture non-monotonic relationships. This problem in most part is because the commonly-used measures of correlation, which are used to quantify the strength of dependence between two random variables, are unable to differentiate non-monotonic dependence from independence.
4. In general, there may exist different dependence structures between each pair of variables. Therefore, in such cases assigning a unique copula for modeling random vectors ($d \geq 3$), which applies the same dependence structure to every pair of variables, is not technically correct. Although the vine copula method has been introduced to circumvent this problem by factorizing multivariate copulas,[22] it has its own drawbacks including high computational cost and the restrictions imposed by the ordinary parametric copulas discussed above.

Addressing these fundamental issues in copulas is a major motivation for developing the rolling pin method.

4.3. Rolling Pin Method

4.3.1. Variable Monotonization

As conventional parametric copula families in their original form are unable to describe joint probability distributions of variables with non-monotonic relationships, a variable transformation is proposed.

Definition 1. Continuous variables Y_i and Y_j are said to have strictly-increasing monotonic relationships if

$$\forall y_i \in \Omega_{Y_i},\ \frac{\partial y_j}{\partial y_i} > 0, \quad i, j = 1, \cdots, d, i \neq j \tag{4.12}$$

where Ω_{Y_i} denotes the domain of Y_i.

Monotonization transformation. Consider continuous variables $X_1, \ldots, X_d$ with arbitrary (monotonic or non-monotonic) and generally unknown relationships. The *monotonization transformation* transforms these variables to new variables $Y_1, \ldots, Y_d$ that have strictly-increasing monotonic relationships to X_r, a *reference variable* that is selected systematically from $X_1, \ldots, X_d$. The *monotonized variable* Y_i is defined as:

$$Y_i = (1 - \alpha_i)X_i + \alpha_i X_r \tag{4.13}$$

where $\alpha_i \in [0, 1]$ is a parameter, called the *monotonization parameter* of variable X_i. Obviously, $\alpha_i = 0$ yields $Y_i = X_i$, and $\alpha_i = 1$ yields $Y_i = X_r$. Furthermore, $\forall \alpha_r \in [0, 1]$, $Y_r = X_r$. Considering this, we simply set $\alpha_r = 0$. A sufficiently large value of α_i, results in a Y_i that has an increasingly monotonic relationship with X_r, regardless of the type of the dependence of X_i on X_r. This statement is always true for $\alpha_i = 1$, as $Y_i = Y_r$. Once a sufficiently large value of $\alpha_i < 1$ is found, an appropriate parametric copula can be used to model the pair (Y_i, Y_r).

Although the monotonization transformation guarantees monotonicity between each Y_i and Y_r, to model multivariate ($d \geq 3$) probability distribution functions using parametric copulas, we first need to prove that all pairs $(Y_i, Y_j),\ i, j \in \{1, \dots, d\}, i \neq j$ have monotonic relationships. The following theorem establishes the sufficient condition for having such relationships.

Theorem 2. If the pair Y_i and Y_r and the pair Y_j and Y_r have strictly-increasing monotonic relationships, then the pair Y_i and Y_j have a strictly-increasing monotonic relationship.

Proof. According to Def.1 and Eq.(4.11):

$$\frac{\partial y_i}{\partial y_r} = \frac{\partial y_i}{\partial x_r} > 0,\quad \frac{\partial y_j}{\partial y_r} = \frac{\partial y_j}{\partial x_r} > 0,\quad \forall y_r \in \Omega_{X_r} \tag{4.14}$$

where Ω_{X_r} is the domain of the reference variable. These imply that:

$$(1 - \alpha_i)\frac{\partial x_i}{\partial x_r} + \alpha_i > 0,\ \left(1 - \alpha_j\right)\frac{\partial x_j}{\partial x_r} + \alpha_j > 0,\quad \forall y_r \in \Omega_{X_r} \tag{4.15}$$

$$\frac{\partial x_i}{\partial x_r} + \frac{\alpha_i}{1-\alpha_i} > 0,\quad \frac{\partial x_j}{\partial x_r} + \frac{\alpha_j}{1-\alpha_j} > 0,\ \forall y_r \in \Omega_{X_r} \tag{4.16}$$

$$\frac{\partial y_j}{\partial y_i} = \left(1 - \alpha_j\right)\frac{\partial x_j}{\partial y_i} + \alpha_j \frac{\partial x_r}{\partial y_i} = \left(1 - \alpha_j\right)\frac{\partial x_j}{\partial x_r}\frac{\partial x_r}{\partial y_i} + \alpha_j \frac{\partial x_r}{\partial y_i} = \left(1 - \alpha_j\right)\left(\frac{\partial x_j}{\partial x_r} + \frac{\alpha_j}{1-\alpha_j}\right)\frac{\partial x_r}{\partial y_i}$$

(4.17)

Eq.(4.15) with Eqs.(4.12) and (4.14) implies that $\forall y_i \in \Omega_{Y_i}$,

$$\frac{\partial y_j}{\partial y_i} > 0 \tag{4.18}$$

■

4.3.2. Rolling Pin Distribution

In this section, the monotonized variables defined earlier will be utilized in combination with the copula method to develop probability distributions that are capable of modeling non-monotonic relationships and complex dependence structures.

Let the vectors of continuous random variables $\mathbf{X} = (X_1, \dots, X_d)^T$ and $\mathbf{Y} = (Y_1, \dots, Y_d)^T$ be defined according to Eq.(4.1) and (4.13) and the vector of optimal monotonizing parameters be $\boldsymbol{\alpha_m} = (\alpha_{1,m}, \dots, \alpha_{d,m})^T$ which assures the pairwise increasingly monotonic relationships between the components of **Y**. Note that the relationship between **X** and **Y** is one-by-one and is therefore invertible. The functionality of every pair (X_i, X_j) can take on any unknown form.

As the relationship of every pair (Y_i, Y_j) is strictly-increasing monotonic, one can model accurately the joint CDF of $\boldsymbol{Y}$ using an appropriate copula function:

$$F_{\mathbf{Y}}(y_1, \dots, y_d) \; = \Pr(Y_1 \leq y_1, \dots, Y_d \leq y_d) \; = \mathsf{C}\big(F_{Y_1}(y_1), \dots, F_{Y_d}(y_d)\big) \qquad (4.19)$$

where $F_{Y_1}, \dots, F_{Y_d}$ are the marginal CDFs of **Y**, and C denotes a parametric copula.

Eqs.(4.13) and (4.19) provide the mathematical basis for modeling arbitrary (including non-monotonic) relationships among the components of **X** using copulas. Let $\mathbf{y} = (y_1, \dots, y_d)^T$, $\mathbf{x} = (x_1, \dots, x_d)^T$, and $\boldsymbol{J}$ be the Jacobean matrix $\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$. Without loss of generality, assume $X_d = X_r$ (i.e., $X_1, \dots, X_d$ are arranged such that the last variable is the reference variable). Then, for the linear monotonization transformations of Eq. (4.13):

$$\boldsymbol{J} = \begin{bmatrix} 1-\alpha_{1,m} & 0 & 0 & \dots & 0 & \alpha_{1,m} \\ 0 & 1-\alpha_{2,m} & 0 & \dots & 0 & \alpha_{2,m} \\ 0 & 0 & \ddots & \dots & \vdots & \vdots \\ \vdots & \vdots & & 1-\alpha_{d-2,m} & 0 & \alpha_{d-2,m} \\ 0 & 0 & \dots & 0 & 1-\alpha_{d-1,m} & \alpha_{d-1,m} \\ 0 & 0 & \dots & 0 & 0 & 1 \end{bmatrix}$$

and

$$\det(\boldsymbol{J}) = \prod_{i=1}^{d-1}\left(1-\alpha_{i,m}\right) > 0$$

which confirms that the monotonization transformations of Eq.(4.13) are one-to-one.

Because of this one-to-one property and the differentiability of the monotonization transformations, the following equality holds:

$$f_{\mathbf{X}}(\mathbf{x}) = f_{\mathbf{Y}}(\mathbf{y})|\det(\boldsymbol{J})| \tag{4.20}$$

Given Eq.(4.20), the probability density function of the random vector **X**, $f_{\mathbf{X}}$ is derived as follows:

$$f_{\mathbf{X}}(x_1,\dots,x_d) = f_{\mathbf{Y}}(y_1,\dots,y_d)\prod_{i=1}^{d}\left(1-\alpha_{i,m}\right) = \mathrm{c}\left(F_{Y_1}(y_1),\dots,F_{Y_d}(y_d)\right)\prod_{i=1}^{d} f_{Y_i}(y_i)(1-\alpha_{i,m}) = \frac{\partial^d \mathrm{C}(F_{Y_1}(y_1),\dots,F_{Y_d}(y_d))}{\partial F_{Y_1}(y_1)\dots\partial F_{Y_d}(y_d)}\prod_{i=1}^{d} f_{Y_i}(y_i)(1-\alpha_{i,m}) \tag{4.21}$$

where $y_i = \left(1-\alpha_{i,m}\right)x_i + \alpha_{i,m}x_r$ and c denotes the copula density function. It will be shown later that in most cases a specific type of copula can be used as C without any need for a systematic way to explore the space of the available parametric copula families. However, such a systematic way will also be presented in Section 4.3.3.4, if a greater level of accuracy is needed.

As the monotonization transformation addresses the major shortcomings of conventional parametric copulas, the rolling pin method is a powerful tool in modeling complex, nonlinear and non-monotonic continuous joint probability distributions for the following reasons:

- The rolling pin method resolves the most important drawback of the conventional parametric copulas in a very natural way. As conventional parametric copulas can capture monotonic interactions only, the monotonization step of the rolling pin method first transforms the original variables to monotonized variables, utilizable by conventional parametric copula functions. Such a copula, C, may be either directly transformed back to $f_{\mathbf{X}}$ through Eq. (4.20) or can be sampled first and the samples will be transformed then back to the samples of the original random vector using the inverses of the invertible monotonization transformations. More details on sampling from copulas can be found in Ref.[23]
- The rolling pin method benefits from the low level of computational complexity borrowed from the parametric copulas. In addition to the advantage made by the copula definition in reducing the number of parameters, many parametric copulas can be trained using by a small number of parameters. For example, elliptical copulas can be defined using the pairwise (rank) correlation coefficients of the variables. For example, Spearman's rank correlation for elliptical copulas is defined as:[24]

$$\rho_s(Y_i, Y_j) = \rho_s(Y_j, Y_i) \ = \frac{\mathrm{Cov}(F_{Y_i}, F_{Y_j})}{\left(\mathrm{Var}\left(F_{Y_i}\right)\mathrm{Var}(F_{Y_j})\right)^{1/2}} \qquad (4.22)$$

  This implies only $\binom{d}{2} = \frac{d(d-1)}{2}$ correlation parameters are required for completing the correlation matrix of the parametric copula C. In addition to this number, $(d-1)$ parameters of the vector of monotonizing parameters $\boldsymbol{\alpha_m}$ and d smoothing parameters (if the kernel method is applied for modeling marginal CDFs. This number will be $s.d$ if Eq.(4.6) is used, where s is the number of states) have to be estimated. In this case, the total number of $\left(\frac{(d+4)(d-1)}{2}+1\right)$ parameters enables a rolling pin distribution to model many non-monotonic behaviors and dependence structures.

- For the cases when $\alpha_{i,m}$, $i \in \{1, \dots, d\}$ can be set as close to 1 as possible, the linearity of Y_i with respect to Y_r (and therefore each Y_j) leads to the idea that pairwise selection of random variables (Y_i, Y_j), $i, j \in \{1, \dots, d\}$ may be treated by an *approximate dependence structure*, which is the dependence structure of random variables $(Y_r + \varepsilon_i, Y_r + \varepsilon_j)$ (equivalently $(F_{Y_r+\varepsilon_i}, F_{Y_r+\varepsilon_j})$), where

$$\varepsilon_i = \frac{1-\alpha_{i,m}}{\alpha_{i,m}} X_i \tag{4.23}$$

$$\mu(\varepsilon_i) = \left(\frac{1-\alpha_{i,m}}{\alpha_{i,m}}\right) \mu(X_i) = 0 \tag{4.24}$$

$$\mathrm{Var}(\varepsilon_i) = \left(\frac{1-\alpha_{i,m}}{\alpha_{i,m}}\right)^2 \mathrm{Var}(X_i) = \left(\frac{1-\alpha_{i,m}}{\alpha_{i,m}}\right)^2 \ll 1 \tag{4.25}$$

$$\lim_{\alpha_{i,m \to 1}} \mathrm{Var}(\, \varepsilon_i) = 0 \tag{4.26}$$

Eqs. (4.23)-(4.26) suggest that as $\alpha_{i,m}$ approaches 1, the effect of ε_i is gradually eliminated from Y_i. On the other hand, since the random variables $F_{Y_r+\varepsilon_i}$ and $F_{Y_r+\varepsilon_j}$ are uniformly distributed as $\mathrm{U}(0,1)$ and behave much alike each other because of the small effects of ε_i and ε_j, the dependence structure of $(Y_r + \varepsilon_i, Y_r + \varepsilon_j)$ and therefor $(F_{Y_r+\varepsilon_i}, F_{Y_r+\varepsilon_j})$ are close to that of (Y_r, Y_r). Therefore, the dependence structure of (Y_i, Y_j) may be approximated by a symmetric copula such as the comonotonicity copula $\mathrm{C}_M = \min\,(u_i, u_j)$ or the Gaussian copula. As a result, whatever dependence structure of (Y_i, Y_j) is, it can be approximated by a simple copula as above. The advantage of such an approximation is fourfold. First, there is no longer a need for a systematic way to explore the space of possible copulas to select the most appropriate candidate. Second, a unique multivariate copula ($d \geq 3$) such as the comonotonicity copula or Gaussian copula can be applied to model the joint probability distribution of random variable **Y** and later **X**, with less concern about the difference between the pairwise dependence structures. Third, such simple parametric copulas can be easily simulated or sampled. Fourth, unknown and new dependence structures without an exact closed-form mathematical formula can be modeled by this method.

The next section presents four approaches of estimating an optimal vector of monotonizing parameters, $\boldsymbol{\alpha_m}$.

4.3.3. Selection of $\boldsymbol{\alpha_{i,m}}$

Although it is obvious that in general $\alpha_1, \cdots, \alpha_d$ should be large enough to ensure strictly-increasing monotonic relations between the components of **Y**, $\alpha_1, \cdots, \alpha_d$ should not be unnecessarily large. As α_i approaches 1, Y_i converges to X_r; that is, the relative information contribution of X_i decreases as α_i increases, and with $\alpha_i = 1$, $Y_i = X_r$. The loss of information will be more serious if the memory assigned to store values is not adequate to include all meaningful digits of X_i and therefore it is likely that X_i is eliminated from Y_i as a result of the round-off processes and mathematical operations performed on data. Hence, the selection of appropriate values for the monotonizing parameters is a trade-off problem in which increasing the monotonicity is accompanied by the information loss. Remember that the transformation decreases the contribution of

X_i by $(1-\alpha_i)$. Therefore, in a decimal system, the number of digits shifted rightward, N_{r_i}, depends on α_i:

$$N_{r_i} = -\operatorname{fl}(\log_{10}(1-\alpha_i)) \tag{4.27}$$

where $\operatorname{fl}(x)$ rounds x to the nearest integer less than or equal x. It indicates that, for example, when α_i is 0.98, N_{r_i} is 2 and all digits of X_i will be shifted rightward by two digits after being multiplied by $(1-\alpha_i)$. This argument provides a mathematical basis for calculating the information loss due to the round-off process.

Let g and h_i denote the number of subunit digits allowed by the computer being used and the number of meaningful subunit digits of X_i. The Information loss measure $\beth_i$ is defined as

$$\beth_i = h_i + N_{r_i} - g \tag{4.28}$$

when $\beth_i \leq 0$, no meaningful digits of the original variable is eliminated as a result of the monotonization process, and when $\beth_i > 0$, $\beth_i$ digits are irreversibly lost in the transformation. $\beth_i$ along with some criteria will be utilized in the next sections to specify each $\alpha_{i,m}$.

4.3.3.1. Selection of $\alpha_{i,m}$ based on Linear Correlation Coefficient

The linear (Pearson) correlation coefficient ρ of two random variables U and V is defined as

$$\rho(U,V) = \frac{\mathrm{Cov}(U,V)}{(\mathrm{Var}(U)\mathrm{Var}(V))^{1/2}} \tag{4.29}$$

Assuming $\mathrm{Var}(U)$ and $\mathrm{Var}(V)$ are non-zero, ρ is a measure of how linearly correlated two variables are. According to the definition of Y_i, $\rho\ (X_r, Y_i)$:

$$\rho(X_r, Y_i) = \frac{\mathrm{Cov}(X_r,Y_i)}{\left(\mathrm{Var}(\ X_r)\mathrm{Var}(Y_i)\right)^{1/2}} \tag{4.30}$$

Since $\mathrm{Cov}(U,U) = \mathrm{Var}(U)$ and $\mathrm{Cov}(U, aU + bV) = a\mathrm{Var}(U) + b\mathrm{Cov}(U,V)$, and $\mathrm{Var}(X_i) = \mathrm{Var}(X_r) = 1$ [according to the definition of X_i in Eq.(4.1)], Eq.(4.30) becomes

$$\rho(X_r, Y_i) = \frac{(1-\rho_{i,0})\alpha_i + \rho_{i,0}}{\left(2(1-\rho_{i,0}){\alpha_i}^2 - 2(1-\rho_{i,0})\alpha_i + 1\right)^{1/2}} \tag{4.31}$$

where $\rho_{i,0} = \mathrm{Cov}(X_r, X_i) = \rho(X_r, X_i)$. $\rho(X_r, Y_i)$ is a strictly-increasing continuous function of α_i, and approaches 1 as α_i increases to 1 (Figure 4.1a). Eq.(4.31) suggests that $\rho(X_r, Y_i)$ depends on $\rho_{i,0}$ and α_i only, but not on the exact relationship between the variables.

This measure, $\rho(X_r, Y_i)$, is used here to define a criterion for setting α_i such that the pair (X_r, Y_i) have a strictly-increasing monotonic relationship. Let ρ_m represent the lowest value of $\rho(X_r, Y_i)$ that assures the pair X_r and Y_i have a strictly-increasing monotonic relationship. Using Eq. (4.31), then we can calculate the value of α_i, denoted by $\alpha_{i,m}$, that is given by ρ_m:

$$\alpha_{i,m} = \frac{\left({\rho_{i,0}}^2{\rho_m}^4 - {\rho_{i,0}}^2{\rho_m}^2 - {\rho_m}^4 + {\rho_m}^2\right)^{1/2} + {\rho_{i,0}}^2 + \rho_{i,0}{\rho_m}^2 - \rho_{i,0} - {\rho_m}^2}{{\rho_{i,0}}^2 + 2\rho_{i,0}{\rho_m}^2 - 2\rho_{i,0} - 2{\rho_m}^2 + 1} \tag{4.32}$$

Eq. (4.28) can be used to find each $\alpha_{i,m}$ that minimizes the information loss while keeping the degree of linearity of the relationship of X_r and Y_i high enough; that is, a large enough ρ_m (e.g. $\rho_m \geq 0.9$) is selected such that $\beth_i$ remains negative. For example,

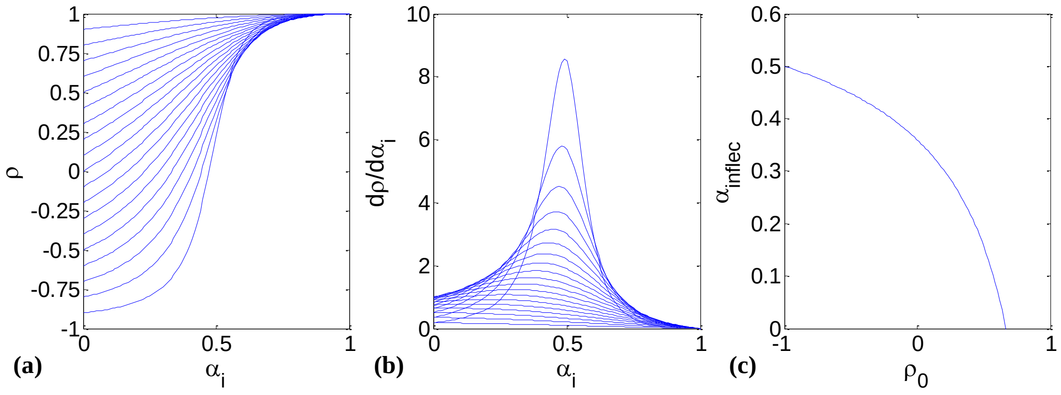

Figure 4.1: (a) Correlation coefficient as a function α_i for different values of ρ_0 ranging from -0.9 to 0.9. (b) Derivative of the correlation coefficient for different values of ρ_0. (c) α_{inflec} as a function of ρ_0.

when $\rho_m = 0.9$, for $\rho_{i,0} = 0.1$ and $\rho_{i,0} = 0.5$, Eq. (4.32) yields $\alpha_{i,m} = 0.66$ and $\alpha_{i,m} = 0.56$, respectively. When $g = 8$ and $h_i = 7$, the corresponding $\beth_i$s will be 0 and 0, repectively. It implies that g and h_i allow achieving $\rho_m = 0.9$ without information loss due to the monotonization. For cases when these conditions ($\rho_m \geq 0.9$ and $\beth_i \leq 0$) cannot be satisfied at the same time, a bi-objective optimization should be performed, where

$$\alpha_{i,m} = \arg\min_{\alpha_i}\left(-\frac{(1-\rho_{i,0})\alpha_i+\rho_{i,0}}{\left(2(1-\rho_{i,0}){\alpha_i}^2-2(1-\rho_{i,0})\alpha_i+1\right)^{\frac{1}{2}}}\ ,-fl(\log_{10}(1-\alpha_i))\right) \tag{4.33}$$

4.3.3.2. Selection of $\alpha_{i,m}$ based on the Derivatives of Relationships

As discussed in the previous section, the linear correlation coefficient is a measure of how linearly X_r and Y_i are correlated. Being a covariance-based function, the correlation coefficient is very sensitive to non-monotonicity of the relationship between the two variables; a small degree of non-monotonicity gives rise to a near zero value of the correlation coefficient (Figure 4.1b). According to Eq.(4.31), the correlation coefficient is an increasing function of α_i; i.e.,

$$\frac{\partial\rho(X_r,Y_i)}{\partial\alpha_i} = \frac{({\rho_{i,0}}^2-1)(\alpha_i-1)}{\left(-2\left(\rho_{i,0}-1\right)\alpha_i^2+2\left(\rho_{i,0}-1\right)\alpha_i+1\right)^{3/2}} \geq 0, \quad \forall\alpha_i \in [0,1],\ \forall\rho_{i,0} \in [-1,1] \tag{4.34}$$

$\rho(X_r, Y_i)$ is an S-shape function of α_i for every $\rho_{0,i}$ with an inflection point at

$$\alpha_{i,inflec} = \frac{\left({\rho_{i,0}}^2 - 18\rho_{i,0} + 17\right)^{1/2} + 7\rho_{i,0} - 7}{8(\rho_{i,0} - 1)} \qquad (4.35)$$

as shown in Figure 4.1c. This behavior suggests that even if $\frac{\partial\rho(X_r,Y_i)}{\partial\alpha_i}$ is nonnegative over the domain of α_i, its value decreases to 0 as α_i goes to 1. In other words, a value close to 1 for α_i blocks the information of X_i, but it does not increase the linearity of Y_i with respect to Y_r substantially. Such a property is used to find an optimal value of α_i using:

$$\alpha_{i,m} = \sup\{\alpha_i | \frac{\partial\rho(X_r,Y_i)}{\partial\alpha_i} \geq s_m,\ \alpha_i \geq \alpha_{i,inflc}\}, \quad i \in \{1, \dots, d\} \qquad (4.36)$$

where s_m is a design parameter set by the user, serving as a means to prevent α_i from getting excessively close to 1. Such an $\alpha_{i,m}$ can be calculated by finding the root of

$$\frac{\partial\rho(X_r,Y_i)}{\partial\alpha_i} - s_m = \frac{({\rho_{i,0}}^2 - 1)(\alpha_i - 1)}{\left(-2\left(\rho_{i,0} - 1\right){\alpha_i}^2 + 2\left(\rho_{i,0} - 1\right)\alpha_i + 1\right)^{3/2}} - s_m = 0 \qquad (4.37)$$

A similar argument about the information loss can be made here. A large enough $\alpha_{i,m}$ calculated using Eq.(4.37) that does not yield a positive $\beth_i$ is an acceptable choice. Otherwise, an optimization such as

$$\alpha_{i,m} = \arg\min_{\alpha_i}\left(\frac{({\rho_{i,0}}^2 - 1)(\alpha_i - 1)}{\left(-2\left(\rho_{i,0} - 1\right){\alpha_i}^2 + 2\left(\rho_{i,0} - 1\right)\alpha_i + 1\right)^{3/2}}\ , -fl(\log_{10}(1 - \alpha_i))\right) \qquad (4.38)$$

subject to $\alpha_{i,m} \geq \alpha_{i,inflc}$ must be performed.

4.3.3.3. Data-Based Selection of $\alpha_{i,m}$

The key point to this method of specifying $\alpha_{i,m}$ is provided by Eq. (4.13). The necessary and sufficient condition for a pair (Y_i, Y_r) to have a strictly-increasing monotonic relation is:

$$\frac{\partial y_i}{\partial y_r} = \frac{\partial y_i}{\partial x_r} > 0, \qquad \forall y_r \in \Omega_{Y_r} \tag{4.39}$$

which implies that:

$$(1-\alpha_i)\frac{\partial x_i}{\partial x_r} + \alpha_i > 0 \tag{4.40}$$

and

$$\frac{\partial x_i}{\partial x_r} > \frac{\alpha_i}{1-\alpha_i}, \quad i \in \{1, \dots, d\},\ i \neq r, \forall x_r \in \Omega_{X_r} \tag{41}$$

Since the function describing the relationship of X_r and X_i is unknown in general, it is not possible to calculate $\frac{\partial x_i}{\partial x_r}$ analytically. In such cases, find the lowest numerically-calculated value of $\frac{\partial x_i}{\partial x_r}$ based on the data, and select $\alpha_{i,m}$ such that $\frac{\alpha_{i,m}}{1-\alpha_{i,m}}$ is lower than the lowest value. For example, $\alpha_{i,m}$ can be calculated using $\alpha_{i,m} = \frac{\delta_{i,m}}{1+\delta_{i,m}}$, where $\delta_{i,m} = e_m \min\left(\frac{\partial x_i}{\partial x_r}\right)$, where $e_m \in [0,1]$ is a design parameter.

4.3.3.4. Maximum Likelihood Estimation of $\alpha_{i,m}$

Maximum Likelihood Estimation (MLE) has been a mainstay for calculating parameters of probabilistic models. It can be applied to a wide range of problems from parameter estimation to model selection.[25] This section proposes an MLE-based method for optimal selection of $\alpha_{1,m}, \dots, \alpha_{d,m}$.

The likelihood of a parameter θ (a random variable) given a random variable π is the conditional probability of π given θ

$$L(\theta|\pi) = \Pr(\pi|\,\theta) = \frac{\Pr(\pi,\theta)}{\Pr(\theta)} \tag{4.42}$$

where $L(.|.)$ and $\Pr(.\,|\,.)$ denote the likelihood function and conditional probability, respectively. For the case of continuous random variables, probability values will be replaced with probability density functions. π or θ may not be indeed random variables (with no joint probability), but they are considered random quantities since their actual states or values are not definitely known to the user.

The MLE method states that the parameter θ which maximizes the likelihood function, has most likely given rise to the observed distribution of π. Therefore, finding an MLE requires a strategy to maximize the likelihood function with respect to its parameter or vector of parameters, even though the likelihood function does not have a closed-form mathematical formula. Using such a strategy should assure finding the global optimum whether analytically or numerically, as the likelihood function may have several local optima.

Let $\boldsymbol{\alpha} = (\alpha_1, \dots., \alpha_d)^T$. Then, the likelihood function of $\boldsymbol{\alpha}$ given historical data $\mathbf{D}$ is:

$$L(\boldsymbol{\alpha}|\mathbf{D}) = f(\mathbf{D}|\boldsymbol{\alpha}) = \prod_{k=1}^{n} f(D_k|\boldsymbol{\alpha}) \tag{4.43}$$

where f denotes the probability density function. It is assumed that the data matrix $\mathbf{D}$ of the size $(d \times n)$ consists of n independent and identically distributed (i.i.d) data vectors D_k with the dimension d, each of which being considered as a realization of the random vector $\mathbf{X}$. Using Eq.(4.21) to replace $f(D_k|\boldsymbol{\alpha})$ yields

$$L(\boldsymbol{\alpha}|\mathbf{D}) = \prod_{k=1}^{n} \left\{ \mathrm{c}\left(F_{Y_1}\big((1-\alpha_1)\chi_{1,k} + \alpha_1\chi_{r,k}\big), \dots, F_{Y_d}\big((1-\alpha_d)\chi_{d,k} + \alpha_d\chi_{r,k}\big) \right) \prod_{i=1}^{d} f_{Y_i}\big((1-\alpha_i)\chi_{i,k} + \alpha_i\chi_{r,k}\big)(1-\alpha_i) \right\} \tag{4.44}$$

where c is a fixed copula function and $\chi_{i,k}$ is the value of the k-th sample of random variable X_i. It is usually more favorable to take the logarithm of the likelihood function. This function, which is called the *log-likelihood*, has the same optima as of the likelihood function besides being expressed in terms of the summations rather than products. Therefore, the log-likelihood function is:

$$\ln\left(L(\boldsymbol{\alpha}|\mathbf{D})\right) = \sum_{k=1}^{n} \ln\left(\mathrm{c}\left(F_{Y_1}\left((1-\alpha_1)\chi_{1,k} + \alpha_1\chi_{r,k}\right), \dots, F_{Y_d}\left((1-\alpha_d)\chi_{d,k} + \alpha_d\chi_{r,k}\right)\right)\right) + \sum_{k=1}^{n}\sum_{i=1}^{d} \ln\left(\ f_{Y_i}\left((1-\alpha_i)\chi_{i,k} + \alpha_i\chi_{r,k}\right)\right) + n\sum_{i=1}^{d}\ln(1-\alpha_i) \qquad (4.45)$$

The optimal values of $\alpha_1, \dots, \alpha_{r-1}, \alpha_{r+1}, \dots, \alpha_d$ are then obtained by solving the $(d-1)$ algebraic equations:

$$\frac{\partial \ln\left(L(\boldsymbol{\alpha}|\mathbf{D})\right)}{\partial \alpha_i} = \sum_{k=1}^{n}\left(\chi_{r,k} - \chi_{i,k}\right)\left(\frac{1}{\mathrm{c}\left(F_{Y_1}(\gamma_{1,k}),\dots,F_{Y_d}(\gamma_{d,k})\right)}\frac{\partial \mathrm{c}\left(F_{Y_1}(\gamma_{1,k}),\dots,F_{Y_d}(\gamma_{d,k})\right)}{\partial F_{Y_i}} f_{Y_i}\left(\gamma_{i,k}\right) + \frac{1}{f_{Y_i}(\gamma_{i,k})}\frac{\partial f_{Y_i}(\gamma_{i,k})}{\partial \gamma_{i,k}}\right) - \frac{n}{1-\alpha_i} = 0, \qquad i = 1, \dots, d,\ i \neq r \qquad (4.46)$$

where $\gamma_{i,k} = (1-\alpha_i)\chi_{i,k} + \alpha_i\chi_{r,k}$. Because of the joint probability of (X_r, X_i) (which is being estimated by the rolling pin method) is unknown, analytical expressions for $\frac{\partial F_{Y_i}}{\partial y_i}$ and $\frac{\partial f_{Y_i}}{\partial y_i}$ cannot be derived. However, considering that $f_{Y_i} = \frac{\partial F_{Y_i}}{\partial y_i}$ and F_{Y_i} can be calculated non-parametrically from data using Eqs. (4.6) or (4.7), it is possible to solve the system of $(d-1)$ equations in Eq.(4.46) numerically to find an optimal $\boldsymbol{\alpha_m}$. Note that the notion of minimizing the information loss is already included in calculating the maximum likelihood estimation, as $\boldsymbol{\alpha_m}$ represents an optimal quantity which gives rise to a distribution that models the historical data best. An initial guess for $\boldsymbol{\alpha_m}$ can be estimated using one of the approaches presented in Sections 4.3.3.1, 4.3.3.2 and 4.3.3.3.

Generally, any numerical global optimization algorithm may be employed to find the MLE of $\boldsymbol{\alpha}_{\mathbf{m}}$ from the objective function of Eq.(4.45), particularly when the computational cost is not a transcendent factor. This same procedure can be used to find an optimal parametric copula from a set of candidate parametric copulas. In this case, the following optimization problem has to be solved:

$$(\mathrm{c}_{opt,}, \boldsymbol{\alpha}_{\boldsymbol{m}}^{T}, \Theta_{opt})^{T} = \mathrm{argmax}_{\mathrm{c}\in\mathbb{C},\boldsymbol{\alpha}\in[0,1]^{d},\Theta}\left(\ln\left(L(\mathrm{c}, \boldsymbol{\alpha}_{\boldsymbol{m}}, \Theta|\mathbf{D})\right)\right) =$$

$$\mathrm{argmax}_{\mathrm{c}\in\mathbb{C},\boldsymbol{\alpha}\in[0,1]^{d},\Theta}\left(\sum_{k=1}^{n}\left(\ln\left(\mathrm{c}\left(F_{Y_1}(\gamma_{1,k}), \dots, F_{Y_d}(\gamma_{d,k})\right)\right) + \sum_{i=1}^{d}\left(\ln f_{Y_i}\left(\gamma_{i,k}\right)\right)\right)\right) +$$

$$n\sum_{i=1}^{d}\ln(1-\alpha_i) \qquad (4.47)$$

where $\mathbb{C}$ is the set of candidate parametric copulas and Θ is the parameter vector corresponding to c.

4.3.3.5. Comparison of the Approaches of α_m Selection

This section briefly compares the four approaches of selecting $\boldsymbol{\alpha}_{\boldsymbol{m}}$. Although the maximum likelihood approach provides a rigorous mathematical way to find an optimal $\boldsymbol{\alpha}_{\boldsymbol{m}}$, it is considerably of higher computational cost. This computational cost is a symptom of two causes. First, a global maximum has to be found. Second, since in general a closed-form mathematical expression cannot be derived for F_{Y_i}, the maximization problem should be solved numerically. On the other hand, the computational costs of the other three approaches are significantly less. Although the optimality of their estimated $\boldsymbol{\alpha}_{\boldsymbol{m}}$ values cannot be shown systematically, their estimated $\boldsymbol{\alpha}_{\boldsymbol{m}}$ values are acceptable as long as they yield transformed variables with strictly-increasing monotonic relationships and their information losses are adequately low.

4.3.4. Selection of the Reference Variable

It may first appear that the reference variable, X_r, can be arbitrarily any of the **X** components. However, as the choice of an appropriate copula function greatly depends on the reference variable, the choice of X_r affects the quality of the joint probability estimation. In this section, several methods are introduced for selecting the reference variable more selectively.

4.3.4.1. Dependence Structure Approach

According to Section 4.3.2, as $\boldsymbol{\alpha_m}$ approaches $(1)_{d\times 1}$, the pairwise dependence structure of each (Y_i, Y_j) can be approximated by the dependence structure (copula) of the (Y_r, Y_r) pair, which is the same dependence structure as of (X_r, X_r). Therefore, selecting X_r such that the dependence structure of (X_r, X_r) is known will help to choose the copula in a more effective and informative manner. For example, if X_r is known to have a Gaussian distribution, the Gaussian copula will be an appropriate approximation of the dependence structure of the random vector **Y**.

4.3.4.2. Witness Variable Approach

There are cases in which none of the variables possess a simple and known distribution describable by a known parametric copula. In such cases a variable called the *witness variable,* X_w, is introduced as the $(d + 1)$-st component of **X**. The witness variable should have the following characteristics: i) it should have a simple distribution with a copula function available, e.g. a Gaussian distribution with a mean of 0 and a variance of 1, and ii) it should be independent of each X_i. This variable serves as X_r. These properties guarantee that the dependence structure of (Y_i, Y_j) can always be approximated by a predetermined copula as of (X_w, X_w). For such an independent witness variable, the correlation function of Eq.(4.31) becomes

$$\rho(X_w, Y_i) = \frac{\alpha_i}{(2\alpha_i{}^2 - 2\alpha_i + 1)^{1/2}} \tag{4.48}$$

and consequently

$$\alpha_{i,m} = \frac{\rho_m\left(1-\rho_m{}^2\right)^{1/2} - \rho_m{}^2}{1-2\rho_m{}^2} \tag{4.49}$$

4.3.4.3. Maximum Likelihood approach

A similar approach like what employed in Section 4.3.3.4 may be used to find an optimal reference variable, such that

$$(X_{r,opt}, \boldsymbol{\alpha}_{\boldsymbol{m}}^T)^T = \operatorname{argmax}_{X_r, \boldsymbol{\alpha} \in [0,1]^d}\left(\ln\left(L(X_r, \boldsymbol{\alpha}_{\boldsymbol{m}} | \mathbf{D})\right)\right) \tag{4.50}$$

such a maximization problem requires to calculate the likelihood each time with a new X_r selected from the set of X_i's, with $\boldsymbol{\alpha}_{\boldsymbol{m}}$ estimated with each selected X_r optimally, where the corresponding copula function is selected from the knowledge on the marginal distribution of X_r or optimally through Eq.(4.47). In this most general case, all adjustable features of the rolling pin distribution may be find optimally using a global optimization scheme; that is:

$$(X_{r,opt}, \mathrm{c}_{opt}, \boldsymbol{\alpha}_{\boldsymbol{m}}^T, \Theta_{opt})^T = \operatorname{argmax}_{X_r, \mathrm{c} \in \mathbb{C}, \boldsymbol{\alpha} \in [0,1]^d, \Theta}\left(\ln\left(L(X_r, \mathrm{c}, \boldsymbol{\alpha}_{\boldsymbol{m}}, \Theta | \mathbf{D})\right)\right) \tag{4.51}$$

4.4. Examples

This section shows the application and performance of the rolling pin method in estimating joint probability distributions using two examples.

4.4.1. Mathematical Example

Consider the following system with three random variables:

$$X_1 \sim N(0,1) \tag{5.52}$$

$$X_2 = \cos(X_1) + \varepsilon_1 \tag{5.53}$$

$$X_3 = \sin^2(X_2) + \varepsilon_2 \tag{5.54}$$

where $N(\mu,\sigma)$ denotes the normal distribution with a mean of μ and a standard deviation of $\sigma \geq 0$, and ε_1 and ε_2 are white noise represented by $N(0,0.1)$ and $N(0,0.05)$, respectively. Eqs.(4.52)-(4.54) offer that the causal structure of the system is $X_1 \rightarrow X_2 \rightarrow X_3$. As the relationships between X_1 and X_2, X_2 and X_3 and X_1 and X_3 are nonlinear and non-monotonic, conventional copulas cannot model this system.

We assume that only historical data (1000 samples) from the system is available; that is, the causal structure of the variables is unknown. To generate the samples, first 1,000 samples are taken from the distribution of X_1 described by Eq.(4.52). Samples of X_2 are generated by adding the cosine of each X_1 sample to a random sample drawn from the distribution of ε_1. A similar procedure is followed to generate 1,000 X_3 samples from X_2 samples. Figures 4.2a, 4.2b and 4.2c represent the sampled data points and the marginal probability densities of X_1, X_2 and X_3 are shown in Figures 4.2d, 4.2e and 4.2f, respectively. Probabilistic models are developed based on the 1000 samples (triplet data points).

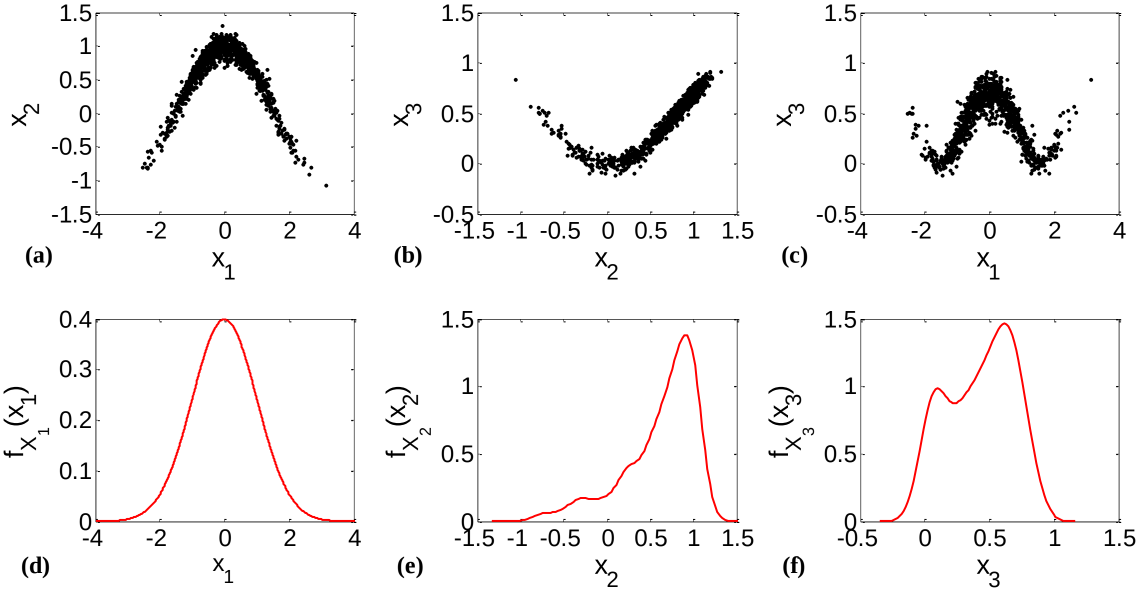

Figure 4.2: Scatter plot of 1,000 training samples of (a) X_2 vs. X_1, (b) X_3 vs. X_2, (c) X_3 vs. X_1. Marginal probability density of (d) X_1, (e) X_2 and (f) X_3.

To model the system with a Bayesian network, one DAG should first be selected from 18 possible DAGs for the triplet (X_1, X_2, X_3). However, the rolling pin method does not require knowing the true causal structure of the system. As the goal here is to model the joint probability distribution of (X_1, X_2, X_3), the choice of the reference variable is arbitrary. Here, X_1 is selected as X_r. Because X_r has a Gaussian distribution with a very well-known copula function, the pairwise dependence structure of the variables will converge to that of Gaussian copula with a right selection of $\boldsymbol{\alpha_m}$ (Figures 4.3d, 4.3e and 4.3f). With $\rho(X_1, X_2) = -0.0486$ and $\rho(X_1, X_3) = 0.005,$ and using$\rho_m = 0.995$, the linear correlation coefficient approach using Eq.(4.32) yields $\alpha_{2,m} = 0.82$, $\alpha_{3,m} = 0.73$, and $\alpha_{r,m} = \alpha_{1,m} = 0$. These values are slightly lower than the optimal values estimated by the method of maximum likelihood. This is because the linear correlation approach satisfies the linearity criterion only, while greater $\alpha_{i,m}$ may be achieved with negligible information loss. This fact is shown comparatively in Table 4.1, where monotonizing parameters derived by the methods described in Section 4.3.3 are compared .The values of the transformed variables (Y_1, Y_2, Y_3) are plotted in Figures 4.3a, 4.3b and 4.3c. The dataset of the monotonized random variables (Y_1, Y_2, Y_3) is used the copula modeling step. After converting the data series into their probability integral transformed form (shown in Figures 4.3d, 4.3e and 4.3f) through the empirical CDFs (F_{Y_i}), a Gaussian copula is applied with the spearman's rank correlation matrix, with elements calculated by Eq.(4.22):

$$\begin{bmatrix} 1.0000 & 0.9997 & 0.9997 \\ 0.9997 & 1.0000 & 1.0000 \\ 0.9997 & 1.0000 & 1.0000 \end{bmatrix}$$

Table 4.1: Monotonizing parameters of the first example derived by different methods in Section 4.3.3.

Method	**Design Parameter**	$\boldsymbol{\alpha_{2,m}}$	$\boldsymbol{\alpha_{3,m}}$
Correlation Coeff.	$\rho_m = 0.995$	0.82	0.73
Correlation Derivative	$s_m = 0.25$	0.84	0.84
Data-Based	$e_m = 0.9$	0.82	0.75
MLE	---	0.86	0.79

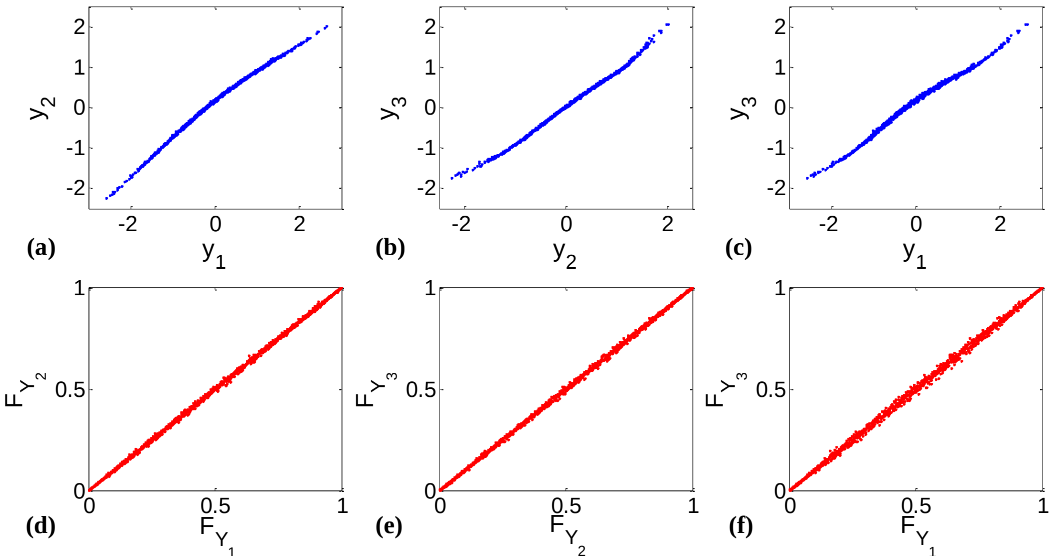

Figure 4.3: Scatter plot of the transformed data (a) Y_2 vs. Y_1 , (b) Y_3 vs. Y_2,(c) Y_3 vs. Y_1, (d) F_{Y_2} vs. F_{Y_1}, (e) F_{Y_3} vs. F_{Y_2}, (f) F_{Y_3} vs. F_{Y_1}.

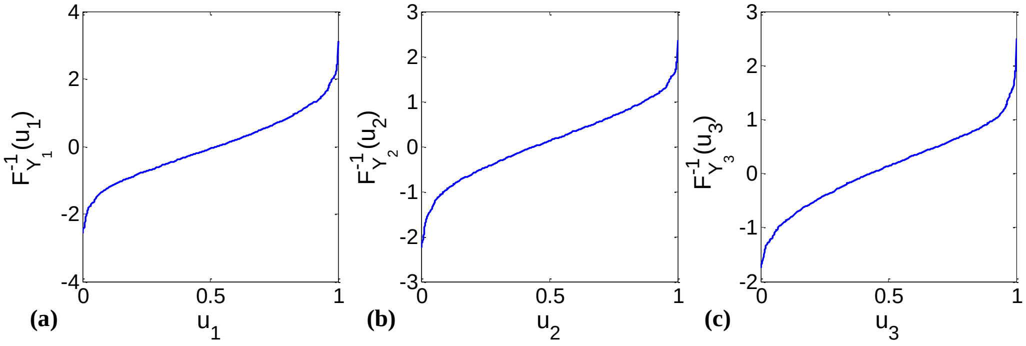

Figure 4.4: Empirical quantile function of (a) Y_1, (b) Y_2, (c) Y_3.

The spearman's correlation matrix validates the initial assumption of the existence of a high level of linear relationship between the random variables (Y_1, Y_2, Y_3), as they are dominated by the information content of the reference variable X_1. This allows one to approximate the copula by the Gaussian copula, as in the case all components of the random vector **Y** are the same dependence structure as (Y_1, Y_1). As it can be seen, with only $8\ ((3-1)+\binom{3}{2}+3)$ parameters and a correct choice of the copula function, the joint probability density of the vector **Y** is fully specified using Eq.(4.21). The joint probability distribution of the original random vector **X** will then be easily estimated by inverting the variable transformation $X_i = \frac{Y_i - \alpha_{i,m} X_1}{1-\alpha_{i,m}}, i = 1, 2, 3$. Figures 4.5a, 4.5b and 4.5c show the contour plots of the estimated joint probability density of **X**, marginalized with respect to the variables X_3, X_1 and X_2, respectively, to calculate the pairwise bivariate probability density functions. Note that the rolling pin method has been able to estimate the skew probability density of Figure 4.5b, even though the non-skew Gaussian copula is used to estimate the dependence structure of the system.

On the other hand, one may desire to take samples from the model distribution of **X** instead. To do so, samples are first taken from the copula function with the rank correlation matrix trained using the transformed data as described above. The procedure to sample the copula mostly depends on the family it belongs to.[26] Samples from the copula then undergo a two-stage transformation. The first transformation converts the samples from the CDF space to **Y** space using the quantile functions of Y_i's. The empirical quantile functions of Y_1, Y_2 and Y_3 are shown in Figures 4.4a, 4.4b and 4.4c. The second transformation converts Y_i's to X_i's. 10,000 samples of each X_i generated by this way are shown in Figures 4.5d, 4.5e and 4.5f. It can be observed that the estimated

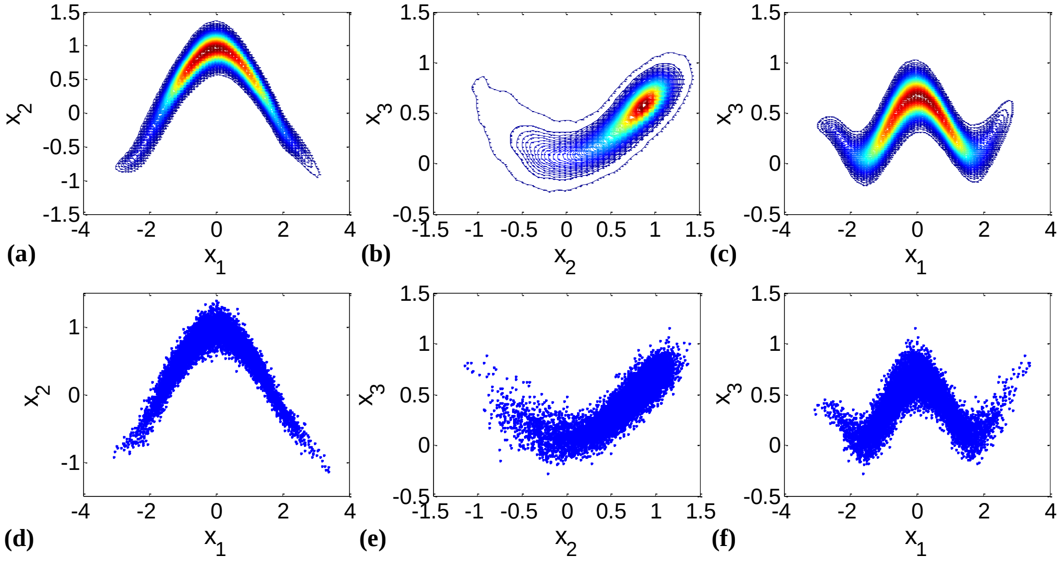

Figure 4.5: Contour plots of the estimated rolling pin probability density of (a) X_2 vs. X_1, (b) X_3 vs. X_2 and (c) X_3 vs. X_1. 10,000 samples from the rolling pin estimated distribution of (a) X_2 vs. X_1, (b) X_3 vs. X_2 and (c) X_3 vs. X_1.

joint probability density function very well represents the non-monotonic relationships among the system variables as shown by the sampled data in Figures 4.2a, 4.2b and 4.2c. It should also be noted that since the training data size is relatively small (1,000 points), the randomness effect caused a slightly longer right tail of the observed data of in the pair (X_1, X_3) (Figure 4.2c), regardless of the symmetry of the distribution. This longer tail is exactly captured by the rolling-pin distribution in Figure 4.5c, and as a result more samples are generated in the right tail of 10,000 samples taken from the rolling pin distribution (Figure 4.5f). This suggests that the random effects can be avoided by using larger data.

4.4.2. Process Example

Consider a continuous stirred tank reactor wherein a first-order exothermic reaction $A \rightarrow B$ takes place. The steady-state behavior of this process is described by:

$$0 = -Z\exp\left(-\frac{E_a}{RT'}\right)C_A' + \frac{C_{A_i} - C_A'}{\tau} \quad (4.55)$$

$$0 = \frac{(-\Delta H)Z}{\rho c}\exp\left(-\frac{E_a}{RT'}\right)C_A' + \frac{T_i - T'}{\tau} + \frac{Q}{\rho c V} \quad (4.56)$$

where Q, T' and C_A' denote the rate of heat removal, the steady-state reaction temperature and steady-state concentration of reactant A . Figures 4.6a, 4.6b and 4.6c depict Q vs. T', Q vs. C_A' and T' vs. C_A', respectively. Here, we assume the system is stochastic, i.e. Q is distributed as $N(4,0.5)$ and

$$T = T' + \varepsilon_1 \quad (4.57)$$

$$C_A = C_A' + \varepsilon_2 \quad (4.58)$$

where $T$and C_A are the measured steady-state temperature and concentration, respectively. ε_1 and ε_2 are white noise variables distributed as $N(0{,}2)$ and $N(0{,}0.1)$. N is defined as in the first example. The model parameter values are listed in Table 4.2.

Similar to the first example, to perform the rolling pin method 1,000 sample were generated by first sampling 1,000 data points from the distribution of Q and then calculating the corresponding samples of $T$and C_A using Eqs.(4.55)-(4.58). Figures 4.7a, 4.7b and 4.7c depict Q vs. T, Q vs. C_A and T vs. C_A sampled data, respectively. As can be seen in Figure 4.6, for each value of Q in the assigned domain, there are three steady-state values for T' and C_A'. On the other hand, for each steady-state T' and C_A', there is only one corresponding Q. Furthermore, T' and C_A' are related monotonically. Therefore, since the rolling pin method is best applicable to functions, we estimate the probabilistic functionality of Q on T or C_A by choosing T or C_A as the reference variable. An important problem here is that both of these variables have complicated unknown marginal distributions. This leads to a pitfall; that is, if one tries to monotonize the remaining two variables with respect to T or C_A, very large values for the monotonizing parameters have to be selected to ensure that the copula takes the form of a comonotonicity copula, leading to a considerable information loss.

To address these problems, we transform the reference variable (here, T) to a new variable, T_N, defined by:

$$T_N = \Phi^{-1}(\hat{F}_T(T)) \tag{4.59}$$

where Φ^{-1} and $\hat{F}_T$ are the inverse standard normal CDF and the empirical CDF of T derived by Eq.(4.6). T_N has a $N(0{,}1)$ distribution and is used as the reference variable. This transformation provides a natural way to make sure T_N has the same order of magnitude as Q and C_A. Also, the use of T_N as the reference variable enables us to capture

Table 4.2: Parameter values of the second example.

Parameter	**Value**	**Unit**
R	8.314	$kJ.kmol^{-1}.K^{-1}$
Z	20,000	min^{-1}
E_a	56,000	$kJ.kmol^{-1}$
$-\Delta H$	50,000	$kJ.kmol^{-1}$
ρ	900	$kg.m^{-3}$
c	2.2	$kJ.kg^{-1}.K^{-1}$
T_i	295.2	K
τ	180	min
V	0.01	m^3
C_{A_i}	10	$kmol.m^{-3}$

the dependence structure of the monotonized variables by the Gaussian copula according to section 4.3.2. In general transformation of Eq.(4.59) can be performed using any inverse CDF, as long as the outcome variable results in a known dependence structure that can be captured by a parametric copula. As $\Phi^{-1}\hat{F}_T(.)$ is a one-by-one function, it is invertible and as a result. Therefore, once the rolling distribution of $(T_N, Q, C_A)^T$ is calculated, deriving the probability distribution of $(T, Q, C_A)^T$ will be straight forward.

After transforming the T data as described above, the monotonizing parameters of Q and C_A are calculated using the maximum likelihood method: $\alpha_{Q,m} = 0.95$ and $\alpha_{C_A,m} = 0.87$. The corresponding rank correlation matrix is then calculated:

$$\begin{bmatrix} 1.0000 & 0.9998 & 0.9999 \\ 0.9998 & 1.0000 & 0.9997 \\ 0.9999 & 0.9997 & 1.0000 \end{bmatrix}$$

The rest of the calculations follow those given in Example 4.1 using the Gaussian copula. Once the rolling pin distribution of $(T_N, Q, C_A)^T$ is obtained, it can be converted to the distribution of $(T, Q, C_A)^T$ by applying the inverse transformation of Eq.(4.59). Figures 4.8a, 4.8b and 4.8c show the corresponding contour plots of the joint probability density generated by this approach. It can be seen that the non-monotone and monotone functionalities between each pair of the variables $(T, Q, C_A)^T$ is very clearly reflected by the probability densities. Although one may find it more appropriate to select a process input variable as the reference variable, in cases where process input variables give rise to multiple values for process output variables or none of the variables have known marginal distributions, the approach described above makes it possible to apply the rolling pin method to model complicated non-monotone relationships.

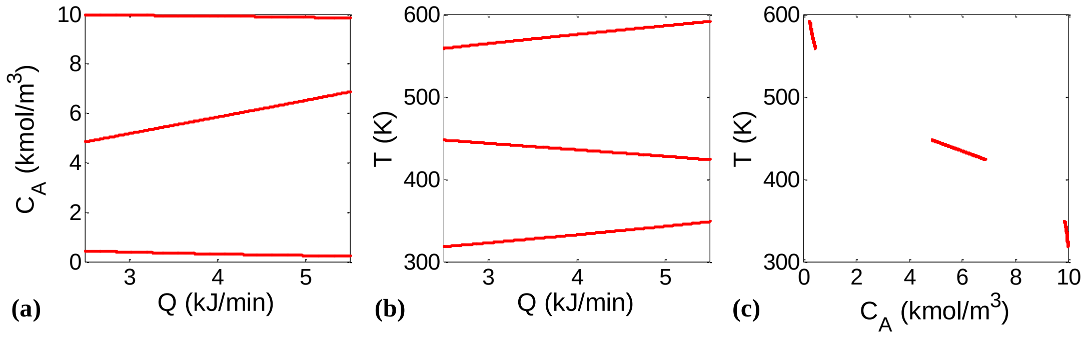

Figure 4.6: Deterministic behavior of (a) C_A' vs. Q, (b) T' vs. Q and (c) T' vs. C_A'.

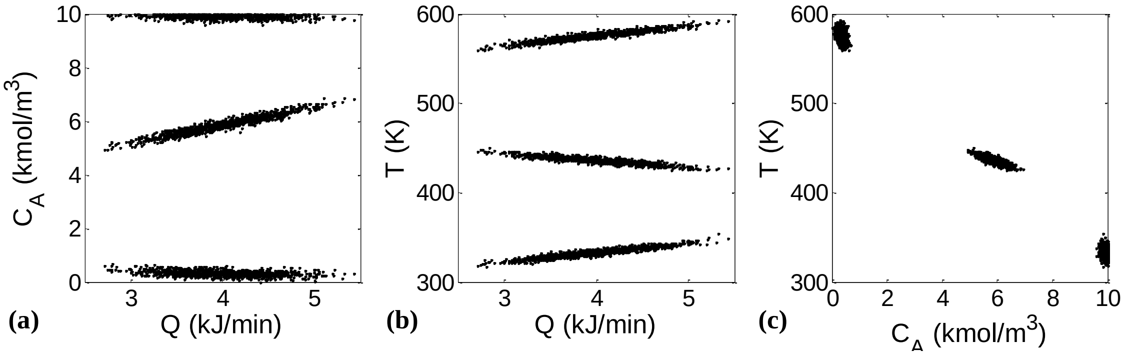

Figure 4.7: 1,000 sampled data of (a) C_A vs. Q, (b) T vs. Q and (c) T vs. C_A.

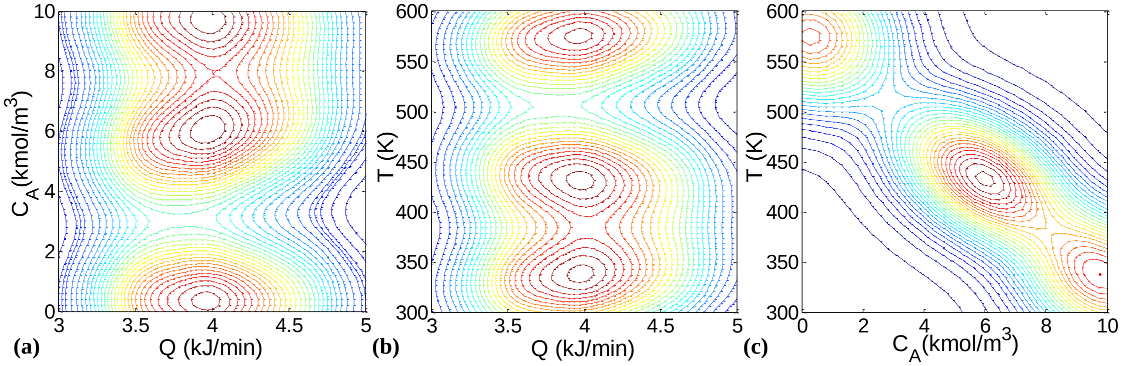

Figure 4.8: Contour plots of the rolling pin-estimated probability density of (a) C_A vs. Q, (b) T vs. Q and (c) T vs. C_A.

4.5. Conclusions

This chapter introduced a novel computationally-efficient and flexible method, named the rolling pin method, of estimating joint probability distribution of highly nonlinear and non-monotonic systems of continuous random variables. There is a broad range of applications for this method in probabilistic modeling and inference of systems with stochastic behavior. As discussed in detail in this chapter, the rolling pin method offers many advantages over its well-known counterparts such as the original parametric copula method, moment-based density estimation and nonparametric techniques of joint probability estimation. The method combines a novel transformation technique with the copula method; this combination offers a powerful tool in modeling multivariate joint probability distributions with arbitrary and not necessarily known pairwise dependence structures among the variables. This implies that the rolling pin method needs no knowledge of the exact dependence structure and its pairwise sameness throughout the system variables. More importantly, the method empowers the copula method to be employed in modeling non-monotonic interactions, which cannot be modeled by the conventional parametric copulas. In summary, the rolling pin method offers the following advantages: 1) the rolling pin method does not require any knowledge of the causal structure of variables, 2) it performs parameter learning significantly fast, with a computational complexity of $O(d^2)$, 3) unlike conventional copulas, the rolling pin method is capable of modeling non-monotonic interactions among variables through the monotonization step, 4) it enables the user to model joint probability distributions over multiple ($d \geq 3$) random variables using a fixed parametric family of copula, regardless of possible differences in the variables pairwise dependence structures, 5) it allows one to model unknown dependence structures with a known one, 6) since it treats random variables as continuous attributes, its estimated probability densities are suitable to model rare events by evaluating the probability values over the states with no historical information.

References

1. Maybeck, P. S. *Stochastic Models, Estimation and Control*; Academic Press: New York, 1982.

2. Spanos, A. *Probability theory and statistical inference: econometric modeling with observational data*; Cambridge University Press: Cambridge, UK, 1999.

3. Pariyani, A.; Seider, W. D.; Oktem, U. G.; Soroush, M. Dynamic risk analysis using alarm databases to improve process safety and product quality: Part II—Bayesian analysis. *AIChE Journal* **2012**, *58*, 826-841.

4. Scott, D. W. *Multivariate density estimation: theory, practice, and visualization (Vol. 383)*; John Wiley & Sons: New York, 2009.

5. Gentle, J. E. Estimation of Probability Density Functions Using Parametric Models. In *Computational Statistics* (pp 475-485); Springer: New York, 2009.

6. Härdle, W.; Müller, M.; Sperlich, S.; Werwatz, A. *Nonparametric and semiparametric models*; Springer: New York, 2004.

7. Horowitz, J. L. *Semiparametric methods in econometrics*; Springer: New York, 1998.

8. Fang, K. T.; Kotz, S.; Ng, K. W. *Symmetric multivariate and related distributions* (pp. 1-220); Chapman and Hall: London, 1990.

9. Johnson, N. L.; Kotz, S.; Balakrishnan, N. *Continuous Multivariate Distributions (volume 1), Models and Applications (Vol. 59)*; John Wiley & Sons: New York:, 2002.

10. Sklar, A. Random variables, distribution functions, and copulas: a personal look backward and forward. In *Distributions with fixed marginals and related topics* (pp 1-14)*;* Institute of Mathematical Statistics: Hayward, Ca, 1996.

11. Ahooyi, T. M.; Soroush, M.; Arbogast, J. E.; Seider, W. D.; Oktem, U. G. Maximum-likelihood maximum-entropy constrained probability density function estimation for prediction of rare events. *AIChE Journal* **2014**, *60*, 1013-1026.

12. Mohseni Ahooyi, T.; Abrogast, J. E.; Oktem, U.; Seider, W. D.; Soroush, M. Estimation of Complete Discrete Multivariate Probability Distributions from Scarce Data with Application to Risk Assessment and Fault Detection. *Industrial & Engineering Chemistry Research* **2014**, *53*, 7538-7547.

13. Scott, D. W. *Multivariate density estimation: theory, practice, and visualization (Vol. 383)*; John Wiley & Sons: New York, 2009.

14. Epanechnikov, V. A. Non-parametric estimation of a multivariate probability density. *Theory of Probability & Its Applications* **1969**, *14*, 153-158.

15. Terrell, G. R.; Scott, D. W. Variable kernel density estimation. *The Annals of Statistics* **1992**, 1236-1265.

16. Olkin, I.; Spiegelman, C. H. A semiparametric approach to density estimation. *Journal of the American Statistical Association* **1987**, *82*, 858-865.

17. Kim, G., Silvapulle; M. J.; Silvapulle, P. Comparison of semiparametric and parametric methods for estimating copulas. *Computational Statistics & Data Analysis* **2007**, *51*, 2836-2850.

18. Sklar, M. *Fonctions de répartition à n dimensions et leurs marges*. Université Paris 8, 229-231, 1959.

19. Nelsen, R. B. *An introduction to copulas*; Springer: New York, 1999.

20. Park, B. U.; Marron, J. S. Comparison of data-driven bandwidth selectors. *Journal of the American Statistical Association* **1990**, *85*(409), 66-72.

21. Trivedi, P. K.; Zimmer, D. M. *Copula modeling: an introduction for practitioners*; Now Publishers Inc.: Boston, 2007 .

22. Bedford, T.; Cooke, R. M. Vines: A new graphical model for dependent random variables. *Annals of Statistics* **2002**, *30*, 1031-1068.

23. Joe, H. *Monographs on Statistics and Applied Probability* (Vol. 73), *Multivariate Models and Dependence Concepts*; Chapman & Hall: London, 1997.

24. Sall, J., Lehman, A.; Stephens, M. L.; Creighton, L. *JMP start statistics: a guide to statistics and data analysis using JMP*; SAS Institute, 2012.

25. Eliason, S. R. *Maximum likelihood estimation: Logic and practice* (No. 96); Sage: Newbury Park, CA, 1993.

Chapter 5: Rolling Pin Method: Efficient General Method of Joint Probability Modeling

5.1. Introduction

Bayesian networks (BNs) (also known as Bayesian belief networks,[2] influence diagrams,[3] and causal networks[4]) have attracted a lot of attention in modeling uncertain knowledge and stochastic systems because of their flexibility, interpretability and natural extension of the human reasoning. The probabilistic inference methods introduced by Pearl[5] and Lauritzen and Spiegelhalter,[6] turned BNs to the mainstay for performing reasoning under uncertainty. Today, BNs have found a broad range of applications in science and technology, including, but not limited to, financial forecasting,[7] weather prediction,[8] medical diagnosis,[9] instrument fault detection and identification,[10,11] and hardware troubleshooting.[12]

Despite their unique capabilities, BNs suffer from multiple issues. First, BNs need an accurate topological structure called the directed acyclic graph (DAG) to be able to properly factorize joint probabilities. Any inaccuracy in the DAG will render the predictions unreliable. There are applications for which the DAG should be extracted from data. Despite efforts made to advance the BN structure learning,[13] available algorithms for learning general Bayesian structures from data are computationally expensive,[14,15] and their generated structures may be unreliable for large and dense networks. Second, both exact and approximate BN inference algorithms are computationally expensive.[16-18] Although algorithms have been developed for performing local inference,[19] the specific structure of DAGs in combination with the BN inference algorithms prevents updating the probability distribution of individual variables given an evidence. Some intermediate variables are often updated in addition to the query variables (nodes).[20] Third, in most of previous studies the attributes have traditionally been assumed to be discrete or multinomial.[21] As a result, in many real-world applications continuous data should be discretized prior to be utilized by the network. The discretization poses multiple problems to the modeling task such as the loss of information due to coarse discretization,[22] and the exponential increase of parameters (conditional probabilities) as finer discretization (higher number of states) is used. This exponential increase translates to exponentially higher computational cost of structure learning, parameter learning and inference. The trade-off between a finer discretization and its resulting computational cost indicates that discretization cut-points should be selected optimally. Such an optimization problem is computationally demanding by its own. Fourth, parameter learning in such discrete networks is usually conducted by the relative frequency method. As a result, states with lack of samples may be left untrained and unused, even though the system is physically realizable in such states.

This chapter introduces a new method that circumvents the BN issues listed above. The method conducts probabilistic inference using a rolling pin joint probability distribution.[1] The rolling pin method uses monotonizing variable-transformations in combination with a parametric copula function. Advantages of this new method of inference over BNs are as follows. First, unlike BNs, this method does not require any knowledge of the causal structure among the variables. Second, it performs the parameter learning and probabilistic inference with computational complexities of $O(d^2)$ and $O(d)$, respectively, which is much faster than its BN counterparts (d denotes the number of system variables). Third, the method allows one to perform probabilistic inference for query variables of interest instead of the entire or unnecessarily large part of the network. Fourth, the rolling pin method is capable of modeling arbitrary joint distributions with non-monotonic interactions among variables. Fifth, the method treats random variables as continuous entities, so no information loss occurs because of the discretization, and there is no need for finding an optimal discretization method. Therefore, this proposed method is not suitable for discrete variables with a few states (such as categorical variables), as the ‘coarse’ discrete nature of these variables cannot be captured by the continuous models. Sixth, the method helps to predict single-variable rare events and complex rare events, where an unlikely event in some variables may lead to an extremely unlikely event of some other variables.

The chapter proceeds as follows. Sections 5.2 and 5.3 briefly review BN modeling and the rolling pin method, respectively. Section 5.4 presents the new method of probabilistic inference using the rolling pin distribution and thoroughly compares inference via the proposed method and BNs. Section 5.5 considers two examples and compares simulation results from inference using the rolling pin method and BNs. Section 5.6 presents some concluding remarks.

5.2. BN Modeling

A BN is commonly considered as a simplified representation of joint probability distributions. This simplification is a result of the ability of BNs in factorizing high dimensional joint probability distributions of the domain variables. This factorization leads to a significant reduction of the parameters (probability values) to be estimated to fully specify joint distributions. BNs borrow the factorization capability from the way their graphical structure; i.e., the directed acyclic graph (DAG), is defined. DAG of a BN is a topological structure, which encodes conditional independence among the variables. This hierarchy of conditional independence may be viewed as a means to specify the causal structure among the variables, where every arc (link) represents a direct cause-and-effect relationship traveling from a cause variable (parent node) to an effect variable (child node). If there is no direct causality in the system, the BN is called an independence map (*I-map*). On the other hand, if every arc of a DAG represents a direct causality, the network is called a dependence map (*D-map*). A network which is both I-map and D-map is called a *perfect-map*. Each node in the network may be descendent (child) of multiple parent nodes. On the other hand, each node may be parent of multiple child nodes. Nodes with no parents are called *root nodes* and nodes with no children are *leaf nodes*.[20] The above lines can be mathematically explained as follows. Let $\Pr(X_1 = x_1, \dots, X_d = x_d) = \Pr(x_1, \dots, x_d)$ denote the joint probability of a random vector $\mathbf{X} = (X_1, \dots, X_d)^T$. Throughout the chapter the random variables are shown by capital letters and their numerical values by small letters. Every joint probability distribution can be factorized using the chain rule; i.e.,

$$\Pr(x_1, \dots, x_d) = \Pr(x_1)\Pr(x_2|x_1) \dots \Pr(x_d|x_1, \dots, x_{d-1}) = \prod_{i=1}^{d} \Pr(x_i|x_1, \dots, x_{i-1}) \quad (5.1)$$

where $\Pr(.\,|.)$ denotes a conditional probability. The BN conditional independence states that given the values of parents, the child node becomes independent of the rest of the network; that is, given a node ordering as above

$$\Pr(x_i | x_1, \dots, x_{i-1}) = \Pr(x_i | \aleph(X_i)) \tag{5.2}$$

where $\aleph(X_i)$ denotes the state of the set of parents of X_i. Therefore, the joint probability of Eq.(5.1) can be written as:

$$\Pr(x_1, \dots, x_d) = \prod_{i=1}^{d} \Pr(x_i | \aleph(X_i)) \tag{5.3}$$

which is the mathematical foundation of BNs. In fact, the DAG is encoded in Eq.(5.3) in the way by which the set of parents of each node is determined. Another important component of BNs appears in Eq.(5.3) as well; i.e., the conditional probability of each node given the state (value) of its parents. Knowing the DAG of a BN in combination with the corresponding conditional probabilities is the sufficient condition to calculate the joint distribution. The advantage offered by this factorization is the reduction in the number of parameters needed to fully specify a joint distribution. For example, determining a joint distribution over binary random variables $(X, Y, Z)^T$ requires estimating $(2^3 - 1 = 7)$ parameters (probabilities), while knowing that the DAG is $X \rightarrow Y \rightarrow Z$ reduces this number to $1 + 2 + 2 = 5$. This difference grows exponentially as the number of variables and their states grow.

However, BNs suffer from many disadvantages. Besides the high computational cost of the inference using BNs, particularly for large-scale systems, there are many cases for which the DAG should be learned from data. Not only is the task of data-driven BN structure learning time consuming, but the available methods often fail to estimate the true causal structure of large and dense networks. Conclusively, the DAG which is considered the greatest strength of BNs may become the most problematic weakness of BNs.

To address these issues, a method is sought in this chapter to estimate joint probability distributions over domain variables with unknown causal relationships and arbitrary functionality between each pair of variables, including highly nonlinear and non-monotonic interactions. The method should be more affordable computationally than BNs, in both parameter learning and probabilistic inference steps. The method should also allow for inference for certain query variables, rather than the entire system. Such a method offers a probabilistic modeling framework that can replace BNs (if the domain variables are continuous). The next section describes the method that has these appealing features.

5.3. The Rolling Pin Method: a Review

As a standard practice throughout this chapter, we first normalize the original random variables $W_1, \dots, W_d$ using

$$X_i = \frac{W_i - \mu(W_i)}{\sqrt{\mathrm{Var}(W_i)}} = \frac{W_i - \mu(W_i)}{\sigma(W_i)} \tag{5.4}$$

where $\sigma(W_i)$ is the empirical standard deviation of W_i and $\mathrm{Var}(W_i)$ denotes the a finite empirical variance of W_i:

$$\mathrm{Var}(W_i) = \frac{1}{n}\sum_{k=1}^{n}\left(w_{i,k} - \mu(W_i)\right)^2 \tag{5.5}$$

n is the number of samples of $w_{i,k}$ for W_i, and $\mu(W_i)$ is the empirical mean of W_i:

$$\mu(W_i) = \frac{1}{n}\sum_{k=1}^{n}(w_{i,k}) \tag{5.6}$$

Therefore, the data of each X_i has a mean value equal to 0 and a variance equal to 1.

Given a vector of normalized continuous random variables $\mathbf{X} = (X_1, \dots, X_d)^T$, let $\mathbf{Y} = (Y_1, \dots, Y_d)^T$ be defined such that

$$Y_i = (1-\alpha_{i,m})X_i + \alpha_{i,m}X_r, \qquad i = 1, \dots, d \tag{5.7}$$

where $\alpha_{i,m} \in [0, 1]$ is a constant parameter, called the *monotonizing parameter* of variable X_i, and X_r is the *reference variable* that is selected systematically from $X_1, \dots, X_d$. As it has been shown in Chapter 4, with an appropriate selection of $\boldsymbol{\alpha_m} = (\alpha_{1,m}, \dots, \alpha_{d,m})^T$, every pair (Y_i, Y_j) is monotonically related and there is a one-by-one correspondence between $\mathbf{X}$ and $\mathbf{Y}$. The elements of the vector of monotonizing parameters $\boldsymbol{\alpha_m} = (\alpha_{1,m}, \dots, \alpha_{d,m})^T$ and X_r are specified using the algorithms given in Chapter 4.

As the relationship of every pair (Y_i, Y_j) has strictly-increasing monotonic relationship, one can model accurately the multivariate cumulative distribution function (CDF) of **Y**, $F_{\mathbf{Y}}$, using a copula function:

$$F_{\mathbf{Y}}(y_1, \dots, y_d) = \mathrm{C}\big(F_{Y_1}(y_1), \dots, F_{Y_d}(y_d)\big) = \Pr(Y_1 \le y_1, \dots, Y_d \le y_d) \tag{5.8}$$

where $F_{Y_1}, \dots, F_{Y_d}$ are the univariate marginal CDFs of $Y_1, \dots, Y_d$, and C denotes an appropriate copula function. Let $\mathbf{y} = (y_1, \dots, y_d)^T$, $\mathbf{x} = (x_1, \dots, x_d)^T$, and $\boldsymbol{J}$ be the Jacobean matrix $\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$, then

$$\det(\boldsymbol{J}) = \prod_{i=1}^{d-1}\big(1-\alpha_{i,m}\big) > 0 \tag{5.9}$$

which confirms that the monotonization transformations of are one-to-one. Because of this one-to-one property and the differentiability of the monotonization transformations, the following equality holds:

$$f_{\mathbf{X}}(\mathbf{x}) = f_{\mathbf{Y}}(\mathbf{y})|\det(\boldsymbol{J})| \tag{5.10}$$

which holds irrespective of the relationships between each pair (X_i, X_j). The preceding multivariate distribution has been called the rolling pin distribution.[1] Here, $F_{\mathbf{X}}, F_{\mathbf{Y}}: \mathbb{R}^d \rightarrow [0,1]$ denote the multivariate CDFs of **X** and **Y**, respectively, and $\mathrm{C}: [0,1]^d \rightarrow [0,1]$ represents a parametric copula as described in Chapter 4. The probability density of **X** is then defined as:

$$f_{\mathbf{X}}(x_1, \dots, x_d) = f_{\mathbf{Y}}(y_1, \dots, y_d) \prod_{i=1}^{d}(1 - \alpha_{i,m}) = \mathrm{c}\big(F_{Y_1}(y_1), \dots, F_{Y_d}(y_d)\big) \prod_{i=1}^{d}(1 - \alpha_{i,m}) f_{Y_i}(y_i) = \frac{\partial^d \mathrm{C}(F_{Y_1}(y_1), \dots, F_{Y_d}(y_d))}{\partial F_{Y_1}(y_1) \dots \partial F_{Y_n}(y_n)} \prod_{i=1}^{d}(1 - \alpha_{i,m}) f_{Y_i}(y_i) \tag{5.11}$$

where $\mathrm{c}: [0,1]^d \rightarrow \mathbb{R}^+ \cap \{0\}$, $f_{\mathbf{X}}, f_{\mathbf{Y}}: \mathbb{R}^d \rightarrow \mathbb{R}^+ \cap \{0\}$ and $f_{X_i}, f_{Y_i}: \mathbb{R} \rightarrow \mathbb{R}^+ \cap \{0\}$denote the copula density, joint density and marginal density functions, respectively. By convention, $\alpha_{r,m}$ is always set equal to 0.

5.4. Performing Probabilistic Inference Using the Rolling Pin Distribution

Although developing a joint probability distribution describing the stochastic interconnections of continuous random variables with general functionality is the main goal of the rolling pin method, the application of the rolling pin distribution in probabilistic inference reveals that the method is indeed a powerful machine learning technique. Probabilistic inference or reasoning refers to set of (mathematical) operations allowing updating the probabilities of a group of random variables, given information (evidence) on other variables. A successful inference usually requires a joint probability distribution of the ensemble of two groups of variables mentioned above, in addition to a well-defined procedure to update probabilities, either analytically or numerically and in the shortest possible time.

The outcomes of the inference process can take on two forms: i) updated (posterior) joint probability distribution of the variables of interest, and ii) updated univariate marginal probabilities of the variables. The second form can be derived directly (through sampling from the updated joint distribution) or by marginalizing the posterior joint probability distribution (through applying analytical or numerical integration). Finding the posterior joint probability distribution of the query variables completely depends of the availability of the joint probability function over (evidence plus query) variables. In the rolling pin method, this joint probability is given by a rolling pin distribution. To perform the inference, all calculations are made in the space of the transformed variables (**Y**) and $\left(F_{Y_1}, \dots, F_{Y_d}\right)^T$, and the results are then transformed back to the original variables (**X**) using the inverse transformations. Let $\mathrm{C}(v_1, \dots, v_d)$ denote a copula distribution function over the variables $(V_1, \dots,\ V_d)^T = \left(F_{Y_1}, \dots, F_{Y_d}\right)^T$. The m-dimensional margin of C, denoted by C_m, is given by

$$\mathrm{C}_m(v_1, \dots, v_m) = \mathrm{C}(v_1, \dots, v_m, 1, \dots, 1) \tag{5.12}$$

Setting $v_i = 1$ here is equivalent to integrating the copula density function with respect to v_i; i.e.,

$$\mathrm{c}_m(v_1, \dots, v_m) = \int_0^1 \dots \int_0^1 \mathrm{c}(v_1, \dots, v_m, v_{m+1} \dots v_d)\, dv_{m+1} \dots dv_d = \frac{\partial^m \mathrm{C}_m}{\partial v_1 \dots \partial v_m} \tag{5.13}$$

Using these definitions, the conditional probability distribution of V_i given the values of a set of evidence variables $(V_j = F_j(y_j), \dots, V_{j+q} = F_{j+q}(y_{j+q})), \mathrm{C}_{cond}\left(v_i | v_j, \dots, v_{j+q}\right)$, is given by:

$$\mathrm{C}_{cond}\left(v_i|v_j,\dots,v_{j+q}\right)=\Pr\left(V_i\le v_i\middle|V_j=v_j,\dots,V_{j+q}=v_{j+q}\right)=$$

$$\left.\frac{\partial^{q+1}\mathrm{C}_{q+2}(v_i,v_j,\dots,v_{j+q})}{\partial v_j,\dots,\partial v_{j+q}}\middle/\frac{\partial^{q+1}\mathrm{C}_{q+1}(v_j,\dots,v_{j+q})}{\partial v_j,\dots,\partial v_{j+q}}\right. \tag{5.14}$$

In terms of the copula density, the conditional density is calculated using:

$$\mathrm{c}_{cond}\left(V_i=v_i|v_j,\dots,v_q\right)=\frac{\mathrm{c}_{q+2}(v_i,v_j,\dots,v_{j+q})}{\mathrm{c}_{q+1}(v_j,\dots,v_{j+q})} \tag{5.15}$$

where $\mathrm{c}_{q+2}\left(v_i,v_j,\dots,v_{j+q}\right)$ and $\mathrm{c}_{q+1}\left(v_j,\dots,v_{j+q}\right)$ are calculated according to Eq. (5.13). Similarly, joint conditional probabilities of a set of variables $(V_i,\dots,V_{i+r})^T$ given $(V_j,\dots,V_{j+q})^T$ are calculated using:

$$\mathrm{C}_{cond}\left(v_i,\dots,v_{i+r}|v_j,\dots,v_{j+q}\right)=\left.\frac{\partial^{q+1}\mathrm{C}_{r+q+2}(v_i,\dots,v_{i+r},v_j,\dots,v_{j+q})}{\partial v_j,\dots,\partial v_{j+q}}\middle/\frac{\partial^{q+1}\mathrm{C}_{q+1}(v_j,\dots,v_{j+q})}{\partial v_j,\dots,\partial v_{j+q}}\right. \tag{5.16}$$

$$\mathrm{c}_{cond}\left(v_i,\dots,v_{i+r}|v_j,\dots,v_q\right)=\frac{\mathrm{c}_{r+q+2}(v_i,\dots,v_{i+r},v_j,\dots,v_{j+q})}{\mathrm{c}_{q+1}(v_j,\dots,v_{j+q})} \tag{5.17}$$

To avoid the unnecessary computational cost required by Eqs.(5.11) and (5.13), we use the cumulative form of the copula function, C, to carry out the inference step. For cases for which the copula CDF does not have a closed form (as in elliptical copulas), the derivatives in Eqs.(5.14) and (5.16) should be calculated numerically. Once for each query variable $\mathrm{C}_{cond}\left(V_i|V_j,\dots,V_{j+q}\right)$ is calculated, these conditional probabilities are used to generate samples for the variable V_i given $V_j,\dots,V_{j+q}$ using conventional sampling techniques. The generated samples for $V_i|V_j,\dots,V_{j+q}$ are then transformed back to $Y_i|Y_j,\dots,Y_{j+q}$ using the quantile function (inverse of CDF) of Y_i:

$$\gamma_{i,k} = F_{Y_i}{}^{-1}\left(\upsilon_{i,k}\right), k = 1, \dots, n. \tag{5.18}$$

and finally, the samples of $X_i|X_j, \dots, X_{j+q}$ are calculated using the inverse of the transformation of Eq.(5.7):

$$\chi_{i,k} = \frac{\alpha_{i,m}(\gamma_{r,k}) - \gamma_{i,k}}{(1-\alpha_{i,m})}, \quad k = 1, \dots, n. \tag{5.19}$$

where $\upsilon_{i,k}$, $\gamma_{i,k}$ and $\chi_{i,k}$ denote the k-th sample of variables V_i, Y_i and X_i, respectively. The samples derived in such a way are later used to estimate the updated probability density functions of the query variables using a parametric or non-parametric density estimation method such as histogram or kernel methods.

5.4.1. A Probabilistic Inference Approach to Determine the Reference Variable

If the final objective of using the rolling pin method is to conduct probabilistic inference, the choice of an appropriate X_r is even more crucial. Consider the case when an arbitrary variable is selected as X_r. If the evidence is given for a set of variables excluding X_r, then none of the variables $Y_i = (1 - \alpha_{i,m})X_i + \alpha_{i,m}X_r$ can be calculated (since X_r remains a random quantity). As a result, the evidence provides no numerical input to the inference process and no updated (conditional) probabilities can be calculated. To address this problem, one of the following proposed solutions may be employed:

1. X_r is selected such that its value can always be determined from the evidence. In other words, although X_r is intrinsically a random variable, but since it is easily measurable at each moment, its status can always be an input to the inference problem.

2. X_r is not selected a priori; it is selected when an evidence becomes available. This means X_r is selected from the variables for which evidence has become available.
3. X_r is selected using the witness variable approach[1].

Although these approaches yield equally good results, the first approach is preferable as it is more computationally favorable.

5.4.2. Comparison with BNs

While BNs have been the most popular framework to carryout probabilistic inference where probabilities of the query variables are updated when evidences are entered, the rolling pin distribution provides a powerful alternative to BNs. In the following paragraphs, major advantages of the rolling pin method over BNs are discussed.

5.4.2.1. Causal Structure

BNs in fact present a factorization of high-dimensional probability distributions, based on the conditional independence and causal structure among the variables. This reduces the number of model parameters (elements of the conditional probability arrays), but it gives rise to the difficult problem of BN *structure learning* from data. When the exact dependence structure of the variables is unknown and cannot be determined from the available knowledge, it is imperative to extract such a structure from data, to build the corresponding BN. There is a wide range of techniques proposed to solve this problem including score-and-search,[23] conditional independence,[24] hybrid,[25] and heuristic techniques.[26] However, none of them provide a general and computationally tractable way to derive the true BN structure from the data, particularly for large-scale networks. Furthermore, inaccuracies present in these techniques make the results unreliable when dealing with large-scale and dense networks. On the other side, unlike BNs, a rolling pin distribution model uses a joint probability distribution constructed over all domain variables without any need to underlying causal structure across the variables; that is, it is not necessary to know anything about the cause-and-effect status of any pair of variables prior to the construction of joint probability distributions. For this reason, the rolling pin method has the advantage of not requiring the time-consuming and somewhat unreliable structure learning step of BNs.

5.4.2.2. Computational Cost

Generally, the implementation of BNs includes three major steps:

i. *Structure learning*: As described earlier, if the exact graph structure of a BN cannot be obtained from the knowledge of the domain variables, it should be estimated from data. It has been proven that the exact structure learning problem is NP-hard in general.[14,15] Finding the exact structure of the trees is polynomial,[27] while learning polytrees is shown to be NP-hard.[28] The computational complexities of a variety of BN structure learning schemes are listed in Table 5.1. Table 5.1 compares the computational costs of different steps of the BN learning and implementation with the equivalent steps of learning the joint probability distribution using the rolling pin method and performing probabilistic inference with it.

ii. *Parameter learning*: This step involves calculating the elements of the conditional probability arrays of the discrete variables of the network. The number of parameters to be learnt (calculated) is proportional to the number of the variables (nodes), number of the states of each node, and the degree of connectivity of individual nodes with their parents. The last two are particularly responsible for the exponential increase of the computational complexity of the parameter learning, with the number of the parameters. For example, the conditional probability array of an S-state variable that has n_{pa} S-state parents has $S(S^{n_{pa}} - 1)$ parameters. It is noteworthy that as the size of the parameter learning increases, the computational cost of the data-based structure learning increases considerably.

iii. *Probabilistic Inference*: Both exact and approximate inferences in BNs are NP-hard problems[16-18] in general. The computational complexities of some well-known BN inference algorithms are as follows. Pearl's *message passing* algorithm has polynomial complexity as a function of the number of domain variables. The computational cost of the *loop cutset conditioning* method for multiply connected networks is exponential in the size of the loop cutest,[29] and also minimizing the size of the loop cutset is NP-hard.[30] The complexity of Lauritzen's *clique-tree propagation* or *clustering method*[6] increases exponentially with the size of the largest clique, and the method becomes very slow for dense network. The *variable elimination* method[31] is NP-hard in optimizing the ordering of the elimination process.

On the other hand, as the rolling pin method estimates a joint probability distribution with no factorization, it eliminates the structure learning step. As mentioned in [1], the rolling pin method has N_{rp} parameters that should be estimated from data, where $N_{rp} = (d-1) + \frac{d(d-1)}{2} + d = \frac{(d+4)(d-1)}{2} + 1$ and d is the number of variables. The

Table 5.1: Computational complexity of the rolling pin method compared to some well-known BN algorithms.

	Parameter learning		Structure learning		Inference	
	method	complexity	method	complexity	method	complexity
BNs	MLE	NP	General exact	NP-hard	General exact	NP-hard
	MAP	NP-hard	Exact tree	Polynomial	General approx.	NP-hard
	EM	NP-hard	Exact polytree	NP-hard	Message passing	Polynomial
					Cut-set conditioning	Exponential in size of largest cut-set
					Clique-tree	Exponential in size of largest clique
					variable elimination	NP-hard
Rolling Pin	$O(d^2)$		Not required		$O(de)$	

parameters are the monotonizing parameters, the correlation parameters and the smoothing parameters (if the marginal kernel densities are used. If the empirical distribution is used, this number will be proportional to d again). The N_{rp} functionality does not depend on the denseness of the causal network. Therefore, the parameter-estimation computational complexity of the rolling pin method is of $O(d^2)$. Finally, as suggested by Section 5.4, inference using the rolling pin method has the computational complexity of $O(de)$, where e denotes the number of evidence variables. A comparison of the computational complexities of learning and inference steps of the rolling pin method and BNs is given in Table 5.1.

5.4.2.3. Inference Over Certain Variables

A basic component of BNs is their DAG structure. The graph determines the conditional independence and direct casualties among variables. It also plays an important role in determining the node ordering by which the probability distributions of the query variables are updated given the evidences. Although several local inference algorithms have been developed,[32] updating the entire BN probabilities is still a common practice. Furthermore, because of BN belief propagation rules, updating probability distribution of a query node is possible only when at least all nodes on the shortest path between the query node and evidence node are also updated. This implies that when updating the belief about the desired query nodes, usually some non-query variables have to be updated. This situation may be computationally problematic, especially when the number of variables grows and/or the inference should be conducted in real-time. On the other hand, according to Eq.(5.14), the rolling pin method enables the user to selectively update the desired query nodes and calculate the posterior probability of each query variable independently of other query or non-query variables. This selective updating reduces the computational complexity significantly to $O(Qe)$, where Q denotes the number of the query variables.

5.4.2.4. Variable Discretization

In many real-world applications, variables are continuous. BNs usually require discretization of continuous variables so that the cost of the computational steps involved BN modeling becomes manageable. Moreover, many of widely-used Bayesian update rules can handle discrete random variables only.[20] The variable discretization partitions continuous variables into ranges (bins), and then BNs consider each interval as a class or category. The discretization has several drawbacks. First, the discretization is always accompanied by an intrinsic irreversible loss of information.[33] Such an information loss is more serious when less number of partitions is used to approximate a continuous variable. On the other hand, increasing the number of partitions gives rise to higher computational complexity (e.g. when estimating the associated probabilities, deriving the posterior distribution or even finding the BN structure). Therefore, there is a trade-off between discretization quality and computational cost. Despite efforts made to reduce the discretization computational cost for BNs,[21] finding an optimal discretization is a hard problem in general.[34] This is mainly because it involves search for an optimal partitioning in a multidimensional space, as in most cases variables are best discretized when their causal interconnections are taken into account. Many such optimization schemes also search for an optimal Bayesian structure simultaneously, which renders the optimal discretization problem even more computationally expensive. On the other hand, as the rolling pin method considers variables in their original continuous form, it does not require discretization, making the method more computationally-efficient and accurate.

5.4.2.5. Problem of Rare Events

BNs mostly rely on the relative-frequency-based techniques (e.g., the maximum likelihood-based methods) to learn the conditional probability values, so they are susceptible to the cases for which there are no data available for certain regions or ranges (rare states) due to the scarcity of data. Although some methods such as the MLME[35] and ME methods[36] have been proposed to estimate the conditional probabilities over the unobserved regions, using BNs for performing inference for rare events is still limited. At the same time, the rolling pin method presents a natural interpretation of rare states that have their near-zero probabilities, which can be predicted by probability density functions of continuous random variables. For this reason, the rolling pin method is appropriate for modeling rare events.

5.5. Examples

This section shows the application and performance of the rolling pin method through two examples. One example is used to compare the rolling pin method and its equivalent BN in terms of the quality of predictive inference, and the other example to compare the two methods in terms of the quality of diagnostic inference. It should be noted that in both examples the BN structures are assumed to be known, and therefore imperfectness of BNs arising from structure learning is not being taken into account. Despite this assumption, as it will be shown, the rolling pin method provides a superior performance in both cases.

5.5.1. Mathematical Example

Consider a system composed of two continuous random variables with an uncertain relationship. The two random variables are X_1 and X_2 governed by:

$$X_1 \sim N(0, 1) \quad (5.20)$$

$$X_2 = \frac{\sin(3X_1)}{X_1} + \ \varepsilon, \varepsilon \sim N(0, 0.2) \ (5.21)$$

where $N(\mu, \sigma)$ denotes the Gaussian distribution with a mean of μ and a standard deviation of $\sigma \geq 0$, and ε is white noise, representing the uncertainty in the relationship between the variables. These equations suggest that the causal structure $X_1 \rightarrow X_2$, meaning X_1 affects X_2. This causal structure and the prior probabilities calculated using the BN are shown in Figure 5.1. For each variable, five states which completely cover the range of the observed (historical) data, are considered. Prior and conditional probabilities are estimated solely based on 1,000 random samples taken from the actual distribution of (X_1, X_2). First, 1,000 samples are simulated using the marginal distribution of X_1 defined by 5.(20), and then these 1000 samples are used to generate 1,000 samples of X_2 according to Eq.(5.21). The same 1,000 sample pairs of (X_1, X_2) are used to construct the rolling pin distribution. To this end, X_1 is used as the reference variable as its empirical distribution is relatively symmetric and close to normal distribution (a measure of the symmetry is the skewness of the distribution). Through the method of

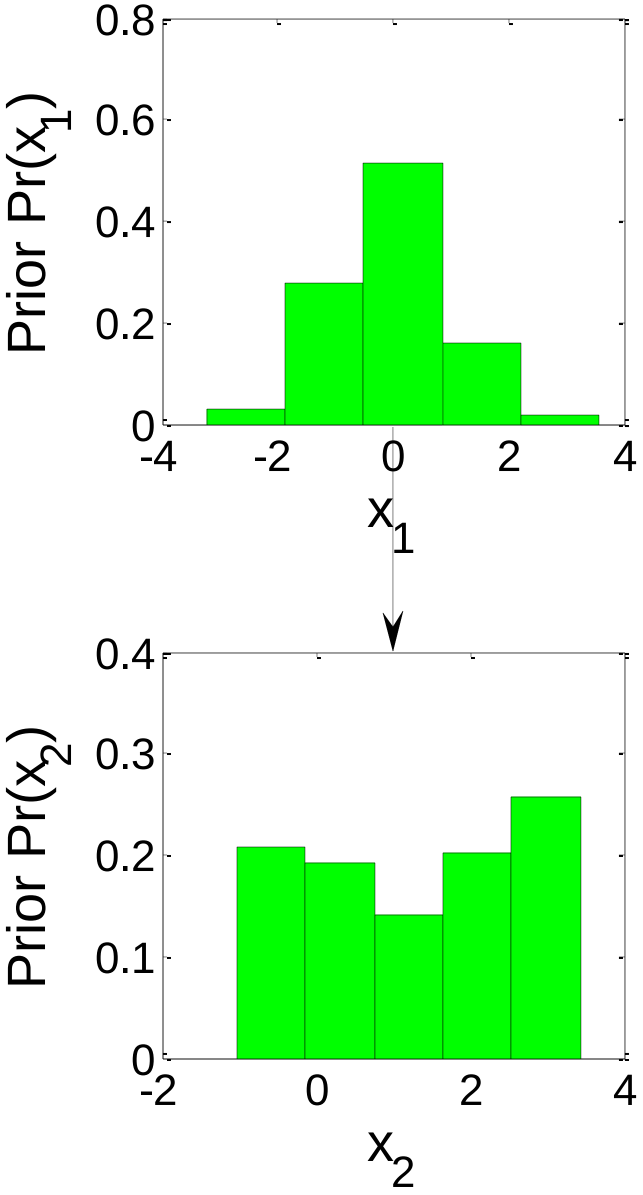

Figure 5.1: BN and prior probabilities of the first example.

correlation coefficient described in Chapter 4, α_m is calculated to be 0.95. 1,000 samples of Y_2 are then calculated using Eq.(5.7). The dependence structure of (Y_1, Y_2) is approximated by the dependence structure of (X_1, X_1) which is represented by the Gaussian copula, with the spearman's rank correlation matrix:

$$\begin{bmatrix} 1.0000 & 0.9999 \\ 0.9999 & 1.0000 \end{bmatrix}$$

Figures 5.2a shows 1,000 samples taken from the actual distribution of (X_1, X_2), and Figure 5.2b depicts the contour plot of the joint probability density function of (X_1, X_2) estimated by the rolling pin method. As can be seen, the rolling pin-method-estimated distribution replicates the behavior observed in the data almost exactly, despite the complex governing equations of the variables. The quality of the estimation will be higher when more data points are available (in vicinity of the mean). In this example, the rolling pin method has 4 parameters and does not need knowledge of the causal structure, while BN has 24 parameters despite the coarse discretization (5 states for each variable). We will show how this coarse discretization will negatively affect the BN inference quality.

Once the joint probabilities are estimated using the BN and the rolling pin distribution, they are compared in terms of inference quality. Here, predictive inference is performed, where the value (state) of the input variable X_1 is given and the goal is to update the belief about the output variable by deriving the posterior probability density (distribution) of X_2. Suppose X_1 is observed at $x_1 = 0.5$, this corresponds to X_1 in its 2nd state in the BN. Given this value (state), the posterior probability of X_2 is calculated.

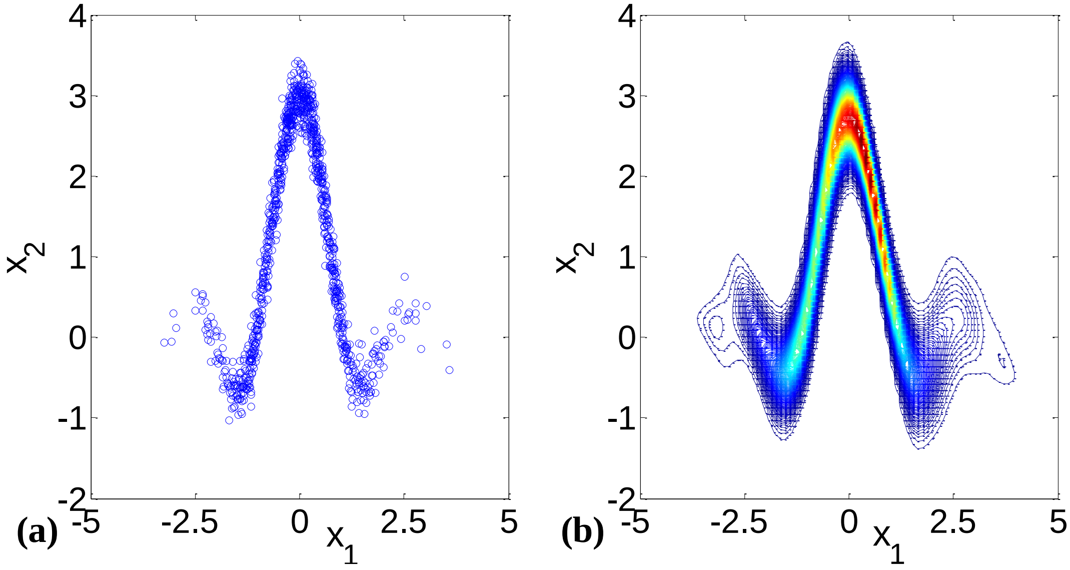

Figure 5.2: (a) 1,000 samples from the distribution of (X_1, X_2), and (b) contour plot of the corresponding rolling pin joint distribution function.

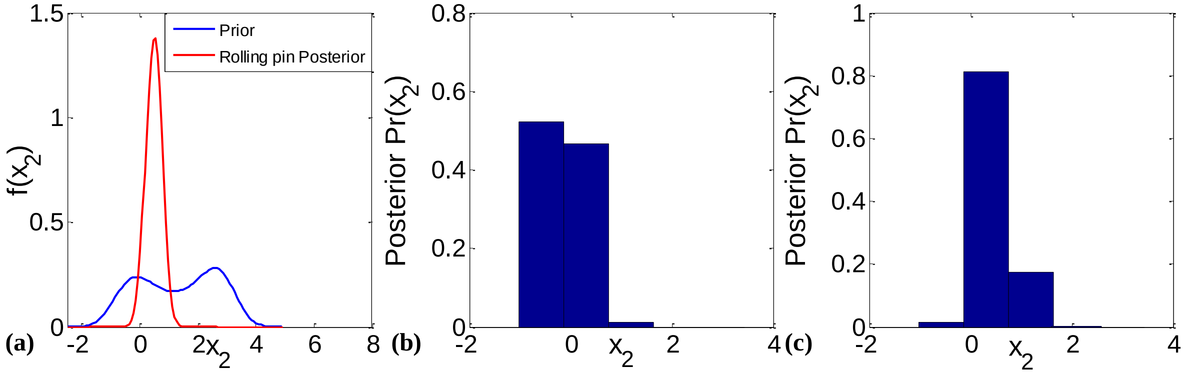

Figure 5.3: (a) Prior and rolling pin-method-calculated posterior density functions of X_2, (b) BN-calculated posterior probability of X_2, and (c) discretized rolling pin-method-calculated posterior distribution of X_2.

Figure 5.3a presents the prior, actual posterior and the rolling pin-method-calculated posterior densities of X_2, Figure 5.3b the discretized posterior probability of X_2, Figure 5.3c the BN-calculated posterior probability distribution of X_2, and Figure 5.3d the discretized rolling pin-method-calculated posterior probability distribution of X_2. The discretization allows comparing the posterior rolling pin- and BN-calculated distributions with the actual posterior probability of X_2. The results indicate that unlike the rolling pin-calculated posterior probability, the BN-calculated posterior probability is an inaccurate representation of the actual posterior probability. This inaccuracy is caused by the discretization of a probability density that bears an irreversible information loss (which increases as less number of states is employed for discretization). Moreover, coarse discretization makes the class labels and attribute values less consistent. As a result, the inference performed given the evidence becomes less reliable as the deviation of the actual value of the evidence from the average value of the corresponding state increases. This trend can be observed in the results; although $x_1 = 0.5$ is in the 2nd state of the variable X_1, as it significantly differs from the state average value, the BN-calculated posterior distribution poorly reflects the effect of this input on the output, rather estimating an average behavior of the output given rise from the entire range of values included in the 2nd state of X_1. On the other hand, increasing the number of states drastically decelerates the BN learning and inference as indicated in Table 5.1.

Finally, an argument can be made about the ability of the BN model in making predictions of the states not shown up in the data. In contrast to the rolling pin method, BN is only able to perform inference for states for which information is available through the historical data. For this reason, since there is no data points observed, say, in the 5th state of X_2 when an instantiation of X_1 is observed in its 2nd states, the posterior probability of X_2 will never shift to the 5th state given the aforementioned evidence, even if it is very unlikely.

5.5.2. Process Example

Consider the stirred tank heating system shown in Figure 5.4. The process at steady state is governed by:

$$\rho_L(F_i - F_o) = 0 \tag{5.22}$$

$$T_o = T_i + Q/\rho_L C_P F_i + \varepsilon_1 \tag{5.23}$$

$$F_o = \frac{h^{1/2}}{R} + \varepsilon_2 \tag{5.24}$$

where ρ_L, C_P, F_i, F_o, T_i, T_o, Q, ε_1, ε_2, h and R denote the liquid density, liquid heat capacity, inlet and outlet flow rates, inlet and outlet temperatures, rate of the thermal energy supplied to the system, two white noise signals, liquid level inside the tank, and the exit pipe resistance, respectively. This model has three applications. First, it is used to generate 1,000 samples representing the process historical data, which will be used as (historical) dataset to train both rolling pin and BN models. Second, the causal structure of the Bayesian model is extracted from the model equations. Third, the model will be used to compare the inference results. Probability distribution functions of the independent variables (roots nodes) and the white noise signals are listed in Table 5.2. The first-principles model parameter values are chosen to be those of water at atmospheric pressure and 25˚C, i.e. $C_P = 4180\,\frac{kJ}{kg^{\circ}\mathrm{C}}$ and $\rho_L = 998\ \frac{kg}{m^3}$. The system's BN and the associated prior probabilities are depicted in Figure 5.5.

Table 5.2: Probability distributions of root nodes (variables) and noise signals of Example 2. $N(\mu, \sigma)$ denotes the Gaussian distribution with a mean of μ and a standard deviation of $\sigma \geq 0$.

Variable (unit)	**Distribution**
$F_i\ (\frac{m^3}{s})$	$N(0.01, 10^{-3})$
T_i (K)	$N(25, 1)$
$Q\ (W)$	$N(10^6,\ 10^5)$
$\varepsilon_1\ (W)$	$N(0, 0.5)$
$\varepsilon_2\ (\frac{m^3}{s})$	$N(0, 0.0002)$

After generating 1,000 random samples using the distributions of Table 5.2 and Eqs.(5.22)-(5.24), the samples are used to train the rolling pin and BN models. If the joint probability distribution trained in this way is marginalized for each of the domain variables in the absence of any evidence, it gives the data-driven prior (normal operation) probability distribution of each node. Two points should be noted here. First, as the rolling pin method treats variables as continuous quantities, it yields probability density functions, unlike BN that yields the probability mass functions for discretized variables. On the other hand, the BN classifies the data into ranges or states. In this case, each variable has three states obtained by dividing the observed data range into equally-sized bins. Since continuous random variables can take infinitely many values, a probability density function is a more natural way to show uncertainty in a variable and allows for higher resolution calculations.

There are two major types of probabilistic inference. The forward inference (prediction) updates the probability of the effect variables given the state of the cause variable. The backward inference (diagnosis) invloves updating the belief about the cause variable given evidence on an effect variable. In this example, the evidence is considered to happen for the outlet temperature T_o, which is an effect variable that has three cause variables F_i, Q and T_i. The objective of performing a probabilistic inference of this kind is to investigate the most probable cause to the observed abnormality in the effect variable. To this end, it is assumed that the abnormal situation is T_o at 60˚C, a value significantly higher than its data-based mean value. Once the evidence is provided, it can be directly fed into the rolling pin inference algorithm described in Section 5.4. Inference is then conducted selectively for the variables of interest. Figures 5.6a, 5.6b and 5.6c

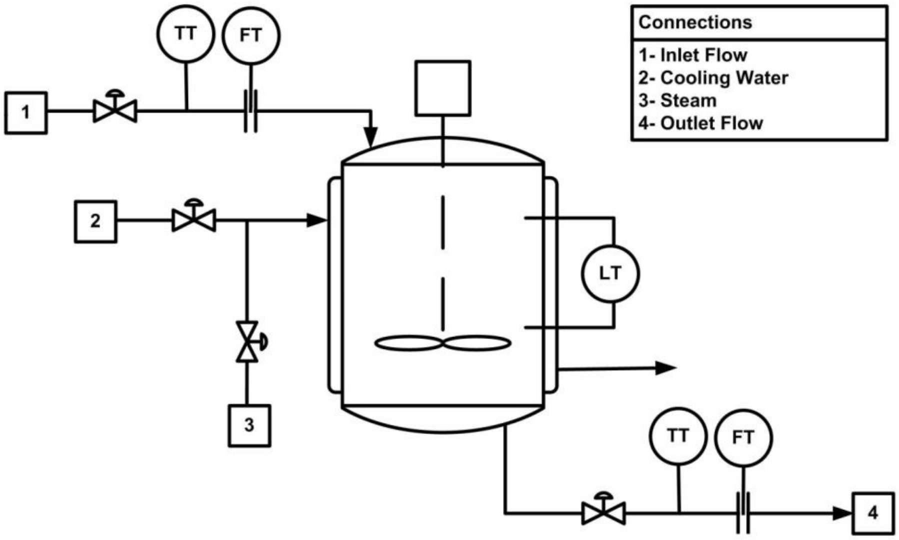

Figure 5.4: Schematic of the heating tank of the second example.

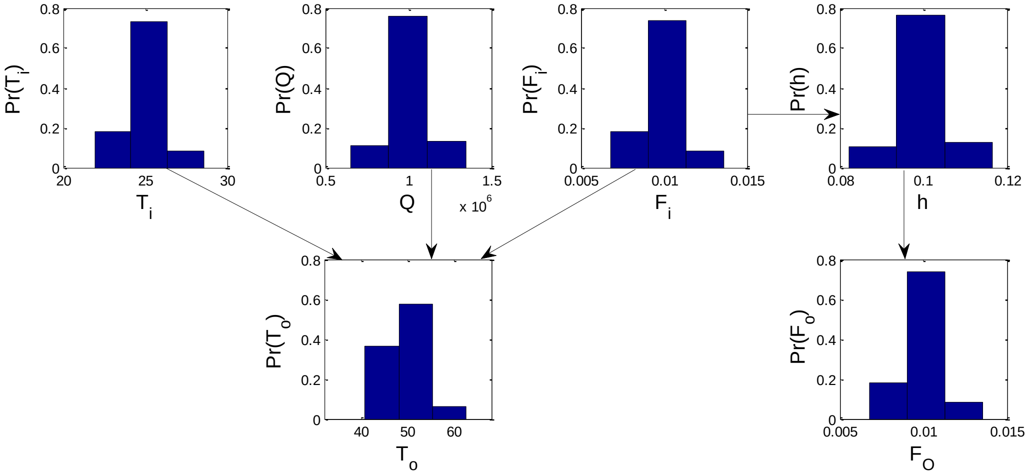

Figure 5.5: BN structure and prior probabilities of the variables of the heating tank example.

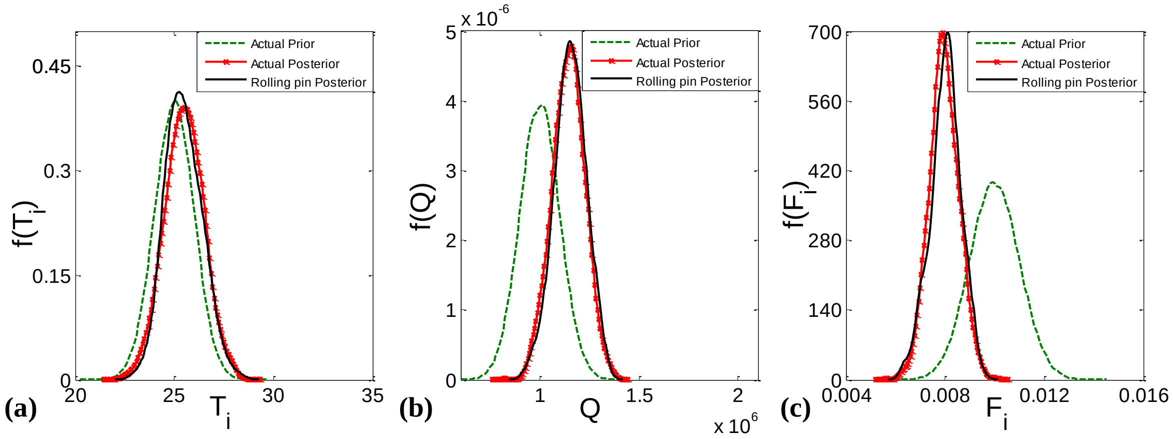

Figure 5.6: Actual prior, actual posterior and rolling pin-method-calculated posterior distributions of (a) T_i, (b) Q and (c) F_i.

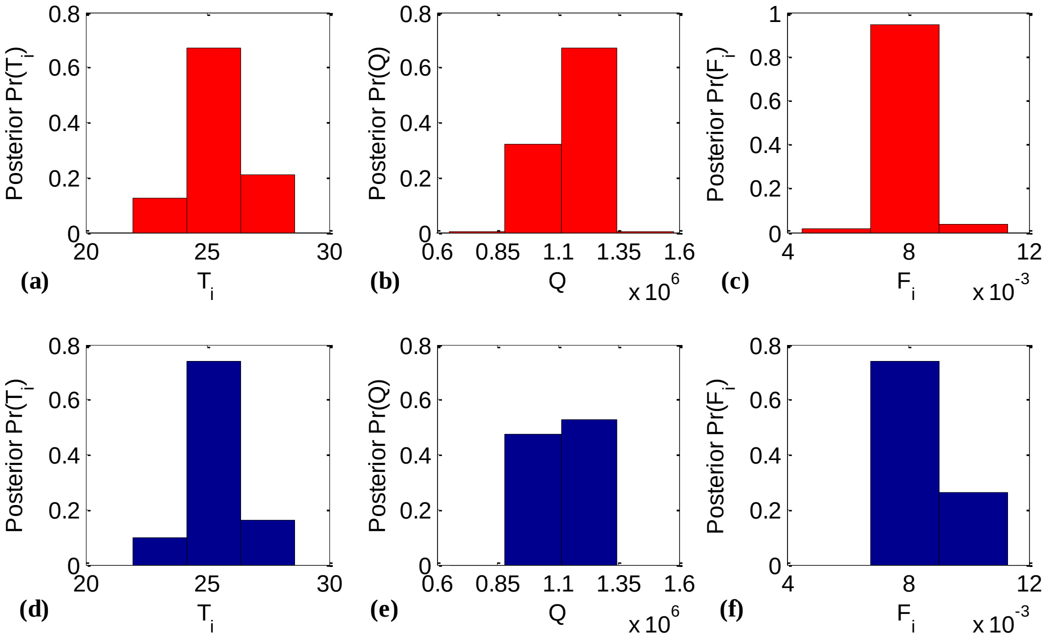

Figure 5.7: Discretized rolling pin-method-calculated posterior probabilities: (a) T_i, (b) Q and (c) F_i. BN-calculated posterior probabilities of the BN: (d) T_i,(e) Q and (f) F_i. Both methods trained by 1,000 samples.

compare the posterior probability density of the query nodes derived by 1,000,000 samples from the model described using Eqs.(5.22)-(5.24) and the rolling pin method with the prior probabilities of the query variables. The discretized posterior probabilities of the rolling pin method trained by the set of 1,000 samples are shown in Figures 5.7a, 5.7b and 5.7c. Here, the reference variable is selected to one of the parent nodes, since they all have Gaussian distributions. It can be seen that the prediction of the rolling pin method almost exactly fits the actual posterior densities, and is consistent with the primary intuitive expectation of an increase in heat rate Q or a decrease of F_i. Surprisingly, the evidence has a small effect on T_i. This suggests that based on the historical behavior of the system, whether T_i is too narrowly distributed or it has had relatively negligible effect on T_o compared to two other parents. Also Q posterior density function demonstrates a more sensible changes than F_i. Analogous to the description above, this implies that Q is a more probable candidate for the observed abnormality of T_o. Figures 5.7d, 5.7e and 5.7f compare the result of the Bayesian diagnostic inference on updating the probability of T_o to their counterparts in Figures 5.7a, 5.7b and 5.7c. The prediction by the BN represents a similar trend in the deviation of the query nodes. However, as can be seen, these probability distributions reflect less detail, due to the coarse discretization. As expected, unlike the rolling pin method the BN does not expand the posterior distributions to less likely states, due to BN's limitation on handling the so-called rare events; i.e., the states where no training data are available for. The final point made here is that the inference using the rolling pin method only requires updating the variables of interest, parents of T_o, rather than what is done by ordinary BNs. In the case of this small example network, the result is significant: only 3 posterior distributions are updated by the rolling pin method given the evidence, compared to 5 nodes updated in the BN. This difference will be much more considerable in large scale network and will result in a highly targeted probabilistic inference.

5.6. Conclusions

This chapter introduced a computationally efficient and flexible framework to perform probabilistic inference over highly nonlinear and non-monotonic systems of random attributes. As discussed in this chapter, the probabilistic inference is performed with the help of a novel joint probability distribution function introduced in our recent paper.[1] The *rolling pin* method combines monotonized random variables with a copula function. This combination allows for modeling multivariate joint probabilities with unknown and not necessarily identical pairwise dependence structures with non-monotonic interactions among the variables. The resulting joint probability distribution replaces the joint distribution constructed by a BN. The method offers many unique advantages over its well-known counterpart, the Bayesian network framework. First, unlike BNs, the rolling pin method does not require any knowledge about the causal structure among the variables, therefore the computational cost and inaccuracies due to the BN structure learning will be eliminated. Second, it performs the parameter learning and probabilistic inference with the computational complexity of $O(d^2)$ and $O(d)$, respectively, which is significantly faster than BNs. Third, the method allows one to perform probabilistic inference for any set of certain query variables of interest with no need to update the intermediate variables or the entire network. Fourth, since the rolling pin method treats random variables as continuous entities, its prior and posterior estimated probability densities may be used to predict single-variable rare events and complex rare events no data available in the historical dataset, where an unlikely event in some variables may lead to an extremely unlikely event of some other variables. Fifth, it does not need discretization of continuous variables, so it decreases information loss and computational cost, and accelerates modeling and inference processes.

References

1. Mohseni Ahooyi, T.; Abrogast, J. E.; Soroush, M. Rolling Pin Method: a Novel Efficient General Method of Joint Probability Modeling. Submitted to *Industrial & Engineering Chemistry Research.*

2. Cheng, J.; Bell, D. A.; Liu, W. An algorithm for Bayesian belief network construction from data. In *proceedings of AI & STAT'97*, 1997; pp 83-90.

3. Shachter, R. D. Evaluating influence diagrams. *Operations research* **1986**, *34*, 871-882.

4. Heckerman, D. A Bayesian approach to learning causal networks. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, 1995; pp 285-295.

5. Pearl, J. *Probabilistic reasoning in intelligent systems: networks of plausible inference*; Morgan Kaufmann: San Francisco, 1988.

6. Lauritzen, S. L.; Spiegelhalter, D. J. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society. Series B (Methodological)* **1988**, *50*, 157-224.

7. Shenoy, C.; Shenoy, P. P. Bayesian network models of portfolio risk and return. In *Computational Finance 1999;* MIT Press: Cambridge, MA, 2000.

8. Kennett, R. J.; Korb, K. B.; Nicholson, A. E. Seabreeze prediction using Bayesian networks. In *Advances in Knowledge Discovery and Data Mining*; Springer: Berlin Heidelberg, 2001.

9. Andreassen, S.; Jensen, F. V.; Olesen, K. G. Medical expert systems based on causal probabilistic networks. *International journal of bio-medical computing* **1991**, *28*, 1-30.

10. Mehranbod, N.; Soroush, M.; Panjapornpon, C. A method of sensor fault detection and identification. *Journal of Process Control* **2005**, *15*, 321-339.

11. Mehranbod, N.; Soroush, M.; Piovoso, M.; Ogunnaike, B. A. Probabilistic model for sensor fault detection and identification. *AIChE Journal* **2003**, *49*, 1787-1802.

12. Barco, R.; Nielsen, L.; Guerrero, R.; Hylander, G.; Patel, S. Automated troubleshooting of a mobile communication network using Bayesian networks. In *Mobile and Wireless Communications Network, 2002. 4th International Workshop on* IEEE, 2002; pp 606-610.

13. Neapolitan, R. E. *Learning Bayesian networks (Vol. 38)*; Prentice Hall: Upper Saddle River, 2004.

14. Chickering, D. M. Learning Bayesian networks is NP-complete. In *Learning from data* (pp 121-130); Springer: New York, 1996.

15. Chickering, D. M.; Heckerman, D; Meek, C. Large-sample learning of Bayesian networks is NP-hard. *The Journal of Machine Learning Research* **2004**, *5*, 1287-1330.

16. Cooper, G. F. The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial intelligence* **1990**, *42*, 393-405.

17. Dagum, P.; Luby, M. Approximating probabilistic inference in Bayesian belief networks is NP-hard. *Artificial intelligence* **1993**, *60*, 141-153.

18. Abdelbar, A. M.; Hedetniemi, S. M. Approximating MAPs for belief networks is NP-hard and other theorems. *Artificial Intelligence* **1998**, *102*, 21-38.

19. Díez, F. J. Local conditioning in Bayesian networks. *Artificial Intelligence* **1996**, *87*, 1-20.

20. Korb, K. B.; Nicholson, A. E. *Bayesian artificial intelligence*; CRC Press: Chicago, 2003.

21. Monti, S.; Cooper, G. F. A multivariate discretization method for learning Bayesian networks from mixed data. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, Morgan Kaufmann Publishers Inc., 1998; pp 404-413.

22. Janssens, D.; Brijs, T.; Vanhoof, K.; Wets, G. Evaluating the performance of cost-based discretization versus entropy-and error-based discretization. *Computers & operations research* **2006**, *33*, 3107-3123.

23. Suzuki, J. Learning Bayesian belief networks based on the MDL principle: an efficient algorithm using the branch and bound technique. *IEICE Transactions on Information and Systems E82-D(2)* **1996**, 356–367.

24. Spirtes, P.; Glymour, C. N.; Scheines, R. *Causation, prediction, and search (Vol. 81)*; MIT press: Cambridge, MA, 2000.

25. Acid, S.; de Campos, L. M. A hybrid methodology for learning belief networks: BENEDICT. *International Journal of Approximate Reasoning* **2001**, *27*, 235-262.

26. Burge, J.; Lane, T. Improving Bayesian network structure search with random variable aggregation hierarchies. In *Machine Learning: ECML 2006*; Springer: Berlin Heidelberg, 2006.

27. Meilă, M.; Jaakkola, T. Tractable Bayesian learning of tree belief networks. In *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence*, Morgan Kaufmann Publishers Inc., 2000; pp 380-388.

28. Dasgupta, S. Learning polytrees. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, Morgan Kaufmann Publishers Inc., 1999; pp 134-141.

29. Pearl, J. Fusion, propagation, and structuring in belief networks. *Artificial intelligence* **1986**, *29*, 241-288.

30. Guo, H.; Hsu, W. A survey of algorithms for real-time Bayesian network inference. In *AAAI/KDD/UAI02 Joint Workshop on Real-Time Decision Support and Diagnosis Systems*, Edmonton, Canada, 2002.

31. Zhang, N. L.; Poole, D. A simple approach to Bayesian network computations. *In Proc. of the Tenth Canadian Conference on Artificial Intelligence*, 1994, pp 171-178.

32. Jensen, F. V.; Lauritzen, S. L.; Olesen, K. G. Bayesian updating in causal probabilistic networks by local computations. *Computational statistics quarterly* **1990**, *4*, 269-282.

33. Dougherty, J.; Kohavi, R.; Sahami, M. Supervised and unsupervised discretization of continuous features. In *ICML*, 1995, pp 194-202.

34. Kotsiantis, S.; Kanellopoulos, D. Discretization techniques: A recent survey. *GESTS International Transactions on Computer Science and Engineering* **2006**, *32*, 47-58.

35. Ahooyi, T. M.; Soroush, M.; Arbogast, J. E.; Seider, W. D.; Oktem, U. G. Maximum-likelihood maximum-entropy constrained probability density function estimation for prediction of rare events. *AIChE Journal* **2014**, *60*(3), 1013-1026.

36. Mohseni Ahooyi, T.; Abrogast, J. E.; Oktem, U. G.; Seider, W. D.; Soroush, M. Estimation of Complete Discrete Multivariate Probability Distributions from Scarce Data with Application to Risk Assessment and Fault Detection. *Industrial & Engineering Chemistry Research* 2014, *53*(18), 7538-7547.

Chapter 6: Rolling Pin Method: Efficient General Method of Joint Probability Modeling

6.1. Introduction

Regression analysis is a statistical approach to describe the quantitative relationship between a set of input variables (also called features, predictive variables, explanatory variables, regressors, etc.) and a set of output variables (also called response variables) based on observational or experimental data. Regression is also known as a supervised machine learning technique whose output variables are usually continuous attributes.[1] Over the past century, numerous regression methods have been introduced, including, but not limited to, linear regression[2], Gaussian process regression[3], nonlinear regression[2], random effect models[4] logistic regression[5], and Bayesian regression.[6] These methods vary considerably in the way they utilize data to develop and train a model. Along with the regression methods, methods of selecting and validating regression models and techniques for measuring goodness-of-fit of regression models[2] have also been developed.

The available regression methods can be divided into three main categories: parametric methods, non-parametric methods, and semi-parametric methods. Parametric methods use a predefined parametric mathematical formula to relate input variables to output variables, where the parameters are estimated from data by optimizing a goodness-of-fit measure.[7,8] Parametric regression models are relatively easy to train and implement. On the other hand, if the model is misidentified, the resulting regression model fails to replicate the actual underlying mechanism that has given rise to the observed data. Although some methods have been introduced to select the parametric model and its validation, this process may be computationally demanding. Also, parametric models may suffer from a huge increase in the number of parameters to be estimated as the system dimensions increase.[9] Non-parametric methods need no predefined mathematical model. They assign a function to each data point and take the mean of these functions as a regression model.[10] Semi-parametric methods combine parametric and nonparametric methods to develop a regression model.[11] In the last two categories, the regression model has to be estimated (identified) fully or partially from data; therefore more data points are required, and the convergence rate is lower compared to parametric methods.[9] However, these two categories are more flexible frameworks for developing a regression model.

One of the methods used for semi-parametric regression is the copula method. A copula is a cumulative joint probability distribution function whose domain is a unit hypercube. Copulas are used to capture the so-called dependence structure of domain variables.[12] Through the Sklar's theorem[13] one can estimate a joint probability distribution using an appropriate copula and the univariate marginal distributions of domain variables. This estimation is considered semi-parametric if either the copula or marginal distributions (but not both) are constructed non-parametrically. The most common semi-parametric estimation is when the margins are defined non-parametrically and copulas parametrically. This will reduce the computational cost while maintain flexibility at an acceptable level. Such a joint probability distribution can be the basis for semi-parametric regression.[14] Gaussian copula regression has been used extensively for cases where the marginal distributions are non-Gaussian while the dependence structure remains Gaussian.[15,16] Non-Gaussian parametric copulas have also been used to estimate (identify) the regression model. Examples include Archimedean copulas.[17,18] A variety of copula-based inference methods have been introduced under the linearity assumptions or using linearization functional.[19] While not explicitly referring to the regression analysis, in several studies;[20,21,22] similar functions have been derived by calculating the copula conditional independence. A mixture of copulas is used to create more complex non-Gaussian copulas describing the feature-response relationship.[23] Constructing the copulas based on the affine generalized hyperbolic distributions is also considered as a way of generating more complicated parametric copulas applicable to the regression models.[24]

Although parametric copulas along with non-parametrically-determined marginal distributions offer a reasonably flexible framework to perform a tractable regression analysis, their application is limited by the following facts: 1) ordinary parametric copulas cannot capture non-monotonic relationships between features and response, 2) there are a limited number of parametric copula families in the literature that are, of course, not representative of every possible dependence structure, and 3) every single parametric copula assigns the same dependence structure to each pairwise combination of the domain variables, which does not necessarily hold in reality.

This work presents a new copula-based semi-parametric method of identifying regression models. This method uses the *rolling pin* method[22,25] to calculate joint probability distributions. As the rolling pin method-estimated joint probabilities can capture non-monotonic interactions, the regression method is capable of modeling non-monotonic behavior appearing in observed data. The method is semi-parametric as it combines parametric copulas with non-parametrically-estimated univariate marginal distributions. It provides regression models with a relatively low number of parameters, which grows quadratically with the number of input variables.[22] This last property is particularly appealing when a regression model is to be identified for a large-scale system. Furthermore, the method can be easily applied to the systems with input-dependent noise terms. The rolling pin distribution provides a well-defined mathematical background for estimating the confidence intervals and other statistical properties of the regression model.

The chapter proceeds as follows. Section 6.2 presents a brief review of the rolling pin method of joint probability estimation. The proposed regression method is described in Section 6.3. Section 6.4 shows the application and performance of the proposed method using two examples. The chapter ends with concluding remarks in Section 6.5.

6.2. Preliminaries and a Brief Review of the Rolling Pin (RP) Method

Throughout this chapter, every random variable is normalized using its empirical mean and variance. Given the samples $w_{i,k}$ of a random variable W_i, its corresponding normalized random variable X_i is defined as:

$$X_i = \frac{W_i - \mu(W_i)}{\sqrt{\mathrm{Var}(W_i)}} \tag{6.1}$$

where $\mu(W_i)$ is the empirical mean of W_i:

$$\mu(W_i) \ = \frac{1}{n}\sum_{k=1}^{n}\left(w_{i,k}\right) \tag{6.2}$$

and $\mathrm{Var}(W_i)$ is the empirical variance of W_i

$$\mathrm{Var}(W_i) = \frac{1}{n}\sum_{k=1}^{n}\left(w_{i,k} - \mu(W_i)\right)^2 \tag{6.3}$$

which is assumed to take a finite ad non-zero value, and n denotes the number of samples of the random variable W_i. Therefore, X_i has an empirical mean value of 0 and an empirical variance of 1.

Given a vector of normalized continuous random variables $\mathbf{X} = (X_1, \dots, X_d)^T$, the random vector $\mathbf{Y} = (Y_1, \dots, Y_d)^T$ is defined by the one-to-one *monotonization transfromation*

$$Y_i = (1 - \alpha_i)X_i + \alpha_i X_r, \quad i = 1, \dots, d \tag{4}$$

where $\alpha_i \in [0,1)$ is the *monotonizing parameter* of the random variable X_i, and X_r is the *reference* variable, selected optimally from the elements of $\mathbf{X} = (X_1, \dots, X_d)^T$. $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_d)^T$ will be referred to as the vector of the *monotonizing parameters* of random vector $\mathbf{X}$. It is shown in[25] that, with an appropriate selection of $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_d)^T$, denoted by $\boldsymbol{\alpha_m}$, every pair of variables (Y_i, Y_j), $i, j \in \{1, \dots, d\}$ are monotonically related.The elements of the vector of monotonizing parameters $\boldsymbol{\alpha_m} = (\alpha_{1,m}, \dots, \alpha_{d,m})^T$ and X_r are selected according to the guidelines presented in Chapter 4.[25]

As the relationship between every Y_i and Y_j is a strictly-increasing monotonic relationship, the multivariate cumulative density function (CDF) of $\mathbf{Y}$, $F_{\mathbf{Y}}$, can be modeled using an appropriate parametric copula function, $\mathrm{C}: [0,1]^d \to [0,1]$:

$$F_{\mathbf{Y}}(y_1, \dots, y_d) = \Pr(Y_1 \le y_1, \dots, Y_d \le y_d) = \mathrm{C}\left(F_{Y_1}(y_1), \dots, F_{Y_d}(y_d)\right) \tag{6.5}$$

where $F_{Y_i}: \mathbb{R} \to [0,1]$ represents the marginal CDFs of Y_i. Since the monotonization transformation is one-to-one it is shown in our paper[25],

$$f_{\mathbf{X}}(\mathbf{x}) = f_{\mathbf{Y}}(\mathbf{y})|\det(\boldsymbol{J})| \tag{6.6}$$

where $f_{\mathbf{X}}: \mathbb{R}^d \to \mathbb{R}^+ \cup \{0\}$ and $f_{\mathbf{Y}}: \mathbb{R}^d \to \mathbb{R}^+ \cup \{0\}$ denote the joint density functions at

$\mathbf{y} = (y_1, \dots, y_d)^T$, $\mathbf{x} = (x_1, \dots, x_d)^T$, respectively, and $\boldsymbol{J}$ is the Jacobian matrix $\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$. Therefore the probability density of **X** is then defined as:

$$f_{\mathbf{X}}(x_1, \ \dots, \ x_d) = f_{\mathbf{Y}}(y_1, \ \dots, \ y_d) \prod_{i=1}^{d} (1 - \alpha_{i,m}) \tag{6.7}$$

or in terms of the copula density

$$f_{\mathbf{X}}(x_1, \dots, \ x_d) = \mathrm{c}\big(F_{Y_1}(y_1), \dots, \ F_{Y_d}(y_d)\big) \prod_{i=1}^{d} (1 - \alpha_{i,m}) f_{Y_i}(y_i) =$$

$$\frac{\partial^d \mathrm{C}(F_{Y_1}(y_1), \dots, F_{Y_d}(y_d))}{\partial F_{Y_1}(y_1) \dots \partial F_{Y_d}(y_d)} \prod_{i=1}^{d} (1 - \alpha_{i,m}) f_{Y_i}(y_i) \tag{6.8}$$

where $\mathrm{c}: [0, 1]^d \to \mathbb{R}^+ \cup \{0\}$ is the copula density function, $f_{X_i}: \mathbb{R} \to \mathbb{R}^+ \cup \{0\}$ and $f_{Y_i}: \mathbb{R} \to \mathbb{R}^+ \cup \{0\}$ denote the marginal density functions of **X** and **Y**, respectively. By convention $\alpha_{r,m}$ always equals to 0.

The joint probability distribution obtained using the rolling pin (RP) method possesses several advantages compared to those obtained using other probability estimation methods:

- **Modeling non-monotonic relationships:** In general, ordinary parametric copulas are unable to capture non-monotonic relationships amongst variables. The monotonization transformation enables parametric copulas to model non-monotonic behavior observed in data without making any changes in the mathematical definition of copula functions.
- **Modeling unknown and unidentified dependence structure:** The rolling pin method monotonizes the original variables with respect to a reference variable. Since the reference variable becomes a dominant part of each of the monotonized variables, it enables the modeling scheme to make a selection from the set of symmetric

parametric copulas, which is eventually used to approximate the dependence structure of the monotonized variables. Such a symmetric copula substitutes the original dependence structure of the variables and makes it possible to estimate complicated, unknown or even unidentified dependence structures with a simplified and known parametric copula.

- **Modeling systems with different pairwise dependence structures:** When a specific parametric copula (such as Gaussian copula, Frank copula, etc.) is used to estimate the dependence structure of multiple ($d \geq 3$) variables, it assigns the same dependence structure to each pair of the variables, even though the pairwise dependence structures are not the same in general. Vine copulas[27] have been introduced to address this problem by expressing the target joint ($d \geq 3$) copula as the product of some factorized lower-dimensional copulas (mainly bivariate copulas). However, these copulas still require finding (a) an appropriate factorization of the main copula and (b) the right copulas describing the pairwise dependence structures (c) the corresponding optimal copula parameters. For these reasons, Vine copulas become computationally expensive and less reliable for large d's. On the other hand, the monotonization transformations enable the rolling pin method to model joint copulas with different pairwise dependence structures with a single parametric copula selected from a limited set of symmetric copulas.
- **Computational efficiency:** The rolling pin method uses parametric copulas, which determine dependence structures based on the correlation or association coefficients of variables.[25] This allows for defining joint distributions using minimum number of parameters. Moreover, sampling from parametric copulas has already been studied

extensively, and numerous efficient sampling methods are available in the literature. These features allow one to model a wide range of dependence structures with a relatively low computational cost.

In the next section we will show how this joint probability estimation method can help identify regression models.

6.3. Regression Model Identification

This section describes how the RP method can be used to efficiently and reliably find regression models relating one set of variables to another set. Let $\mathbf{X} = (X_1, \ \ldots, X_d)^T$ and Z denote the d-dimentional vectors of d continuous input random variables and the continuous output random variable, respectively. Let Z be related to **X** according to:

$$Z = g(\mathbf{X}) + \varepsilon \tag{6.9}$$

where g is a deterministic function, and ε is a noise term. Assumptions made about the behavior of ε (represented by the probability distribution function of ε) significantly affect the choice of the method of finding the regression function $m(\boldsymbol{X})$. A common practice in statistics is to assume that the conditional mean of ε equal to 0; that is:

$$\forall \mathbf{x} \in \Omega(\mathbf{X}), \ E(\varepsilon|\mathbf{x}) = \int_{-\infty}^{+\infty} \epsilon f_{\varepsilon|\mathbf{X}}(\epsilon|\mathbf{x}) d\epsilon = 0 \tag{6.10}$$

where $\Omega(\mathbf{X})$ and $E(.)$ denote the domain of **X** and expectation (mean) function, respectively, and $f_{\varepsilon|\mathbf{X}}$ is the conditional probability density function of the noise on **X**. It can be implied that ε is not considered independent of **X**, i.e. the target probability density can be heteroskedastic. This leads to:

$$E(Z|\mathbf{x}) = E(g(\mathbf{X})|\mathbf{x}) + E(\varepsilon|\mathbf{x}) = E\big(g(\mathbf{X})\big) = g(\mathbf{X}) \tag{6.11}$$

Eq.(6.11) implies that given the aforementioned assumptions about ε, $E(Z|\mathbf{x}) = g(\mathbf{x})$. The conditional expectation of Z given $\mathbf{X}$ can always be calculated, if the true joint probability distribution of the input and output variables is available, then:

$$\forall \mathbf{x} \in \Omega(\mathbf{X}),\ g(\mathbf{x}) = E(Z|\mathbf{x}) = \int_{-\infty}^{+\infty} z f_{Z|\mathbf{X}}(z|\mathbf{x}) dz \tag{6.12}$$

One can write $f_{Z|\mathbf{X}}$, the conditional density of Z given $\mathbf{X}$, in terms of a joint probability distribution which probabilistically connects the input and output variables, leading to:

$$\forall \mathbf{x} \in \Omega(\mathbf{X}),\ g(\mathbf{x}) = E(Z|\mathbf{x}) = \int_{-\infty}^{+\infty} z \frac{f_{\mathbf{X},Z}(x_1,\dots,x_d,z)}{f_{\mathbf{X}}(x_1,\dots,x_d)} dz \tag{6.13}$$

where $f_{\mathbf{X},Z}$ and $f_{\mathbf{X}}$ denote the joint probability distributions of $(\mathbf{X}^T, Z)^T$ and $\mathbf{X}$, respectively. Eq. (6.13) implies that the quality of the identified regression model strongly depends on the strategy of finding $f_{\mathbf{X},Z}(x_1, \dots, x_d, z)$. Since (a) the probabilistic relationship between $\mathbf{X}$ and Z may take any form, including highly nonlinear and non-monotonic relationships, and (b) the dependence structure of $(\mathbf{X}^T, Z)^T$ can be very complex and unknown, particularly as d grows, a method should first be used to accurately estimate $f_{\mathbf{X},Z}$. In general, it is desired to use a method that has a low computational cost when applied to high-dimensional systems. The rolling pin method, briefly reviewed in Section 6.2, allows one to model complex joint probability distributions and therefore can be used to identify regression models using Eq.(6.13). The next paragraphs explain how this can be achieved.

Let $\mathbf{Y} = (Y_1, \dots, Y_d, Y_Z)^T$ be the vector of monotonized variables derived through the application of the monotonization transformation to $(\mathbf{X}^T, Z)^T$ using the vector of monotonizing parameters $\boldsymbol{\alpha_m} = (\alpha_{1,m}, \dots, \alpha_{d,m}, \alpha_{Z,m})^T$. Hence, according to Eq.(6.8), the estimated joint density function of the system of inputs and output, $\hat{f}_{\mathbf{X},Z}$ is defined as

$$\hat{f}_{\mathbf{X},Z}(x_1, \dots, x_d, z) = f_{\mathbf{Y}}(y_1, \dots, y_d, y_z)(1-\alpha_{Z,m})\prod_{i=1}^{d}(1-\alpha_{i,m}) = \mathrm{c}\big(F_{Y_1}(y_1),$$
$$\dots, F_{Y_d}(y_d), F_{Y_Z}(y_z)\big)(1-\alpha_{Z,m})\big)f_{Y_Z}(y_z)\prod_{i=1}^{d}(1-\alpha_{i,m})f_{Y_i}(y_i) =$$
$$\frac{\partial^{d+1}\mathrm{C}\big(F_{Y_1}(y_1),\dots,F_{Y_d}(y_d), F_{Y_Z}(y_z)\big)}{\partial F_{Y_1}(y_1)\dots\partial F_{Y_d}(y_d)\partial F_{Y_Z}(y_Z)}(1-\alpha_{Z,m})f_{Y_Z}(y_z)\prod_{i=1}^{d}(1-\alpha_{i,m})f_{Y_i}(y_i) \tag{6.14}$$

where $F_{Y_i}(y_i)$ and $f_{Y_i}(y_i)$, the marginal CDF and marginal density function of Y_i, can be estimated non-parametrically from data. C and c denote the parametric-copula CDF and density function that are chosen to model the dependence structure of the components of the random $\mathbf{Y}$, respectively. Conditioned on the choice of the reference variable X_r from $X_1, \dots, X_d$, a simple parametric copula such as the normal copula or comonotonicity copula will be appropriate to approximate the pairwise and joint dependence structures of $\mathbf{Y}$. The estimated joint probability density function of $\mathbf{X}$, $f_{\mathbf{X}}$, can be defined in a similar way:

$$\hat{f}_{\mathbf{X}}(x_1, \dots, x_d) = f_{\mathbf{Y}\backslash\{Y_Z\}}(y_1, \dots, y_d)\prod_{i=1}^{d}(1-\alpha_{i,m}) =$$
$$\mathrm{c}_d\big(F_{Y_1}(y_1), \dots, F_{Y_d}(y_d)\big)\prod_{i=1}^{d}(1-\alpha_{i,m})f_{Y_i}(y_i) \tag{6.15}$$

where $f_{\mathbf{Y}\backslash\{Y_Z\}}$ is the joint density function of $(Y_1, \dots, Y_d)$, and $\mathrm{c}_d\colon [0,1]^d \to \mathbb{R}^+$ is the marginalized copula density given by:

$$\mathrm{c}_d\big(F_{Y_1}(y_1), \dots, F_{Y_d}(y_d)\big) = \frac{\partial^d \mathrm{C}_d\big(F_{Y_1}(y_1),\dots,F_{Y_d}(y_d)\big)}{\partial F_{Y_1}(y_1)\dots\partial F_{Y_d}(y_d)} \tag{6.16}$$

where C_d is a d-copula derived from C by marginalizing C with respect to F_{Y_z}:

$$\mathrm{C}_d\big(F_{Y_1}(y_1),\dots,F_{Y_d}(y_d)\big) = \mathrm{C}(F_{Y_1}(y_1),\ \dots,F_{Y_d}(y_d),1) \tag{6.17}$$

Using Eqs. (6.14)-(6.17), the regression function, $m(\mathbf{x})$, can be defined as:

$$m(x_1,\ \dots,x_d) = \int_{-\infty}^{+\infty} z\hat{f}_{Z|\mathbf{X}}(z|\mathbf{x})dz =$$

$$\int_{-\infty}^{+\infty}\left[(1-\alpha_{Z,m})zf_{Y_Z}(y_z)\frac{\mathrm{c}\big(F_{Y_1}(y_1),\dots,F_{Y_d}(y_d),F_{Y_Z}(y_z)\big)}{\mathrm{c}_d\big(F_{Y_1}(y_1),\dots,F_{Y_d}(y_d)\big)}\right]dz =$$

$$\int_{-\infty}^{+\infty}\left[\big(1-\alpha_{Z,m}\big)zf_{Y_Z}(y_z)\mathrm{c}_{F_{Y_Z}|F_{Y_1}\dots F_{Y_d}}\big(F_{Y_Z}(y_z)|F_{Y_1}(y_1),\dots,F_{Y_d}(y_d)\big)\right]dz \tag{6.18}$$

where $\mathrm{c}_{F_{Y_Z}|F_{Y_1}\dots F_{Y_d}}$ is the conditional copula density. For a given and constant x_r, we have

$$y_z = \big(1-\alpha_{Z,m}\big)z + \alpha_{Z,m}x_r \Rightarrow \frac{dy_z}{dz} = \big(1-\alpha_{Z,m}\big) \tag{6.19}$$

and therefore

$$\begin{aligned} &m(x_1,\dots,\ x_d) \\ &= \int_{-\infty}^{+\infty}\left[zf_{Y_Z}(y_z)\mathrm{c}_{F_{Y_Z}|F_{Y_1}\dots F_{Y_d}}\big(F_{Y_Z}(y_z)|F_{Y_1}(y_1),\dots,F_{Y_d}(y_d)\big)\right]dy_z \\ &= \int_{-\infty}^{+\infty}\left[\left(\frac{y_z-\alpha_{Z,m}y_r}{1-\alpha_{Z,m}}\right)\mathrm{c}_{F_{Y_Z}|F_{Y_1}\dots F_{Y_d}}\big(F_{Y_Z}(y_z)|F_{Y_1}(y_1),\dots,F_{Y_d}(y_d)\big)\right]f_{Y_Z}(y_z)dy_z \\ &= E\left[\Upsilon(Y_z)\mathrm{c}_{F_{Y_Z}|F_{Y_1}\dots F_{Y_d}}\big(F_{Y_Z}(Y_z)|F_{Y_1}(y_1),\dots,F_{Y_d}(y_d)\big)\right] \end{aligned} \tag{6.20}$$

where $\Upsilon(Y_z) = \left(\frac{Y_z-\alpha_{Z,m}y_r}{1-\alpha_{Z,m}}\right)$. With the assumptions of Eq.(6.10) and $\alpha_{Z,m} \neq 1$, Eq.(6.20) states that the regression model, $m(\mathbf{x})$, is the expectation (mean) of the random variable $\Upsilon(Y_z)\mathrm{c}_{F_{Y_Z}|F_{Y_1}\dots F_{Y_d}}\big(F_{Y_Z}(Y_z)|F_{Y_1}(Y_1),\dots,F_{Y_d}(Y_d)\big)$. The calculation of this function requires analytical (if possible) or numerical integration of Eq.(6.20). However, one can compute the expectation function empirically, i.e. using the samples of the random variable Y_z:

$$\bar{m}(x_1, \ldots, x_d) = \frac{1}{n}\sum_{k=1}^{n} \Upsilon\left(\gamma_{z,k}\right) \mathrm{c}_{F_{Y_Z}|F_{Y_1}\ldots F_{Y_d}}\left(F_{Y_Z}(\gamma_{z,k})|F_{Y_1}(y_1), \ldots, F_{Y_d}(y_d)\right) \quad (6.21)$$

where $\bar{m}$ and n denote the empirical regression model and number of samples in the dataset, respectively. $\gamma_{z,k}$ =$(1-\alpha_{Z,m})\sigma_k + \alpha_{Z,m}\chi_{r,k}$, where $\gamma_{z,k}$, σ_k and $\chi_{r,k}$ are the k-th samples of Y_z, Z and X_r, respectively. When calculating Υ, care must be taken to set the term $\alpha_{Z,m}y_r$ with respect to X_r but not to $\chi_{r,k}$.

6.3.1. Using the Copula CDF to Identify m

There are cases that the copula (CDF) function is available instead of the copula density or it is easier to work with. In such cases the regression model m can be estimated as follows. First, the conditional copula CDF of the transformed output variable F_{Y_Z} given $F_{Y_1}(y_1), \ldots, F_{Y_d}(y_d)$ is defined as:

$$\mathrm{C}(F_{Y_Z}|F_{Y_1}(y_1), \ldots, F_{Y_d}(y_d)) = \int_0^{F_{Y_Z}} \mathrm{c}_{F_{Y_Z}|F_{Y_1}\ldots F_{Y_d}}(u|F_{Y_1}(y_1), \ldots, F_{Y_d}(y_d))du =$$

$$\int_0^{F_{Y_Z}} \frac{\mathrm{c}\left(F_{Y_1}(y_1),\ldots,F_{Y_d}(y_d),u\right)}{\mathrm{c}_d\left(F_{Y_1}(y_1),\ldots,F_{Y_d}(y_d)\right)} du = \left. \frac{\partial^d \mathrm{C}\left(F_{Y_1},\ldots,F_{Y_d},u\right)}{\partial F_{Y_1},\ldots,\partial F_{Y_d}} \middle/ \frac{\partial^d \mathrm{C}_d\left(F_{Y_1},\ldots,F_{Y_d}\right)}{\partial F_{Y_1},\ldots,\partial F_{Y_d}} \right. \quad (22)$$

$\mathrm{C}(F_{Y_Z}|F_{Y_1}(y_1), \ldots, F_{Y_d}(y_d))$ can then be sampled at a desired point of $(x_1,\ \ldots,\ x_d)^T$(or equivalently $(y_1,\ \ldots,\ y_d)^T$). These samples are transformed back to Y_Z and Z using the inverse CDF of Y_Z and the inverse monotonization transformation. The empirical mean of these samples give an estimation of the regression function, $m(x_1,\ \ldots,\ x_d)$.

6.4. Confidence Intervals of the Regression Model

As the semi-parametric method of identifying a regression model outlined in the previous sections employs a joint probability model to estimate $m(\mathbf{x})$, it provides a natural basis for treating the confidence intervals of the regression model. The confidence interval $I(\mathbf{x}) = [l(\mathbf{x}), u(\mathbf{x})]$ is defined by a lower and an upper bound and is used to measure how narrowly $\hat{f}_{Z|\mathbf{X}}(z|\mathbf{x})$ is distributed around its expected value, $m(\mathbf{x})$. A narrower density $\hat{f}_{Z|\mathbf{X}}(z|\mathbf{x})$ (lower variance) is indicator of a more accurate estimate of $m(\mathbf{x})$. Therefore, $l(\mathbf{x})$ and $u(\mathbf{x})$ are defined as:

$$l(\mathbf{x}) = \sup\{z| \int_{-\infty}^{z} \hat{f}_{Z|\mathbf{X}}(z|\mathbf{x})dz \leq \aleph\} \qquad (6.23)$$

$$u(\mathbf{x}) = \inf\{z| \int_{-\infty}^{z} \hat{f}_{Z|\mathbf{X}}(z|\mathbf{x})dz \geq \beth\} \qquad (6.24)$$

where $\hat{f}_{Z|\mathbf{X}}(z|\mathbf{x}) = (1-\alpha_{Z,m}) f_{Y_Z}(y_z) \mathrm{c}_{F_{Y_Z}|F_{Y_1}\ldots F_{Y_d}}\left(F_{Y_Z}(y_z)|F_{Y_1}(y_1), \ldots, F_{Y_d}(y_d)\right)$. $\aleph$ and $\beth$ are fixed lower and upper probability bounds and independent of **X**. For example, the user may set these values equal to 0.05 and 0.95, respectively.

6.5. Properties of the Method of Regression Model Identification

Modeling highly nonlinear and non-monotonic relationships: The regression model identification (estimation) method provides a semi-parametric tool for modeling highly nonlinear and non-monotonic relationships among input and output variables. This is performed by adjusting the monotonizing parameters to some adequately large values. According to this approach, the only requirement for modeling more nonlinear relationships is to use large enough monotonizing parameter values which assure the strictly increasing transformed variables, rather than the convention of utilizing more intricate models with numerous parameters. For example, to derive the regression model with one input variable and one output variable, only one monotonizing parameter and a total of four parameters are required, regardless of the complexity of the input-output relationship.

Model selection: A great deal of work on the development of regression models has been dedicated to parametric models, which use a particular mathematical equation to relate the input and output variables. Such an equation always has parameters that should be trained with respect to the data using a goodness-of-fit measure. In general, more sophisticated models are required for describing more complicated systems. As a result, a model selection scheme must be applied prior to the model training step. Besides the uncertainties carried by such a model selection step, it tends to be computationally expensive and may lead to unnecessarily complex models, particularly when the model is black-box. On the other hand, the proposed method uses a probabilistic model (the rolling pin joint probability distribution) to identify a regression model. It has been shown in Chapter 4 that with appropriate values of the monotonizing parameters, the actual multidimensional dependence structure can be approximated by a single parametric copula, chosen from the set of symmetric parametric copulas.

Tractability: The computational tractability of the proposed regression model identification method is appealing. A rolling pin distribution is modeled with a number of parameters of the order as low as $O(d^2)$. This number arises from d monotonizing parameters, $d + 1$ smoothing parameters of the marginal distributions (when the kernel method is used. This number will be linearly proportional to d if empirical distribution is employed) and $\binom{d+1}{2}$ correlation or association parameters of the copula function. This feature becomes more appealing when d grows. Not only is it useful when training the joint probability model, but it also renders of the proposed regression model identification method less computationally expensive.

Convergence rate: although there is no generally accepted method to find the minimum number of observations n with respect to the number of the independent variables d, one may use the rule of thumb proposed by reference 26 to calculate the sample size as $n = m^{(d+1)/2}$, where m is the sample size to model a bivariate system with one input variable.

Modeling the probabilistic behavior of a regression model: Unlike many conventional regression-model identification methods which offer a point estimate of the regression model, the proposed method provides a unified framework to simultaneously identify a regression model and estimate the model statistical characteristics, through the conditional probability density $\hat{f}_{Z|\mathbf{X}}(z|\mathbf{x})$. This function helps to investigate the quality of the regression model from a fully statistical point of view; that is, a complete probabilistic profile of the behavior of the output variable Z is derived at each point $(X_1, \dots, X_d)^T$ by means of the information encoded in $\hat{f}_{Z|\mathbf{X}}(z|\mathbf{x})$. As a result, statistical characteristics such as the variance, confidence intervals, skewness, kurtosis, etc. of this function may be computed easily whenever necessary.

6.6. Examples

This section aims at demonstrating the application and performance of the proposed regression-model identification method using two examples, a mathematical example and a realistic biological system. The quality of the identified regression models is assessed both qualitatively (through visually comparing the predictions of the models with the actual data/function) and quantitatively (by means of evaluating the residuals and confidence bounds). Also, the cases of univariate and multivariate inputs are considered in these examples.

6.6.1. Mathematical Example

Consider a bivariate system where the input variable X affects the output variable Z through a function $g(X)$ and a random noise function ε:

$$Z = g(X) + \varepsilon \tag{25}$$

where $g(X) = \cos(X) + 0.2\sin(5X)$ and $\varepsilon \sim N(0, 0.05)$. $N(\mu, \sigma)$ denotes a Gaussian distribution with the mean μ and standard deviation σ. X has a $\text{Gamma}(1, 1)$ distribution, where $\text{Gamma}(\vartheta, \theta)$ denotes a Gamma distribution with the shape and scale factors equal to ϑ and θ, respectively.

Assume that 500 samples are available for the pair of random variables X and Z. First, 500 samples of X are generated using the $\text{Gamma}(1, 1)$ distribution. Second, 500 samples of ε are generated using the $N(0, 0.05)$ distribution. Third, 500 samples of Z are generated using Eq. (6.25). These 500 X and Z samples are shown in Figure 6.1, where the solid line represents $Z = g(X)$. Drawing the X samples from such a Gamma distribution is a way to exemplify cases where obtaining samples from the predictors becomes increasingly infeasible with the increase in the predictor magnitude.

The 500 $(X,\ Z)^T$ samples are then used to identify the regression model governing their dependence. To develop the needed rolling pin joint distribution, the following assumptions are made. First, the marginal probability densities are estimated using a nonparametric method, which is the Gaussian kernel method here. Second, a specific parametric copula and its corresponding parameters are used to approximate the input-output dependence structure. Third, the noise function is assumed to have a mean of zero. Fourth, it is assumed that the X samples have been collected with no error. Note that no assumption is made about the family to which the noise function belongs to or about the noise (error) and the input variable relationship.

The first step in applying the rolling pin method is to select the reference variable. This variable can be found systematically using the methods described in Chapter 4. Here we use X as the reference variable. Then, samples of Z are monotonized with respect to the X samples using the monotonizing parameter.

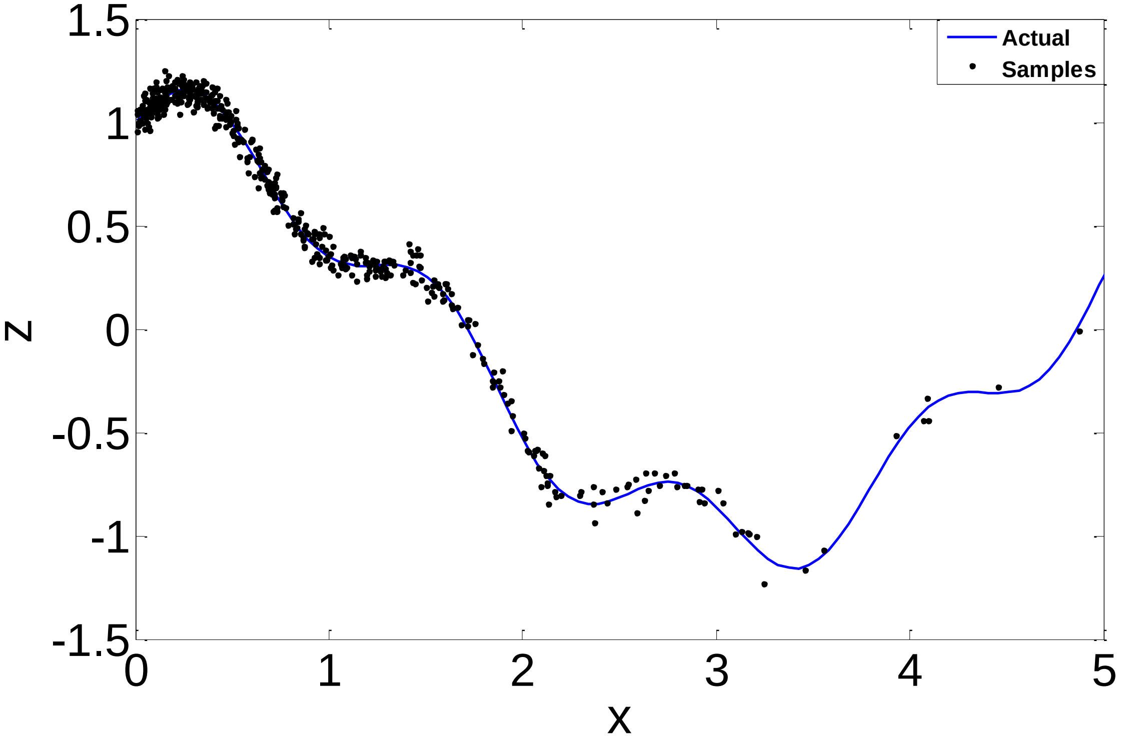

Figure 6.1: Actual function (with no additive noise) and 500 samples of Example 4.1.

The selection of α_m can be carried out through multiple ways. In this example, this value is determined optimally using the maximum likelihood estimation (MLE) method described in Chapter 4. Applying the MLE method requires the user to select the parametric copula function beforehand. Alternatively, one can use the correlation-based methods to find and appropriate monotonizing parameter according to Chapter 4. To illustrate the effect of the selected copula on the final regression model, we use two well-known parametric copulas from the elliptical family, i.e. the Gaussian and student's t copulas. These copulas are symmetrical, so they are appropriate for our purpose which is to model the dependence structure of the monotonized variables. Using such copulas allows us to study the tail dependence effect (which the Gaussian copula lacks and the t copula possesses) on the identified regression model behavior, especially at the extremes (very low and very high values of the input variable). The rest of the MLE procedure is as follows. For each α, the samples of the transformed variable $Y_Z = (1 - \alpha)Z + \alpha X$ are calculated. The marginal probability densities of X and Y_Z are then estimated using the nonparametric Gaussian kernels. The marginal distributions are then used to transform X and Y_Z data to the data in the F_X and F_{Y_Z} space, where the copula function is applied. As both Gaussian and t copulas are elliptical, the strength of the dependence is determined by the Spearman's rank correlation matrix. When the copulas are trained by this rank correlation matrix, together with the estimated marginal densities they are used to compute the likelihood of α given the data of X and Y_Z. Eventually, the value of α which maximizes the likelihood function globally is adopted as the MLE α_m, and the corresponding joint distribution is the rolling pin distribution. Taking these steps leads to the MLE monotonizing parameters $\alpha_m = 0.87$ and $\alpha_m = 0.96$ for the Gaussian and t copulas, respectively. α_m's can also be calculated using non-MLE-based methods in Chapter 4.

We then apply Eq. (6.20) or Eq.(6.21) to obtain the deterministic regression model $m(x)$. The calculated regression models and their 99% confidence bounds are depicted in Figures 6.2 and 6.3 for the Gaussian and t copula-derived rolling pin distributions. It can be observed that the t copula-derived regression function models the underlying function g more accurately. This can be confirmed quantitatively by a measure of error such as the sum of square errors (SSE). Table 6.1 compares the goodness-of-fit of the Gaussian and t copula-derived regression models in terms of their SSE values. For both cases, it can be seen that the identified regression models can capture the $g(X)$ behavior, and the quality of the predictions is higher in the regions with higher number of data points. As the t copula has the property of tail dependence, in the regions where the data density decreases the quality of the prediction is higher than that of the Gaussian copula which lacks this property. Therefore, it can be concluded that for cases where the input variable(s) data is not distributed uniformly, as in this case, the t copula may offer a better estimate of a regression model.

We also compared the identified (estimated) regression models with some traditional parametric regression models, in terms of prediction quality and number of model parameters. Figure 6.4 compares predictions of the regression model obtained using the proposed method with those of a polynomial of order 9 model, a Gaussian model, and a rational regression model:

Table 6.1: Comparison of the number of parameters and goodness-of-fit measures of the different regression methods used in the first example.

Method	No. of parameters	Goodness of fit (SSE)
Rolling pin method-based (Gaussian copula)	4	1.132
Rolling pin method-based (t copula)	4	0.535
9th-order polynomial	10	4.461
5 degree numerator 5 degree denominator rational	11	2.908
8-term Gaussian	24	2.756

Polynomial of order 9:

$$Z = \sum_{i=0}^{9} a_i X^i \tag{6.26}$$

Gaussian model:

$$Z = \sum_{i=1}^{8} a_i \exp\left(-\left(\frac{X-b_i}{c_i}\right)^2\right) \tag{6.27}$$

Rational regression function:

$$Z = \frac{p_1 X^5 + p_2 X^4 + p_3 X^3 + p_4 X^2 + p_5 X + p_6}{X^5 + q_1 X^4 + q_2 X^3 + q_3 X^2 + q_4 X + q_5} \tag{6.28}$$

It can be seen that our proposed copula-based regression model presents the closest fit to $g(X)$. As listed in Table 6.1, this is indicated by the SSE values for the regression models. An argument should also be made about the computational complexity of the parameter learning step of the above regression models. All the parametric models utilized in this example suffer from an exponential growth of the number of parameters with the system dimension, d. This is not only a serious problem when training the models for high-dimensional systems; it also restricts the applicability of the parametric models for high-dimensional and complicated systems. This restriction is twofold. First, an exponential growth in the number of parameters will significantly decelerate the model selection process; that is, greater number of parameters considerably slows down the quantification of a candidate model at different points in the input variable domain which is required for the model evaluation. On the other hand, such a high-dimensional regression model is difficult to quantify at a desired point in the input variable space. This fact makes these parametric models less computationally favorable for online

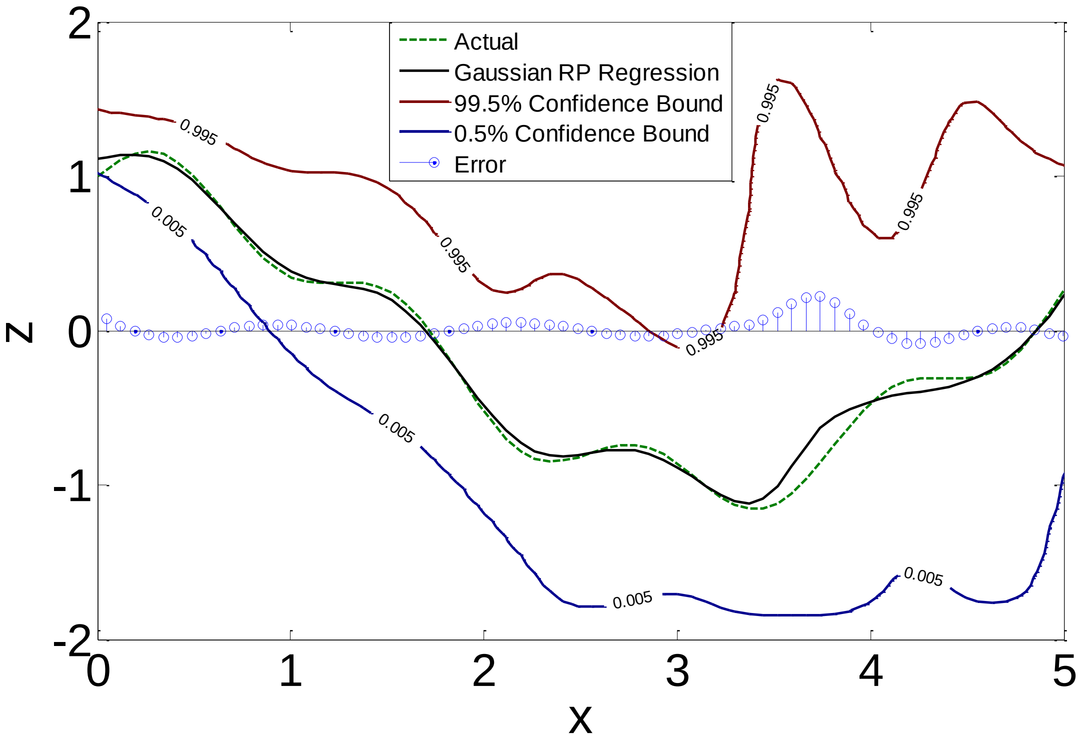

Figure 6.2: Actual function, Gaussian copula-based RP regression model, 99% confidence interval, and prediction error for Example 4.1.

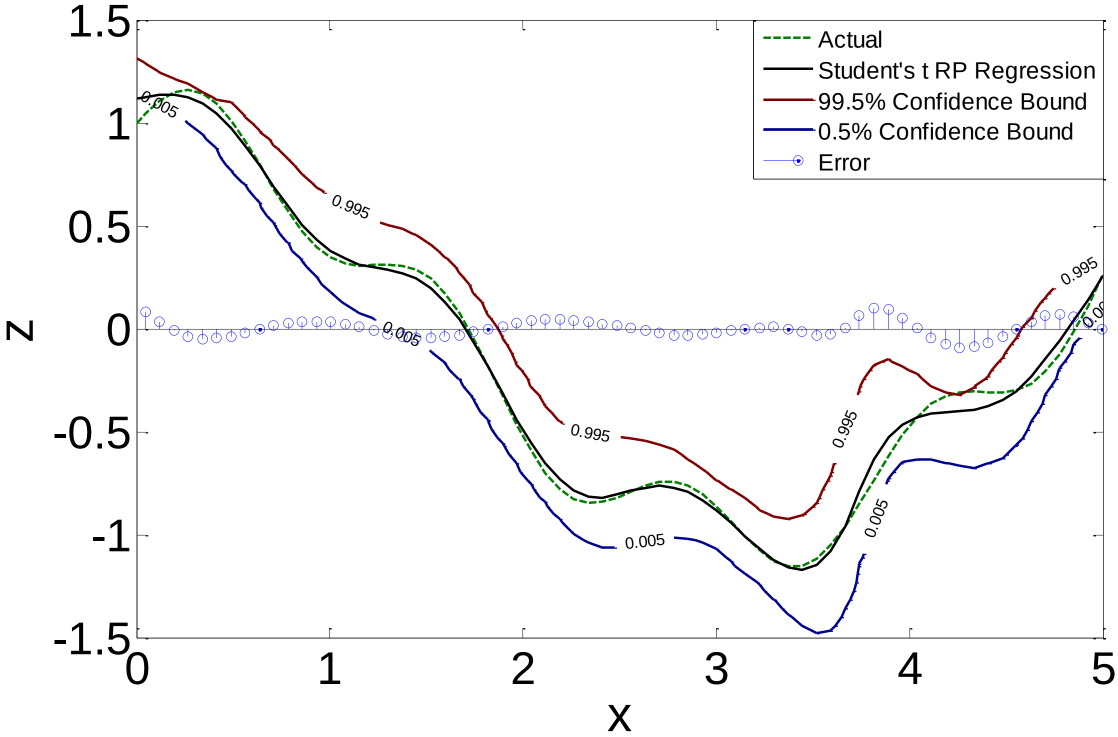

Figure 6.3: Actual function, student's t copula-based RP regression model, 99% confidence interval, and prediction error for Example 4.1.

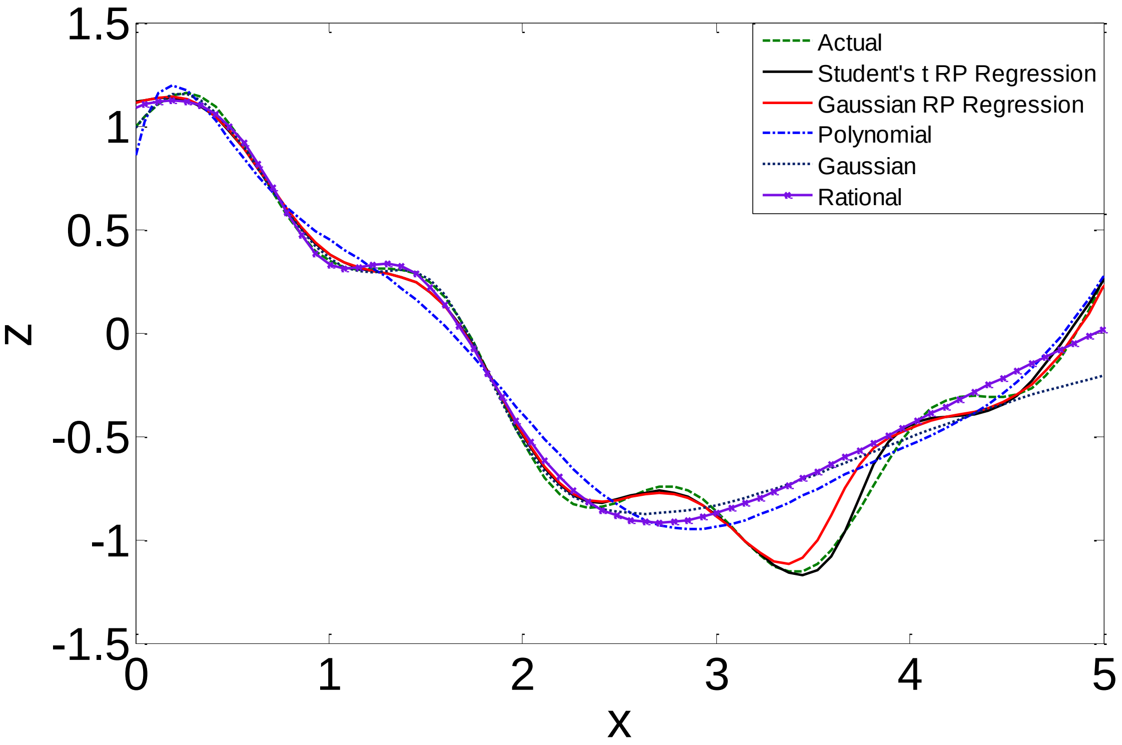

Figure 6.4: Comparison of the predictions of the RP-based regression models and several linear and nonlinear parametric regression models.

applications. However, it should be mentioned that since our method *is semi-parametric*, its convergence rate to the actual underlying function is slower than parametric models with respect to the number of samples n, as part of the information encoded in the data has to be utilized to determine the non-parametric marginal densities "*forms*". This is a common pitfall of the non-parametric estimation techniques; however the fact that the proposed model has a parametric part renders this less problematic than purely non-parametric methods. Therefore, the proposed method combines the tractability of parametric methods and the flexibility of non-parametric methods, besides its intrinsic capabilities resulting from the monotonization transformation.

6.6.2. Biological System Example

This biological system consists of a common enzymatic reaction.[28] The objective is to quantify the dependence of the enzyme activity on the environment temperature and pH. Consider a case where the enzyme E participates in a two-stage protonation reaction. It is assumed that the substrate is available in an excess amount and its effect on the enzyme activity is insignificant. The protonation reactions are reversible and undergo the chemical equilibrium:

$$EH^{2+} \overset{K_1}{\leftrightarrow} EH^{+} + H^{+} \overset{K_2}{\leftrightarrow} E + 2H^{+} \tag{6.29}$$

where K_1 and K_2 denote the equilibrium constants for the first and second deprotonation reactions. Assuming the enzyme reacts with the substrate only in its EH^{+} form, the rate of this reaction is defined as

$$v = -\frac{d(Substrate)}{dt} = \frac{K_1[H^+]}{K_2K_1+K_1[H^+]+[H^+]^2}v_{max} = \frac{K_1 10^{-pH}}{K_2K_1+K_1 10^{-pH}+10^{-2pH}}v_{max} \tag{6.30}$$

where v_{max} (Peterson et al. (2007)), K_1, and K_2 are given by:

$$v_{max} = \frac{k_B T\exp\left(-\frac{\Delta G_{cat}}{RT}\right)[E_0]\exp\left(\frac{-k_B T\exp\left(-\frac{\Delta G_{inact}}{RT}\right)\Psi(T)t}{h\left(1+\Psi(T)\right)}\right)}{h\left(1+\Psi(T)\right)} \tag{6.31}$$

$$\Psi(T) = \exp(-\frac{\Delta H_{eq}(\frac{1}{T_{eq}}-\frac{1}{T})}{R}) \tag{6.32}$$

$$K_1 = c_1\exp(-\frac{\Delta E_1}{RT}) \tag{6.33}$$

$$K_2 = c_2\exp(-\frac{\Delta E_2}{RT}) \tag{6.34}$$

Definitions, values and units of the constants used in Eqs. 6.31-6.34 are given in Table 6.2 based on reference 28. It can be seen from these equations that $v = v(T, \mathrm{pH}, t)$, such that this functionality is non-monotonic with respect to T and pH. We assume that:

$$T \sim N(325, 8) \tag{6.35}$$

$$\mathrm{pH} \sim N(3.6, 0.5) \tag{6.36}$$

Since $v(T, \mathrm{pH}, t)$ is a deterministic function, we add some uncertainty to the dependence of v:

$$v' = v(T, pH, 1) + \varepsilon \tag{6.37}$$

where $\varepsilon \sim N(0, 0.05)$ is a white noise.

To generate a noise-included data set, we simulated 1,000 samples of the triplet $(T, \mathrm{pH}, v')^T$ at $t = 1\,sec$ using an approach similar to the one employed in the mathematical example. Figures 6.5a-6.5c depict samples of $(T, \mathrm{pH}, v')^T$ normalized according to Section 6.2. The non-monotonic behavior of the function $v(T, \mathrm{pH}, 1)$ is clearly seen from the dome-shaped surface of v shown in Figure 6.6a.

Table 6.2: Parameters and constants of Example 6.6.2.

Parameter	Definition	Value (Unit)
ΔG_{cat}	Gibbs free energy of enzyme catalysis	68.9 (kJ/mol)
ΔG_{inact}	Gibbs free energy of enzyme inactivation	93.7 (kJ/mol)
ΔH_{eq}	enthalpy of the equilibrium	138.2 (kJ/mol)
T_{eq}	temperature at which active and inactive enzyme concentration are equal	325 (K)
R	gas universal constant	8.314 $(J/mol.\mathrm{K})$
$[E_0]$	initial enzyme concentration	5.5×10^{-2} (mol/m^3)
h	Planck constant	6.62606957×10^{-34} (m^2kg/s)
k_B	Boltzmann constant	1.3806488×10^{-23} $(m^2kg/s^2\mathrm{K})$
ΔE_1	activation energy difference	17 (kJ/mol)
ΔE_2	activation energy difference	28.5 (kJ/mol)

Using this set of 1,000 samples to identify the regression model $m(T, \mathrm{pH})$ is pretty similar the first example. Note that to avoid any round-off error resulting from the difference in the orders of magnitude of T, pH and v', all these calculations are performed using the normalized data of these variables, as described in Section 6.2. The following steps are then taken to obtain m:

1. The first step is to select the reference variable. As the input variables are considered independent, it makes more sense to monotonize all variable with respect to a variable that already known to be connected to other variables in some way; i.e. the output variable, v'.
2. With v' as the reference variable and using the MLE method and the t copula, the monotonizing parameters of the normalized $T$and pH are obtained to be $\alpha_{T,m} = 0.95$ and $\alpha_{pH,m} = 0.96$, respectively. Figures 6.5d-6.5f shows the data of the monotonized variables $\left(Y_T,\ Y_{pH}, Y_{v\prime}\right)^T$.
3. Once the monotonized variables data become available, using the probability integral transform (which is applied via the marginal densities of Y_T, Y_{pH} and $Y_{v\prime}$ derived by the non-parametric Gaussian kernel method) transforms the data into the space where the copula is applied. Figures 6.5g-6.5i show the data points of these transformed variables F_{Y_T}, $F_{Y_{pH}}$ and $F_{Y_{v\prime}}$.
4. Since the Gaussian copula belongs to the elliptical family, its parameters can be estimated as the elements of the Spearman's rank correlation matrix, which are

$$\rho_s = \begin{bmatrix} 1.0000 & 0.9972 & 0.9987 \\ 0.9972 & 1.0000 & 0.9986 \\ 0.9987 & 0.9986 & 1.0000 \end{bmatrix}$$

5. At this stage, $\hat{f}(T, pH, v')$ is calculated using Eq. (6.8).

6. $\hat{f}(T, \text{pH}, v')$ can now be employed to identify the regression model $m(T, \text{pH})$ using Eq. (6.21).

Figure 6.6b shows the resulting regression-model predictions. To visually compare $m(T, \text{pH})$ with $v(T, \text{pH}, 1)$, the absolute difference between $v(T, \text{pH}, 1)$ and the regression function predictions are shown vs. T and pH in Figure 6.6c. As expected, the magnitude of the error increases as the input variables approach the boundary of the observed data set. This makes a perfect sense; the uncertainty of the estimation significantly increases in regions where less data points are available. One can see that such a deviation is inevitable, even though a copula with strong tail dependence such as the t copula is employed as the parametric part of the model. The performance of such models will be improved if the data used are distributed uniformly. This is often possible through active learning, where the input variables can be manipulated such that a uniform distribution of their data is applied to the system (compare that with Gaussian distribution of T and pH that we used in this example).

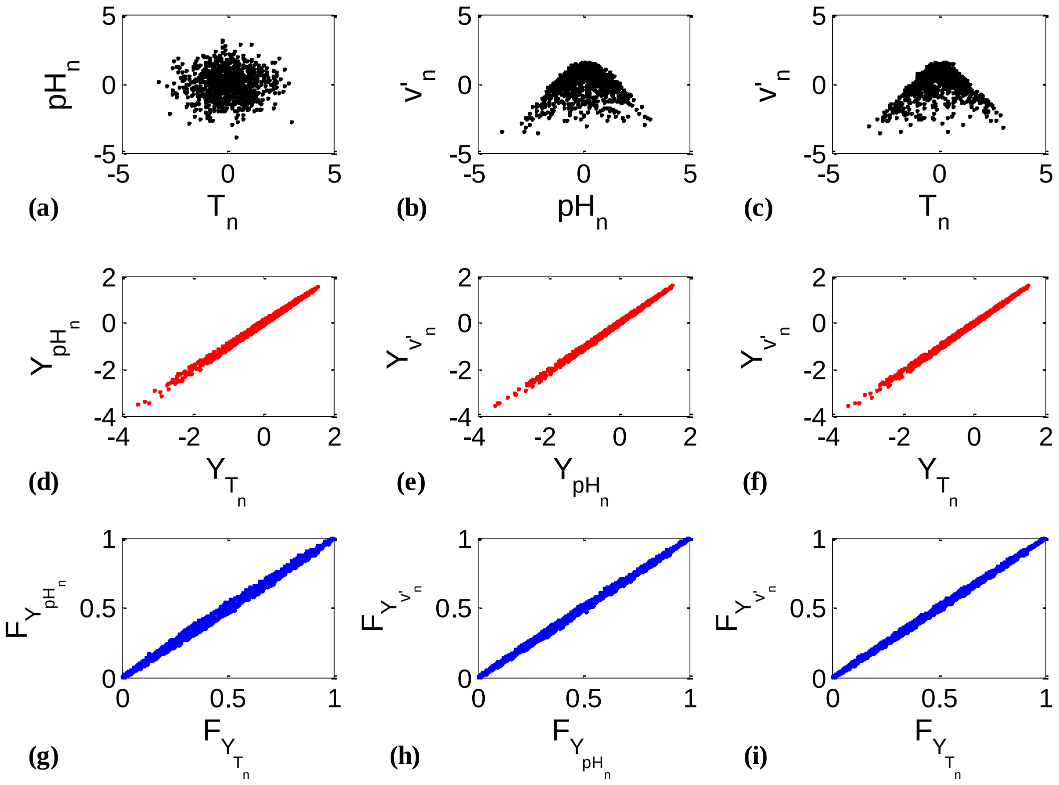

Figure 6.5: Normalized variables data: (a) pH_n vs. T_n, (b) v'_n vs. pH_n, and (c) v'_n vs. T_n. Monotonized data: (d) Y_{pH_n} vs. Y_{T_n}, (e) $Y_{v'_n}$ vs. Y_{pH_n}, and (f) $Y_{v'_n}$ vs. Y_{T_n}. Probability integral transforms of the data: (d) $F_{Y_{pH_n}}$vs. $F_{Y_{T_n}}$, (e) $F_{Y_{v'_n}}$ vs. $F_{Y_{pH_n}}$, and (f) $F_{Y_{v'_n}}$ vs. $F_{Y_{T_n}}$.

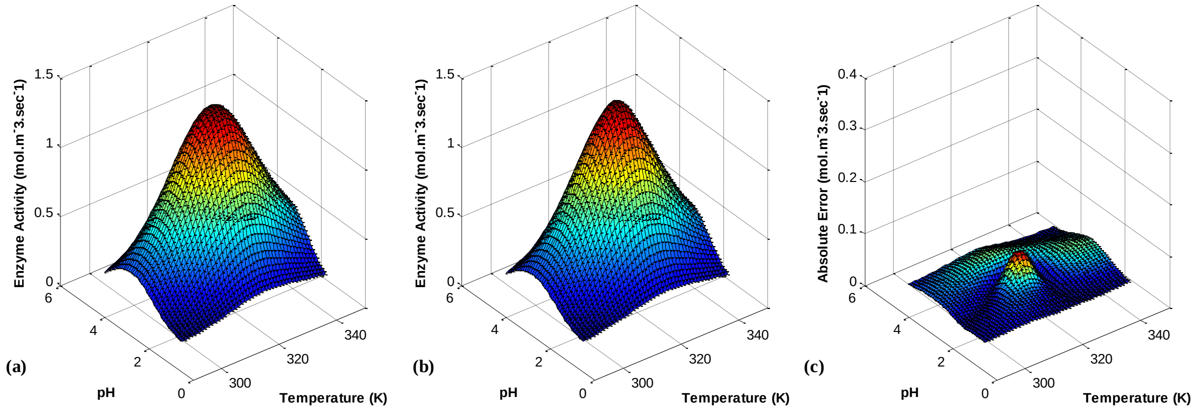

Figure 6.6: (a) Actual enzyme activity in Example 4.2 vs. temperature and pH, (b) the identified-regression-model predictions of the enzyme activity, and (c) absolute error in the predicted enzyme activity.

6.7. Conclusions

In this chapter we proposed a new method of regression model identification based on a joint probability distribution of the data estimated using our previously-introduced *rolling pin* method. This regression model identification method has several appealing features. First, it is capable of modeling nonmonotonicity, and it does so without adding complexity to the functional form of the regression model. Second, as the rolling pin method does not assume any limitations on the heteroskedasticity of the data, the identified regression models can take into account more complicated noise terms whose probability density depends on the input variables. Third, a single parametric copula is used to model unidentified dependence structure of the input and output variables. Fourth, since the regression model is obtained using a joint probability distribution, this joint probability distribution can be used readily to calculate the confidence intervals of the model identification (estimation). Fifth, different pairwise dependence structure of the domain variables can be modeled using a single symmetric parametric copula. Sixth, the modeling computational complexity grows rather slowly by $O(d^2)$, where d is the dimension of the input-variable vector.

References

1. Hastie, T., Tibshirani, R., Friedman, J., 2009. *The elements of statistical learning* (Vol. 2, No. 1). Springer, New York.

2. Cohen, J., Cohen, P., West, S.G., Aiken, L.S., 2013. *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge, New York.

3. Williams, C. K. 1998. Prediction with Gaussian processes: From linear regression to linear prediction and beyond. In *Learning in graphical models* (pp. 599-621). Springer, Netherlands.

4. Hedeker, D., Gibbons, R. D., 2006. *Longitudinal data analysis* (Vol. 451). John Wiley & Sons, New York.

5. Freedman, D., 2009. *Statistical models: theory and practice*. Cambridge University Press.

6. Sahu, S. K., Dey, D. K., Branco, M. D., 2003. A new class of multivariate skew distributions with applications to Bayesian regression models. *Canadian Journal of Statistics 31*(2), 129-150.

7. Ahooyi, T.M., Soroush, M., Arbogast, J.E., Seider, W.D., Oktem, U. G., 2014. Maximum-likelihood maximum-entropy constrained probability density function estimation for prediction of rare events. *AIChE Journal 60* (3), 1013-1026.

8. Mohseni Ahooyi, T.; Abrogast, J. E.; Oktem, U. G.; Seider, W. D.; Soroush, M., 2014. Estimation of Complete Discrete Multivariate Probability Distributions from Scarce Data with Application to Risk Assessment and Fault Detection. *Industrial & Engineering Chemistry Research 53* (18), 7538-7547.

9. Härdle, W., Müller, M., Sperlich, S., Werwatz, A., 2004. *Nonparametric and semiparametric models*. Springer, New York.

10. Linton, O., Nielsen, J. P. 1995. A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika*, 93-100.

11. Ichimura, H., 1993. Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics 58*(1), 71-120.

12. Nelsen, R.B., 1999. *An introduction to copulas*. Springer, New York.

13. Sklar, M., 1959. *Fonctions de répartition à n dimensions et leurs marges*. *Université Paris 8*, 229-231.

14. Noh, H., Ghouch, A.E., & Bouezmarni, T., 2013. Copula-based regression estimation and inference. *Journal of the American Statistical Association 108*(502), 676-688.

15. Oakes, D., Ritz, J., 2000. Regression in a bivariate copula model. *Biometrika 87*(2), 345-352.

16. Pitt, M., Chan, D., Kohn, R., 2006. Efficient Bayesian inference for Gaussian copula regression models. *Biometrika 93*(3), 537-554.

17. Genest, C., Rivest, L.P., 1993. Statistical inference procedures for bivariate Archimedean copulas. *Journal of the American statistical Association 88*(423), 1034-1043.

18. Frees, E.W., Valdez, E.A., 1998. Understanding relationships using copulas. *North American actuarial journal 2*(1), 1-25.

19. Cuadras, C. M., 1992. Probability distributions with given multivariate marginals and given dependence structure. *Journal of multivariate analysis 42*(1), 51-66.

20. Leong, Y. K., Valdez, E. A., 2005. Claims prediction with dependence using copula models. *Insurance: Mathematics and Economics*.

21. Käärik, M., Selart, A., Käärik, E., & Liivi, J., 2011. The use of copulas to model conditional expectation for multivariate data. In *58th World Statistics Congress of the International Statistical Institute (ISI 2011)*, 21-26.

22. Mohseni Ahooyi, T., Arbogast, J. E., Soroush, M., 2014. Applications of the Rolling Pin Method. 1. an Efficient Alternative to Bayesian Network Modeling and Inference. *Industrial & Engineering Chemistry Research*, DOI: 10.1021/ie503585m.

23. Escarela, G., Mena, R.H., Castillo-Morales, A., 2006. A flexible class of parametric transition regression models based on copulas: application to poliomyelitis incidence. *Statistical Methods in Medical Research 15*(6), 593-609.

24. Schmidt, R., Hrycej, T., Stützle, E., 2006. Multivariate distribution models with generalized hyperbolic margins. *Computational statistics & data analysis 50*(8), 2065-2096.

25. Mohseni Ahooyi, T., Arbogast, J. E., Soroush, M., 2014. Rolling Pin Method: Efficient General Method of Joint Probability Modeling. *Industrial & Engineering Chemistry Research 53* (52), 20191–20203.

26. Good, P.I., Hardin, J.W., 2009. *Common Errors in Statistics* (3rd ed.). Wiley, New Jersey.

27. Kurowicka, D., Joe, H., 2011. *Dependence Modeling: Vine Copula Handbook*. World Scientific.

28. Peterson, M., Daniel, R., Danson, M., Eisenthal, R., 2007. The dependence of enzyme activity on temperature: determination and validation of parameters. *Biochem. J 402*, 331-337.

Chapter 7: Concluding Remarks and Future Directions

The way we treat variables in our surrounding world profoundly affects our understanding of the phenomena and analyses and decision making processes. One major way of such treatments is to divide variables into deterministic and probabilistic (stochastic) quantities. By definition, a deterministic variable is specified using a single characteristic (numerical or qualitative value) while probabilistic (random) variables are fully specified only through their probability distributions. Deterministic and probabilistic variables give rise to two totally different classes of models, called deterministic and probabilistic models, respectively. Whether deterministic or probabilistic, all models aim to provide most accurate description of the system under study. To this end, models gather facts and speculations about the system structure and integrate them into a meaningful and well-defined framework. Examples of such facts are the laws of physics and the way inputs (causes) and outputs (effects) interact. In that sense, one can consider a model as a qualitative or quantitative mathematical framework to represent a system or describe its cause-effect relationships.

A deterministic model assigns so called point estimates to the filed variables. While being the basis of modeling for centuries, this assumption is acceptable only when there is no uncertainty associated with the variables. Any uncertainty of any type would dramatically reduce the credibility and reliability of the deterministic models and their predictions. Comparatively, probabilistic models have a relatively shorter history; they appeared about a century ago and their widespread use was not practical until a few decades ago due to the lack of sufficient theoretical background and computational power. A probabilistic model assigns a probability distribution to each (random) variable instead of a point estimate. This probability distribution encodes important information about the likelihood of each of the possible values (states) of the variable. In other words, from the perspective of a probabilistic model the random variable can probably take any of its states, with the condition of higher chance for the states with higher probability.

Probabilistic models are essential modeling tools in the modern day. This arises from different factors which more or less are related to a change of perspective which has led to recognizing the real-world variables as probabilistic rather than deterministic. First, when dealing with real-world variables, we often have to rely on the sensor measurements. There are multiple sources of uncertainty associated with sensors including random errors and systematic biases. These will render the measured variables a random variable. Second, even though a system could be modeled deterministically, an accurate representation requires the model to account for so many details. For example, deterministic modeling of throwing a dice requires one to consider accurate measurements of the initial conditions along with complicated motion equations and precise knowledge of air flow patterns around the dice. Since this level of accuracy is impossible and the simplified deterministic model would bear a great deal of uncertainty in it, a probabilistic model seem to be an appropriate replacement to represent such a system. Third, human knowledge is not unlimited; that is, there is always a level of uncertainty tied with any fact or bit of knowledge. For example, waiting for a bus, no one is sure about the exact arrival time of the next bus. This type of uncertainty is usually called epistemic uncertainty.

Whether deductive or empirical, probabilistic models strongly rely on probability distributions to work properly and accurately. With an increase in popularity of probabilistic modeling, probability distribution estimation has become an active area in information technology and knowledge engineering. Obviously, accurate probability estimation becomes more important and more challenging as the dimensions of the system grow. When the probabilistic model of a multidimensional system $(d \geq 2)$ is sought, the corresponding probability distribution is called *multivariate or joint probability distribution*.

In mathematical jargon, a probabilistic model is defined as the pair (Ξ, P) where Ξ is the sample space, the set of all possible events (values or states) the model can take, and P is the probability distribution over the sample space. P encodes the essential information needed to develop a probabilistic model and perform probabilistic inference as it captures the qualitative and quantities aspects of the interaction among the system components. In a few cases the true probability distribution of the system can be identified, but usually P represents an approximation or simplification of the actual system probabilistic behavior. If Ξ is multidimensional, the corresponding P is called the joint probability distribution of the system. These dimensions may arise from different quantities present in the system (such as temperature, pressure, etc.), multiple spatial dimensions (as in 2-dimentional or 3-dimensional systems) or presence of dynamic behavior in the system under study. Therefore, using an appropriate P , one can model the spatial distributions, quantify relationships among variables or predict the time evolution of dynamic systems. Considering the importance of estimation of high-dimensional distributions in a meaningful, interpretable and accurate manner, a large variety of estimation methods has emerged in the literature. Joint probability distributions can be estimated parametrically, non-parametrically or semi-parametrically.

This research has introduced novel computationally-efficient and flexible methods of estimating joint probability distribution of highly nonlinear and non-monotonic systems of continuous and discrete random variables. There is a broad range of applications for these methods in probabilistic modeling and inference in systems with stochastic behavior. As discussed in detail in this treatise, these methods offer many advantages over their well-known counterparts such as the original parametric copula method, Bayesian networks and nonparametric techniques of joint probability estimation. The methods offer a powerful tool in modeling multivariate joint probability distributions with arbitrary and not necessarily known pairwise dependence structures among the variables. This implies that the methods need no knowledge of the exact dependence structures and the pairwise sameness throughout the system variables. More importantly, the methods are capable of modeling non-monotonic interactions, which cannot be modeled by the conventional parametric copulas.

Future Directions

As discussed in this treatise, so far my research has mainly focused on addressing the most important problems with the Bayesian Networks and copulas. This includes introducing several methods for rare event probability estimation, developing methods to estimate joint probability distributions of variables with unknown causal structures and/or complex non-monotonic relationships, reducing the computational cost, etc.. Each of these newly developed joint probability distribution estimation methods exhibit higher flexibility, interpretability and tractability compared to their original BN and copula counterparts. As shown by examples in our published journal papers, these methods can be readily used to do probabilistic modeling and perform risk analysis for systems operating in steady-states mode.

My recommended future direction is to apply these methods to perform probabilistic modeling and inference for time-varying process systems, which includes:

1) Defining a general mathematical framework to employ the developed joint probability modeling methods to construct dynamic stochastic models (such as Markov processes).
2) Applying the constructed model to some well-known and important process systems such as fluidized beds, bubbling beds, CSTRs and PFRs, etc.
3) Performing risk analysis for the stochastically modeled dynamic systems.

Also, the following topics have a good potential as the basis to extend the research presented in this thesis:

1) Generalizing the rolling pin method to discrete random variables and the mixture of continuous and discrete random variables
2) Applying the rolling pin methods to non-functions: the rolling pin method in its current definition is best applied to the functions; that is, the systems where for each input entry there is only one output (such as $\mathrm{y} = x^2$). Also the method works properly for non-functions whose inverse is a function (e.g. $\mathrm{y} = \pm\sqrt{x}$). Algorithms to apply the method to the systems which are not of the types above (such as $\mathrm{y}^2 + x^2 = a^2$) will be of research interest.

Vita

Taha Mohseni Ahooyi

Education

- **Jan 2011– Dec 2015**: Ph.D. in Chemical Engineering, Drexel University, Philadelphia, PA, USA
- **Sep 2007- Feb 2010**: M.Sc. in Chemical Engineering, Sharif University of Technology, Tehran, Iran
- **Sep 2003- Sep 2007**: B.Sc. in Chemical Engineering, Sharif University of Technology, Tehran, Iran

Awards & Honors

- AIChE Computing & Systems Technology Division (CAST) Student Presentation Travel Award, 2015
- William Casey Scholarship for Outstanding Ph.D. Research Achievements & Publications, 2015
- Drexel CBE Outstanding Ph.D. Student Award, 2014
- Ranked 3rd in the Iranian Nationwide Chemical Engineering Graduate Program Entrance Exam, 2007
- Ranked among Top 0.1% in the Iranian Nationwide Undergraduate Program Entrance Exam, 2003

Membership in Professional & Honorary Societies

- American Institute of Chemical Engineers (AIChE)
- AIChE Computing & Systems Technology Division (CAST)