Construction, Simulation and Testing of Causal Probabilistic Networks

Ulrich G. Oppel, Alexander Hierle,
Masoud Noormohammadian

Mathematisches Institut der Ludwig–Maximilians–Universität München
Theresienstr. 39, D 80333 München, Germany
Email: ulrich@oppel2.mathematik.uni-muenchen.de

ABSTRACT

From the probabilistic point of view a complex system subject to vagueness, randomness and uncertainty may be characterized by a multivariate probability distribution. Such a multivariate distribution may be approximated by the sequence of empirical distributions of a properly chosen sample of realizations of the system, e.g. obtained from Monte Carlo simulations of the system or by properly collected data. But how to obtain these Monte Carlo simulations?

Another way of characterizing such a system is to use an ancestral causal probabilistic network (CPN). Such a CPN is a directed graph and a family of Markov kernels. The graph describes the dependencies of the subsystems qualitatively and the Markov kernels describe them quantitatively. The conditional probabilities of the Markov kernels may be interpreted as stochastic cause–effect relations. To such a CPN a directed Markov field is associated which is the joint multivariate distribution of the system.

The representation of multivariate distribution by a CPN has some advantages: it makes even complicated multivariate distributions storable and operable, it allows for Bayesian learning by introducing and propagating of evidence, it may serve for calculation of marginal distributions, and it may be used for Monte Carlo simulations of the system. These properties make it possible to evaluate the description of the system by the multivariate distribution and the CPN. This evaluation is based on comparison of marginal distributions of the multivariate distribution with empirical distributions obtained from Monte Carlo simulations and from data. The comparison may be based on properly chosen symmetric or asymmetric distances or statistical tests. We give some examples.

1 Introduction

In many fields of application such as physics, technology, biology, medical science, and economy very often we have to deal with complex systems which are subject to randomness, uncertainty, imprecision, incompleteness and vagueness due to variabilty between and within its components or due to the kind of its observation and description. Several methods have been developed to deal with such systems.

Probability theory was the first such method and it has been successful in many fields. According to any of the many systems of axioms of probability (e.g. see Kolmogorov [11], de Finetti [5], [6], Richter [26], [27], Carnap [1], Savage [29], Stegmüller [32]) such a complex system may be described mathematically by a probability space $(\Omega, \mathfrak{K}, p)$ and the components v of such a complex system may be characterized by random variables $X_v : \Omega \to S_v$ assuming their values in rather general state spaces S_v with σ–algebras $\mathfrak{S}_v$. From the probabilistic point of view the qualitative and quantitative aspects of the total system are completely determined by the joint distribution $\mathbb{P} := p \circ X^{-1}$ of the family $X := (X_v : v \in V)$ of random variables representing its components. The joint distribution is a probability measure

$$\begin{aligned} \mathbb{P} : \mathfrak{S} \to [0,1] \quad \text{with} \quad B \mapsto \mathbb{P}(B) &:= p \circ X^{-1}(B) := p(X^{-1}(B)) \\ &:= p\Big(\{\omega \in \Omega : (X_v(\omega) : v \in V) \in B\}\Big) \end{aligned}$$

on the product–σ–algebra $\mathfrak{S} := \bigotimes_{v \in V} \mathfrak{S}_v$ on the product state space $S := \prod_{v \in V} S_v$ belonging to the family of state spaces $((S_v, \mathfrak{S}_v) : v \in V)$. The system is completely determined by the probability space $(S, \mathfrak{S}, \mathbb{P})$. However, to find $\mathbb{P}$ is often very difficult. Usually $\mathbb{P}$ may be determined at most only approximately by theoretical or statistical procedures.

Very often in fields of application like medical science the systems are so extremely complicated and the information is such poor that it seems to be impossible to determine $\mathbb{P}$, neither totally nor partially. Therefore, new concepts for the description of such complex systems have been developed, for example belief, plausibility, possibility, propositional calculus, compositional systems, graphical models, neural networks, and fuzzy sets and logics; e.g. see Hajek et al. [8], Kruse et al. [12], [13] and Pearl [25]. All of these concepts and theories have some important advantages, but probability theory has some advantages, too.

The most important advantage of probability theory is that it is a rich theory. Many concepts, theorems, and procedures have been developed during its long history. To mention just a few of them: Methods of composition (integration; Fubini's and Ionescu–Tulcea's theorem) and decomposition (desintegration; Radon–Nikodym theorem, conditioning, Bayes' theorem), concepts of convergence (in probability, almost everywhere; L_p; weak, vage and uniform), inequalities (Tschebyshev's, Kolmogorov's), limit theorems (law of large numbers, ergodic theorem, Glivenko–Cantelli–Tucker theorem; cen-

tral limit theorem; stationary distributions), and statistical methods (estimates for parameters or intervals of parameters; test procedures). This rich theory may be used to obtain methods for estimating the probability measure $\mathbb{P}$ from empirical knowledge, for checking $\mathbb{P}$ against data, and for simulating $\mathbb{P}$.
One way of finding $\mathbb{P}$ partially or totally is to calculate the empirical distribution from properly collected data. Here a data point x is the point of the joint state space which is obtained as the family $X(\omega) := (X_v(\omega) : v \in V)$ of states $X_v(\omega)$ of the components v for a realization ω of the complex system; i.e. $x = (X_v(\omega) : v \in V) \in S$. If we have a sequence $(x_{(k)} : k \in \mathbb{N})$ of data points which are obtained from a sequence $(X_{(k)} : k \in \mathbb{N})$ of independent (or ergodic) repetitions of X, then the law of large numbers tells us that for every $B \in S$ the relative frequency $h_k(\omega, B)$ of the occurance of B in the sequence $(X_{(k)}(\omega) : k \in \mathbb{N})$ tends with $k \to \infty$ for p–almost all $\omega \in \Omega$ to the probability $\mathbb{P}(B)$ $(= p(\{\omega \in \Omega : X(\omega) \in B\}))$; i.e.

$$h_k(\omega, B) \;:=\; 1/k \sum_{i=1}^{k} 1_B(X_{(i)}(\omega)) \;\to\; \mathbb{P}(B)$$
$$\text{for } k \to \infty \text{ for } p\text{–almost all } \omega \in \Omega.$$

If we assume that all state spaces $(S_v, \mathfrak{S}_v)$ are polish spaces S_v and $\mathfrak{S}_v$ are the σ–algebras of Borel subsets and that V is countable, then S is polish, too, and $\mathfrak{S}$ the Borel–σ–algebra of S. (A topological space Z is polish, if it is separabel and completely metrizable. The σ–algebra of Borel subsets of Z is generated by the open subsets of Z.) In applications we always may make this assumption. Then the theorem of Glivenko–Cantelli (e.g. see Parthasarathy [24]) tells us that the sequence $(\mu_{k,\omega} : k \in \mathbb{N})$ of empirical distributions obtained from the sequence $(X_{(k)}(\omega) : k \in \mathbb{N})$ converges weakly to $\mathbb{P}$ for p–almost all $\omega \in \Omega$, i.e.

$$\mu_{k,\omega} \;:=\; 1/k \sum_{i=1}^{k} \delta_{X_{(i)}(\omega)} \;\Longrightarrow\; \mathbb{P} \qquad \text{for } k \to \infty \text{ for } p\text{–almost all } \omega \in \Omega.$$

(A sequence $(\mu_k : k \in \mathbb{N})$ of probability measures on the polish space S converges weakly to a probability measure μ if the sequence $(\int_S f d\mu_k : k \in \mathbb{N})$ of real numbers converges to $\int_S f d\mu$ for every bounded continuous function $f : S \to \mathbb{R}$. For $x \in S$ we denote by $\delta_x : \mathfrak{S} \to [0,1]$ with $B \mapsto \delta_x(B) := 1_B(x)$ the normed Dirac measure which is concentrated in x.)
If in addition S is some finite dimensional Euklidean space $\mathbb{R}^m$, then we obtain even uniform convergence of the associated cumulative distribution functions $F_{k,\omega}(x) := \mu_{k,\omega}(\{y \in \mathbb{R}^m : y \leq x\}$ and $F(x) := \mu(\{y \in \mathbb{R}^m : y \leq x\})$, i.e.

$$\lim_{k\to\infty} \sup_{x\in\mathbb{R}^m} \left|F_{k,\omega}(x) - F(x)\right| \;=\; 0 \qquad \text{for } p\text{–almost all } \omega \in \Omega.$$

The empirical distribution $\mu_{k,\omega}$ is concentrated on the sample $\{X_{(1)}(\omega), \ldots, X_{(k)}(\omega)\}$ of k points $X_{(i)}(\omega)$ in the joint state space.
For some (mathematical and interpretational) purposes this "hard concentration" has disadvantages. If $S = \mathbb{R}^m$ for example, we can switch from $\mu_{k,\omega}$ to a "soft" empirical distribution by convoluting $\mu_{k,\omega}$ with some continuously distributed probability ν_k (e.g. a normal distribution) such that the sequence $(\nu_k : k \in \mathbb{N})$ converges weakly to the Dirac measure in $0 \in \mathbb{R}^m$. Because of the continuity of the convolution with respect to the weak convergence a in such a way modified theorem of the Glivenko–Cantelli type still holds.
Another way of getting more information about $\mathbb{P}$ would be to formulate some hypotheses about $\mathbb{P}$ and to apply some appropriate statistical test procedures to draw conclusions on $\mathbb{P}$. Because of the usually high dimension of the joint state space, first, it is in general not obvious how to formulate these hypotheses, and second, the usually small sample size (relative to the dimension of the joint state space) will not allow for sufficiently powerful tests. However, for lower dimensional marginal distributions this will work.
Finally, there is a way of constructing $\mathbb{P}$ using causal probabilistic networks. We shall study this method. We shall define causal probabilistic networks, point out some important properties, and present some methods for the construction, simulation and testing of causal probabilistic networks.
This will also yield a method of generating random samples of high dimensional multivariate distributions. Hence, the method of describing $\mathbb{P}$ by a causal probabilistic network will also provide us with a method for Monte Carlo simulation of $\mathbb{P}$ or of some marginal of $\mathbb{P}$ and therefore with empirical distributions which converge to $\mathbb{P}$ or its marginal, respectively.

2 Causal probabilistic networks and their associated multivariate distribution

A **causal probabilistic network (CPN)** is a directed graph $G := (V, E)$ with $E \subset V \times V$ and $(u, v) \notin E$ for $(v, u) \in E$ and a family $\mathcal{P} := (\mathcal{P}_v : v \in V)$ of Markov kernels

$$\begin{array}{rcl} \mathcal{P}_v \ : \ S(Pa(v)) \times \mathfrak{S}_v & \to & [0,1] \quad \text{with} \\ \Big((x_u : u \in Pa(v)), B\Big) & \mapsto & \mathcal{P}_v\Big((x_u : u \in Pa(v)); B\Big) \end{array}$$

where $Pa(v) := \{u \in V : (u, v) \in E\}$ is the set of the parents of v in G. For $\emptyset \neq U \subset V$

$$S(U) := \prod_{u \in U} S_u \qquad \text{is the product set and}$$

$$\mathfrak{S}(U) := \bigotimes_{u \in U} \mathfrak{S}_u \qquad \text{is the product–}\sigma\text{–algebra}$$

of the family $((S_u, \mathfrak{S}_u) : u \in U)$ of state spaces $(S_u, \mathfrak{S}_u)$.
Every node $v \in V$ represents a random variable X_v with the state space $(S_v, \mathfrak{S}_v)$, E describes the dependency of $(X_v : v \in V)$ qualitatively, and $\mathcal{P}$ describes this dependency quantitatively.
(A Markov kernel $\mathcal{P}$ from a measurable space $(Y, \mathfrak{Y})$ to a measurable space $(Z, \mathfrak{Z})$ is a mapping $P : Y \times \mathfrak{Z} \to [0,1]$ with $(y, B) \mapsto P(y; B)$ such that for fixed $y \in Y$ the mapping $P(y; \cdot) : \mathfrak{Z} \to [0,1]$ is a probability measure and for each fixed B the mapping $P(\cdot\,; B) : (Y, \mathfrak{Y}) \to [0,1]$ is measurable.)
The Markov kernel $\mathcal{P}_v$ is considered to be a family of conditional probabilities $\mathcal{P}_v((x_u : u \in Pa(v)); B)$ which are supposed to describe "the" probability of a (measurable) subset B of the state space of the node v given the state x_u of the state space of the node u for every parent node u of v. In other words, $\mathcal{P}_v$ is supposed to be "the" conditional distribution of X_v given $(X_u : u \in Pa(v))$. In applications such a Markov kernel is used to describe stochastic cause–effect relations. Therefore CPNs may be used effectively to model stochastic knowledge in expert systems; e.g. Pearl [25] and Neapolitan [18].
A very natural question now arising is the following: Is there any probability measure $\mathbb{P} : \mathfrak{S}(V) \to [0,1]$ on the σ–algebra $\mathfrak{S}(V)$ of the joint state space of the family $(X_v : v \in V)$ of random variables X_v which is the joint distribution of the random variables X_v and which has the given Markov kernels $\mathcal{P}_v$ as conditional distributions of X_v given the values of the "parent" variables X_u? And if so, is it uniquely determined.
We call a graph $G := (V, E)$ **ancestral** if there exists an enumeration $V := \{v_n : n \in N\}$ with $N \subset \mathbb{N}$ such that no descendant is enumerated before one of his ancestors. Such an enumeration we call ancestral, too.
Some examples: If G is finite and acyclic, then G is ancestral. If G is cyclic, then G is not ancestral. $\mathbb{Z}$ with the natural ordering and linear graph structure is not ancestral. The graph $G := (V, E)$ with $V := \mathbb{N} \times \mathbb{N}$ and $E := \{((k,l),(m,n)) \in V \times V :$ either $m = k+1$ and $n = l$ or $m = k$ and $n = l+1\}$ is ancestral.
For any CPN with an ancestral graph $G := (V, E)$ there exists a uniquely determined probability measure $\mathbb{P} : \mathfrak{S}(V) \to [0,1]$ which is the joint distribution of the random variables X_v, has the given Markov kernels $\mathcal{P}_v$ as conditional distributions of X_v given the values of the "parent" variables X_u, and is a directed Markov field. We call this probability measure $\mathbb{P}$ the **multivariate joint probability** or the **multivariate distribution associated to this CPN**.

For finite graphs this can be shown by iterative integration:
Find an ancestral enumeration $V = \{v_i : i = 0, \ldots, n\}$ of V and define

$$S^{(i)} := S\Big(\{v_j : j = 0, \ldots, i\}\Big) \qquad \text{to be the product set and}$$

$$\mathfrak{S}^{(i)} := \bigotimes_{j=0}^{i} \mathfrak{S}_{v_j} \qquad \text{to be the product } \sigma\text{–algebra for } 0 \leq i \leq n.$$

Defining Markov kernels for $i = 0, \ldots, n$ by

$$\begin{array}{rcl} P_i \ : \ S^{(i-1)} \times \mathfrak{S}_{v_i} & \to & [0,1] \quad \text{with} \\ \Big((x_{v_0}, \ldots, x_{v_{i-1}}), B\Big) & \mapsto & P_i(x_{v_0}, \ldots, x_{v_{i-1}}; B) := \mathcal{P}_{v_i}\Big((x_u : u \in Pa(v_i)); B\Big) \end{array}$$

The probability measure $\mathbb{P} : \mathfrak{S}(V) \to [0,1]$ is obtained by iterative integration:

$$\begin{aligned} \mathbb{P}(A) \ := \ \int\limits_{S_{v_0}} \int\limits_{S_{v_1}} \cdots \int\limits_{S_{v_n}} & 1_A(x_{v_0}, \ldots, x_{v_n}) \quad P_n(x_{v_0}, \ldots, x_{v_{n-1}}; dx_{v_n}) \quad \ldots \\ & \ldots \quad P_1(x_{v_0}; dx_{v_1}) \quad P_0(dx_{v_0}) \end{aligned} \tag{2.1}$$

for $A \in \mathfrak{S} := \mathfrak{S}^{(n)} = \mathfrak{S}(V)$. For infinite graphs the existence of the probability measure $\mathbb{P} : \mathfrak{S}(V) \to [0,1]$ may be obtained from C. Ionescu–Tulcea's theorem; e.g. see Neveu [19]. Its finite dimensional marginal measures are given by iterative integrations of the form of equation 2.1.

The probability measure $\mathbb{P} : \mathfrak{S}(V) \to [0,1]$ determined by equation 2.1 does not depend on the chosen enumeration and it is a directed Markov field (in the sense of Lauritzen et al. [14]; see Oppel [20] or [22]).

The iterative construction of equation 2.1 is continuous for several norms and topologies; see Matthes et al. [17]:

Let us consider a sequence $(\mathcal{C}_n : n \in \mathbb{N})$ of CPNs $\mathcal{C}_n$ with a finite acyclic directed graph $G := (V, E)$ and sequence $(\mathcal{P}_{(n)} : n \in \mathbb{N})$ of families $\mathcal{P}_{(n)} := (\mathcal{P}_{n,v} : v \in V)$ of Markov kernels

$$\begin{array}{rcl} \mathcal{P}_{n,v} : S(Pa(v)) \times \mathfrak{S}_v & \to & [0,1] \quad \text{with} \\ \Big((x_u : u \in Pa(v)), B\Big) & \mapsto & \mathcal{P}_{n,v}\Big((x_u : u \in Pa(v)); B\Big) \end{array}$$

and associated multivariate distributions

$$\mathbb{P}_n : \mathfrak{S} \ \to \ [0,1]$$

and a CPN $\mathcal{C}$ with the same graph G and the family $\mathcal{P} := (\mathcal{P}_v : v \in V)$ of Markov kernels

$$\begin{array}{rcl} \mathcal{P}_v : S(Pa(v)) \times \mathfrak{S}_v & \to & [0,1] \quad \text{with} \\ \Big((x_u : u \in Pa(v)), B\Big) & \mapsto & \mathcal{P}_v\Big((x_u : u \in Pa(v)); B\Big) \end{array}$$

and the associated multivariate distribution

$$\mathbb{P} : \mathfrak{S} \ \to \ [0,1].$$

If the sequence $(\mathcal{P}_{(n)} : n \in \mathbb{N})$ converges uniformly to $\mathcal{P}$, i.e.

$$\begin{aligned} \lim_{n\to\infty} \|\mathcal{P}_{n,v} - \mathcal{P}_v\| &= 0 \qquad \text{for all } v \in V \text{ where} \\ \|\mathcal{P}_{n,v} - \mathcal{P}_v\| &:= \sup\Big\{ |\mathcal{P}_{n,v}(\vec{x}; B) - \mathcal{P}_v(\vec{x}; B)| : \\ &\qquad (\vec{x}; B) := \Big((x_u : u \in Pa(v)); B\Big) \in S(Pa(v)) \times \mathfrak{S}_v\Big\}, \end{aligned}$$

then the sequence $(\mathbb{P}_n : n \in \mathbb{N})$ of associated multivariate distributions $\mathbb{P}_n$ converges uniformly to $\mathbb{P}$, i.e.

$$\begin{aligned} \lim_{n\to\infty} \|\mathbb{P}_n - \mathbb{P}\| &= 0 \qquad \text{where} \\ \|\mathbb{P}_n - \mathbb{P}\| &:= \sup\Big\{ |\mathbb{P}_n(A) - \mathbb{P}(A)| : A \in \mathfrak{S}\Big\}. \end{aligned}$$

The topologies of uniform convergence of Markov kernels and of measures are very fine topologies. If all the state spaces are polish and the σ–algebras are the σ–algebras of Borel subsets, then we may consider instead of these very fine topologies the much coarser topologies of pointwise weak convergence of Markov kernels and measures.
However, if we reduce the assumption of uniform convergence of the sequences $(\mathcal{P}_{n,v} : n \in \mathbb{N})$ to the Markov kernel $\mathcal{P}_v$ to pointwise weak convergence of the sequences $(\mathcal{P}_{n,v} : n \in \mathbb{N})$ to the Markov kernel $\mathcal{P}_v$ [i.e. the weak convergence of the sequences $(\mathcal{P}_{n,v}((x_u : u \in Pa(v)); \cdot) : n \in \mathbb{N})$ of probability measures to the probability measure $\mathcal{P}_v((x_u : u \in Pa(v)); \cdot)$ for all $(x_u : u \in Pa(v)) \in S(Pa(v))$], then the weak convergence of the sequence $(\mathbb{P}_n : n \in \mathbb{N})$ of the associated multivariate distributions to $\mathbb{P}$ needs not to be true:
Let be $G := (V, E)$ with $V := \{0, 1\}$ and $E := \{(0, 1)\}$, $S_0 := \{0\} \cup \{1/n : n \in \mathbb{N}\}$ with the topology induced by the usual topology on $\mathbb{R}$, $S_1 := \{0, 1\}$ with the discrete topology, and

$$\mathcal{P}_0 : \mathfrak{S}_0 \to [0, 1] \quad \text{with} \quad A \mapsto \mathcal{P}_0(A) := 1/2\ \mu(A) + 1/2\ \delta_0(A),$$

$$\mathcal{P}_1 : S_0 \times \mathfrak{S}_1 \to [0, 1] \quad \text{with} \quad (x_0, B) \mapsto \mathcal{P}_1(x_0; B) := \begin{cases} \delta_0(B) & \text{for } x_0 = 0, \\ \delta_1(B) & \text{for } x_0 \neq 0, \end{cases}$$

$$\mathcal{P}_{n,0} : \mathfrak{S}_0 \to [0, 1] \quad \text{with} \quad A \mapsto \mathcal{P}_{n,0}(A) := 1/2\ \mu(A) + 1/2\ \delta_{1/n}(A),$$

$$\mathcal{P}_{n,1} := \mathcal{P}_1, \quad \text{where} \quad \mu(\{1/n\}) := 2^{-(n+1)} \text{ for } n \in \mathbb{N}.$$

The associated multivariate distributions are $\mathbb{P} : \mathfrak{S} \to [0, 1]$ and $\mathbb{P}_n : S \to [0, 1]$ with

$$\mathbb{P}(C) := 1/2\ \nu(C) + 1/2\ \delta_{(0,0)}(C) \qquad \text{and}$$

$$\begin{aligned} \mathbb{P}_n(C) &:= 1/2\ \nu(C) + 1/2\ \delta_{(1/n,1)}(C) \qquad \text{where} \\ \nu(C) &:= \int\limits_{S_0} 1_C(x_0, 1)\ \mu(dx_0) \qquad \text{for } C \in \mathfrak{S}. \end{aligned}$$

Obviously, the sequences $(\mathcal{P}_{n,0} : n \in \mathbb{N})$ and $(\mathcal{P}_{n,1} : n \in \mathbb{N})$ converge pointwise weakly to $\mathcal{P}_0$ and $\mathcal{P}_1$, respectively, but the sequence $(\mathbb{P}_n : n \in \mathbb{N})$ does not converge weakly to $\mathbb{P}$.

The continuity assertion is true only under additional conditions such as: all Markov kernels are Feller kernels, have the strong Feller convergence property **(SFCP)**, and have the compact uniform tightness property **(CUTP)**; again see Matthes et al. [17]. For finite state spaces (each with the discrete topology) all these conditions are fulfilled and the pointwise weak convergence coincides with the uniform convergence.

What about the variation of the graph in such a limit? The links between nodes indicate dependencies. Missing links indicate the stochastic independence or the conditional stochastic independence of the variables represented by the nodes. Conditional stochastic independence means stochastic independence given the states of the latest common ancestors. Stochastic independence is measure theoretically described by product measures. Since the formation of product measures is continuous with respect to the considered topologies (of uniform or pointwise weak convergence), independence or conditional independence will be preserved in the limit. Hence, new links will not show up in the graph. Eventually, links will disappear; examples are easy to construct.

As we have seen, by iterative integration we may obtain a multivariate probability distribution from a CPN. We have mentioned some continuity properties of this iterated integration procedure. Indeed, under very mild assumptions we may obtain from a multivariate probability distribution a CPN, or even many.

Let V be countable, $\mathbb{P} : \mathfrak{S}(V) \to [0, 1]$ be a probability distribution, all state spaces S_v be polish, and S_v be their Borel–σ–algebras. Choose any enumeration $V := \{v_n : n \in N\}$ with $N \subset \mathbb{N}$ of V and let $S^{(i)}$ and $\mathfrak{S}^{(i)}$ be defined as above. Furthermore, define $\pi_i : S(V) \to S^{(i)}$ and $\pi_{i,j} : S^{(i)} \to S^{(j)}$ for $0 \leq j \leq i \in N$ to be the canonical projections and $P_i : S^{(i-1)} \times \mathfrak{S}^{(i)} \to [0, 1]$ be the desintegration kernel of the marginal measure $\mathbb{P} \circ \pi_i^{-1} : \mathfrak{S}^{(i)} \to [0, 1]$ with respect to $\pi_{i,i-1}$. Applying the factorization lemma for measurable functions, we obtain from $(P_i : i \in N)$ an acyclic graph and a family of Markov kernels, i.e. we obtain a CPN. See also Oppel [20] or [22].

One important advantage of a CPN is the possibility to store a highdimensional multivariate distribution (in principle completely) which we never could store explicitely. For example, the multivariate joint distribution of a system with 1000 variables with two values each would need about 10^{300} bytes of storage.

Another important advantage of a CPN is that we may use it for Bayesian learning. Bayesian learning is deductive learning from the general knowledge (given by the multivariate distribution $\mathbb{P}$) by conditioning with the given evidence. To calculate

conditional probabilities is a very simple task, in principle. If the system is very large, however, this is a very difficult task in practice. For the calculation of $\mathbb{P}$ or marginals of $\mathbb{P}$ and for introducing and propagating of evidence into such a system there are very efficient algorithms. For finite state spaces we have available the Lauritzen–Spiegelhalter algorithm (e.g. in the shell HUGIN) which is a combination of local and global backward and forward calculations based on Fubini's and Bayes' theorem; e.g. see Lauritzen et al. [15] or Jensen et al. [10]. For more general state spaces there are a number of Monte Carlo algorithms available; e.g. see Henrion [9], Chavez et al. [2], Chin et al. [3], Fung et al. [7], Shachter et al. [31]. We use variance reduced Monte Carlo methods based on importance sampling and on equation 2.1.
Unfortunately, conditioning has severe discontinuity properties with respect to all kinds of topologies. (This demands for caution in applications of expert systems with a knowledge base represented as a CPN.) For example:
Let be $G := (V, E)$ with $V := \{0, 1\}$ and $E := \{(0, 1)\}$, $S_0 := \{0\} \cup \{1/k : k \in \mathbb{N}\}$ and $S_1 := \{0, 1\}$ with the discrete topology, and

$$\begin{aligned}
\mathcal{P}_0 : \mathfrak{S}_0 \to [0,1] \quad &\text{with} \quad A \mapsto \mathcal{P}_0(A) := \delta_0(A), \\
\mathcal{P}_1 : S_0 \times \mathfrak{S}_1 \to [0,1] \quad &\text{with} \quad (x_0, B) \mapsto \mathcal{P}_1(x_0; B) := \delta_1(B), \\
\mathcal{P}_{n,0} : \mathfrak{S}_0 \to [0,1] \quad &\text{with} \quad A \mapsto \mathcal{P}_{n,0}(A) := (1 - 1/n)\ \delta_{1/n}(A) + 1/n\ \delta_0(A), \\
\mathcal{P}_{n,1} : S_0 \times \mathfrak{S}_1 \to [0,1] \quad &\text{where} \quad (x_0, B) \mapsto \mathcal{P}_{n,1}(x_0; B) := \begin{cases} \delta_0(B) & \text{for } x_0 = 0, \\ \delta_1(B) & \text{for } x_0 \neq 0. \end{cases}
\end{aligned}$$

Then the sequence $(\mathcal{P}_{n,v} : n \in \mathbb{N})$ of Markov kernels $\mathcal{P}_{n,v}$ converge uniformly to $\mathcal{P}_v$ for every $v \in V$. The sequence $(\mathbb{P}_n : n \in \mathbb{N})$ of associated multivariate distributions

$$\mathbb{P}_n \quad = \quad (1 - 1/n) \cdot \delta_{(1/n,1)} + 1/n \cdot \delta_{(0,0)}$$

of G and $(\mathcal{P}_{n,0}, \mathcal{P}_{n,1})$ converges weakly to the associated multivariate distribution $\mathbb{P} = \delta_{(0,1)}$ of G and $(\mathcal{P}_0, \mathcal{P}_1)$, but the sequence $(Q_n(0; \cdot) : n \in \mathbb{N})$ of conditional distributions

$$Q_n(0; \cdot) \quad := \quad \mathbb{P}_n(\,\cdot\,|X_0 = 0) \circ X_1^{-1} \quad = \quad \delta_0$$

does not converge weakly to the conditional distribution

$$Q(0; \cdot) \quad := \quad \mathbb{P}(\,\cdot\,|X_0 = 0) \circ X_1^{-1} \quad = \quad \delta_1.$$

As we mentioned above, CPNs may be used effectively to model stochastic knowledge in expert systems. The knowledge base of such expert systems containing uncertain and vague knowledge may be represented by a CPN. Such a representation is most adequate especially then, if the knowledge is given or attainable in form of stochastic

cause–effect relations. Stochastic cause–effect relations are usually given as conditional probabilities: if certain conditions are given, then they will imply a certain event with a certain probability. These conditional probabilities may be combined to tables of conditional probabilities, i.e. to Markov kernels. The dependencies define a directed graph.

To find the directed graph is very often not difficult. The graph structure may be found by qualitative considerations of experts. Much more difficult is the problem of finding the Markov kernels. To determine the Markov kernels in non trivial applications, many conditional probabilities have to be estimated. E.g. for the Markov kernel belonging to one node with 10 states and with 5 parent nodes having 10 states each we have to find 10^6 conditional probabilities. Usually, the available data is insufficient to estimate that many numbers. To make use of the excellent properties of CPNs, we have to design methods for the construction of the needed Markov kernels. We have to use qualitative and quantitative expert knowledge as much as possible. We have designed and applied several methods for the construction of Markov kernels based on global and local procedures.

Some of the global procedures for the construction of the family of Markov kernels for a CPN are techniques for the transformation of systems of differential equations (e.g. describing compartmental models of metabolic processes or predator–prey systems; see Oppel et al. [21], Salzsieder et al. [28], Oppel [22], [23]) and of rule–based expert systems (e.g. see Liesenfeld et al. [16]). These procedures use the deep "deterministic" knowledge and stochastify it properly to obtain a more realistic stochastic model and to be able to make use of the new powerful computational methods available for CPNs.

Some of the local procedures are methods for the estimation of single Markov kernels based on functional relations, generalized linear models, and neural networks combined with properly chosen stochastifications or based on statistical estimates of (conditional) moments of the (conditional) distributions contained in the Markov kernel; e.g. see Liesenfeld et al. [16].

Nevertheless, in complex applications we are forced to choose many of the Markov kernels somewhat deliberately. Anyway, we have the problem of evaluation of the constructed CPN. For example, we have to check that the associated multivariate distribution $\mathbb{P}$ or at least some of its marginals are consistent with the given data or expert knowledge. To do this, we propose to make Monte Carlo simulations of $\mathbb{P}$ or its marginals, to calculate marginal distributions exactly, to calculate or to simulate "limiting" distributions (if some Markov chain may be embedded into the CPN), and to compare these results with the given data or expert knowledge by statistical testing and by measuring appropriate distances. We shall give some examples.

3 Monte Carlo simulation of a CPN and its associated multivariate distribution

Let us now consider a causal probabilistic network (CPN) with an acyclic directed graph $G := (V, E)$ with $E \subset V \times V$ and $(u, v) \notin E$ for $(v, u) \in E$ and a family $\mathcal{P} := (\mathcal{P}_v : v \in V)$ of Markov kernels

$$\begin{array}{rcl} \mathcal{P}_v \ : \ S(Pa(v)) \times \mathfrak{S}_v & \to & [0,1] \quad \text{with} \\ \Big((x_u : u \in Pa(v)), B\Big) & \mapsto & \mathcal{P}_v\Big((x_u : u \in Pa(v)); B\Big). \end{array}$$

Again for $\emptyset \neq U \subset V$

$$\begin{array}{rl} S(U) := \prod\limits_{u \in U} S_u & \text{is the product set and} \\ \mathfrak{S}(U) := \bigotimes\limits_{u \in U} \mathfrak{S}_u & \text{is the product–}\sigma\text{–algebra} \end{array}$$

of the family $((S_u, \mathfrak{S}_u) : u \in U)$ of state spaces $(S_u, \mathfrak{S}_u)$, respectively. Every node $v \in V$ represents a random variable X_v with the state space $(S_v, \mathfrak{S}_v)$; we may assume that $X_v : S(V) \to S_v$ is the canonical projection. Furthermore, for $\emptyset \neq U \subset W \subset V$ we denote by

$$\begin{array}{rcl} X_U : S(V) & \to & S(U) \quad \text{with} \\ x := (x_v : v \in V) & \mapsto & X_U(x) := (x_v : v \in U) \quad \text{and by} \\ \\ X_{WU} : S(W) & \to & S(U) \quad \text{with} \\ x_W := (x_v : v \in W) & \mapsto & X_{WU}(x_W) := (x_v : v \in U) \end{array}$$

the canonical projections.
We shall describe a very natural method for obtaining independent (or ergodic) realizations of the random variable $X := X_V := (X_v : v \in V)$ which assumes its values in the joint state space $S(= S(V))$ and which is distributed according to $\mathbb{P}$. In other words, we shall describe a method for the Monte Carlo simulation of $\mathbb{P}$, and hence, for the Monte Carlo simulation of the CPN associated to $\mathbb{P}$. This method is similar to the one proposed by Henrion [9] for a CPN with finite state spaces and which is called the probabilistic logic sampling. (Here, we do not want to propagate evidence in a CPN via Monte Carlo simulation. For that purpose there are more efficient algorithms; e.g. see Shachter–Peot [31].)

The proposed Monte Carlo simulation methods works as follows:
First, find some (appropriate) ancestral enumeration $V := \{v_i : i = 0, \ldots, n\}$ of V.

Again, define

$$S^{(i)} := S\Big(\{v_j : j = 0, \ldots, i\}\Big) \qquad \text{to be the product set and}$$

$$\mathfrak{S}^{(i)} := \bigotimes_{j=0}^{i} \mathfrak{S}_{v_j} \qquad \text{to be the product } \sigma\text{–algebra for } 0 \leq i \leq n.$$

Again, defining Markov kernels for $i = 0, \ldots, n$ by

$$\begin{array}{rcl} P_i \ : \ S^{(i-1)} \times \mathfrak{S}_{v_i} & \to & [0,1] \quad \text{with} \\ \Big((x_{v_0}, \ldots, x_{v_{i-1}}), B\Big) & \mapsto & P_i(x_{v_0}, \ldots, x_{v_{i-1}}; B) := \mathcal{P}_{v_i}\Big((x_u : u \in Pa(v_i)); B\Big) \end{array}$$

The joint distribution $\mathbb{P} : \mathfrak{S}(V) \to [0,1]$ associated to the given CPN is obtained by iterative integration:

$$\begin{aligned} \mathbb{P}(A) \ := \ \int_{S_{v_0}} \int_{S_{v_1}} \cdots \int_{S_{v_n}} & 1_A(x_{v_0}, \ldots, x_{v_n}) \quad P_n(x_{v_0}, \ldots, x_{v_{n-1}}; dx_{v_n}) \quad \ldots \\ & \ldots \quad P_1(x_{v_0}; dx_{v_1}) \quad P_0(dx_{v_0}) \end{aligned} \tag{3.2}$$

for $A \in \mathfrak{S} := \mathfrak{S}^{(n)} = \mathfrak{S}(V)$. This equation is the basis for a Monte Carlo simulation of $\mathbb{P}$. A Monte Carlo simulation of $\mathbb{P}$ (i.e. of the associated acyclic CPN) yields a point $x := (x_v : v \in V) \in S(V)$ which is a random realization of the family of random variables $(X_v : v \in V)$ on $(S(V), \mathfrak{S}(V), \mathbb{P})$.
The flow chart for the Monte Carlo simulation of $\mathbb{P}$ is based on equation 3.2:

Let $V := \{v_i : i = 0, \ldots, n\}$ be the chosen ancestral enumeration of V.

0. Choose $x_{v_0} \in S_{v_0}$ at random according to the probability distribution $P_0 = \mathcal{P}_{v_0}$. (The Markov kernel P_0 is degenerate!)
1. Choose $x_{v_1} \in S_{v_1}$ at random according to the probability distribution $P_1(x_{v_0}; \cdot)$.
2. Choose $x_{v_2} \in S_{v_2}$ at random according to the probability distribution $P_2(x_{v_0}, x_{v_1}; \cdot)$.

   $\vdots$

n. Choose $x_{v_n} \in S_{v_n}$ at random according to the probability distribution $P_n(x_{v_0}, \ldots, x_{v_{n-1}}; \cdot)$.

Flow chart:

$$\xrightarrow{P_0} \ x_{v_0} \ \xrightarrow{P_1(x_{v_0}; \cdot)} \ x_{v_1} \ \xrightarrow{P_2(x_{v_0}, x_{v_1}; \cdot)} \ x_{v_2} \ \xrightarrow{P_3(x_{v_0}, x_{v_1}, x_{v_2}; \cdot)} \ x_{v_3} \ \longrightarrow \ \ldots$$

Remember: $P_k(x_{v_0}, \ldots, x_{v_{k-1}}; \cdot) = \mathcal{P}_{v_k}\Big((x_u : u \in Pa(v_k)); \cdot\Big)$.

Let us now assume that every state space S_v is polish and S_v is its Borel–σ–algebra and that $(x_{(m)} : m \in \mathbb{N})$ with $x_{(m)} := (x_v^{(m)} : v \in V) \in S(V)$ is a sequence of independent (or ergodic) repetitions of Monte Carlo simulations of $\mathbb{P}$. From the Glivenko–Cantelli theorem we know:

$$\mathbb{P}^{\xi,k} \quad := \quad 1/k \sum_{m=1}^{k} \delta_{x_{(m)}} \quad \Longrightarrow \quad \mathbb{P} \qquad \text{for } k \to \infty$$

i.e. the empirical distributions $\mathbb{P}^{(\xi,k)}$ associated to the sequence $\xi := (x_{(m)} : m \in \mathbb{N})$ of independent (or ergodic) repetitions of Monte Carlo simulations of $\mathbb{P}$ converge weakly to $\mathbb{P}$ for $\mathbb{P}^{\mathbb{N}}$–almost all such repetitions.
Analogue statements are true for marginal distributions $\mathbb{P} \circ X_W^{-1}$ with $\emptyset \neq W \subset V$ of $\mathbb{P}$.

An example of a CPN:
The CPN "NEPHRO_RISK" for the risk of progression of the diabetic nephropathy.

To have a realistic CPN in mind, let us have a look at the CPN "NEPHRO_RISK". This CPN was designed for the prognosis of the risk of progression of diabetic nephropathy. It was constructed in a close cooperation by the mathematician A. Hierle and the physicians B. Liesenfeld and H.–J. Lüddeke as a part of the DIADOQ project "Diabetes mellitus: Optimierte Betreuung durch wissensbasierte Qualitätssicherung" of the research program MEDWIS sponsored by the German Minister of Science and Technology (BMFT).
The nodes of the CPN "NEPHRO_RISK" are representing the following variables:

age, anamnesis, ACE, smoking, HbA 1c;
albumin: 2 years ago, today;
blood pressure (Riva–Rocci): 2 years ago, today;
alpha–1–microglobulin: 2 years ago, today;
risk of progression of diabetic nephropathy.

The selection of the variables and the structure of the graph of this CPN was the result of a balancing consideration taking into account the possibly important dependencies, the physician's availability of the data, and the attainability of knowledge needed for the determination of the Markov kernels.
The quantitative knowledge needed for the Markov kernels was "extracted" from medical experts and data pools of clinics. The method was based on a linear model describing expected values of nodes given the states of the parent nodes as a weighted linear

combination of scored expected values given the state of the single parent node. The conditional distribution was obtained by a stochastification with a properly chosen and discretized normal distribution centered at the conditional expectation given the states of the parent nodes. For more details see Liesenfeld et al. [16]; a detailed description of the procedure will be published soon.

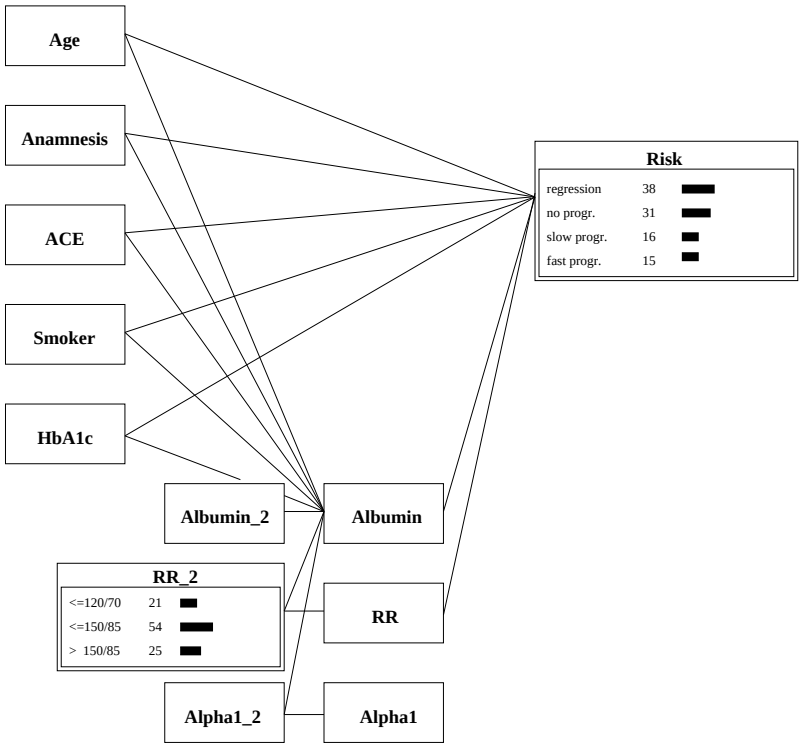

Figure 1: The directed graph of CPN "NEPHRO_RISK". The windows of the nodes "RR_2" and "Risk" are open and displaying the distributions of their variables.

As a result of the construction procedure described above, a first version of this CPN was obtained. It was checked by physicians comparing their experience with the prognosis of the HUGIN driven expert system based on this CPN. It passed this type of evaluation. Furthermore, this CPN was checked using data from the data pools of a clinic using statistical methods (such as L_1– and L_2–norms, Kullback–Leibler divergence, and chi–square fitting tests; see below). The first version of this CPN failed this statistical evaluation. Of course, the available data was at least as questionable as the experience of the physicians. But we had the feeling, that we should adapt the CPN more to the data without loosing the acception by the physicians.
Such a CPN is a very complex and delicate system. It is like a mobile: if you

touch it somewhere slightly, it may change far away wildly. Hence, any changes have to be supervised by properly designed controlling. To adapt the constructed CPN "NEPHRO_RISK" to data, a random search optimization procedure was applied. The scores and, hence, the Markov kernels were changed at random (by a properly chosen selection probability). Then the L_2–norm distance between some marginals of the old associated multivariate distribution and the same marginals of the empirical distribution of the data was compared with the L_2–norm distance between the marginals of the new associated multivariate distribution and the marginals of the empirical distribution of the data. If the random variation of the CPN resulted in an improvement, this variation was taken as the starting point for the next step. If not, the old CPN was taken as the starting point again. Indeed, this procedure resulted in some improvement. But we are not completely satisfied until now. To obtain better results we designed Monte Carlo simulations for CPNs and methods for calculation, simulation and comparison of marginal distributions of the associated multivariate distributions of CPNs. Below, we shall describe such methods and report on some results.
Figure 1 shows the graph of the CPN "NEPHRO_RISK". The windows of two nodes are displaying their onedimensional marginal distributions of the associated multivariate distribution. For sake of simplicity, we shall only consider these two nodes and their one– and twodimensional distributions.

Examples of Monte Carlo simulations:
Simulations of a twodimensional marginal distribution of the associated multivariate distribution of the CPN "NEPHRO_RISK".

For the CPN "NEPHRO_RISK" we consider the joint distribution of the two variables "RR_2" and "Risk". The variable "RR_2" has the three states

$$\begin{aligned} 1 &: \text{“}\leq 120/70\text{”}, \\ 2 &: \text{“}120/70 < \ldots \leq 150/85\text{”}, \\ 3 &: \text{“}> 150/85\text{”}\ [mm\ Hg]; \end{aligned}$$

it is describing the Riva–Rocci blood pressure of the patient two years ago. The variable "Risk" is describing the risk of progression of diabetic nephropathy; it has the four states

$$\begin{aligned} 1 &: \text{“regression”}, \\ 2 &: \text{“no progression”}, \\ 3 &: \text{“slow progression”}, \\ 4 &: \text{“fast progression”}. \end{aligned}$$

Table 1 to 5 give examples of Monte Carlo simulations of the CPN "NEPHRO_RISK". They show the empirical probabilities (which here are the relative frequencies), the

relative errors in percent (which here are the quotients of the standard deviation and the expectation of the sample divided by the squareroot of the size of the sample), and the absolute frequencies (counts) for different sizes of samples.

RR_2	Risk	probability	rel. error in %	counts
1	1	0.0760	22.1	19
2	1	0.2240	11.8	56
3	1	0.0480	28.2	12
1	2	0.0560	26.0	14
2	2	0.2040	12.5	51
3	2	0.0520	27.1	13
1	3	0.0200	44.4	5
2	3	0.0920	19.9	23
3	3	0.0800	21.5	20
1	4	0.0200	44.4	5
2	4	0.0760	22.1	19
3	4	0.0520	27.1	13

Table 1: **250** Monte Carlo simulations of the CPN "NEPHRO_RISK".

RR_2	Risk	probability	rel. error in %	counts
1	1	0.1060	9.2	106
2	1	0.2140	6.1	214
3	1	0.0550	13.1	55
1	2	0.0660	11.9	66
2	2	0.1720	6.9	172
3	2	0.0720	11.4	72
1	3	0.0160	24.8	16
2	3	0.0920	9.9	92
3	3	0.0630	12.2	63
1	4	0.0100	31.5	10
2	4	0.0760	11.0	76
3	4	0.0580	12.8	58

Table 2: **1000** Monte Carlo simulations of the CPN "NEPHRO_RISK".

RR_2	Risk	probability	rel. error in %	counts
1	1	0.1197	2.7	1197
2	1	0.2083	1.9	2083
3	1	0.0465	4.5	465
1	2	0.0616	3.9	616
2	2	0.1672	2.2	1672
3	2	0.0865	3.2	865
1	3	0.0183	7.3	183
2	3	0.0843	3.3	843
3	3	0.0551	4.1	551
1	4	0.0108	9.6	108
2	4	0.0743	3.5	743
3	4	0.0674	3.7	674

Table 3: **10000** Monte Carlo simulations of the CPN "NEPHRO_RISK".

RR_2	Risk	probability	rel. error in %	counts
1	1	0.1235	0.8	12346
2	1	0.2049	0.6	20492
3	1	0.0471	1.4	4709
1	2	0.0605	1.2	6054
2	2	0.1702	0.7	17017
3	2	0.0848	1.0	8483
1	3	0.0193	2.3	1926
2	3	0.0885	1.0	8848
3	3	0.0535	1.3	5347
1	4	0.0096	3.2	958
2	4	0.0704	1.1	7036
3	4	0.0678	1.2	6784

Table 4: **100000** Monte Carlo simulations of the CPN "NEPHRO_RISK".

RR_2	Risk	probability	rel. error in %	counts
1	1	0.1226	0.3	122604
2	1	0.2067	0.2	206731
3	1	0.0477	0.4	47667
1	2	0.0610	0.4	60951
2	2	0.1695	0.2	169477
3	2	0.0838	0.3	83799
1	3	0.0189	0.7	18913
2	3	0.0901	0.3	90054
3	3	0.0531	0.4	53111
1	4	0.0100	1.0	9970
2	4	0.0689	0.4	68943
3	4	0.0678	0.4	67780

Table 5: **1000000** Monte Carlo simulations of the CPN "NEPHRO_RISK".

4 Construction of marginal distributions of the associated multivariate distribution of a CPN

Let $\mathbb{P} : \mathfrak{S}(V) \to [0,1]$ be the associated multivariate distribution of a given CPN with the graph $G := (V, E)$ and the family of Markov kernels $(\mathcal{P}_v : v \in V)$. Again we assume that all state spaces are polish.
For $v \in V$ and $\emptyset \neq W \subset V$ with $v \notin W$ the conditional distribution of X_v given $X_W := (X_w : w \in W)$ exists and is denoted by $\mathcal{P}_{v|W} : S(V) \times \mathfrak{S}_v \to [0,1]$. We have:

1. For fixed $B \in \mathfrak{S}_v$ the function $\mathcal{P}_{v|W}(\,\cdot\,; B) : \Big(S(V), X_W^{-1}(\mathfrak{S}(W))\Big) \to [0,1]$ with $x \mapsto \mathcal{P}_{v|W}(\,\cdot\,; B)$ is measurable.

2. For each $B \in \mathfrak{S}_v$ and $C \in \mathfrak{S}(W)$ we have
$$\mathbb{P}\Big(X_v^{-1}(B) \cap X_W^{-1}(C)\Big) \;=\; \int\limits_{X_W^{-1}(C)} \mathcal{P}_{v|W}(x; B)\ \mathbb{P}(dx).$$

Applying the factorization lemma, we get a function $\mathbb{P}_{v|W} : S(W) \times \mathfrak{S}_v \to [0,1]$ such that $\mathcal{P}_{v|W}(x; B) = \mathbb{P}_{v|W}(X_W(x), B)$ for all $x \in S(V)$ and all $B \in \mathfrak{S}_v$.
The Markov kernel $\mathbb{P}_{v|W} : S(W) \times \mathfrak{S}_v \to [0,1]$ is a desintegration kernel of the marginal distribution

$$\mathbb{P}_U \;:=\; \mathbb{P} \circ X_U^{-1} \qquad \text{for } U := W \cup \{v\}$$

with respect to the projection $X_{UW} : S(U) \to S(W)$; i.e. we have:

$$\mathbb{P}_U(C) \;=\; \int\limits_{S(W)} \int\limits_{S_v} 1_C(x_W, x_v)\ \mathbb{P}_{v|W}(x_W, dx_v)\ \mathbb{P}_W(dx_W)$$

where $x_W := (x_u : u \in W)$. For discrete state spaces $\mathbb{P}_{v|W}$ may be calculated easily via HUGIN (and API) and $\mathbb{P}_U$ may be calculated iteratively via HUGIN + API in combination with additional programs. Future versions of HUGIN will contain such algorithms.
For general state spaces $\mathbb{P}_U$ may be determined approximately via Monte Carlo simulation and $\mathbb{P}_{v|W}$ may be determined approximately via variance reduced Monte Carlo simulation.
Let us now describe a method to calculate $\mathbb{P}_{v|W}$ using HUGIN: For discrete state spaces we have $\mathbb{P}_{v|W}((x_w : w \in W); B) = \mathbb{P}(X_v \in B | X_w = x_w \text{ for } w \in W)$, and hence:

1. Select $(x_w : w \in W) \in S(W)$.

2. Open the windows of each of the nodes $w \in W$.

3. Introduce evidence "$X_w = x_w$" for all $w \in W$ in any order and propagate.

4. Open the window of node v: the distribution displayed at this window is $\mathbb{P}_{v|W}\big((x_w : w \in W); \cdot\big)$.

This procedure may be implemented as an automatic procedure using the HUGIN API. Let us now describe a method to calculate $\mathbb{P}_U$ iteratively using HUGIN+API and additional programs:

1. For $W := \{w\}$ the onedimensional marginal $\mathbb{P}_W$ is displayed at the window of node w.

2. For $W := \{w_1, w_2\}$ calculate the desintegration kernel $\mathbb{P}_{w_2,\{w_1\}}$ and take the marginal distribution $\mathbb{P}_{\{w_1\}}$ from the window of node w_1.
Obtain $\mathbb{P}_{\{w_1,w_2\}}$ from:

$$\mathbb{P}_{\{w_1,w_2\}}(C) = \int\limits_{S(\{w_1\})} \int\limits_{S_{w_2}} 1_C(x_{w_1}, x_{w_2})\ \mathbb{P}_{w_2,\{w_1\}}(x_{w_1}; dx_{w_2})\ \mathbb{P}_{\{w_1\}}(dx_{w_1})$$

(which is a finite sum only). (4.3)

$\vdots$

$k+1$. For $W := \{w_1, w_2, \ldots, w_{k+1}\}$ calculate the desintegration kernel $\mathbb{P}_{w_{k+1},\{w_1,\ldots,w_k\}}$ and take the marginal distribution $\mathbb{P}_{\{w_1,\ldots,w_k\}}$ from the preceeding calculation.
Obtain $\mathbb{P}_{\{w_1,\ldots,w_{k+1}\}}$ from $\mathbb{P}_{\{w_1,\ldots,w_k\}}$ and $\mathbb{P}_{w_{k+1},\{w_1,\ldots,w_k\}}$ analogue to equation 4.3.

Example: Calculation of a twodimensional joint distribution

We take the CPN "NEPHRO_RISK" designed for the prognosis of the risk of progression of diabetic nephropathy which has been introduced above. We shall show how to calculate the joint distribution of the following two variables:

RR_2 := (Riva–Rocci–) blood pressure two years ago
Risk := risk of progression of diabetic nephropathy

This calculation will be done in three steps:

1. Calculation of onedimensional marginal distribution of "RR_2".

2. Calculation of conditional probabilities of "Risk" given "RR_2".

3. Summation of the products of marginal and conditional probabilities.

a) Calculation of onedimensional marginal distribution of "RR_2"

We use HUGIN to calculate the probability distribution of the variable "RR_2". This distribution is the onedimensional marginal distribution of the multivariate distribution which is associated to the CPN "NEPHRO_RISK". The algorithm of HUGIN will calculate all the onedimensional marginal distributions of the associated multivariate distribution. To get to know it, we only have to open the window of "RR_2". This window will display this distribution immediately graphically. (The precise numbers of this distribution may be obtained by using the program HUGIN–API which may be accessed by the programming language ANSI C.) This is shown in Figure 1. The window of "Risk" displays the distribution of the variable "Risk". This distribution again is a onedimensional marginal of the multivariate distribution associated to the CPN "NEPHRO_RISK".

b) Calculation of conditional probabilities of "Risk" given "RR_2"

We use HUGIN to introduce each of the possible three states of the variable "RR_2" and to propagate this evidence. To do this, we open the window of "RR_2" and the window of "Risk". Then we set one of the states of the window of "RR_2"; this state will show the 100% bar. After propagation the window of "Risk" will display the conditional distribution of "Risk" given the introduced evidence in the window of "RR_2".

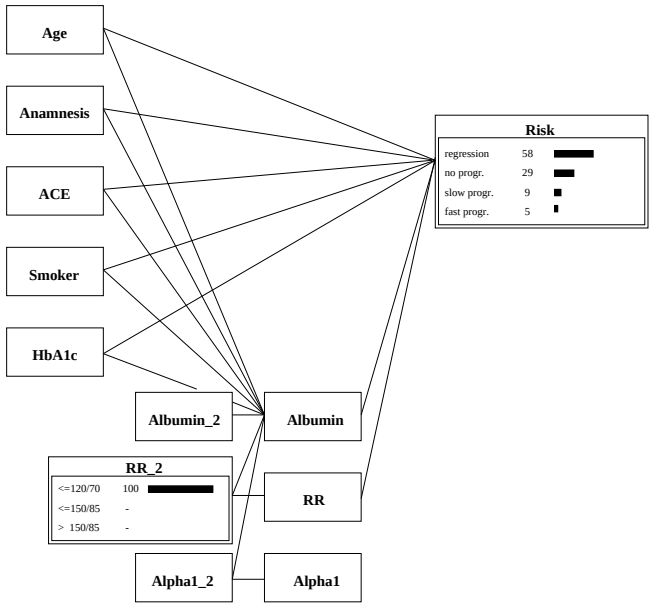

Figure 2: Introduce evidence "RR_2 $\leq$ 120/70" and propagate: Now the window of "Risk" displays the conditional distribution of "Risk" given "RR_2 $\leq$ 120/70".

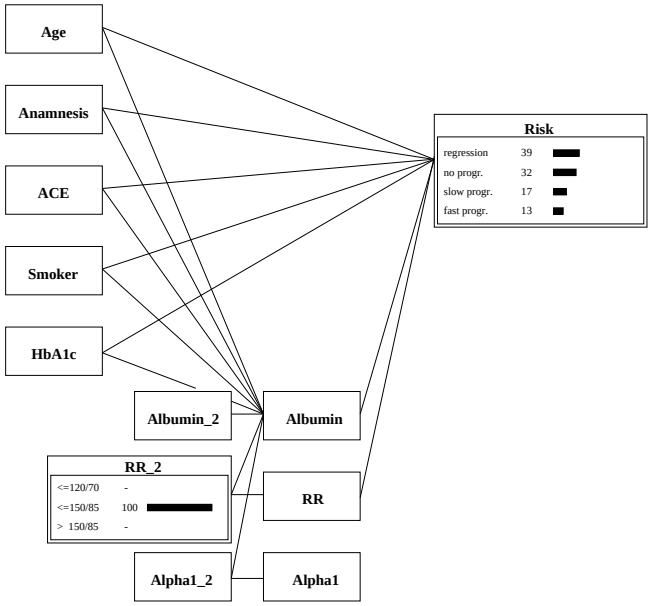

Figure 3: Introduce evidence "$120/70 < \text{RR\_2} \leq 150/85$" and propagate: Now the window of "Risk" displays the conditional distribution of "Risk" given "$120/70 < \text{RR\_2} \leq 150/85$".

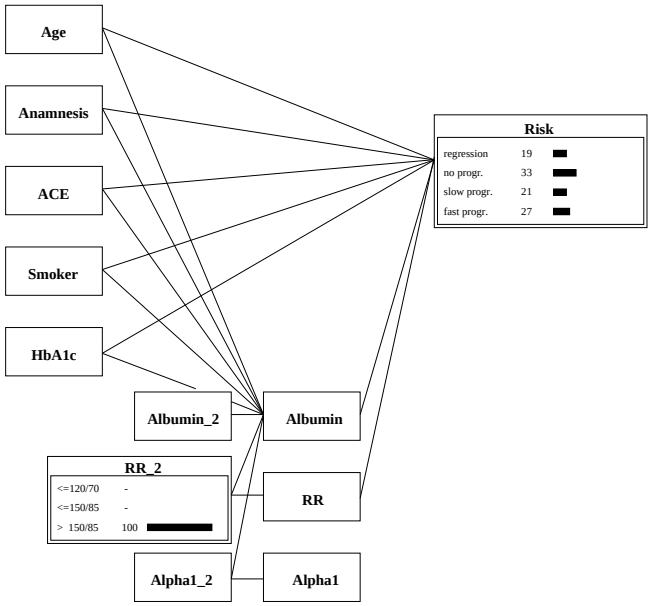

Figure 4: Introduce evidence "$150/85 < \text{RR\_2}$" and propagate: Now the window of "Risk" displays the conditional distribution of "Risk" given "$150/85 < \text{RR\_2}$".

c) Summation of the products of marginal and conditional probabilities

The summation of the products of the marginal and the conditional probabilities results in the ("calculated") twodimensional joint distribution of the variables "RR_2" and "Risk". We call it the exact twodimensional joint distribution of the variables "RR_2" and "Risk" obtained from the CPN "NEPHRO_RISK".

RR_2	Risk	probability
1	1	0.12251
2	1	0.20676
3	1	0.04734
1	2	0.06098
2	2	0.16915
3	2	0.08408
1	3	0.01876
2	3	0.08999
3	3	0.05337
1	4	0.00988
2	4	0.06945
3	4	0.06772

Table 6: The exact twodimensional joint distribution of the variables "RR_2" and "Risk" obtained from the CPN "NEPHRO_RISK".

5 Methods for evaluation of a CPN

Let us consider a CPN with an acyclic finite graph $G := (V, E)$ and a nonempty subset W of V. The associated multivariate distribution $\mathbb{P} : \mathfrak{S}(V) \to [0,1]$ of this CPN has the marginal distribution $\mathbb{P}_W := \mathbb{P} \circ X_W^{-1} : \mathfrak{S}(W) \to [0,1]$. Furthermore, let

$$\mathbb{P}^{\xi,k} \quad := \quad 1/k \sum_{m=1}^{k} \delta_{x_{(m)}} : \mathfrak{S}(V) \to [0,1]$$

be the empirical distribution associated to the sequence $\xi := (x^{(m)} : m \in \mathbb{N})$ obtained from independent or ergodic repetitions of Monte Carlo simulations of $\mathbb{P}$ or from data. From the Glivenko–Cantelli theoren we know that for independent or ergodic

repetitions of Monte Carlo simulations of $\mathbb{P}$ the sequence of empirical distributions converges weakly to $\mathbb{P}$ with probability one. If the data is collected properly, we may assume this convergence, too. (Unfortunately, very often the data is not collected properly. In many medical data pools we have strongly dependent repetitions; e.g. one person produces many data sets which are not recognizable as being generated by the same person. In such a case we should reduce the weight of the data corresponding to the multiplicity.) If the sequence of empirical distributions converges weakly to $\mathbb{P}$, then also the sequence $(\mathbb{P}_W^{\xi,k} : k \in \mathbb{N})$ of the marginals

$$\mathbb{P}_W^{\xi,k} \quad := \quad \mathbb{P}^{\xi,k} \circ X_W^{-1} : \mathfrak{S}(W) \to [0,1]$$

of $\mathbb{P}^{\xi,k}$ converges weakly to the marginal $\mathbb{P}_W$ of $\mathbb{P}$. We shall now discuss some methods for the evaluation of a CPN by evaluating instead of its associated multivariate distribution some of its marginal distributions. Because of very frequent lack of data we may only evaluate those marginal distributions for which enough data is available. We propose to apply symmetric measures of distance like the norm of total variation, L_p–norms or other metrics (which coincide for finite state spaces) or asymmetric measures of similarity or coincidence like the Kullback–Leibler divergence or some test for fitting (e.g. Kolmogorov, Kolmogorov–Smirnov, chi–square fitting).
We may compare exact marginal distributions with empirical distributions obtained from data or we may compare empirical distributions obtained from Monte Carlo simulations of the associated multivariate distribution with empirical distributions obtained from data.
Now, we shall demonstrate these procedures by applying them to our example, the CPN "NEPHRO_RISK". Again, we shall consider the two variables "RR_2" and "Risk" and their twodimensional joint distribution. We compare this distribution to the ones obtained by Monte Carlo simulation and from data. We also compare these with each other. The comparison is based on the L_1–norm, the L_2–norm, the Kullback–Leibler divergence and the chi–square fitting. The first variable has three states, the second variable has four states, and hence, the joint state space has 12 states. If we take the counting measure on this joint state space, the L_p–norm of the density f_μ of some measure μ on the joint state space with respect to this counting measure will be

$$\|\mu\|_p \quad := \quad \|f_\mu\|_p \quad := \quad \Big(\sum_{i=1}^{3} \sum_{j=1}^{4} \Big| \mu\Big(\{(i,j)\}\Big) \Big|^p \Big)^{1/p}.$$

The (symmetric) L_p–distance of two measures μ and ν is defined by $\|\mu - \nu\|_p$. If two measures μ and ν are positive for each singleton, the (asymmetric) Kullback–Leibler divergence from μ to ν is defined by

$$D_{KL}(\mu;\nu) \quad := \quad \sum_{i=1}^{3} \sum_{j=1}^{4} \nu\Big(\{(i,j)\}\Big) \cdot \log\Big(\nu\Big(\{(i,j)\}\Big) \;/\; \mu\Big(\{(i,j)\}\Big) \Big).$$

The (asymmetric) chi–square estimate for (the unsymmetric) testing of a measure μ against an empirical distribution ν belonging to a sample of size k is

$$D_{\chi^2}(\mu;\nu) \;:=\; k \cdot \sum_{i=1}^{3}\sum_{j=1}^{4} \Big(\nu\big(\{(i,j)\}\big) - \mu\big(\{(i,j)\}\big)\Big)^2 \;/\; \mu\big(\{(i,j)\}\big).$$

Table 8 shows some of the results of the comparison of one distribution against the other distribution. We write for short:

μ_1 := exact joint distribution,
μ_2 := empirical distribution obtained from 250 MC simulations,
μ_3 := empirical distribution obtained from 1.000 MC simulations,
μ_4 := empirical distribution obtained from 1.000.000 MC simulations,
μ_5 := empirical distribution obtained from 264 patients with repetitions,
μ_6 := empirical distribution obtained from 53 patients without repetitions.

Finally, we denote by "μ_i/μ_j" the "...distance from μ_i to μ_j" or "μ_i against μ_j".

RR_2	Risk	from 264 pairs	from 53 pairs
1	1	0.193	0.1698
2	1	0.197	0.2075
3	1	0.03	0.0377
1	2	0.133	0.1321
2	2	0.121	0.1509
3	2	0.015	0.0189
1	3	0.034	0.0189
2	3	0.057	0.0566
3	3	0.019	0.0189
1	4	0.027	0.0189
2	4	0.098	0.0943
3	4	0.076	0.0755

Table 7: The twodimensional empirical joint distributions of the variables "RR_2" and "Risk" obtained from 264 pairs of data of not necessarily different patients and from 53 pairs of data of different patients, respectively.

We had available a set of 264 pairs of data of "RR_2" and "Risk". This set was taken from the data pool of the clinic Krankenhaus München Bogenhausen. The empirical

distribution of this sample is presented in Table 7. A careful inspection of these pairs of data revealed that the 264 pairs of data came from 53 patients only. Hence, a large part of this sample is strongly dependent. At random we selected only one pair of data for each of the 53 patients. The empirical distribution of this sample is presented in Table 7.

	L_1–distance	L_2–distance	KL–divergence	chi–square	sample size k
μ_1/μ_2	0.1985	0.0763	0.0350	—	—
μ_1/μ_3	0.0821	0.0283	0.0051	—	—
μ_1/μ_4	0.0024	0.0008	-0.0001	—	—
μ_1/μ_5	0.4233	0.1460	0.1733	—	—
μ_1/μ_6	0.3218	0.1226	0.1296	—	—
μ_2/μ_1	0.1985	0.0763	0.0336	16.7444	250
μ_2/μ_3	0.1460	0.0548	0.0188	9.7702	250
μ_2/μ_4	0.1983	0.0763	0.0335	16.7016	250
μ_2/μ_5	0.522	0.1875	0.2066	128.9095	250
μ_2/μ_6	0.4234	0.1573	0.1603	101.7737	250
μ_5/μ_1	0.4233	0.1460	0.1460	77.0654	264
μ_5/μ_2	0.522	0.1875	0.1989	117.6430	264
μ_5/μ_6	0.104	0.0438	0.0127	7.3118	264
μ_6/μ_1	0.3218	0.1226	0.1078	11.0418	53
μ_6/μ_2	0.4234	0.1573	0.1518	17.6455	53
μ_6/μ_5	0.104	0.0438	0.0119	1.2198	53

Table 8: A comparison of symmetric and asymmetric distances between the twodimensional distributions obtained from the CPN "NEPHRO_RISK" and from clinical data pools.

From Table 8 we learn that the missing independence within the sample from the 264 patients causes severe problems: The hypotheses that the CPN "NEPHRO_RISK" is true will be rejected for reasonable levels of significance (e.g. $\alpha = 0.01$, 0.05 and 0.10). This is also suggested by the distance measures (L_1, L_2, KL). The opposite is true for the independent sample from the 53 patients.

Acknowledgement: This research was supported by the German Minister of Science and Technology (BMFT) and done as part of the project "DIADOQ: Diabetes Mellitus, optimierte Betreuung durch wissensbasierte Qualitätssicherung".

References

[1] Carnap, R.: Logical Foundations of Probability. The University of Chicago Press: 1950.

[2] Chavez, R.M.; Cooper, G.F.: An empirical evaluation of a randomized algorithm for probabilistic inference. In: Henrion, M.; Shachter, R.D.; Kanal, L.N.; Lemmer, J.F. (eds.): Uncertainty in Artificial Intelligence 5. Elsevier Science Publishers B.V. (North–Holland), 1990.

[3] Chin, H.L.; Cooper, G.F.: Bayesian belief network inference using simulation. In: Kanal, L.N.; Levitt, T.s.; Lemmer, J.F. (eds.): Uncertainty in Artificial Intelligence 3. Elsevier Science Publishers B.V. (North–Holland), 1989.

[4] Feller, W.: An Introduction to Probability Theory and its Applications. J. Wiley: New York, 1950. Volume I. J. Wiley and Sons: New York, 1960. Volume II. J. Wiley and Sons: London, 1966.

[5] Finetti, B. de: Probability: Sulle funzioni a incremento aleatorio. Rend. Acad. Lincei Cl. Sci. Fis. Mat. 10 (6), 1929.

[6] Finetti, B. de: Theory of Probability. Volume 1. J. Wiley and Sons: London–New York–Sydney–Toronto, 1974.

[7] Fung, R.; Chang, K.–C.: Weighing and integrating evidence for stochastic simulation in Bayesian networks. In: Henrion, M.; Shachter, R.D.; Kanal, L.N.; Lemmer, J.F. (eds.): Uncertainty in Artificial Intelligence 5. Elsevier Science Publishers B.V. (North–Holland), 1990.

[8] Hájek, P.; Havránek, T.; Jirouš̆ek,R.: Uncertain Information Processing in Expert Systems. CRC Press: Boca Raton–Ann Arbor–London–Tokyo, 1992.

[9] Henrion, M.: Propagating uncertainty in Bayesian networks by probabilistic logic sampling. In: Lemmer, J.F.; Kanal, L.N. (eds.): Uncertainty in Artificial Intelligence 2. Elsevier Science Publishers B.V. (North–Holland), 1988.

[10] Jensen, F.V.; Lauritzen, S.L.; Oleson, K.G.: Bayesian updating in causal probabilistic networks by local computations. Computational Statistics Quaterly 4 (1990), 269-282.

[11] Kolmogorov, A.: Grundbegriffe der Wahrscheinlichkeitsrechnung. Springer: Berlin, 1933. (Reprint: Springer: Berlin–Heidelberg–New York, 1973.)

[12] Kruse, R; Schwecke, E.; Heinsohn, J.: Uncertainty and Vagueness in Knowledge Based Systems: Numerical Methods. Springer: New York–Berlin–Heidelberg–London–Paris–Tokyo, 1991.

[13] Kruse, R; Gebhardt, J.; Klawonn, F.: Fuzzy Systeme. Teubner: Stuttgart, 1993.

[14] Lauritzen, S.L.; Dawid, A.P.; Larsen, B.N.; Leimer, H.–G.: Independence properties of directed Markov fields. Networks 20 (1990), 491-505.

[15] Lauritzen, S.L.; Spiegelhalter, D.: Local Computations with Probabilities on Graphical Structures and Their Application to Expert Systems. J. Roy. Stat. Soc. B, 50 (2) (1988), 157-224.

[16] Liesenfeld, B.; Hierle, A.: A new clinical decision support tool for differential diagnosis, likelihood of development and progression of diabetic nephropathy in insulin–dependent diabetics. Proceedings of the International Conference on Neural Networks and Expert Systems in Medicine and Healthcare, Plymouth, August 24-26, 1994.

[17] Matthes, R.; Oppel, U.G.: Convergence of causal probabilistic networks. In: Bouchon–Meunier, B., Valverde, L.; Yager, R.R.(eds.): Intelligent Systems with Uncertainty. North–Holland: Amsterdam–London–New York–Tokyo, 1993.

[18] Neapolitan, R.E.: Probabilistic Reasoning in Expert Systems. Theory and Algorithms. J. Wiley and Sons: New York–Chichester–Brisbane–Toronto–Singapore, 1990.

[19] Neveu, J.: Mathematische Grundlagen der Wahrscheinlichkeitstheorie. Oldenburg: M"unchen, 1969.

[20] Oppel, U.G.: Every Complex System Can be Determined by a Causal Probabilistic Network without Cycles and Every Such Network Determines a Markov Field. In:Kruse, R.; Siegel, P. (eds.): Symbolic and Quantitative Approaches to Uncertainty. Lecture Notes of Computer Science 548. Springer: Berlin–Heidelberg–New York, 1991.

[21] Oppel, U.G.; Hierle, A.; Janke, L.; Moser, W.: Transformation of Compartmental Models into Sequences of Causal Probabilistic Networks. In: Andreassen, S.; Engelbrecht, R.; Wyatt, J. (eds.): Artificial Intelligence in Medicine. IOS Press: Amsterdam–Oxford–Washington–Tokyo, 1993.

[22] Oppel, U.G.: Causal probabilistic networks and their application to metabolic processes. To appear in: Mammitzsch, V.; Schneeweiß, H.: Proceedings of the Second Gauss Symposium, Munich, August 1993.

[23] Oppel, U.G.: Series of causal probabilistic networks induced by systems of differential equations for diagnosis and prognosis of metabolic processes. Proceedings of the Fifth International Conference on Information Processing and Management of Uncertainty in Knowledge–Based Systems (IPMU), July 4-8, 1994, Paris, France.

[24] Parthasarathy, K.R.: Probability Measures on Metric Spaces. Academic Press, New York–London, 1967.

[25] Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann: San Matteo, 1988.

[26] Richter, H.: Zur Grundlegung der Wahrscheinlichkeitstheorie. Math. Ann. 125, 129-139, 223-234, 335-343 (1953); 126, 362-374 (1953); 128, 305-339 (1954).

[27] Richter, H.: Wahrscheinlichkeitstheorie. Springer: Berlin–Heidelberg, 1956.

[28] Salzsieder, E.; Fischer, U.; Hierle, A.; Oppel, U.G.: A Causal Probabilistic Network Associated to the Karlsburg Model of the Glucose–Insulin Metabolism for Approximate Assessment of Parameters, Diagnosis and Prognosis. Proceedings of the International Conference on Neural Networks and Expert Systems in Medicine and Healthcare, Plymouth, August 24-26, 1994.

[29] Savage, L.R.: The Foundations of Statistics. J. Wiley and Sons: New York, 1954. Dover Publications: New York, 1972.

[30] Shachter, R.D.: A linear approximation method for probabilistic inference. In: Shachter, R.D.; Levitt, T.S.; Kanal, L.N.; Lemmer, J.F. (eds): Uncertainty in Artificial Intelligence. Elsevier Science Publishers B.V. (North–Holland), 1990.

[31] Shachter, R.D.; Peot, M.A.: Simulation approaches to general probabilistic inference on belief networks. In: Henrion, M.; Shachter, R.D.; Kanal, L.N.; Lemmer, J.F. (eds.): Uncertainty in Artificial Intelligence 5. Elsevier Science Publishers B.V. (North–Holland), 1990.

[32] Stegmüller, W.: Personelle und statistische Wahrscheinlichkeit. I and II. Springer: Berlin–Heidelberg–New York, 1973.