International Journal of Approximate Reasoning 100 (2018) 115–134

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Scalable importance sampling estimation of Gaussian mixture posteriors in Bayesian networks

Darío Ramos-López [a,*], Andrés R. Masegosa [a], Antonio Salmerón [a], Rafael Rumí [a], Helge Langseth [c], Thomas D. Nielsen [b,a], Anders L. Madsen [d,b]

[a] *Department of Mathematics, University of Almería, Spain*
[b] *Department of Computer Science, Aalborg University, Denmark*
[c] *Department of Computer and Information Science, The Norwegian University of Science and Technology, Norway*
[d] *HUGIN EXPERT A/S, Aalborg, Denmark*

ARTICLE INFO

Article history:
Received 11 January 2018
Received in revised form 8 June 2018
Accepted 9 June 2018
Available online 13 June 2018

Keywords:
Importance sampling
Bayesian networks
Conditional linear Gaussian models
Scalable inference
Gaussian mixtures

ABSTRACT

In this paper we propose a scalable importance sampling algorithm for computing Gaussian mixture posteriors in conditional linear Gaussian Bayesian networks. Our contribution is based on using a stochastic gradient ascent procedure taking as input a stream of importance sampling weights, so that a mixture of Gaussians is dynamically updated with no need to store the full sample. The algorithm has been designed following a Map/Reduce approach and is therefore scalable with respect to computing resources. The implementation of the proposed algorithm is available as part of the AMIDST open-source toolbox for scalable probabilistic machine learning (http://www.amidsttoolbox.com).

1. Introduction

Bayesian networks (BNs) [11,27] provide a well-founded and principled approach for Bayesian reasoning in complex domains endowed with uncertainty. A prominent feature of BNs as a framework for representing uncertain knowledge is the possibility for defining efficient algorithms for performing probabilistic inference (e.g. computation of the posterior distribution of a target variable). These algorithms are typically designed to take advantage of the independence properties implied by the structure of the Bayesian network [10,17,18,36].

Even though most of the methodological development around Bayesian networks has focused on discrete variables, there are plenty of problems in which discrete and continuous variables coexist. A BN is called *hybrid* if it contains discrete and continuous variables simultaneously. The most established approach for explicitly handling hybrid BNs came with the definition of the conditional linear Gaussian (CLG) model [16]. This model class is based on the assumption of normality over the continuous variables, and the structural restriction that prevents discrete variables from having continuous parents. If these modeling assumptions are not congruent with the domain being modeled, one may instead opt to discretize the continuous variables [12,25,26], thus transforming the hybrid BN to a standard discrete BN. Unfortunately, such transformations typically also result in a loss of information.

Mixtures of truncated basis functions [14] provide a generalization of standard discretization and do not impose any structural restriction on the model nor do they make distributional assumptions like the normality assumption imposed by

[*] Corresponding author.
E-mail address: dramoslopez@ual.es (D. Ramos-López).

https://doi.org/10.1016/j.ijar.2018.06.004
0888-613X/© 2018 Elsevier Inc. All rights reserved.

CLG models. Furthermore, they are compatible with exact probabilistic inference algorithms as, for instance, the Shenoy–Shafer architecture [38] and the variable elimination scheme [43]. However, the complexity of probabilistic inference in these models often renders them inappropriate when dealing with a large number of variables and a limited response time [37].

2. Motivation and contribution

In this paper, we are interested in approximate probabilistic inference methods for hybrid BNs satisfying the following two requirements: (i) they should be able to scale with the computational resources available in order to provide results after a short computing time; and (ii) the provided output should be an explicit probability density, rather than just a set of quantiles or moments of the distribution. A widely applicable scenario where both requirements are needed comes up when processing data streams at high speed, and for each item in the stream, we need to know the result of processing the item through probabilistic inference in a hybrid BN. The result of the inference process should be quickly available (before the next data item arrives). At the same time, the availability of the explicit form of the posterior density facilitates subsequent analysis. For instance, as a basis for expected utility calculations or as a tool for anomaly detection from the input data stream.

For the former reason, we focus our analysis on CLG models, instead of less restrictive alternatives such as mixtures of truncated basis functions, as inference in the latter models is in general more time consuming [33]. Another advantage of CLG models is that it is known that the posterior distribution on a continuous variable in a CLG network is always a mixture of Gaussians [15].

Many approaches can be used to perform approximate inference in CLG models. Deterministic approximations include (mean field) variational methods [40] and expectation propagation [22]. The problem is that these methods are iterative in nature and, as a consequence, difficult to parallelize and scale up. Scalable alternatives exist [9,20], but they are not designed for general BNs, but for restricted plate models oriented to learning problems. Furthermore, the approximations provided by these models are often expressed by a single Gaussian distribution in order to make the methods computationally efficient, but, as we will show in the experimental section, this approximation is usually not sufficiently accurate.

Monte-Carlo methods define another widely used class of approximate inference approaches which could be used in this setting. An important group of them are based on the importance sampling technique, that provides a flexible approach for constructing *anytime* probabilistic reasoning algorithms [4,23,41,42], where the term *anytime* means that the accuracy of the results provided by an algorithm is proportional to the time it is allowed to run [29]. The advantage of importance sampling methods is that they are embarrassingly parallelizable, as shown in [34]. However, the plain application of importance sampling yields an empirical distribution that approximates the posterior, rather than an explicit density (a mixture of Gaussians, for instance).

In this paper we extend the method in [34] enabling it to compute mixture of Gaussians posterior densities. Our contribution is based on using a stochastic gradient ascent procedure taking as input a stream of importance sampling weights, so that a mixture of Gaussians is dynamically updated with no need to store the full sample. The algorithm has been designed following a Map/Reduce approach and is therefore scalable with respect to computing resources. The implementation of the algorithm is available as part of the AMIDST open-source toolbox for scalable probabilistic machine learning (http://www.amidsttoolbox.com) [21].

3. Preliminaries

Consider a set of N random variables $\mathbf{X} = \{X_1, \ldots, X_N\}$. A BN over $\mathbf{X}$ is composed of a directed acyclic graph, where each node represents a variable in $\mathbf{X}$, and a set of conditional probability distributions such that the joint distribution over $\mathbf{X}$ factorizes as

$$p(\mathbf{X}) = \prod_{i=1}^{N} p_i(X_i|\text{pa}(X_i)), \tag{1}$$

where $\text{pa}(X_i)$ denotes the set of parents of X_i in the graph representation.

We will use lowercase letters to refer to values or configurations of values, so that x denotes a value of X and boldface $\mathbf{x}$ is a configuration of the variables in $\mathbf{X}$. Given a set of observed variables $\mathbf{X}_E \subset \mathbf{X}$ and a set of variables of interest $\mathbf{X}_I \subset \mathbf{X} \setminus \mathbf{X}_E$, *probabilistic inference*, also called *belief update*, consists of computing the posterior distribution $p(x_i|\mathbf{x}_E)$ for each $i \in I$; here we allow X_i to be either discrete or continuous.[1]

If we denote by $\mathbf{X}_C$ and $\mathbf{X}_D$ the set of continuous and discrete variables not in $\{X_i\} \cup \mathbf{X}_E$, and by $\mathbf{X}_{C_i}$ and $\mathbf{X}_{D_i}$ the set of continuous and discrete variables not in $\mathbf{X}_E$, the goal of probabilistic inference can generally be formulated as computing

[1] In this paper we only consider inference w.r.t. the posterior marginal distribution of a variable and not joint distributions over several variables.