

# Exact Inference on Conditional Linear $\Gamma$ -Gaussian Bayesian Networks

Ivar Simonsson

Petter Mostad

*Department of Mathematical Sciences*

*Chalmers University of Technology and University of Gothenburg*

*Gothenburg*

SIMONSSI@CHALMERS.SE

MOSTAD@CHALMERS.SE

## Abstract

Exact inference for Bayesian Networks is only possible for quite limited classes of networks. Examples of such classes are discrete networks, conditional linear Gaussian networks, networks using mixtures of truncated exponentials, and networks with densities expressed as truncated polynomials. This paper defines another class with exact inference, based on the normal inverse gamma conjugacy. We describe the theory of this class as well as exemplify our implemented inference algorithm in a practical example. Although generally small and simple, we believe these kinds of networks are potentially quite useful, on their own or in combination with other algorithms and methods for Bayesian Network inference.

**Keywords:** Hybrid Bayesian networks; inference in hybrid Bayesian networks; variable elimination algorithm.

## 1. Introduction

A Bayesian network (BN) is a directed acyclic graph whose nodes are associated with a set of random variables  $\mathcal{X}$  and where the graph encodes independencies. A central task is to compute from such networks the marginal distribution of some variables given fixed values for some other variables. General results can be expressed in analytical formulas, i.e., be done exact, only for some classes of BNs, most notably discrete BNs, Gaussian BNs, and conditional linear Gaussian BNs, see e.g. (Lauritzen and Jensen, 2001). There are additional examples of classes of BNs with exact inference such as BNs where variables are mixtures of truncated exponentials (Moral et al., 2001) or densities are expressed with truncated polynomials (Shenoy and West, 2011). The main usage of BNs from these two last classes may be as approximations of other BNs.

In this paper, we instead study BNs where conjugacies between variables are a natural part of the BN. In particular, we focus on using the conjugacy between the precision of the normal distribution and the gamma distribution. This will enable us to define  $\Gamma$ -Gaussian networks, and also conditional linear  $\Gamma$ -Gaussian networks. The family of factors we define on these networks is closed under all relevant factor operations except for one type of marginalization. However, this marginalization also yields an analytic closed form, and can be delayed until the last step of the algorithm, hence exact results can always be obtained. Using this, we describe how the variable elimination algorithm (sometimes simply called *the algorithm* in this paper) can be implemented on  $\Gamma$ -Gaussian networks, basing our presentation on the terminology and notation of (Koller and Friedman, 2009).

In Section 2 we review the variable elimination algorithm in a general setting by describing the local operations needed and the algorithm itself. We also mention how the factors of Gaussian BNs are parameterized as canonical forms. In Section 3 we describe how one can modify this parameterization to work with BNs in which we include a gamma distributed variable that models

the precision of the normal variables. Based on this we define  $\Gamma$ -Gaussian BNs, and also conditional linear  $\Gamma$ -Gaussian BNs in which we allow finite-valued nodes. We describe how the local operations act on the new factors. Unfortunately, our new class of factors is not completely closed under the local operations, and we thus adjust implementation of inference accordingly.

In Section 4 we present a concrete example of a  $\Gamma$ -Gaussian BN and show how it may be used as a model in an application from forensic statistics. Based on data from such an application, we present results from our implemented algorithm applied to the example.

Finally, in Section 5, we discuss the usefulness of  $\Gamma$ -Gaussian BNs. We also discuss the possibilities for further classes of BNs on which exact inference is possible, as well as the possibilities for extending algorithms for such classes of BNs to more general BNs by combining them with other algorithms and methods.

## 2. Review of the Variable Elimination Algorithm

The variable elimination algorithm is conveniently formulated in terms of a series of operations on smaller components of the network, called *factors*. Formally, a factor over a set of random variables is simply defined as a real valued function of the range of the variables. In general we will use calligraphic letters, (e.g.  $\mathcal{X}$ ,  $\mathcal{Y}$ ,  $\mathcal{Z}$ ) to denote sets of random variables and boldface letters (e.g.  $\mathbf{X}$ ,  $\mathbf{Y}$ ,  $\mathbf{Z}$ ) to denote the associated random vectors. We use  $Val(X)$  to denote the set of possible values a random variable  $X$  can attain and for a random vector  $\mathbf{X} = (X_1, \dots, X_n)$  we write  $Val(\mathbf{X}) = (Val(X_1), \dots, Val(X_n))$ .

**Definition 1** A factor  $\phi$  over a set of random variables  $\mathcal{X} = \{X_1, \dots, X_n\}$  is a function from  $Val(\mathbf{X})$  to  $\mathbb{R}$ . The set  $\mathcal{X}$  is called the scope of  $\phi$ .

These factors are initially created directly from the conditional distribution of each node given its parents. If  $Val(X)$  is finite for each  $X$  in the scope of the factor, then the factor is called *finite* and will simply be a finite collection of numbers, each representing the value of the factor at a particular configuration of its scope.

### 2.1 Local Operations

After constructing the factors, the first step is to update the network in order to incorporate the evidence, and this is done locally on each factor. The procedure for this is called *factor reduction* and is formally defined as follows.

**Definition 2 (Factor reduction)** Let  $\phi$  be a factor over a set of variables  $\mathcal{X}$  and assume that we have evidence on some other set of variables  $\mathcal{Y} = \{Y_1, \dots, Y_m\} \subseteq \mathcal{X}$ , so that  $\mathbf{Y} = \mathbf{y}$  for some  $\mathbf{y} \in Val(\mathbf{Y})$ . The factor reduction of  $\phi$  with respect to the evidence  $\mathbf{Y} = \mathbf{y}$  is a new factor  $\phi'$  over  $\mathcal{Z} = \mathcal{X} \setminus \mathcal{Y}$  such that  $\phi'(\mathbf{z}) = \phi(\mathbf{z}, \mathbf{y})$  for each  $\mathbf{z} \in Val(\mathbf{Z})$ .

After performing factor reduction on each affected factor, it is time to perform *factor multiplication*.

**Definition 3 (Factor multiplication)** The factor multiplication of two factors,  $\phi_1$  over  $\mathcal{X}$  and  $\phi_2$  over  $\mathcal{Y}$ , is a new factor  $\phi$  over  $\mathcal{Z} = \mathcal{X} \cup \mathcal{Y}$  such that  $\phi(\mathbf{z}) = \phi_1(\mathbf{x})\phi_2(\mathbf{y})$  for each  $\mathbf{z} \in Val(\mathbf{Z})$ , where  $\mathbf{x}$  and  $\mathbf{y}$  are the restrictions of  $\mathbf{z}$  to  $\mathcal{X}$  and  $\mathcal{Y}$ , respectively.

Note that we include what can be called *factor extension* into factor multiplication.

The last operation is *factor marginalization*.

**Definition 4 (Factor marginalization)** Let  $\phi$  be a factor over  $\mathcal{Z}$  and  $\{\mathcal{X}, \mathcal{Y}\}$  be a partition of  $\mathcal{Z}$ . The factor marginalization of  $\phi$  with respect to  $\mathcal{Y}$  is another factor  $\phi'$  over  $\mathcal{X}$ , such that

$$\phi'(\mathbf{X}) = \int_{Val(\mathbf{Y})} \phi(\mathbf{X}, \mathbf{Y}) d\mathbf{Y}. \quad (1)$$

When performing factor marginalization, we will say that we marginalize out  $\mathbf{Y}$  from  $\phi$ .

Note that if  $\mathbf{Y}$  is discrete, summation is performed in (1), not integration.

## 2.2 The Algorithm

As input to the algorithm, we need a Bayesian network  $\mathcal{B}$ , defined by all the conditional probability distributions (CPDs). We also need the set of query variables,  $\mathcal{X}$  which we are ultimately interested in, and an optional set of evidence variables  $\mathcal{Y}$ , with associated evidence  $\mathbf{y}$ . Given this input, the algorithm looks like this:

1. Construct factors  $\Phi = \{\phi_1, \dots, \phi_n\}$  from the CPDs of  $\mathcal{B}$ .
2. Partition the set of variables into three sets, the set of query variables  $\mathcal{X}$ , the set of evidence variables  $\mathcal{Y}$ , and the set of elimination variables  $\mathcal{Z}$ .
3. For each factor, if there is evidence in its scope, reduce the factor with respect to this evidence by performing factor reduction.
4. Choose an ordering  $Z_1, \dots, Z_N$  of the variables in  $\mathcal{Z}$ , in which the elimination should be done. (This important step is not focused on in this paper).
5. For each variable  $Z_i \in \mathcal{Z}$ , according to the ordering determined in Step 4:
  - (a) Identify all the factors  $\phi_1, \dots, \phi_k$  whose scope includes  $Z_i$ .
  - (b) Compute the product  $\psi = \prod_{j=1}^k \phi_j$  by performing factor multiplication.
  - (c) Compute the integral (or the sum)  $\tau = \int_{Val(Z_i)} \psi$  by performing factor marginalization.
  - (d) Update  $\Phi$  by replacing  $\phi_1, \dots, \phi_k$  with  $\tau$ .
6. Normalize the resulting factor(s).

When we are done with all these steps, we will have obtained the conditional distribution of  $\mathbf{X}$  given  $\mathbf{Y} = \mathbf{y}$ . Step 5 is the variable elimination step. A more detailed description of the algorithm can be found in (Koller and Friedman, 2009) where it is described in pseudocode.

If all the variables in the network are finite, factors can be represented as vectors of real numbers, and the algorithm boils down to taking sums and products. In order to extend the algorithm to incorporate variables with infinite range, i.e. continuous random variables, we need a unified way to parameterize the corresponding factors. Moreover, we need this parameterization to be closed under the local operations. Thus we can not expect the algorithm to be useful on all types of Bayesian networks and we have to restrict ourselves to subclasses of Bayesian networks in a suitable way.

### 2.3 Gaussian Bayesian Networks

The algorithm above can be implemented relatively smoothly on the class of *Gaussian* Bayesian networks. A Gaussian Bayesian network (GBN) is a network in which the distribution of each variable is Gaussian with fixed variance and a mean that is an affine combination of its parents. This is done by Cowell et al. (1999) and Koller and Friedman (2009), in which the factors defined are called *canonical forms* and are defined as follows.

**Definition 5** Let  $\mathbf{X}$  be a random vector and let  $\phi$  be a factor over  $\mathcal{X}$ . We say that  $\phi$  is a canonical form, denoted by  $\mathcal{C}(\mathbf{X}; K, \mathbf{h}, g)$  (or simply  $\mathcal{C}(K, \mathbf{h}, g)$ ), if it can be written as

$$\phi(\mathbf{X}) = \mathcal{C}(\mathbf{X}; K, \mathbf{h}, g) = \exp\left(-\frac{1}{2}\mathbf{X}^T K \mathbf{X} + \mathbf{h}^T \mathbf{X} + g\right) \quad (2)$$

where  $K \in \mathbb{R}^{n \times n}$  is symmetric,  $\mathbf{h} \in \mathbb{R}^n$  and  $g \in \mathbb{R}$ .

Explicit formulas for the local operations on canonical forms are presented in (Koller and Friedman, 2009). Note that when performing the local operations, the form on the resulting factor is completely determined by Definitions 2-4 and the input factor.

As hinted earlier, the reason that the algorithm is well suited for GBNs is that the family of canonical forms is closed under the local operations, meaning that as long as the input factors are canonical forms, the output factors will also be canonical forms.

### 2.4 Including Both Finite and Gaussian Variables

In (Koller and Friedman, 2009), it is described how the algorithm can be used on Bayesian networks that includes both Gaussian variables and finite variables. These are called *conditional linear Gaussian* networks, and has the restriction that no finite nodes can have continuous parents. For each fixed configuration of its finite parents, a continuous node must then be normally distributed with fixed variance and a mean that is an affine combination of its continuous parents, much like the case for GBNs. The factors used in these cases can be described as *canonical tables*, the entries of which are canonical forms over the normal variables.

Koller and Friedman (2009) also discuss the computational problems that occurs when doing exact inference on conditional linear Gaussian networks, and they present an approximate inference alternative using *weak marginalization*. However, in many applications the networks are fairly small and exact inference could be used nonetheless. Therefore, we choose to implement the exact version, which requires yet another extensions of the factors. Since exact marginalization over finite variables of a canonical table results in a weighted sum of canonical forms, we must allow for these as well.

## 3. Extension to $\Gamma$ -Gaussian Bayesian Networks

One way of explaining why the algorithm works on GBNs is that the normal density is the conjugate prior for the mean of the normal density, when the variance is known. One could also look at priors for the variance. Recall the normal-gamma distribution, which is the joint distribution for  $(X, Z)$  when  $X|Z \sim \mathcal{N}(\mu, (\lambda Z)^{-1})$  and  $Z \sim \Gamma(\text{shape} = \alpha_0, \text{rate} = \beta_0)$ , see for example (DeGroot, 2005). It is known that the normal-gamma distribution is the conjugate prior for the normal density when both the mean and the variance are unknown.

Now assume that  $X|Y, Z$  has a normal distribution with expectation  $\alpha + \beta Y$  and variance  $\sigma^2/Z$  for some fixed parameters  $\alpha, \beta$ , and  $\sigma^2$ . A factor  $\phi$  corresponding to the density  $\pi(X | Y, Z)$  can be written as

$$\phi(x, y, z) = \exp\left(-\frac{z}{2\sigma^2}(x - (\alpha + \beta y))^2 + \frac{1}{2}\log(z) - \frac{1}{2}\log(2\pi\sigma^2)\right), \quad (3)$$

which can be rewritten as

$$\phi(x, y, z) = \exp\left(z\left(-\frac{1}{2}\mathbf{x}^T K \mathbf{x} + \mathbf{h}^T \mathbf{x} + g\right) + a \log(z) + b\right), \quad (4)$$

where  $\mathbf{x} = (x, y)$  and

$$\begin{cases} K &= \frac{1}{\sigma^2}\boldsymbol{\beta}\boldsymbol{\beta}^T \\ \mathbf{h} &= -\frac{\alpha}{\sigma^2}\boldsymbol{\beta} \\ g &= -\frac{\alpha^2}{2\sigma^2} \\ a &= \frac{1}{2} \\ b &= -\frac{1}{2}\log(2\pi\sigma^2) \end{cases} \quad (5)$$

where  $\boldsymbol{\beta} = (-1, \beta)$ . Since the normal-gamma distribution is the conjugate prior for the normal distribution, we now know that if we multiply the factor  $\phi(x, y, z)$  with another factor  $\phi'$  representing the normal-gamma distribution for  $(Y, Z)$ , then the resulting factor can be written in the same form as  $\phi$  above. However, if  $\phi'$  does not include the gamma variable  $Z$ , then we will not be able to use the conjugacy described earlier. This is reflected in the factor parameterization (4) in that  $z$  has a very specific role. One obvious way of dealing with this is to insist on that *every* Gaussian variable in the network must have variance that is inversely proportional to the same gamma variable. For the small network over  $(X, Y, Z)$  explained above, this means that  $Y$  *must* have  $Z$  as a parent and that the variance of  $Y$  *must* be proportional to  $1/Z$ .

In order to generalize these thoughts, we want to define a family of factors from (4), much like what was done with canonical forms. Our aim should be that the local operations should be easily identified and performed on these factors and that the family is closed under these operations. However, it turns out that this is a little bit too much to hope for. Nevertheless, there are ways of using the algorithm of Section 2.2 to perform inference on these networks. Before trying this, we need to restrict the networks to those we can operate.

**Definition 6** *Let  $\mathcal{B}$  be a BN over a set of random variables  $\mathcal{X} \cup \{Z\}$ . We say that  $\mathcal{B}$  is  $\Gamma$ -Gaussian (gamma Gaussian) if  $Z \sim \Gamma(\alpha_0, \beta_0)$  for some positive constants  $\alpha_0$  and  $\beta_0$ , and if for each  $X \in \mathcal{X}$ , we have that  $Z \in \text{Pa}_X$  and*

$$X|\text{Pa}_X \sim \mathcal{N}\left(\alpha + \sum_{i=1}^m \beta_i Y_i, \frac{\sigma^2}{Z}\right) \quad (6)$$

for real valued constants  $\alpha, \beta_1, \dots, \beta_m$  and  $\sigma^2$ . The variables  $Y_1, \dots, Y_m$  are together with  $Z$  the parents of  $X$ .

Here, and henceforth, the notation  $\text{Pa}_X$  is used to denote the set of parents of  $X$ . We will call the variables in  $\mathcal{X}$  *normal* and we will call  $Z$  the *gamma variable*. Note that a  $\Gamma$ -Gaussian BN contains exactly one gamma variable.

**Definition 7** We say that a factor  $\phi$  over a set of variables  $\mathcal{X} \cup \{Z\}$  is a  $\Gamma$ -canonical form if it can be written as

$$\phi(\mathbf{X}, Z) = \exp \left( Z \left( -\frac{1}{2} \mathbf{X}^T K \mathbf{X} + \mathbf{h}^T \mathbf{X} + g \right) + a \log(Z) + b \right) \quad (7)$$

where  $K \in \mathbb{R}^{n \times n}$  is symmetric,  $\mathbf{h} \in \mathbb{R}^n$  and  $g, a, b \in \mathbb{R}$ .

When  $K = 0$  and  $\mathbf{h} = 0$  we get the following special case:

**Definition 8** We say that a factor,  $\phi$ , over a single random variable,  $Z$ , is a gamma factor if it can be written as

$$\phi(Z) = \exp (Zg + a \log(Z) + b). \quad (8)$$

In order to create the initial factors of a  $\Gamma$ -Gaussian BN, we need the following result.

**Proposition 9** Let  $X$  be a normal random variable in a  $\Gamma$ -Gaussian BN and let  $Y_1, \dots, Y_m, Z$  be its parents, i.e.  $X | \text{Pa}_X$  is given by (6). Then the factor  $\phi(X, Y_1, \dots, Y_m, Z) = \pi(X | Y_1, \dots, Y_m, Z)$  is a  $\Gamma$ -canonical form with parameters given by (5) with  $\beta = (-1, \beta_1, \dots, \beta_m)$ .

**Proof** The conditional distribution  $X | \text{Pa}_X$  is normal with mean  $\alpha + \sum_{i=1}^m \beta_i Y_i$  and variance  $\sigma^2 / Z$ , hence we have that

$$\pi(X | Y_1, \dots, Y_m, Z) = \exp \left( -\frac{Z}{2\sigma^2} \left( X - \alpha - \sum_{i=1}^m \beta_i Y_i \right)^2 + \frac{1}{2} \log(Z) - \frac{1}{2} \log(2\pi\sigma^2) \right). \quad (9)$$

Letting  $\mathbf{X} = (X, Y_1, \dots, Y_m)$ , we want to compare this expression with (7). The parameters  $a$  and  $b$  of (7) are directly identifiable and the formulas for  $K$ ,  $\mathbf{h}$  and  $g$  follows by rewriting the quadratic term of (9) on matrix form.  $\blacksquare$

The results shows how the initial factors corresponding to normal variables in a  $\Gamma$ -Gaussian network can be written in  $\Gamma$ -canonical form. Additionally, one can easily check that the factor corresponding to the gamma distributed node  $Z \sim \Gamma(\alpha_0, \beta_0)$  can be written as a gamma factor using parameters  $g = -\beta_0$ ,  $a = \alpha_0 - 1$  and  $b = \alpha_0 \log(\beta_0) - \ell\Gamma(\alpha_0)$ , where  $\ell\Gamma(\cdot) = \log(\Gamma(\cdot))$ .

We continue by describing how the local operations act on  $\Gamma$ -canonical forms.

### 3.1 Factor Reduction

The factor reduction is done differently depending on whether we have evidence on the gamma node or on a normal node. In the former case we have the following result.

**Proposition 10** Let  $\phi$  be a  $\Gamma$ -canonical form with scope  $\mathcal{X} \cup \{Z\}$  and parameters  $K$ ,  $\mathbf{h}$ ,  $g$ ,  $a$  and  $b$ . Reducing  $\phi$  with respect to the evidence  $Z = z$  will result in a canonical form  $\phi'$  over  $\mathcal{X}$  with parameters  $K' = zK$ ,  $\mathbf{h}' = z\mathbf{h}$  and  $g' = zg + a \log(z) + b$ .

**Proof** Definition 2 tells us that the new factor  $\phi'$  over  $\mathcal{X}$  must satisfy  $\phi'(\mathbf{x}) = \phi(\mathbf{x}, z)$  for all  $\mathbf{x} \in \text{Val}(\mathbf{X})$ . But for  $\mathbf{x} \in \text{Val}(\mathbf{X})$ , we have that

$$\begin{aligned} \phi(\mathbf{x}, z) &= \exp \left( z \left( -\frac{1}{2} \mathbf{x}^T K \mathbf{x} + \mathbf{h}^T \mathbf{x} + g \right) + a \log(z) + b \right) \\ &= \exp \left( -\frac{1}{2} \mathbf{x}^T (zK) \mathbf{x} + (z\mathbf{h})^T \mathbf{x} + zg + a \log(z) + b \right), \end{aligned}$$

which proves the result. ■

Proposition 10 tells us that whenever we have evidence on the gamma variable, we can reduce the whole network to a purely Gaussian BN since all variances will be fixed. All resulting factors will therefore be canonical forms and we can simply use the variable elimination algorithm on these.

We now turn to the case where reduction is done with respect to a normal variable. If the reduction is done with respect to the *only* normal variable, we have the following result, which is as straightforward to confirm as Proposition 10.

**Proposition 11** *Let  $\phi$  be a  $\Gamma$ -canonical form with scope  $\{Y, Z\}$  and parameters  $K, h, g, a$  and  $b$ . Reducing  $\phi$  with respect to the evidence  $Y = y$  will result in a gamma factor,  $\phi'$ , over  $Z$  with parameters  $g' = -\frac{1}{2}Ky^2 + hy + g, a' = a$  and  $b' = b$ .*

On the other hand, if there are still normal variables left in the scope of the factor, we have the following result, which is similar to the factor reduction of canonical forms as presented in Chapter 14 of (Koller and Friedman, 2009).

**Proposition 12** *Let  $\phi$  be a  $\Gamma$ -canonical form with scope  $\mathcal{X} \cup \{Y, Z\}$ , where  $Z$  is the gamma variable, and parameters  $K, \mathbf{h}, g, a$  and  $b$ . Reducing  $\phi$  with respect to the evidence  $Y = y$  will result in a new  $\Gamma$ -canonical form  $\phi'$  over  $\mathcal{X} \cup \{Z\}$  with parameters  $K' = K_{\mathbf{X}\mathbf{X}}, \mathbf{h}' = \mathbf{h}_{\mathbf{X}} - K_{\mathbf{X}Y}y, g' = g + \mathbf{h}_Y^T y - \frac{1}{2}K_{YY}y^2, a' = a$  and  $b' = b$ .*

Here the subscript of the matrices and vectors denotes the submatrices and subvectors associated with those subscript. We will use this notation in general.

**Proof** So we have normal variables  $\mathbf{X}$  and  $Y$  and gamma variable  $Z$ . For fixed values  $\mathbf{x}, y$  and  $z$ , we can rewrite the quadratic polynomial of  $\phi(\mathbf{x}, y, z)$  as

$$-\frac{1}{2}(\mathbf{x}, y)^T K(\mathbf{x}, y) + \mathbf{h}^T(\mathbf{x}, y) + g = -\frac{1}{2}\mathbf{x}^T K_{\mathbf{X}\mathbf{X}}\mathbf{x} - K_{Y\mathbf{X}}y\mathbf{x} - \frac{1}{2}K_{YY}y^2 + \mathbf{h}_{\mathbf{X}}^T\mathbf{x} + \mathbf{h}_Y^T y + g.$$

Rearranging the right hand side of this equation now yields the result. ■

### 3.2 Factor Multiplication

The factor multiplication step is probably the easiest one to figure out. For example, let  $\phi_1$  and  $\phi_2$  be two  $\Gamma$ -canonical forms with the same scope and with parameters  $K_i, \mathbf{h}_i, g_i, a_i$  and  $b_i$  for  $i = 1, 2$ . Then the multiplication of  $\phi_1$  and  $\phi_2$  will result in a new  $\Gamma$ -canonical form with parameters  $K = K_1 + K_2, \mathbf{h} = \mathbf{h}_1 + \mathbf{h}_2, g = g_1 + g_2, a = a_1 + a_2$  and  $b = b_1 + b_2$ . This is not hard to confirm.

### 3.3 Factor Marginalization

The factor marginalization is a bit tricky and, as one would expect, the resulting factor will have different forms depending on whether the marginalization is done over a normal variable or the gamma variable.

**Proposition 13** *Let  $\phi$  be a  $\Gamma$ -canonical form with scope  $\mathcal{X} \cup \{Z\}$  and parameters  $K, \mathbf{h}, g, a$  and  $b$ . If  $\mathcal{X} = \{X\}$ , marginalizing  $\phi$  with respect to  $X$  will result in a gamma factor  $\phi'$  over  $Z$  with*

parameters  $g' = g + \mathbf{h}^2/2K$ ,  $a' = a - 1/2$  and  $b' = b + \log(2\pi)/2 - \log(K)/2$ . On the other hand, if  $|\mathcal{X}| > 1$ , marginalizing  $\phi$  with respect to  $Y \in \mathcal{X}$  will result in a new  $\Gamma$ -canonical form,  $\phi'$ , over  $\mathcal{X} \cup \{Z\} \setminus \{Y\}$  with parameters

$$\begin{cases} K' &= K_{-Y,-Y} - K_{-Y,Y}K_{Y,-Y}/K_{YY} \\ \mathbf{h}' &= \mathbf{h}_{-Y} - \frac{\mathbf{h}_Y}{K_{YY}}K_{Y,-Y} \\ g' &= g + \frac{\mathbf{h}_Y^2}{2K_{YY}} \\ a' &= a - 1/2 \\ b' &= b + \frac{1}{2}\log(2\pi) - \frac{1}{2}\log(K_{YY}). \end{cases}$$

The subscript with minus sign is used to denote all rows (or columns) except that particular one. Hence for example  $K_{-Y,Y}$  is the column vector consisting of the column of  $K$  corresponding to  $Y$ , but without the row corresponding to  $Y$ . The product  $K_{-Y,Y}K_{Y,-Y}$  is thus an outer product.

**Proof** Assume first that  $\mathcal{X} = \{X\}$ , then we have that

$$\begin{aligned} \phi'(Z) &= \int \exp\left(Z\left(-\frac{1}{2}KX^2 + \mathbf{h}X + g\right) + a\log(Z) + b\right) dX \\ &= \exp\left(gZ + a\log(Z) + b + \frac{\mathbf{h}^2}{K}Z\right) \int \exp\left(-\frac{1}{2}ZK\left(X - \frac{\mathbf{h}}{K}\right)^2\right) dX. \end{aligned}$$

The integrand in this last expression is proportional to the normal density with mean  $\mathbf{h}/K$  and variance  $1/(ZK)$ , hence the value of the integral is equal to  $\sqrt{2\pi/ZK}$ . Sorting out the term in the resulting expression yields the result. To prove the second part, i.e. the case when  $|\mathcal{X}| > 1$ , the technique is exactly the same but the computations are more tiresome, hence we omit them. ■

We see that as long as marginalization on  $\Gamma$ -canonical forms occurs with respect to a variable in  $\mathcal{X}$ , the resulting factor will either be a  $\Gamma$ -canonical form or a gamma factor. However, marginalization done over the gamma variable  $Z$  will not result in such factors. In fact, it is known that if  $(X, Z)$  is normal-gamma distributed, then  $X$  is non-standardized Student- $t$  distributed, again see (DeGroot, 2005). This also holds if  $X$  is multivariate, which it is in our case. In general, the multivariate Student- $t$  distribution has density

$$C_0 \left(1 + \frac{(\mathbf{x} - \delta)^T \Sigma^{-1} (\mathbf{x} - \delta)}{\nu}\right)^{-\frac{\nu+m}{2}} \quad (10)$$

where  $C_0$  is a normalizing constant,  $\delta$ ,  $\Sigma$ ,  $\nu$  are parameters of the distribution and  $m$  is the dimension of the random vector, see for example (Kotz and Nadarajah, 2004). The following result tells us what the output factor looks like when we marginalize a  $\Gamma$ -canonical form with respect to  $Z$ .

**Proposition 14** *Let  $\phi$  be a  $\Gamma$ -canonical form with scope  $\mathcal{X} \cup \{Z\}$  and parameters  $K$ ,  $\mathbf{h}$ ,  $g$ ,  $a$  and  $b$ , with  $K$  positive definite. Marginalizing  $\phi$  with respect to  $Z$  will result in a new factor  $\phi'$  over  $\mathcal{X}$  that is proportional to the density of a multivariate Student- $t$  distribution with parameters*

$$\begin{cases} \Sigma &= -\frac{1}{\nu} (\mathbf{h}^T K^{-1} \mathbf{h} + 2g) K^{-1} \\ \delta &= K^{-1} \mathbf{h} \\ \nu &= 2(a + 1) - m. \end{cases} \quad (11)$$



**Proof** Recall that  $\phi$  can be written as

$$\phi(\mathbf{x}, z) = \exp \left( z \left( -\frac{1}{2} \mathbf{x}^T K \mathbf{x} + \mathbf{h}^T \mathbf{x} + g \right) + a \log(z) + b \right).$$

The general identity

$$\int_0^\infty \exp(-\beta z + (\alpha - 1) \log(z)) dz = \exp(\ell\Gamma(\alpha) - \alpha \log(\beta)) \quad (12)$$

can be derived by using that the gamma density integrates to unity. Using (12), we see that if we integrate  $\phi(\mathbf{x}, z)$  over  $z$ , we get that

$$\phi'(\mathbf{x}) \propto \exp \left( -(a + 1) \log \left( \frac{1}{2} \mathbf{x}^T K \mathbf{x} - \mathbf{h}^T \mathbf{x} - g \right) \right). \quad (13)$$

We want to identify the parameters in this expression with the parameters of (10), which is proportional to

$$\exp \left( -\frac{\nu + m}{2} \log \left( 1 + \frac{(\mathbf{x} - \delta)^T \Sigma^{-1} (\mathbf{x} - \delta)}{\nu} \right) \right). \quad (14)$$

The result follows by comparing the parameters of (13) to those of (14). ■

For us this result means that marginalizing out  $Z$  from a  $\Gamma$ -canonical form will result in a factor that is proportional to the multivariate Student- $t$  distribution. This is unfortunate in the sense that the corresponding factor is not a  $\Gamma$ -canonical form. The good news is that we still have an analytic expression for the factor and that we can still associate it with a known probability distribution. In the light of this discussion, we need the gamma variable to be eliminated last in the algorithm, so that all factors will be  $\Gamma$ -canonical forms (or gamma factors) for as long as possible. This means that we in Step 4 of the algorithm must make sure that this is done.

### 3.4 Extension to Conditional Linear $\Gamma$ -Gaussian Bayesian Networks

Just as in Section 2.4, the definition of an algorithm for  $\Gamma$ -Gaussian networks can be extended to networks that also contain finite variables. Again we need the restriction that no finite node has continuous parents. Extending previous nomenclature, we call these networks conditional linear  $\Gamma$ -Gaussian Bayesian networks.

## 4. An Example

One of the many application areas of Bayesian Networks is forensics (Taroni et al., 2006). Here we will look at a forensic comparison question using a BN including one gamma variable, one 0-1 discrete variable and many normal variables.

More specifically, consider a case where a burglary has been committed, there is a broken window on the crime scene and glass fragments are found on a suspect. A forensic expert is consulted to assess the evidential value of the two collections of glass. When analyzing different types of glass, one can measure the *refractive index* (RI) (Zadora et al., 2013), in which case repeated measurements are done on each source of glass. We will assume that such repeated measurements are normally distributed around the underlying parameter, i.e., the true refractive index of the glass. In

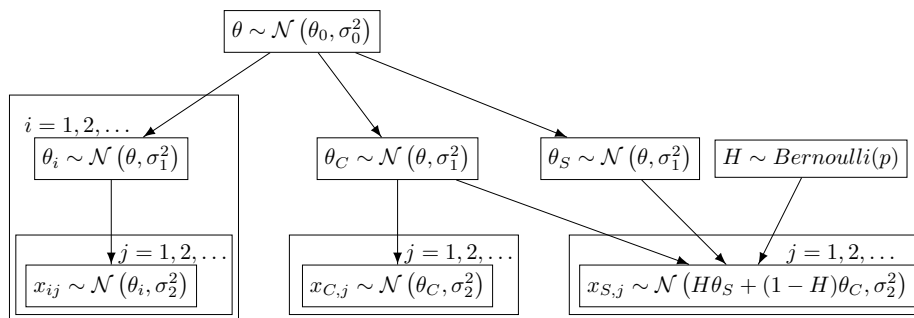


Figure 1: A conditional linear Gaussian model for glass data comparisons.

turn, we will assume that these true RIs are normally distributed around some population mean. The prosecutor’s hypothesis may be that the RIs of the two sets of glass fragments are the same, while the defense may claim that the two are independently sampled from the relevant population.

Our data comes from a database of 405 glass sources, where between 4 and 14 RI measurements have been performed on each source<sup>1</sup>. We denote the data from the crime scene by  $x_{C,1}, \dots, x_{C,n_C}$ , the data from the suspect by  $x_{S,1}, \dots, x_{S,n_S}$ , and the data from the data base sources by  $x_{i,1}, \dots, x_{i,n_i}$ . Figure 1 shows a simple model for our situation, in which  $H$  is our query node. To make computations with the network in Figure 1, values for  $\sigma_0$ ,  $\sigma_1$ , and  $\sigma_2$  need to be fixed. It may be quite natural to set  $\sigma_0 \approx \infty$ , using a flat prior for  $\theta$ . However, values for  $\sigma_1$  and  $\sigma_2$  must be estimated from the database; indeed, the estimation of these will be the main way in which the database influences the distribution of  $H$ . Using the full database, we obtain  $\sigma_1 = 2.4 \cdot 10^{-3}$  and  $\sigma_2 = 4.2 \cdot 10^{-5}$ . With our large database, there is little uncertainty in these estimates. However, similar comparison problems may occur for types of materials where the database is much smaller. In such cases, considerable uncertainty will remain for the variance parameters, and more accurate results will be obtained if this uncertainty is taken into account in the network. To do so, a natural model might be to use independent nodes with uninformative priors for  $\sigma_1$  and  $\sigma_2$ . However, this would lead to a BN for which exact inference is not available. Instead we use the  $\Gamma$ -Gaussian BN shown in Figure 2. The advantage relative to the Gaussian BN is that some uncertainty stemming from variance estimation is propagated to the final result. A disadvantage is that the variances are not independent, and reasonable values for  $C_1$  and  $C_2$  must be estimated.

#### 4.1 Computational Results

We tested the BNs from both Figures 1 and 2 using 4 different subsets of the database, containing 3, 10, 50, and 100 glass sources respectively. We used 5 different pairs from the complete database for which we computed  $\Pr(H = 1 | \text{data})$ . Pair 1 are the extremes, i.e. the sample with the largest and smallest mean, Pair 2 are the samples with means on the 10% and the 90% empirical quantiles, Pair 3 the 25% and the 75% quantiles and Pair 4 the 40% and the 60% quantiles. In Pair 5 we have picked two samples whose means are in the middle of the empirical distribution.

For each database subset, we produced estimates of  $\sigma_1$  and  $\sigma_2$  and used them with the Gaussian BN. As for the  $\Gamma$ -Gaussian BN, we let  $\alpha = \beta = 10^{-3}$ , and we produced the estimate for  $C_2$  as

1. The database has been kindly provided by the Swedish National Forensic Center

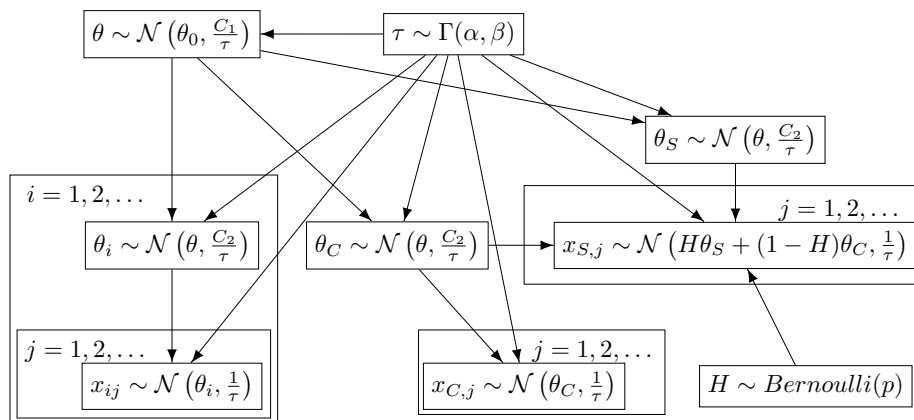


Figure 2: A conditional linear  $\Gamma$ -Gaussian model for glass data comparisons.

$\hat{\sigma}_1^2/\hat{\sigma}_2^2$ . Since we want uninformative priors for  $\theta$ , we let  $\sigma_0^2 = C_1 = 10^{20}$ . In all cases we let  $p = 0.5$ . The results of our tests are summarized in Table 1.

We see from Table 1 that when pairs are selected far away from each other in the glass population, the posterior probability in favor of the sources being different is almost exactly equal to 1, and when the pairs are more close, the result is more ambiguous. Results from the  $\Gamma$ -Gaussian BN should then be more trusted, as at least some of the uncertainty coming from variance estimation is taken into account in the result.

$N$	Gaussian BN				$\Gamma$ -Gaussian BN			
	3	10	50	100	3	10	50	100
Pair1	1.000	1.000	1.000	1.000	0.989	1.000	1.000	1.000
Pair2	1.000	1.000	1.000	1.000	0.015	0.023	0.180	0.803
Pair3	1.000	1.000	1.000	1.000	0.010	0.013	0.030	0.076
Pair4	1.000	1.000	1.000	1.000	0.010	0.012	0.014	0.018
Pair5	0.017	0.018	0.020	0.020	0.010	0.011	0.011	0.011

Table 1: Values for  $\Pr(H = 1|\text{data})$  using various database subsets, pairs, and networks.

### 5. Discussion and Conclusion

There are a number of different approaches to inference for models formulated as Bayesian Networks. Most focus on various types of approximations, as exact analytical solutions are possible for only a limited class of networks. For many approximations, such as those based on MCMC simulation, it is difficult to obtain useful and provable bounds for the accuracy of results. In some areas of application of BNs the lack of exact results may be particularly problematic: In forensic statistics, results may be scrutinized by a court, and loose ends such as lack of proof of convergence in computations may be an easy target for legal attack.

Consequently, we believe it is valuable to study the boundaries of the class of BNs for which exact inference is possible. In this paper, we introduce one particular class, called conditional linear  $\Gamma$ -Gaussian BNs, which is based on the conjugacy between the precision parameter for the normal distribution and the gamma distribution. We present theory and an implementation for inference

in such networks, as well as an example computation. However, similar classes of BNs based on additional conjugate pairs should also lend themselves to exact inference, and we hope to extend our theory and algorithms to also include such networks.

When formulating models used in many small statistical investigations as Bayesian Networks, one will see that most resulting BNs will not be of the type for which exact analytical inference is possible. For example, the most natural model for the data used in the computational example of this paper may be one where the two precision parameters  $\tau_2$  and  $\tau_3$  are independent, and then some posterior distributions will not be analytically computable. However, we believe there is great potential in combining the type of theory and algorithms developed in this paper with other algorithms and methods to produce efficient inference algorithms for also these types of networks.

## References

- R. Cowell, A. Dawid, S. Lauritzen, and D. Spiegelhalter. *Probabilistic networks and expert systems*. Springer-Verlag, 1999.
- M. DeGroot. *Optimal statistical decisions*. John Wiley & Sons, 2005.
- D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.
- S. Kotz and S. Nadarajah. *Multivariate t-distributions and their applications*. Cambridge University Press, 2004.
- S. Lauritzen and F. Jensen. Stable local computation with conditional gaussian distributions. *Statistics and Computing*, 11:191–203, 2001.
- S. Moral, R. Rumí, and A. Salmerón. Mixtures of truncated exponentials in hybrid bayesian networks. In *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, pages 156–167. Springer, 2001.
- P. Shenoy and J. West. Inference in hybrid bayesian networks using mixtures of polynomials. *International Journal of Approximate Reasoning*, 52:805–818, 2011.
- F. Taroni, C. Aitken, P. Garbolino, and A. Biedermann. *Bayesian Networks for Evaluating Scientific Evidence*. John Wiley & Sons, Ltd, 2006.
- G. Zadora, A. Martyna, D. Ramos, and C. Aitken. *Statistical Analysis in Forensic Science: Evidential Values of Multivariate Physicochemical Data*. John Wiley & Sons, 2013.