Profile Likelihood in Directed Graphical Models from BUGS Output

Malene Højbjerre
Department of Mathematical Sciences
Aalborg University
Fredrik Bajers Vej 7E
9220 Aalborg
Denmark

Abstract

This paper presents a method for using output of the computer program BUGS to obtain approximate profile likelihood functions of parameters or functions of parameters in directed graphical models with incomplete data. The method also provides a tool to approximate integrated likelihood functions. The prior distributions specified in BUGS do not have a significant impact on the profile likelihood functions and we consider the method as a desirable supplement to BUGS that enables us to do both Bayesian and likelihood based analyses in directed graphical models.

1 Introduction

During the last decade Markov chain Monte Carlo (MCMC) methods have become increasingly popular as a computational tool for approximating high-dimensional complex integrals. The methods are based on the ideas of Metropolis, Rosenbluth, Rosenbluth, Teller & Teller (1953) and Hastings (1970) and the most common one is Gibbs sampling (Geman & Geman 1984). The use of MCMC methods in Bayesian statistics was introduced by Gelfand & Smith (1990), and pure likelihood based analyses have been considered for exponential families in Geyer & Thompson (1992) and for the general case in Geyer (1994). A general discussion on MCMC methods and further references are given in Gilks, Richardson & Spiegelhalter (1996) and a comprehensive tutorial review is given in Brooks (1998).

Directed graphical models, introduced by Lauritzen, Dawid, Larsen & Leimer (1990), represent the conditional independence structure of the model through an appropriate factorization of the joint density w.r.t. a graph. This property can be exploited in MCMC methods, see e.g. Spiegelhalter (1998), and software for performing statistical analyses in Bayesian graphical models by means of MCMC methods is the computer program BUGS (Bayesian inference Using Gibbs Sampling), see Spiegelhalter, Thomas, Best & Gilks (1996a).

This paper presents a method for using posterior samples (here produced in BUGS) to approximate profile likelihood functions of parameters or functions of parameters in directed graphical models with incomplete data. Profile likelihood functions might have a misleading behavior as pointed out in Berger, Liseo & Wolpert (1999). Nevertheless, it represents an aspect of the parameter uncertainty which does not depend on specification of prior distributions.

2 Directed graphical model

A directed graphical model is defined by a directed acyclic graph, $\mathcal{G} = (V, E)$, and a joint probability distribution of $\boldsymbol{v} = (v_v)_{v \in V}$ that is *directed Markov* w.r.t. to $\mathcal{G}$. There are several equivalent ways to define a directed Markov distribution, see e.g. Lauritzen (1996). One way is to assume that the distribution has density w.r.t. a product measure, and that the density admits the *recursive factorization property* given as

$$p(\boldsymbol{v}) = \prod_{v \in V} p(v_v | \boldsymbol{v}_{\mathrm{pa}(v)})$$

where $p(v_v|\boldsymbol{v}_{\mathrm{pa}(v)})$ is the density of v_v given $\boldsymbol{v}_{\mathrm{pa}(v)} = (v_v)_{v \in \mathrm{pa}(v)}$. The term $\mathrm{pa}(v)$ denotes the parent vertices of v in the same meaning as in a genealogical tree. We assume that all the *parent-child densities* are strictly positive. Note that the joint model is totally specified by all the local parent-child distributions.

Traditionally a directed graphical model only includes the data (observed or missing) as vertices in the graph, and it does not include quantities like parameters, latent variables and/or covariates. As in Spiegelhalter (1998) we extend our model to do that, and assume that V is divided into four disjoint subset as

$$V = X \cup Y \cup \Theta \cup C$$

	random (single-edged)	constant (double-edged)
observed (rectangle)	X: observed data □	C: covariates ⧈
unobserved (circle)	Y: missing data/ latent variables ○	Θ: parameters ◎

Table 1: *Different kinds of vertices.*

where each subset and corresponding symbol are further described in Table 1.

The vertices are classified according to random/constant and observed/unobserved. Concerning the missing data we assume that they are missing at random (MAR), see Rubin (1976). The parameters and the latent variables are clearly missing completely at random (MCAR), see Cowell, Dawid, Lauritzen & Spiegelhalter (1999) page 200. The constant vertices (double-edged) are by nature not allowed to have parents and are so-called *founder vertices*.

In notation, let an observed random variable be denoted by $x_v, v \in X$, and a missing data/latent variable by $y_v, v \in Y$. Let $\theta_v, v \in \Theta$, be the parameter corresponding to the vertex v and c_v, $v \in C$, be the covariate corresponding to the vertex v. For a subset $A \subseteq V$, let $\boldsymbol{x}_A = (x_v)_{v \in A}$, $\boldsymbol{y}_A = (y_v)_{v \in A}$, $\boldsymbol{\theta}_A = (\theta_v)_{v \in A}$ and $\boldsymbol{c}_A = (c_v)_{v \in A}$. Furthermore $\boldsymbol{x} = \boldsymbol{x}_X$, $\boldsymbol{y} = \boldsymbol{y}_Y$, $\boldsymbol{\theta} = \boldsymbol{\theta}_\Theta$ and $\boldsymbol{c} = \boldsymbol{c}_C$.

In Bayesian statistics there is a need to calculate high-dimensional complex integrals over the posterior distribution, and MCMC methods is a powerful tool to approximate such complex integrals. BUGS uses a systematic scheme to produce dependent samples from the posterior distribution of the unobserved quantities ($\boldsymbol{\theta}$ and $\boldsymbol{y}$) given the observed quantities ($\boldsymbol{x}$ and $\boldsymbol{c}$) by successively simulating values from the *full conditional* distribution of each unobserved quantity given the current value of all the other quantities. Each full conditional distribution is proportional to the product of the local parent-child distributions of the corresponding *Markov blanket* expressed as

$$p(v_v|\boldsymbol{v}_{V\setminus v}) \; \propto \; p(v_v|\boldsymbol{v}_{\mathrm{pa}(v)}) \prod_{w: v \in \mathrm{pa}(v)} p(v_w|\boldsymbol{v}_{\mathrm{pa}(w)})$$

i.e. the local parent-child distributions are necessary and sufficient in order to produce dependent samples from the posterior distribution. The posterior sample can be used to perform inference about the specific parameters of interest.

In this paper we want to perform likelihood inference about the parameters, $\boldsymbol{\theta}$, and we consider them as unobserved constants in contrast to BUGS where all the quantities are considered as random variables. It is important to realize that the constant vertices can be considered as random variables with a given prior distribution and condition on these.

From the recursive factorization property we get that the joint density factorizes as

$$\begin{aligned} p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{c}) \;&=\; \prod_{v \in X} p(x_v|\boldsymbol{x}_{\mathrm{pa}(v)}, \boldsymbol{y}_{\mathrm{pa}(v)}, \boldsymbol{\theta}_{\mathrm{pa}(v)}, \boldsymbol{c}_{\mathrm{pa}(v)}) \\ &\times \prod_{v \in Y} p(y_v|\boldsymbol{x}_{\mathrm{pa}(v)}, \boldsymbol{y}_{\mathrm{pa}(v)}, \boldsymbol{\theta}_{\mathrm{pa}(v)}, \boldsymbol{c}_{\mathrm{pa}(v)}) \\ &\times \prod_{v \in \Theta} p(\theta_v) \prod_{v \in C} p(c_v) \end{aligned}$$

Due to the conditional independence assumptions specified by the graph, the elements of $\boldsymbol{\theta}$ and $\boldsymbol{c}$ are mutually independent, and therefore conditioning on $\boldsymbol{\theta}$ and $\boldsymbol{c}$ we get

$$p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{c}) = \prod_{v \in X \cup Y} p(z_v|\boldsymbol{z}_{\mathrm{pa}(v)}, \boldsymbol{\theta}_{\mathrm{pa}(v)}, \boldsymbol{c}_{\mathrm{pa}(v)})$$

where z_v is either x_v or y_v and $\boldsymbol{z}_{\mathrm{pa}(v)} = (\boldsymbol{x}_{\mathrm{pa}(v)}, \boldsymbol{y}_{\mathrm{pa}(v)})$.

3 Profile likelihood from BUGS output

The likelihood function of $\boldsymbol{\theta}$ is given as

$$L(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{c}) = \int p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{c}) d\boldsymbol{y} \tag{1}$$

Let θ_i, $i \in \Theta$, be the specific parameter of interest for which we want to approximate the profile likelihood function defined as

$$\hat{L}(\theta_i|\boldsymbol{x}, \boldsymbol{c}) = \sup_{\boldsymbol{\theta}_{\setminus i}} L(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{c})$$

where $\boldsymbol{\theta}_{\setminus i} = \boldsymbol{\theta}_{\Theta \setminus \{i\}}$. We consider two versions of a method to approximate $\hat{L}(\theta_i|\boldsymbol{x}, \boldsymbol{c})$ - one where we initially fix the parameters and one where we sample the parameters.

Version 1: fixed $\boldsymbol{\theta}$

We use the same technique as in Geyer & Thompson (1992). Let $\boldsymbol{\theta}_0$ be an arbitrary fixed value of $\boldsymbol{\theta}$, and consider

$$\frac{p(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{c})}{p(\boldsymbol{x}|\boldsymbol{\theta}_0, \boldsymbol{c})} = \frac{1}{\alpha} L(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{c})$$

where $\alpha = p(\boldsymbol{x}|\boldsymbol{\theta}_0, \boldsymbol{c})$ is a constant. This can also be expressed as

$$\begin{aligned} &L(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{c}) \\ &= \alpha \int \frac{p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{c})}{p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}_0, \boldsymbol{c})} \frac{p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}_0, \boldsymbol{c})}{p(\boldsymbol{x}|\boldsymbol{\theta}_0, \boldsymbol{c})} d\boldsymbol{y} \\ &= \alpha \int \prod_{v \in \mathrm{ch}(\Theta)} \frac{p(z_v|\boldsymbol{z}_{\mathrm{pa}(v)}, \boldsymbol{\theta}_{\mathrm{pa}(v)}, \boldsymbol{c}_{\mathrm{pa}(v)})}{p(z_v|\boldsymbol{z}_{\mathrm{pa}(v)}, \boldsymbol{\theta}_{0\mathrm{pa}(v)}, \boldsymbol{c}_{\mathrm{pa}(v)})} p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}_0, \boldsymbol{c}) d\boldsymbol{y} \end{aligned}$$

where we have used that

$$\begin{aligned}\frac{p(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{c})}{p(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{\theta}_0,\boldsymbol{c})} &= \prod_{v\in X\cup Y}\frac{p(z_v|\boldsymbol{z}_{\mathrm{pa}(v)},\boldsymbol{\theta}_{\mathrm{pa}(v)},\boldsymbol{c}_{\mathrm{pa}(v)})}{p(z_v|\boldsymbol{z}_{\mathrm{pa}(v)},\boldsymbol{\theta}_{0\mathrm{pa}(v)},\boldsymbol{c}_{\mathrm{pa}(v)})}\\ &= \prod_{v\in \mathrm{ch}(\Theta)}\frac{p(z_v|\boldsymbol{z}_{\mathrm{pa}(v)},\boldsymbol{\theta}_{\mathrm{pa}(v)},\boldsymbol{c}_{\mathrm{pa}(v)})}{p(z_v|\boldsymbol{z}_{\mathrm{pa}(v)},\boldsymbol{\theta}_{0\mathrm{pa}(v)},\boldsymbol{c}_{\mathrm{pa}(v)})}\end{aligned}$$

and $\mathrm{ch}(\Theta)$ denotes the set of vertices, which are children of vertices in Θ. When v is not a child of a parameter the terms in the fraction cancel.

This likelihood expression is in general computationally difficult. Therefore draw a sample $\boldsymbol{y}^{(1)},\boldsymbol{y}^{(2)},\ldots,\boldsymbol{y}^{(N)}$ from an ergodic Markov chain with stationary distribution $p\left(\boldsymbol{y}|\,\boldsymbol{x},\boldsymbol{\theta}_0,\boldsymbol{c}\right)$. The Ergodic Theorem yields for large N that the likelihood function can be approximated by

$$\tilde{L}(\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{c})=\frac{\alpha}{N}\sum_{j=1}^{N}\prod_{v\in\mathrm{ch}(\Theta)}\frac{p\big(z_v^{(j)}|\boldsymbol{z}_{\mathrm{pa}(v)}^{(j)},\boldsymbol{\theta}_{\mathrm{pa}(v)},\boldsymbol{c}_{\mathrm{pa}(v)}\big)}{p\big(z_v^{(j)}|\boldsymbol{z}_{\mathrm{pa}(v)}^{(j)},\boldsymbol{\theta}_{0\mathrm{pa}(v)},\boldsymbol{c}_{\mathrm{pa}(v)}\big)}$$

where $z_v^{(j)}=x_v$, when $v\in X$, $z_v^{(j)}=y_v^{(j)}$, when $v\in Y$ and $\boldsymbol{z}_{\mathrm{pa}(v)}^{(j)}=(\boldsymbol{x}_{\mathrm{pa}(v)},\boldsymbol{y}_{\mathrm{pa}(v)}^{(j)})$.

In practice sampling from $p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}_0,\boldsymbol{c})$ can be made in BUGS by running a Gibbs sampler on the model with $\boldsymbol{\theta}$ fixed as $\boldsymbol{\theta}_0$. Note that the quality of the approximation will be best when $\boldsymbol{\theta}$ is not too far from $\boldsymbol{\theta}_0$ (Geyer 1996). We suggest choosing $\boldsymbol{\theta}_0$ as the posterior mean obtained from running an initial Gibbs sampler on the model with prior distributions specified on $\boldsymbol{\theta}$.

We compute $\log\tilde{L}(\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{c})$ up to the constant in a grid of the parameters, where the defining values of the grid is a certain number of quantiles from the initial run of the Gibbs sampler. To get an estimate of the profile log-likelihood function of a specific parameter we maximize it w.r.t. all other parameters over the grid.

Version 2: variable $\boldsymbol{\theta}$

Instead of fixing $\boldsymbol{\theta}$ initially, it might be an idea to simulate values of it. Therefore let $\boldsymbol{\psi}$ be another generic symbol of $\boldsymbol{\theta}$, and consider

$$\int\frac{p(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{c})}{p(\boldsymbol{x}|\boldsymbol{\psi},\boldsymbol{c})}p(\boldsymbol{\psi}|\boldsymbol{x},\boldsymbol{c})d\boldsymbol{\psi}=\frac{1}{\beta}L(\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{c})$$

where $\beta=p(\boldsymbol{x}|\boldsymbol{c})$ is a constant. As in Version 1 this can be expressed as

$$\begin{aligned}&L(\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{c})\\ =&\beta\iint\frac{p(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{c})}{p(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{\psi},\boldsymbol{c})}\frac{p(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{\psi},\boldsymbol{c})p(\boldsymbol{\psi}|\boldsymbol{x},\boldsymbol{c})}{p(\boldsymbol{x}|\boldsymbol{\psi},\boldsymbol{c})}d\boldsymbol{y}d\boldsymbol{\psi}\\ =&\beta\iint\prod_{v\in\mathrm{ch}(\Theta)}\frac{p(z_v|\boldsymbol{z}_{\mathrm{pa}(v)},\boldsymbol{\theta}_{\mathrm{pa}(v)},\boldsymbol{c}_{\mathrm{pa}(v)})}{p(z_v|\boldsymbol{z}_{\mathrm{pa}(v)},\boldsymbol{\psi}_{\mathrm{pa}(v)},\boldsymbol{c}_{\mathrm{pa}(v)})}p(\boldsymbol{y},\boldsymbol{\psi}|\boldsymbol{x},\boldsymbol{c})d\boldsymbol{y}d\boldsymbol{\psi}\end{aligned}$$

which in general is computationally difficult. Draw a sample $(\boldsymbol{y}^{(1)},\boldsymbol{\psi}^{(1)}),(\boldsymbol{y}^{(2)},\boldsymbol{\psi}^{(2)}),\ldots,(\boldsymbol{y}^{(N)},\boldsymbol{\psi}^{(N)})$ from an ergodic Markov chain with stationary distribution $p(\boldsymbol{y},\boldsymbol{\psi}|\boldsymbol{x},\boldsymbol{c})$. Then for large N the likelihood function can be approximated by

$$\tilde{\tilde{L}}(\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{c})=\frac{\beta}{N}\sum_{j=1}^{N}\prod_{v\in\mathrm{ch}(\Theta)}\frac{p\big(z_v^{(j)}|\boldsymbol{z}_{\mathrm{pa}(v)}^{(j)},\boldsymbol{\theta}_{\mathrm{pa}(v)},\boldsymbol{c}_{\mathrm{pa}(v)}\big)}{p\big(z_v^{(j)}|\boldsymbol{z}_{\mathrm{pa}(v)}^{(j)},\boldsymbol{\psi}_{\mathrm{pa}(v)}^{(j)},\boldsymbol{c}_{\mathrm{pa}(v)}\big)}$$

Note that this approximation might be computationally dangerous if the stationary distribution has heavy tails. Then sampling from the tail will cause a few large values of the ratio to dominate the sum.

Sampling from $p(\boldsymbol{y},\boldsymbol{\psi}|\boldsymbol{x},\boldsymbol{c})$ can be made in BUGS by running a Gibbs sampler on the model with prior distributions specified on $\boldsymbol{\theta}$. These prior distributions are only introduced as a computational tool - they should not have any influence on the likelihood results. Profile log-likelihood approximations are produced by the same procedure as in Version 1, but now the defining values of the grid are quantiles from the run of the Gibbs sampler in this version.

4 Profile likelihood of a function from BUGS output

The profile likelihood of a function of the parameters, say $\phi=g(\boldsymbol{\theta})$, is defined as

$$\hat{L}(\phi|\boldsymbol{x},\boldsymbol{c})=\sup_{\boldsymbol{\theta}\in g^{-1}(\phi)}L(\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{c}),\quad \phi\in\Phi$$

where Φ is the image of the parameter space of $\boldsymbol{\theta}$ under g. To approximate $\hat{L}(\phi|\boldsymbol{x},\boldsymbol{c})$ by our method (e.g. Version 1) we form the pairs

$$\big(g(\boldsymbol{\theta}),\tilde{L}(\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{c})\big),\quad \boldsymbol{\theta}\in\widetilde{\nabla}$$

where $\widetilde{\nabla}$ are the grid points considered earlier. We only need to calculate $g(\boldsymbol{\theta})$, whereas we already have $\tilde{L}(\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{c})$. We order these pairs in increasing order w.r.t. $\phi=g(\boldsymbol{\theta})$, partition this ordering into bins each containing the same number of pairs. For each bin we find the pair that has the highest value of $\tilde{L}(\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{c})$. Let $\boldsymbol{\theta}_\nu^*$ be the value of $\boldsymbol{\theta}$ where the maximum values in bin ν is attained. Combining these maximum pairs, $(g(\boldsymbol{\theta}_\nu^*),\tilde{L}(\boldsymbol{\theta}_\nu^*|\boldsymbol{x},\boldsymbol{c}))$, we obtain an approximation of the profile likelihood function of $g(\boldsymbol{\theta})$. A similar procedure can be used for Version 2.

5 Integrated likelihood from BUGS output

Due to the sometimes misleading behavior of the profile likelihood Berger et al. (1999) consider the integrated likelihood function defined as

$$\begin{aligned}
\check{L}(\theta_i|\boldsymbol{x},\boldsymbol{c}) &= \int L(\theta_i,\boldsymbol{\theta}_{\setminus i}|\boldsymbol{x},\boldsymbol{c})p(\boldsymbol{\theta}_{\setminus i}|\theta_i)d\boldsymbol{\theta}_{\setminus i} \\
&= \iint p(\boldsymbol{x},\boldsymbol{y}|\theta_i,\boldsymbol{\theta}_{\setminus i},\boldsymbol{c})p(\boldsymbol{\theta}_{\setminus i}|\theta_i,\boldsymbol{c})d\boldsymbol{y}d\boldsymbol{\theta}_{\setminus i} \\
&= \iint p(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta}_{\setminus i}|\theta_i,\boldsymbol{c})d\boldsymbol{y}d\boldsymbol{\theta}_{\setminus i} \\
&= \int p(\boldsymbol{x},\boldsymbol{w}|\theta_i,\boldsymbol{c})d\boldsymbol{w}
\end{aligned}$$

where $\boldsymbol{w} = (\boldsymbol{y},\boldsymbol{\theta}_{\setminus i})$ and we have used that the elements of $\boldsymbol{\theta}$ and $\boldsymbol{c}$ are mutually independent. This expression is identical to (1) by replacing $\boldsymbol{y}$ with $\boldsymbol{w}$. Therefore by considering $\Theta \setminus \{i\}$ also as latent variables, the method can be used to approximate the integrated likelihood function of θ_i. In this case we get a 1-dimensional grid and the marginalisation is done by summation instead of maximalisation.

Another way to approximate the integrated likelihood function of θ_i is to consider the ratio of the marginal posterior density to the marginal prior density. This can be seen from the following calculations

$$\begin{aligned}
\check{L}(\theta_i|\boldsymbol{x},\boldsymbol{c}) &= \int \frac{p(\theta_i,\boldsymbol{\theta}_{\setminus i}|\boldsymbol{x},\boldsymbol{c})p(\boldsymbol{x},\boldsymbol{c})}{p(\theta_i,\boldsymbol{\theta}_{\setminus i},\boldsymbol{c})}p(\boldsymbol{\theta}_{\setminus i}|\theta_i)d\boldsymbol{\theta}_{\setminus i} \\
&= \int \frac{p(\theta_i,\boldsymbol{\theta}_{\setminus i}|\boldsymbol{x},\boldsymbol{c})p(\boldsymbol{x},\boldsymbol{c})p(\boldsymbol{\theta}_{\setminus i})}{p(\boldsymbol{\theta}_{\setminus i})p(\boldsymbol{c})p\left(\theta_i\right)}d\boldsymbol{\theta}_{\setminus i} \\
&\propto \frac{p(\theta_i|\boldsymbol{x},\boldsymbol{c})}{p(\theta_i)}
\end{aligned}$$

Remark that the integrated likelihood conforms with a Bayesian approach - the weight function is a prior density on $\boldsymbol{\theta}_{\setminus i}$. This means that the integrated likelihood represents the parameter uncertainty, but it depends on the specification of the prior distribution. MCMC integrated likelihood has been considered in Andersen (1997).

6 Example

We consider the same model as in Example 2 of Spiegelhalter, Thomas, Best & Gilks (1996b). The data are taken from George, Makov & Smith (1993), but were originally treated in Gaver & O'Muircheartaigh (1987). The example concerns ten power plant pumps, for which the operation time, c_i, and the number of failures, x_i, are measured. The data are shown in Table 2.

The model is illustrated by the graph in Figure 1, where the repetitive structure of the model is represented by large enclosing boxes. The parent-child distributions are given as follows for $i = 1, 2, \ldots, 10$

$$\begin{aligned}
x_i|\, y_i, c_i &\sim \mathrm{Po}\left(y_i c_i\right),\ x_i \geq 0 \\
y_i|\, \theta_1, \theta_2 &\sim \Gamma\left(\theta_1, \theta_2\right),\ y_i \geq 0,\ \theta_1 > 0,\ \theta_2 > 0
\end{aligned}$$

Pump	1	2	3	4	5
c_i	94.3	15.7	62.9	126	5.24
x_i	5	1	5	14	3
Pump	**6**	**7**	**8**	**9**	**10**
c_i	31.4	1.05	1.05	2.1	10.5
x_i	19	1	1	4	22

Table 2: *The operation time, c_i, and the number of failures, x_i, for ten power plant pumps.*

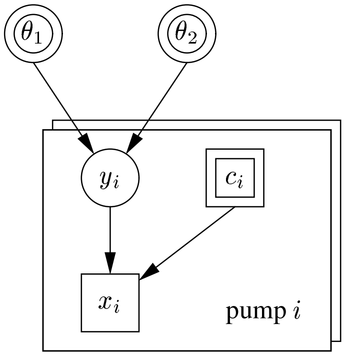

Figure 1: *Directed acyclic graph for the model.*

where y_i is the failure rate of pump i, θ_1 is the shape parameter and θ_2 is the scale parameter of the gamma distribution. All the parent-child densities are strictly positive. Note that the vertices are denoted by the name of the corresponding quantity - in contrast to Section 2, where they are denoted by index.

We need to specify prior distributions on θ_1 and θ_2 in order to find initial values for Version 1 and to run the Gibbs Sampler in Version 2. We try different prior distributions to see whether they have a significant impact on the likelihood results. As in George et al. (1993) we choose exponential and gamma distributions for the family of the prior distribution of θ_1 and θ_2, respectively. We have tried the prior distributions stated in Table 3, where Prior 1 has the smallest variance and Prior 3 has the largest.

We have used BUGS version 0.6 on a UNIX platform to perform Gibbs sampling, and to check for convergence the S-PLUS function CODA (Convergence Diagnosis and Output Analysis Software for Gibbs sampling output) version 0.4 (Best, Cowles & Vines 1996) has been used. We have implemented our own S-PLUS functions to approximate the log-likelihood functions and the results are produced in S-PLUS version 5.1 on a UNIX platform.

	Prior 1	Prior 2	Prior 3
θ_1	$\mathrm{Exp}(10)$	$\mathrm{Exp}(1.0)$	$\mathrm{Exp}(0.01)$
θ_2	$\Gamma(0.1, 1.0)$	$\Gamma(0.01, 0.1)$	$\Gamma(0.001, 0.001)$

Table 3: *Prior distributions on θ_1 and θ_2.*

After a burn-in we simulate samples of size $N = 5.000$. We calculate the approximations in 200×200 grids and maximize over the grids in order to obtain approximate profile log-likelihood functions. We have one approximation for each combination of version and prior distribution, see Figure 2. The log-likelihood functions are standardized such that they have maximum value 0 and are plotted between -4 and 0.

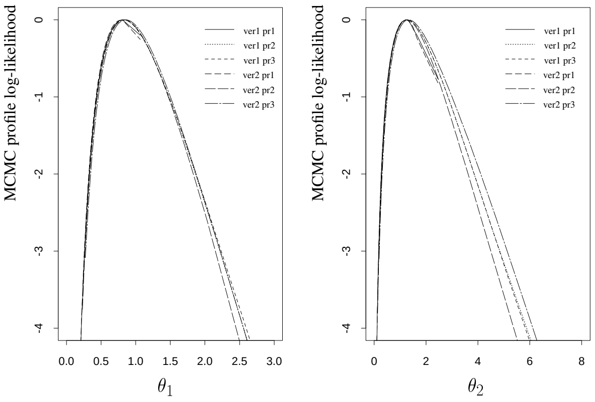

Figure 2: MCMC profile log-likelihood of θ_1 and θ_2 for each version of the method and prior distribution.

Generally the approximations are the same, but the approximations for Prior 1 are cut off close to the maxima. This is caused by the grid being quantiles of the posterior distribution, which is strongly influenced by the very small variance of Prior 1. But keep in mind that these prior distributions are only chosen for illustrative purposes - in practice one would never choose a prior distribution with such a small variance as Prior 1. Even though Prior 3 has a very large variance the method does not have any difficulties identifying the profile log-likelihood approximations for this prior distribution. Finally the approximations differ slightly in the right-hand tails, which might be an effect of the sampling not being concentrated here. We conclude that neither the prior distribution nor the version of the method have a significant impact on the profile log-likelihood approximations.

In this particular model it is possible to find an exact expression of the log-likelihood function, so for comparison reasons we calculate this in the same grid as above and maximize it w.r.t. to θ_1 and θ_2, respectively. This does not give us the exact profile log-likelihood, but we do get an impression of the function which might be a little bit smaller than the true profile log-likelihood. We also approximate the integrated log-likelihood functions of θ_1 and θ_2 by the procedures described in Section 5. All these functions gives us an over-all impression of the parameter uncertainty, and the results for Version 1 Prior 2 are shown in Figure 3.

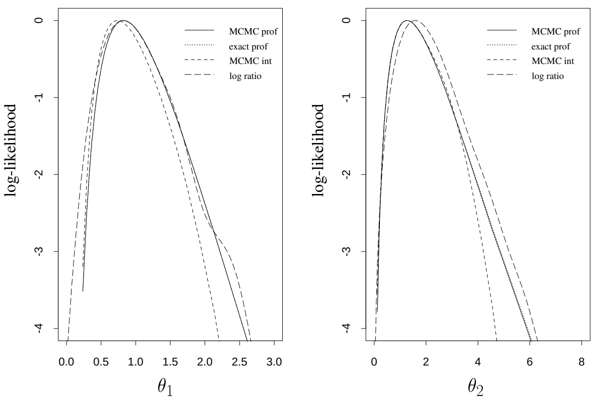

Figure 3: Exact profile log-likelihood, MCMC profile log-likelihood, MCMC integrated log-likelihood and log(posterior/prior) of θ_1 and θ_2 for Version 1 Prior 2.

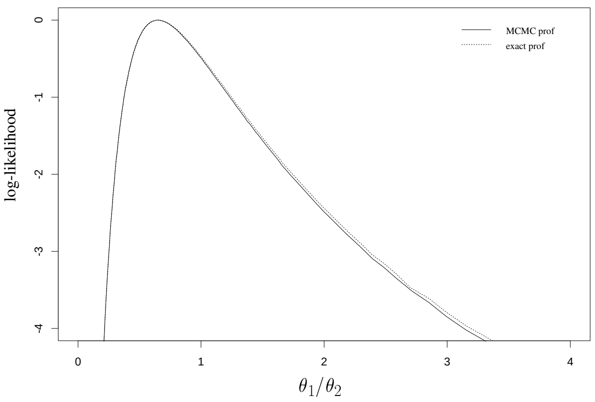

Figure 4: MCMC and exact profile log-likelihood of θ_1/θ_2 for Version 1 Prior 2.

The exact profile log-likelihood and the MCMC approximation are indistinguishable and we deduce that our method generally gives good approximations of the profile log-likelihood function. Furthermore it does not depend on the prior distributions, in contrast to the Bayesian entities.

As described in Section 4 the method can also be used to approximate the profile log-likelihoods of a function of the parameters. Therefore we approximate the profile log-likelihood function of θ_1/θ_2, the mean failure rate. We have performed the same procedure for the exact likelihood surface to get an impression of the exact profile log-likelihood function of θ_1/θ_2. The results are illustrated in Figure 4 for Version 1 Prior 2. Again we see that the method performs a good approximation.

7 Discussion

The method described represents ways of approximating the profile likelihood function of parameters or functions of parameters in complex directed graphical models with incomplete data. The parent-child distributions of the parameter's children are all needed to compute the approximation. It does not require much additional programming and computationally it is not heavy. It exploits the already existing software, BUGS, and the prior distributions specified there do not have a significant impact on the profile likelihood results.

The method is a hybrid between Bayesian and likelihood inference in the sense that the prior information we might have on the parameters is used to compute a credible region for the parameters and then we approximate the profile likelihood in this region. Thus the prior information is not neglected, but the likelihood results do not depend significantly on them in contrast to Bayesian entities like posterior distribution and integrated log-likelihood.

In many cases it may not be satisfactory to base a scientific analysis solely on a pure Bayesian analysis and it is desirable also to be able to do plain likelihood inference to see how much influence the prior distributions have. Therefore we consider the method as a complementary tool to BUGS.

Acknowledgment

The work has been funded by The Danish Natural Science Research Council and part of it was performed while the author was visiting the Department of Statistics, University of Georgia, Athens, Georgia. I am grateful to the department for hosting me. I am indebted to Steffen L. Lauritzen for providing detailed and constructive comments on earlier versions of the manuscript, and to Kim Emil Andersen, Claus Dethlefsen and Søren Lundbye-Christensen for helpful comments.

References

Andersen, H. H. (1997). *Monte Carlo Likelihood in Complex Graphical Models*, PhD thesis, Department of Mathematical Sciences, Aalborg University, Aalborg, Denmark.

Berger, J. O., Liseo, B. & Wolpert, R. L. (1999). Integrated likelihood methods for eliminating nuisance parameters, *Statistical Science* **14**(1): 1–28.

Best, N. G., Cowles, M. K. & Vines, K. (1996). *CODA, Convergence Diagnosis and Output Analysis Software for Gibbs Sampling Output, Version 0.30*, MRC Biostatistics Unit, Cambridge.

Brooks, S. P. (1998). Markov chain Monte Carlo method and its application, *The Statistician* **47**: 69–100.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L. & Spiegelhalter, D. J. (1999). *Probabilistic Networks and Expert Systems*, Statistics for Engineering and Information Science, 1st edn, Springer-Verlag, New York.

Gaver, D. P. & O'Muircheartaigh, I. G. (1987). Robust empirical Bayes analyses of event rates, *Technometrics* **29**: 1–15.

Gelfand, A. E. & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association* **85**(410): 398–409.

Geman, S. & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**(6): 721–741.

George, E. I., Makov, U. E. & Smith, A. F. M. (1993). Conjugate likelihood distributions, *Scandinavian Journal of Statistics* **20**: 147–156.

Geyer, C. J. (1994). On the convergence of Monte Carlo maximum likelihood calculations, *Journal of the Royal Statistical Society, Series B* **56**(1): 261–274.

Geyer, C. J. (1996). Estimation and optimization of functions, *in* W. R. Gilks, S. Richardson & D. J. Spiegelhalter (eds), *Markov Chain Monte Carlo in Practice*, Chapman and Hall, New York, pp. 241–258.

Geyer, C. J. & Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data, *Journal of the Royal Statistical Society, Series B* **54**(3): 657–699.

Gilks, W. R., Richardson, S. & Spiegelhalter, D. J. (eds) (1996). *Markov Chain Monte Carlo in Practice*, Chapman and Hall, New York.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications, *Biometrika* **57**(1): 97–109.

Lauritzen, S. L. (1996). *Graphical Models*, Clarendon Press, Oxford, UK.

Lauritzen, S. L., Dawid, A. P., Larsen, B. N. & Leimer, H.-G. (1990). Independence properties of directed Markov fields, *NETWORKS* **20**: 491–505.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953). Equations of state calculations by fast computing machine, *Journal of Chemical Physics* **21**: 1087–1091.

Rubin, D. B. (1976). Inference and missing data, *Biometrika* **63**(3): 581–592.

Spiegelhalter, D. J. (1998). Bayesian graphical modelling: a case-study in monitoring health outcomes, *Applied Statistics* **47**: 115–133.

Spiegelhalter, D. J., Thomas, A., Best, N. G. & Gilks, W. (1996a). *BUGS 0.5 Bayesian inference Using Gibbs sampling Manual (version ii)*, MRC Biostatistics Unit, Cambridge.

Spiegelhalter, D. J., Thomas, A., Best, N. G. & Gilks, W. (1996b). *BUGS 0.5 Examples Volume 1 (version i)*, MRC Biostatistics Unit, Cambridge.