# Efficient Gradient Computation for Conditional Gaussian Models

**Bo Thiesson**
Microsoft Research
thiesson@microsoft.com

**Christopher Meek**
Microsoft Research
meek@microsoft.com

## Abstract

We introduce Recursive Exponential Mixed Models (REMMs) and derive the gradient of the parameters for the incomplete-data likelihood. We demonstrate how one can use probabilistic inference in Conditional Gaussian (CG) graphical models, a special case of REMMs, to compute the gradient for a CG model. We also demonstrate that this approach can yield simple and effective algorithms for computing the gradient for models with tied parameters and illustrate this approach on stochastic ARMA models.

## 1 Introduction

The computation of parameter gradients given incomplete data is an important step in learning the parameters of a statistical model with missing data. In particular, for gradient based optimization methods, such as the conjugate gradient method, the gradient is used to iteratively adapt the parameters of the model in order to improve the incomplete-data likelihood and, in this way, identify the MLE or local maxima of the incomplete-data likelihood.

In this paper, we derive parameter gradients for a broad class of graphical models called recursive exponential mixed models (REMMs). REMMs generalize the well-known conditional Gaussian (CG) directed graphical models introduced by Lauritzen and Wermuth (1989). While REMMs have modelling advantages over CG models (e.g., allowing discrete variables to have continuous parents) our primary motivation for introducing REMMs in this paper is as a tool to derive the parameter gradient for CG models.

CG models are an important class of graphical models. They generalize discrete and Gaussian Bayesian networks and, importantly, they have efficient exact probabilistic inference algorithms for computing conditional marginal probabilities (Lauritzen and Jensen 2001). We derive the expression for the parameter gradient in CG models and demonstrate that standard probabilistic inference techniques can be used to efficiently compute these gradients.

We also demonstrate that our approach can yield simple and effective algorithms for computing parameter gradients of graphical models with tied parameters. We illustrate our approach to handling tied parameters on stochastic ARMA models (Thiesson *et al.* 2004). The stochastic ARMA model is of particular interest because it is a simple and useful model for modeling time-series data.

Previous papers have derived parameter gradients for graphical models. For instance, Thiesson (1997) and Binder *et al.* (1997) derive the parameter gradient for general classes of graphical models. In addition, both Thiesson (1997) and Binder *et al.* (1997) demonstrate that, for graphical models with only discrete variables, one can compute the parameter gradient using exact probabilistic inference. Binder *et al.* (1997) also discuss computation of the parameter gradient for models that have continuous variables. For such models, they resort to stochastic simulation to compute the gradient; they do so even with conditional Gaussian models for which exact probabilistic inference algorithms exist. Our results extend this work by demonstrating precisely how one can use probabilistic inference to compute the gradient for CG graphical models.

## 2 Recursive exponential mixed models

We consider directed graphical models with both discrete and continuous variables. For a directed graphical model the structural relations between variables $X = (X_v)_{v \in V}$, are represented by a directed acyclic graph (DAG), where each node $v$ represents a variable, $X_v$, and directed edges represent direct influence from variables represented by parent nodes, $X_{pa(v)}$. Markov properties with respect to the graph (Kiiveri, Speed, and Carlin 1984; Lauritzen *et al.* 1990) imply that any distribution, which is structurally defined by a such model, can be represented by (local) conditional

distributions, $p(X_v|X_{pa(v)})$.

Thiesson (1997), defines a class of directed graphical models called recursive exponential models (REMs) for which the focus is on discrete variables. We extend this definition to mixed models with both continuous and discrete variables. We call this class of models for *recursive exponential mixed models* (REMMs).

Both REM and REMM models assume *global variation independence* for the parameters in the models. That is,

$$p(X|\theta) = \prod_{v \in V} p(X_v|X_{pa(v)}, \theta_v), \quad (1)$$

where $\Theta = \times_{v \in V} \Theta_v$, and $\theta_v \in \Theta_v$ completely specifies the relationship between the variable $X_v$ and its conditional set of variables $X_{pa(v)}$.

For mixed models, the conditioning set for distributions on the right-hand side of (1) may have both continuous variables, denoted $X^c_{pa(v)}$, and discrete variables, denoted $X^d_{pa(v)}$. That is, $X_{pa(v)} = (X^c_{pa(v)}, X^d_{pa(v)})$. When the conditioning set contains discrete variables, REMM models will in addition assume *partial local variation independence* between parameters in conditional distributions with different values for the discrete conditioning variables. Let $\Pi^d_v$ denote the set of all configurations for discrete parents of $v$, and let $\pi^d_v \in \Pi^d_v$ denote a particular configuration. By partial local parameter independence, $\Theta_v = \times_{\pi^d_v \in \Pi^d_v} \Theta_{v|\pi^d_v}$, and $\theta_{v|\pi^d_v} \in \Theta_{v|\pi^d_v}$ completely defines the *local model* $p(X_v|X^c_{pa(v)}, \pi^d_v, \theta_{v|\pi^d_v})$. Notice that if the discrete set of parent variables is empty, then the local model $p(X_v|X^c_{pa(v)}, \pi^d_v, \theta_{v|\pi^d_v}) = p(X_v|X_{pa(v)}, \theta_v)$. Hence, REMM models with only continuous variables, will only require global parameter independence. Notice also that if all involved variables are discrete, then partial local parameter independence is the same as local parameter independence, as defined for the REM models.

Given global and partial local parameter independence the likelihood for a single observation factors into *local likelihoods* as follows

$$p(x|\theta) = \prod_{v \in V} p(x_v|x_{pa(v)}, \theta_{v|\pi^d_v}).$$

For a REMM, the local models have to be representable as regular exponential models. Hence, a local likelihood is represented as

$$\begin{aligned} & p(x_v|x_{pa(v)}, \theta_{v|\pi^d_v}) \\ & \quad = \quad b(x_v) \exp\left(\theta_{v|\pi^d_v} t(x_v)' - \phi(\theta_{v|\pi^d_v})\right), \end{aligned} \quad (2)$$

where $b$ is the carrying density, $t$ the canonical statistics, $\phi$ the normalization function, and $'$ denotes transpose. Notice that $b$, $t$, $\phi$ are specific to the distribution, where we condition on the discrete parents $x^d_{pa(v)} = \pi^d_v$.

As described above, a model is a REMM if it is defined in terms of local models represented as regular exponential models and the collection of local models satisfy global and partial local variation independence. Later, in Section 5, we will see that the assumptions of variation independence can easily be relaxed to allow parameters to be tied across local models.

### 2.1 Conditional Gaussian models

The REMMs are particularly designed to generalize the class of conditional Gaussian (CG) models introduced by Lauritzen and Wermuth (1989). The CG models are of particular interest because we can, as we demonstrate below, use the exact inference scheme of Lauritzen and Jensen (2001) to efficiently compute the parameter gradients for these models.

A conditional Gaussian directed graphical models is a graphical model in which (i) the graphical DAG structure has no discrete variable with a continuous parent variable, (ii) the local models for discrete variables are defined by conditional multinomial distributions that can be represented in the usual way via conditional probability tables, and (iii) the local models for continuous variables (given continuous and discrete parents) are defind by conditional Gaussian regressions – one for each configuration of values for discrete parents. In particular

$$\begin{aligned} & p(X_v|X^c_{pa(v)}, \pi^d_v, \theta_{v|\pi^d_v}) \\ & \qquad \sim \quad \mathcal{N}\left(c(\pi^d_v) + \beta(\pi^d_v) X^c_{pa(v)}, \sigma(\pi^d_v)\right). \end{aligned}$$

We have here emphazised that the intercept for the regression, $c$, the linear regression coefficients, $\beta$, and the variance $\sigma$ all depend on the particular configuration for the discrete parents, $\pi^d_v$. To simplify the notation in what follows, we will drop this explicit dependence.

Later, in Sections 4.2 and 4.3, we will see that local conditional multinomial and non-degenerate (or positive) local conditional Gaussian distributions can be represented as exponential models. A CG model assuming global and partial local parameter independence is therefore a REMM.

## 3 The incomplete-data gradient

We consider samples of incomplete observation and assume that the observations are incomplete in an non-informative way (e.g., missing at random; Gelman *et al.* 1995). Let $\mathbf{y} = (y^1, y^2, \ldots, y^L)$ denote a sample of possibly incomplete observations which are mutually independent. Given the mutual independence, the likelihood factorizes as a product over likelihoods for each observation

$$p(\mathbf{y}|\theta) = \prod_{l=1}^{L} p(y^l|\theta).$$

The gradient for the sample log-likelihood can therefore be obtained by simply adding the individual gradients for each observation. That is,

$$\frac{\partial \log p(\mathbf{y}|\theta)}{\partial \theta_{v|\pi_v^d}} = \sum_{l=1}^{L} \frac{\partial \log p(y^l|\theta)}{\partial \theta_{v|\pi_v^d}}. \quad (3)$$

We will in the next section derive the gradient expression for a single observation, knowing that the gradient for a sample can be obtained by simply adding up gradients for each observation, as in (3).

## 4 Single observation gradient

Suppose for a given model that a complete observation $x$ is only observed indirectly through the *incomplete* observation $y$. Denote by $\mathcal{X}(y)$ the set of possible completions that are obtainable by augmenting the incomplete observation $y$. The likelihood for the incomplete observation then becomes

$$\begin{aligned} p(y|\theta) &= \int_{x\in\mathcal{X}(y)} p(x|\theta)\mu(x) \\ &= \int_{x\in\mathcal{X}(y)} \prod_{v\in V} p(x_v|x_{pa(v)}, \theta_{v|\pi_v^d})\mu(x), \quad (4) \end{aligned}$$

where $\mu$ is a generalized measure, which for a CG model is an appropriate combination of the counting measure for discrete variables and the Lebesque measure for continuous variables.

The gradient for the log-likelihood can now be expressed as

$$\begin{aligned} \frac{\partial \log p(y|\theta)}{\partial \theta_{v|\pi_v^d}} &= \frac{1}{p(y|\theta)} \frac{\partial p(y|\theta)}{\partial \theta_{v|\pi_v^d}} \\ &= \frac{1}{p(y|\theta)} \int_{x\in\mathcal{X}(y)} \frac{\partial p(x|\theta)}{\partial \theta_{v|\pi_v^d}} \mu(x), \quad (5) \end{aligned}$$

where the last equality follows from (4) and by using Leibnitz's rule for interchanging the order of differentiation and integration.

Now, consider the local gradient for the complete observation $x$. The chain rule for differentiation implies

$$\begin{aligned} \frac{\partial p(x|\theta)}{\partial \theta_{v|\pi_v^d}} &= \frac{p(x|\theta)}{p(x_v|x_{pa(v)}, \theta_{v|\pi_v^d})} \frac{\partial p(x_v|x_{pa(v)}, \theta_{v|\pi_v^d})}{\partial \theta_{v|\pi_v^d}} \\ &= p(x|\theta) \frac{\partial \log p(x_v|x_{pa(v)}, \theta_{v|\pi_v^d})}{\partial \theta_{v|\pi_v^d}}. \quad (6) \end{aligned}$$

Thus, by the exponential representation in (2), the local gradient for a complete observation becomes

$$\frac{\partial p(x|\theta)}{\partial \theta_{v|\pi_v^d}} = p(x|\theta) I^{\pi_v^d}(x^d_{pa(v)}) \left(t(x_v) - \tau(\theta_{v|\pi_v^d})\right), \quad (7)$$

where

$$\tau(\theta_{v|\pi_v^d}) = \frac{\partial \phi(\theta_{v|\pi_v^d})}{\partial \theta_{v|\pi_v^d}}$$

and $I^{\pi_v^d}(x^d_{pa(v)})$ is the indicator function, which is one for $x^d_{pa(v)} = \pi_v^d$ and zero otherwise.

It is a well-known fact from exponential model theory that the derivative for the normalizing function equals the expected value of the canonical statistics (see, e.g., Schervish 1995). That is

$$\tau(\theta_{v|\pi_v^d}) = \mathbf{E}_{\theta_{v|\pi_v^d}}[t(X_v)].$$

We will later use this fact when deriving the gradient for specific distributions.

Now, by inserting (7) into (5) we get the following expression for the local gradient of the incomplete observation

$$\begin{aligned} \frac{\partial \log p(y|\theta)}{\partial \theta_{v|\pi_v^d}} &= \int_{x\in\mathcal{X}(y)} \frac{p(x|\theta)}{p(y|\theta)} I^{\pi_v^d}(x^d_{pa(v)}) \\ &\quad \times \left(t(x_v) - \tau(\theta_{v|\pi_v^d})\right) \mu(x). \end{aligned}$$

Finally, by applying the fact that

$$p(x|y,\theta) = \begin{cases} \frac{p(x|\theta)}{p(y|\theta)} & \text{for } x \in \mathcal{X}(y) \text{ and } p(y|\theta) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

we obtain the final expression for the local gradient

$$\begin{aligned} &\frac{\partial \log p(y|\theta)}{\partial \theta_{v|\pi_v^d}} \\ &\quad = \int p(x_{v\cup pa(v)}|y,\theta) I^{\pi_v^d}(x^d_{pa(v)}) \\ &\qquad \times \left(t(x_v) - \tau(\theta_{v|\pi_v^d})\right) \mu(x_{v\cup pa(v)}) \\ &\quad = \int p(x_v, x^c_{pa(v)}, \pi_v^d|y,\theta) \\ &\qquad \times \left(t(x_v) - \tau(\theta_{v|\pi_v^d})\right) \mu(x_v, x^c_{pa(v)}) \quad (8) \end{aligned}$$

where the last equation follows from integrating over all discrete parents and exploiting the fact that $I^{\pi_v^d}(x^d_{pa(v)})$ is one for $x^d_{pa(v)} = \pi_v^d$ and zero otherwise.

The incomplete-data log-likelihood gradient expression in (8) apply for any local exponential model. This generality makes the expression appear somewhat complicated. However, as we will see below, in Sections 4.2 and 4.3, the exponential model expression leads to simple expressions for the more specific local conditional multinomials and conditional Gaussians in the CG models.

### 4.1 Re-parameterization

The local gradient for a regular exponential model distribution is a step on the way in deriving the local gradients for the specific local distributions in CG models. For these specific distributions, we will, however, consider specific standard (non-exponential) parameterizations, as we will see in the subsections below. To

obtain the gradient with respect to a new local parameterization $\psi$, we apply the chain rule and multiply the derivative in (7) by the Jacobian $\partial\theta_{v|\pi_v^d}/\partial\psi$. Doing so, we obtain

$$\begin{aligned}\frac{\partial p(x|\theta)}{\partial\psi} &= \frac{\partial p(x|\theta)}{\partial\theta_{v|\pi_v^d}}\frac{\partial\theta_{v|\pi_v^d}}{\partial\psi} \\ &= p(x|\theta)I^{\pi_v^d}(x^d_{pa(v)}) \\ &\quad\times\left(t(x_v)-\tau(\theta_{v|\pi_v^d})\right)\frac{\partial\theta_{v|\pi_v^d}}{\partial\psi}.\end{aligned}$$

Performing the same operations that lead from (7) to (8) is trivial, and we finally obtain the expresion for the local gradient of the incomplete-data log-likelihood with repect to the re-parameterization

$$\begin{aligned}&\frac{\partial\log p(y|\theta)}{\partial\psi} \\ &= \int p(x_v, x^c_{pa(v)}, \pi_v^d|y,\theta) \\ &\quad\times\left(t(x_v)-\tau(\theta_{v|\pi_v^d})\right)\frac{\partial\theta_{v|\pi_v^d}}{\partial\psi}\mu(x_v, x^c_{pa(v)}). \quad (9)\end{aligned}$$

### 4.2 Conditional multinomial local gradient

Let us now take a look at the local gradient for the two specific types of local distributions in a CG model. First consider a *conditional multinomial* distribution for $p(X_v|\pi_v^d)$. As demonstrated in Thiesson (1997), we can obtain an exponential model representation for this distribution as follows. Let $s_0$ denote a value of reference for the discrete variable $X_v$ and let $s_+ = 1,\ldots,S$ be the remaining possible values for $X_v$. If we choose $s_0$ as any value for which $p(s_0|\pi_v^d) > 0$, we can represent the conditional multinomial distribution by an exponential model with probabilities of the form (2) by letting

$$\begin{aligned}\theta^{s_+} &= \log\left[p(s_+|\pi_v^d)/p(s_0|\pi_v^d)\right] \\ t^{s_+}(x_v) &= \begin{cases}1 & \text{for } x_v = s_+ \\ 0 & \text{otherwise}\end{cases} \\ \phi(\theta_{v|\pi_v^d}) &= \log\left(1+\sum_{s_+=1}^{S}\exp(\theta^{s_+})\right) \\ b(x_v) &= 1\end{aligned}$$

where $\theta_{v|\pi_v^d} = (\theta^1,\ldots,\theta^S)$ and $t(x_v) = (t^1(x_v),\ldots,t^S(x_v))$.

The expected value for the canonical statistics in the above exponential model representation is

$$\begin{aligned}\tau(\theta_{v|\pi_v^d}) &= \mathbf{E}_{\theta_{v|\pi_v^d}}[t(X_v)] \\ &= \sum_{x_v} t(x_v)p(x_v|x_{pa(v)},\theta_{v|\pi_v^d}) \\ &= \left(p^1,\ldots,p^S\right)\end{aligned}$$

where $p^{s_+} = p(X_v = s_+|\pi_v^d,\theta_{v|\pi_v^d})$.

We finally obtain the expression for the local gradient with respect to the exponential model parameterization by inserting the above expressions for $t(x_v)$ and $\tau(\theta_{v|\pi_v^d})$ into equation (8). The elements of this vector are

$$\begin{aligned}&\frac{\partial\log p(y|\theta)}{\partial\theta^{s_+}} \\ &= \int p(x_v,\pi_v^d|y,\theta)t^{s_+}(x_v)\mu(x_v) \\ &\qquad -\int p(x_v,\pi_v^d|y,\theta)p^{s_+}\mu(x_v) \\ &= p(s_+,\pi_v^d|y,\theta_{v|\pi_v^d}) \\ &\qquad -p(\pi_v^d|y,\theta_{v|\pi_v^d})p(s_+|\pi_v^d,\theta_{v|\pi_v^d}). \quad (10)\end{aligned}$$

We can now use the Lauritzen and Jensen (2001) propagation scheme for Bayesian networks with CG distributions to efficiently compute the quantities in (10). The propagation scheme enables us to efficiently compute posterior marginal distributions for any family $X_{v\cup pa(v)}$ given evidence $y$. The family for a conditional multinomial distribution in a CG model only involves discrete variables and the distribution is in this case simply represented by the marginal multinomial. The posterior probabilities $p(s_+,\pi_v^d|y,\theta_{v|\pi_v^d})$ and $p(\pi_v^d|y,\theta_{v|\pi_v^d})$ can therefore easily be extracted from this marginal distribution and, hence, a conditional multinomial local gradient can be efficiently computed.

The Lauritzen and Jensen (2001) propagation scheme utilizes the traditional parameterization for the conditional multinomial distribution. This representation is given by the conditional probabilities $(p^0,\cdots,p^S)$, where $p^0 = p(X_v = s_0|\pi_v^d,\theta_{v|\pi_v^d})$ and $p^{s_+}$ is defined as above. Hence, after we update the parameters for the exponential model representation during the line-search in a gradient based optimization method, we will need to switch back into the traditional representation – using (2) – in order to use the propagation scheme to compute the next gradient.

Switching between representations has a minor computational cost. On the other hand, performing the gradient optimization for parameters in the exponential model representation has the benefit that this parameterization automatically enforces the constraints $p^s \geq 0$ and $\sum_s p^s = 1$, which is not the case for gradient optimization using the traditional parameters.

We consider next the alternative gradient for the traditional parameter representation. In order to derive this gradient, we first derive the Jacobian from the exponential model representation to the traditional probability parameterization

$$\frac{\partial\theta_{v|\pi_v^d}}{\partial(p^0,\cdots,p^S)}$$

$$= \begin{bmatrix} -1/p^0 & 1/p^1 & 0 & \dots & 0 \\ & & \vdots & & \\ -1/p^0 & 0 & \dots & 0 & 1/p^S \end{bmatrix}.$$

By insertion into equation (9) we now obtain the local gradient with respect to the traditional representation. The $s^{th}$ $(s = 0, \dots, S)$ element in this gradient is given by

$$\frac{\partial \log p(y|\theta)}{\partial p^s} = \frac{p(s, \pi_v^d|y, \theta)}{p(s|\pi_v^d, \theta_{v|\pi_v^d})} - p(\pi_v^d|y, \theta). \quad (11)$$

Notice that the expression for this gradient differs slightly from the gradient expression in Binder *et al.* (1997) [Equation (4)].

Binder *et al.* (1997) ensure that the constraint $\sum_s p^s = 1$ is satisfied by using a standard method in which one projects the gradient onto the surface defined by this constraint. This method can be used for an optimization method based on our gradient in (11) as well. Still, however, both methods will have to ensure the constraint that $p^s \geq 0$ by inspecting the probability parameterization during a gradient update (i.e., a line-search).

### 4.3 Conditional Gaussian local gradient

Next, let us consider a *conditional Gaussian* (CG) local regression model for the continuous variable $X_v$ given the parents $X_{pa(v)} = (X^c_{pa(v)}, X^d_{pa(v)})$, where the conditioning parent set may contain continuous variables, $X^c_{pa(v)}$, as well as discrete variables, $X^d_{pa(v)}$. Recall that $\pi_v^d$ denotes a particular configuration of values for discrete parents, and to ease notation, we will in this section use $a$ (instead of $\pi_v^c$ or $x^c_{pa(v)}$) to denote a particular configuration of values for continuous parents. The CG regression model defines a set of linear regressions on the continuous parent variables – a regression for each configuration of distrete parent variables. See Lauritzen and Wermuth (1989) for more details on CG models. Let us now consider a particular distribution for $X_v$, given the values $a$ for continuous parents and the configuration of values for discrete parents $\pi_v^d$. The distribution is defined by the mean $\mu = c + \beta a'$ and variance $\sigma$, where $c$ and $\beta$ are respectively the intercept and the coefficients for the regression on continuous parents, and $'$ denotes transpose. Restricting attention to non-degenerate (or positive) Gaussians, where $\sigma > 0$, we can obtain an exponential model representation of the form (2) as follows

$$\begin{aligned} \theta_{v|\pi_v^d} = (\theta_1, \theta_2) &= \left(\frac{\mu}{\sigma}, -\frac{1}{2\sigma}\right) \\ &= \left(\frac{c + \beta a'}{\sigma}, -\frac{1}{2\sigma}\right) \\ t(x_v) &= (x_v, x_v^2) \\ \phi(\theta_{v|\pi_v^d}) &= -\frac{\theta_1^2}{4\theta_2} - \frac{1}{2}\log(-2\theta_2) \\ &= \frac{\mu^2}{2\sigma} + \frac{1}{2}\log\sigma \\ b(x_v) &= (2\pi)^{-1/2}. \end{aligned}$$

Notice that the restriction to positive Gaussians ($\sigma > 0$) ensures that the natural parameters for the exponential representation are defined.

The expected value of the canonical statistics is

$$\begin{aligned} \tau(\theta_{v|\pi_v^d}) &= \mathbf{E}_{\theta_{v|\pi_v^d}}[t(X_v)] \\ &= (\mu, \sigma + \mu^2) \\ &= \left(c + \beta a', \sigma + (c + \beta a')^2\right). \end{aligned}$$

Again, we can use the Lauritzen and Jensen (2001) propagation scheme to efficiently compute the gradient for the parameters of the exponential model. As with the conditional multinomial case, the inference scheme utilizes a parameterization for a conditional Gaussian, that is different from our exponential model. In the case of conditional multinomial models, we can easily switch between parameterizations. This allowed us to use a gradient method (e.g., a line-search) to update the exponential model parameters and then convert the resulting parameterization back into the propagation scheme parameters in order to compute the next gradient. However, in the case of conditional Gaussian models, the propagation scheme requires the parameters $(c, \beta, \sigma)$, and these parameters cannot be obtained from the parameters of the exponential model representation. We therefore compute the gradient for these parameters directly.

By inserting the expressions for $t(X_v)$ and $\tau(\theta_{v|\pi_v^d})$ into the equation (9) and by using the Jacobian

$$\frac{\partial \theta_{v|\pi_v^d}}{\partial(c, \beta, \sigma)} = \begin{bmatrix} \frac{1}{\sigma} & \frac{a}{\sigma} & -\frac{c+\beta a'}{\sigma^2} \\ 0 & 0 & \frac{1}{2\sigma^2} \end{bmatrix}$$

we can derive the following local gradient with respect to the parameterization $(c, \beta, \sigma)$. Let $\mu = c + \beta a'$, then

$$\begin{aligned} &\frac{\partial \log p(y|\theta)}{\partial(c, \beta, \sigma)} \\ &= \frac{\partial \log p(y|\theta)}{\partial \theta_{v|\pi_v^d}} \frac{\partial \theta_{v|\pi_v^d}}{\partial(c, \beta, \sigma)} \\ &= \int p(x_v, a, \pi_v^d|y, \theta) \\ &\qquad \times \begin{bmatrix} \frac{x_v - \mu}{\sigma} \\ \frac{(x_v - \mu)a}{\sigma} \\ \frac{(x_v - \mu)^2 - \sigma}{2\sigma} \end{bmatrix}' \mu(x_v, a) \qquad (12) \\ &= p(\pi_v^d|y, \theta) \int p(x_v, a|\pi_v^d, y, \theta) \\ &\qquad \times \begin{bmatrix} \frac{x_v - \mu}{\sigma} \\ \frac{(x_v - \mu)a}{\sigma} \\ \frac{(x_v - \mu)^2 - \sigma}{2\sigma} \end{bmatrix}' \mu(x_v, a) \qquad (13) \end{aligned}$$

where $p(x_v, a|\pi_v^d, y, \theta) = 0$ for values of discrete parents $\pi_v^d$ not consistent with the incomplete observation $y$. The step from (12) to (13) follows by factoring $p(x_v, a, \pi_v^d|y, \theta)$ into $p(x_v, a|\pi_v^d, y, \theta)$ and $p(\pi_v^d|y, \theta)$ and then pulling the discrete density out from under the integration.

Let $\overline{(x_v, a)}$ denote the expected value for the vector $(X_v, X^c_{pa(v)})$ with respect to the posterior Gaussian distribution for $(X_v, X^c_{pa(v)})$ given $\pi_v^d$ That is,

$$\begin{aligned}\overline{(x_v, a)} &= \mathbf{E}_{\theta_{v|\pi_v^d}}[X_v, X^c_{pa(v)}|y] \\ &= \int p(x_v, a|\pi_v^d, y, \theta)(x_v, a)\mu(x_v, a).\end{aligned}$$

Similarly, let $\overline{(x_v, a)'(x_v, a)}$ denote the expected value for the matrix $\left((X_v, X^c_{pa(v)})'(X_v, X^c_{pa(v)})\right)$. For instance, $\overline{x_v a} = \mathbf{E}_{\theta_{v|\pi_v^d}}[X_v X^c_{pa(v)}|y]$.

The expression for the local gradient in (13) then reduces to

$$\begin{aligned}\frac{\partial \log p(y|\theta)}{\partial c} &= p(\pi_v^d|y, \theta)\left(\overline{x_v} - \beta\overline{a} - c\right)/\sigma \\ \frac{\partial \log p(y|\theta)}{\partial \beta} &= p(\pi_v^d|y, \theta)\left(\overline{x_v a} - c\overline{a} - \beta\overline{a'a}\right)/\sigma \\ \frac{\partial \log p(y|\theta)}{\partial \sigma} &= p(\pi_v^d|y, \theta)\left(\overline{x_v x_v} - 2c\overline{x_v} - 2\beta\overline{x_v a'}\right. \\ &\quad \left. +\beta\overline{a'a}\beta' + 2c\beta\overline{a'} + c^2 - \sigma\right)/2\sigma^2.\end{aligned} \tag{14}$$

We can now use the Lauritzen and Jensen (2001) propagation scheme and this time efficiently compute the gradient for CG regression models. Recall that the propagation scheme allows us to efficiently compute the posterior marginal distribution for any family $X_{v\cup pa(v)}$. This distribution is represented as the product of a marginal distribution for discrete variables, $p(X^d_{pa(v)})$, and conditional distributions for the continuous variables given each cofiguration of the discrete variables in the family $p(X_v, X^c_{pa(v)}|X^d_{pa(v)})$. In many situation the (conditional) continuous distributions will be simple Gaussians, but they are, in general, mixtures of Gaussians. For a mixture of Gaussians, a so-called weakly marginalized distribution can be constructed. This weak marginal is the Gaussian distribution closest in Kullback-Leibler distance to the true (conditional) marginal – but more importantly for our purpose, the weak marginal matches the mean vector and covariance matrix for the true (conditional) marginal (see e.g. Lauritzen and Jensen, 2001). Hence, given a particular configuration for the discrete variables, $\pi_v^d$, the mean vector $\mu$ and covariance matrix $\Sigma$ for this conditional (weak) marginal Gaussian equals

$$\begin{aligned}\mu &= \overline{(x_v, a)} \\ \Sigma &= \overline{(x_v, a)'(x_v, a)} - \mu'\mu.\end{aligned}$$

The expected statistics on the right-hand side of (14) can therefore easily be extracted from the parameterization of the (weak) marginal distribution and hence, the gradient for a CG regression can be efficiently computed.

## 5 Parameter tying

Tying of parameters is an essential feature for some types of models, including, for example, models for stochastic temporal processes and pedigree analysis. We consider parameter tying that relaxes the global variation independence in (1) by assuming that the parameterization for the relationship between the variable $X_v$ and its conditional variables $X_{pa(v)}$ is the same across a set of variables. Let $\tilde{v} \subseteq V$ denote a such set of variables and let $\tilde{V}$ denotes all of such sets. We will let $\theta_{\tilde{v}}$ denote the tied parameterization across all $v \in \tilde{v}$. In this case, the model factorizes as

$$p(X|\theta) = \prod_{\tilde{v}\in\tilde{V}} \prod_{v\in\tilde{v}} p(X_v|X_{pa(v)}, \theta_{\tilde{v}}) \tag{15}$$

where $\Theta = \times_{\tilde{v}\in\tilde{V}}\Theta_{\tilde{v}}$. We call this type of tying for *global parameter tying*. Global parameter tying is, of course, only possible between conditional models that are similar. That is, for all $X_v$, where $v \in \tilde{v}$, the number of discrete and continuous conditioning parent variables must be the same and the set of possible state configurations for discrete parents must be the same. We will let $\pi^d_{\tilde{v}}$ denote a particular configuration of states for discrete parent variables. This configuration will be the same across all $v \in \tilde{v}$.

We are now seeking the incomplete-data log-likelihood with respect to the parameterization $\theta_{\tilde{v}|\pi^d_{\tilde{v}}}$. Similar to (6) and (7), we use the chain rule and exponential model representation to first compute the local gradient for a complete observation

$$\begin{aligned}\frac{\partial p(x|\theta_{\tilde{v}})}{\partial \theta_{\tilde{v}|\pi^d_{\tilde{v}}}} &= \sum_{v\in\tilde{v}} p(x|\theta)\frac{\partial \log p(x_v|x_{pa(v)}, \theta_{\tilde{v}|\pi^d_{\tilde{v}}})}{\partial \theta_{\tilde{v}|\pi^d_{\tilde{v}}}} \\ &= \sum_{v\in\tilde{v}} p(x|\theta) I^{\pi^d_{\tilde{v}}}(x_{pa(v)})\left(t(x_v) - \tau(\theta_{\tilde{v}|\pi^d_{\tilde{v}}})\right).\end{aligned}$$

We then obtain the expression for the local gradient for the incomplete data log-likelihood by the same steps, which lead to (8) and (9). Hence,

$$\begin{aligned}&\frac{\partial \log p(y|\theta)}{\partial \psi} \\ &= \sum_{\tilde{v}\in\tilde{V}}\sum_{v\in\tilde{v}} \int p(x_v, x^c_{pa(v)}, \pi_{\tilde{v}|\pi^d_{\tilde{v}}}|y, \theta)\left(t(x_v) - \tau(\theta_{\tilde{v}|\pi^d_{\tilde{v}}})\right) \\ &\qquad \times \frac{\partial \theta_{\tilde{v}|\pi^d_{\tilde{v}}}}{\partial \psi}\mu(x_v, x^c_{pa(v)}).\end{aligned} \tag{16}$$

Setting $\frac{\partial \theta_{\tilde{v}|\pi^d_{\tilde{v}}}}{\partial \psi} = 1$ gives us the expression for the gradient with respect to the natural parameters in the

exponential model representation.

Notice that the only difference between (9) and (16) is that the gradient in (16) adds the gradients computed at each $v \in \tilde{v}$. In other words, with global parameter tying, the gradient for the incomplete-data log-likelihood can be computed by proceeding as if parameters were not tied and then add up the gradients which are related by tying. That is,

$$\frac{\partial \log p(y|\theta)}{\partial \psi} = \sum_{v \in \tilde{v}} \frac{\partial \log p(y|\theta)}{\partial \psi_{v|\pi_v^d}} \quad (17)$$

where $\psi_{v|\pi_v^d}$ denotes the (artificial) non-tied parameterization for the local model, with $\psi_{v|\pi_v^d} = \psi$ for all $v \in \tilde{v}$.

For simplicity, we will only consider global parameter tying, as describe above. More sophisticated tying schemes are, of course, possible.

### 5.1 Stochastic ARMA models

The stochastic ARMA ($\sigma$ARMA) models of Thiesson *et al.* (2004) is an illustrative example of a stochastic temporal process, where tying of parameters plays an important role. $\sigma$ARMA models are closely related to the classic autoregressive moving average (ARMA) time-series models (see, e.g., Box, Jenkins, and Reinsel 1994 or Ansley 1979). As demonstrated in Thiesson *et al.* (2004), both the ARMA and $\sigma$ARMA models are naturally represented as graphical models with only continuous variables. The $\sigma$ARMA models differs from the ARMA models by replacing the deterministic component of an ARMA model with a Gaussian distribution having a small variance, as we will see below. This variation allow us to smooth the time series model in a controlled way.

A $\sigma$ARMA(p,q) time-series model is defined as follows. We denote a temporal sequence of continuous observation variables by $Y = (Y_1, Y_2, \ldots, Y_T)$. Time-series data is a sequence of values for these variables – some of which may be missing. The models associate a latent "white noise" variable with each observable variable. These latent variables are denoted $E = (E_1, E_2, \ldots, E_T)$.

The $\sigma$ARMA model is now defined by the conditional Gaussian distribution

$$Y_t | Y_{t-p}, \ldots, Y_{t-1}, E_{t-q}, \ldots, E_t \sim \mathcal{N}(\mu_t, \sigma) \quad (18)$$

where the functional expression for the mean $\mu_t$ and the variance $\sigma$ are shared across the observation variables. The variance is fixed at a given (small) value to be specified by the user. The mean is related to the conditional variables as follows

$$\mu_t = c + \sum_{j=0}^{q} \beta_j E_{t-j} + \sum_{i=1}^{p} \alpha_i Y_{t-i} \quad (19)$$

where $c$ is the intercept for the regression, $\sum_{i=1}^{p} \alpha_i Y_{t-i}$ is the autoregressive (AR) part, $\sum_{j=0}^{q} \beta_j E_{t-j}$ is the moving average (MA) part with $\beta_0$ fixed as 1, and $E_t \sim \mathcal{N}(0, \gamma)$ with $E_t$ mutually independent for all $t$. The model therefore involves the free parameters $c$, $(\alpha_1, \ldots, \alpha_p)$, $(\beta_1, \ldots, \beta_q)$, and $\gamma$. These parameters are tied across time steps.

From the above description, one may realize that an ARMA model is the limit of a $\sigma$ARMA model as $\sigma \to 0$. Letting $\sigma \to 0$ will in effect replace the conditional Gaussian distribution in (18) by a deterministic relation, where $Y_t$ equals the right-hand side of (19).

We are interested in computing the gradient for the *conditional log-likelihood model*, where we condition on the first $R = max(p, q)$ variables. Relations between variables for $t \leq R$ can therefore be ignored. The graphical representation for an $\sigma$ARMA(2,2) model is shown in Figure 1. It should be noted that if we artificially extend the time series back in time for R (unobserved) time steps, this model represents what is known in the literature as the *exact likelihood model*. There are alternative methods for dealing with the beginning of a time series (see, e.g., Box, Jenkins, and Reinsel 1994).

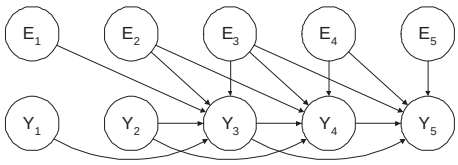


Figure 1: Graphical representation for $\sigma$ARMA(2,2) time-series model with five observations.

Let us first consider the variance parameter $\gamma$, which is tied across all the "white noise" variables $E_R, E_{R+1}, \ldots, E_T$. We intend to use (17) to compute the partial gradient for this parameter and will therefore first derive the expression for a partial gradient under the assumption that the $\gamma$ parameters for these variables are not tied. Notice that $\gamma$ will in this case take the place of the variance parameter $\sigma$ in all of the formulas in Section 4.3. Also, recall that the "white noise" variables do not have any parents, which means that there are no regression coefficients and hence no partial derivative with respect to $\beta$. Because the Gaussian distribution is restricted to have a mean of zero, we invoke the chain-rule once more and multiply the gradient expression in (14) by the Jacobian $[0\ 1]'$ going from the $(c, \gamma)$ parameterization to a parameterization, where $\gamma$ is the only parameter. Notice that because the Jacobian is constant with respect to the integral in (13), the parameter gradient can be computed by simply multiplying the Jacobian and equation (14). As expected, we obtain a partial gradient for the non-tied variance of $E_t$ that equals the partial derivative for the variance parameter in (14) – but it is not quite as

complicated because $c = 0$ and $E_t$ has no parents. Finally, by using (17), we arrive at the expression for the partial gradient with respect to the tied $\gamma$ parameter

$$\frac{\partial \log p(y|\theta)}{\partial \gamma} = \sum_{t=R+1}^{T} \frac{\overline{e_t e_t} - \gamma}{2\gamma^2}. \quad (20)$$

In a similar fashion, we can derive the gradient for the free parameters $c$, $(\alpha_1, \ldots, \alpha_p)$, and $(\beta_1, \ldots, \beta_q)$ associated with the conditional Gaussian distribution for the observation variables. As above, we apply the chain rule to achieve the gradient for the free parameters. Let $A_t = (Y_{t-p}, \ldots, Y_{t-1}, E_{t-q}, \ldots, E_t)$ denote all the parents for the observation variable $Y_t$ and let $Z_t = A_t \setminus E_t$ denote the parents except for the parent $E_t$ associated with the fixed regression coefficient $\beta_0$. We denote all of the regression coefficients by $\beta = (\alpha_1, \ldots, \alpha_p, \beta_0, \beta_1, \ldots, \beta_q)$ and let $\beta_{z_t} = (\alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_q)$ denote the free regression coefficients. The expression for the partial gradient for the tied $c$ and $\beta_{z_t}$ parameters now becomes

$$\frac{\partial \log p(y|\theta)}{\partial (c, \beta_{z_t})} = \sum_{t=R+1}^{T} \begin{bmatrix} (\overline{x_t} - \beta \overline{a_t} - c)/\sigma \\ \left(\overline{x_t z_t} - c\overline{z_t} - \beta_{z_t} \overline{z_t' z_t}\right)/\sigma \end{bmatrix}'.$$

## 6 Discussion and further work

In this paper, we derived the gradient for recursive exponential mixed models, a class of probabilistic models with both discrete and continuous variables. We demonstrated that positive conditional Gaussian models are a specific subclass of the REMMs and that one can use probabilistic inference to compute the parameter gradient for the incomplete-data likelihood for these models. As descried above, one can use this gradient to adapt the parameters in order to improve the incomplete-data likelihood and identify the MLE or local maxima of the likelihood. It is easy to extend this analysis to obtain similar results for MAP estimation by differentiating a prior with respect to the parameters of interest.

Alternative methods for learning parameters that do not directly compute the parameter gradient exist. For instance, the EM algorithm is a general method for improving parameters of a statistical model given incomplete data. In the context of graphical models, the E-step of the EM algorithm is accomplished via probabilistic inference in the graphical model (see, e.g., Lauritzen 1995 for a treatment of the EM algorithm for discrete graphical models). It should be noted, however, that in many situations, one can improve the speed of convergence of the EM algorithm through the use of the gradient (see, e.g., Thiesson 1995). Also, in some situations, the EM algorithm cannot be applied to improve the parameters of the model. In such situations a gradient method can often be used instead.

It is important to note that CG models that involve local degenerate conditional Gaussian models cannot be expressed as REMMs. The requirement for non-degenerate conditional Gaussians, where the variance $\sigma > 0$, can be seen by examining the exponential parameterization of the conditional Gaussian local model in Section 4.3. Unfortunately, some standard models can therefore not be naturally expressed as REMMs. For instance, the ARMA (a stochastic ARMA model in which the variance $\sigma$ is zero) cannot be represented as a CG model. It is an open question as to whether or not probabilistic inference can be used to efficiently compute the gradient for non-positive CG models.

Finally, the class of REMMs has the advantage over CG models of allowing discrete variables to have continuous parents. Given this advantage, it would be worthwhile to investigate efficient methods for computing the parameter gradients for REMMs (i.e., the quantity in equation 8).

## References


Ansley, C. F. (1979). An algorithm for the exact likelihood of a mixed autoregressive-moving average process. *Biometrika*, *66*, 59–65.

Binder, J., Koller, D., Russell, S. J., & Kanazawa, K. (1997). Adaptive probabilistic networks with hidden variables. *Machine Learning*, *29*, 213–244.

Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (1994). *Time series analysis*. New Jersey: Prentice Hall.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. London: Chapman and Hall.

Kiiveri, H., Speed, T. P., & Carlin, J. B. (1984). Recursive causal models. *Journal of the Australian Mathematical Society*, *36*, 30–52.

Lauritzen, S. L. (1995). The EM algorithm for graphical association models with missing data. *Computational Statistics & Data Analysis*, *19*, 191–201.

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

Lauritzen, S. L., & Jensen, F. (2001). Stable local computation with conditional gaussian distributions. *Statistics and Computing*, *11*, 191–203.

Lauritzen, S. L., & Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *The Annals of Statistics*, *17*, 31–57.

Schervish, M. J. (1995). *Theory of statistics*. New York: Springer-Verlag.

Thiesson, B. (1995). Accelerated quantification of Bayesian networks with incomplete data. *Proceedings of First International Conference on Knowledge Discovery and Data Mining* (pp. 306–311). AAAI Press.

Thiesson, B. (1997). Score and information for recursive exponential models with incomplete data. *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence* (pp. 453–463). Morgan Kaufmann Publishers.

Thiesson, B., Chickering, D. M., Heckerman, D., & Meek, C. (2004). ARMA time-series modeling with graphical models. *Proceedings of the Twentieth Conference on Uncertainty in Artificial Intelligence* (pp. 552–560). AUAI Press.