Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 51 (2006) 699–709
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

# Automatic approximation of the marginal likelihood in non-Gaussian hierarchical models

Hans J. Skaug[a,*], David A. Fournier[b]

[a] *Department of Mathematics, University of Bergen, Johannes Brunsgate 12, 5008 Bergen, Norway*
[b] *Otter Research Ltd., P.O. Box 2040 Sidney, Canada V8L 3S3*

Received 3 January 2006; received in revised form 16 March 2006; accepted 16 March 2006
Available online 17 April 2006

**Abstract**

Fitting of non-Gaussian hierarchical random effects models by approximate maximum likelihood can be made automatic to the same extent that Bayesian model fitting can be automated by the program BUGS. The word "automatic" means that the technical details of computation are made transparent to the user. This is achieved by combining a technique from computer science known as "automatic differentiation" with the Laplace approximation for calculating the marginal likelihood. Automatic differentiation, which should not be confused with symbolic differentiation, is mostly unknown to statisticians, and hence basic ideas and results are reviewed. The computational performance of the approach is compared to that of existing mixed-model software on a suite of datasets selected from the mixed-model literature.


*Keywords:* AD Model Builder; Automatic differentiation; Importance sampling; Laplace approximation; Mixed models; Random effects

## 1. Introduction

Hierarchical models have gained widespread use in statistics during the last few decades (Hobert, 2000). It is the invention of new computational algorithms, as well as the increased speed of computers, that has made the use of such models practical. Bayesians and frequentists agree that the heart of the computational problem is the numerical evaluation of a high-dimensional integral. The Laplace approximation has been used both for calculating the normalizing constant of Bayesian posterior distributions (Tierney and Kadane, 1986) and to obtain the (marginal) likelihood in frequentist random effects models (Breslow and Lin, 1995). For complex hierarchical models, calculating by hand the second order derivatives involved in the Laplace approximation is both tedious and error prone. Our main message is that this burden can be lifted from the shoulders of the statistician by the use of a technique called "automatic differentiation" (Griewank, 2000).

The present paper deals with the computational aspects of the Laplace approximation. Let $\boldsymbol{u}$ be a vector of latent random variables, let $\boldsymbol{\theta}$ be a vector of parameters, and let $g(\boldsymbol{u}, \boldsymbol{\theta})$ be their joint penalized loglikelihood (see Section 2).

* Corresponding author. Tel.: +47 55584861; fax: +47 55589672.
*E-mail addresses:* skaug@mi.uib.no (H.J. Skaug), otter@otter-rsch.com (D.A. Fournier)
*URL:* http://otter-rsch.com (D.A. Fournier).

0167-9473/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2006.03.005

Assume that the function $g(\boldsymbol{u}, \boldsymbol{\theta})$ is such that

$$\widehat{\boldsymbol{u}}(\boldsymbol{\theta}) = \underset{\boldsymbol{u}}{\operatorname{argmax}}\ g(\boldsymbol{u}, \boldsymbol{\theta}) \tag{1}$$

and

$$\boldsymbol{H}(\boldsymbol{\theta}) = \frac{\partial^2}{\partial \boldsymbol{u}^2} g(\boldsymbol{u}, \boldsymbol{\theta})\Big|_{\boldsymbol{u}=\widehat{\boldsymbol{u}}(\boldsymbol{\theta})} \tag{2}$$

are well defined on the range of $\boldsymbol{\theta}$. Our goal is to derive an efficient algorithm for maximizing the Laplace approximation

$$L^*(\boldsymbol{\theta}) = |\det\{\boldsymbol{H}(\boldsymbol{\theta})\}|^{-1/2} \exp\left[g\left\{\widehat{\boldsymbol{u}}(\boldsymbol{\theta}), \boldsymbol{\theta}\right\}\right] \tag{3}$$

with respect to $\boldsymbol{\theta}$. A key part of the problem is numerical evaluation of $\boldsymbol{H}(\boldsymbol{\theta})$ by automatic differentiation.

What is automatic differentiation (AD), or "algorithmic differentiation" as it is sometimes called? Somewhat simplistically we may think AD as a black-box that, given as input a computer code for evaluating a function $g(\boldsymbol{u})$, produces a new program which evaluates numerically the derivatives of $g(\boldsymbol{u})$. Hence, AD differs from symbolic differentiation, as performed by for instance the programs Mathematica and Maple, which aims at producing analytic derivative formulae. Further, the fact that the numerical derivatives produced with AD are accurate to machine precession distinguishes AD from the method of "finite differences". More details about AD are given in Section 3.

Direct maximization of the loglikelihood $l^*(\boldsymbol{\theta}) = \log\ L^*(\boldsymbol{\theta})$ is in general impossible, so we must resort to numerical optimization. Accurate derivative information greatly helps numerical optimization algorithms in locating the optimum of $l^*(\boldsymbol{\theta})$. As noted by several authors (Raudenbush et al., 2000; Bell, 2001), the expression for the gradient of $l^*(\boldsymbol{\theta})$ involves up to third order partial derivatives of $g(\boldsymbol{u}, \boldsymbol{\theta})$. We shall device a scheme for evaluating these derivatives by AD. Skaug (2002) employed so-called reverse-mode AD in the context of the Laplace approximation, but used a hand-coded version of $\boldsymbol{H}(\boldsymbol{\theta})$, and AD was used only to obtain the gradient of $l^*(\boldsymbol{\theta})$.

Bayesian hierarchical models have recently been made available to a large group of research workers through the appearance of the software system BUGS. According to Gilks et al. (1994), one of the original goals for the BUGS project was to create a system that could "accommodate a very large class of models". As a result, today we see BUGS being used in a wide range of scientific disciplines. A second goal for BUGS was that it should be "automatic", in the sense of hiding the technical details of the (MCMC) computations used to sample from the posterior distribution. Software with the same level of flexibility and automatization as offered by BUGS is lacking in the frequentist domain, although the SAS procedure NLMIXED offers some degree of flexibility in the formulation of nonlinear mixed-effects models. In the present paper we explore how AD, in conjunction with the Laplace approximation, can be used to fulfill the goals from Gilks et al. (1994) in a frequentist setting, where $\boldsymbol{\theta}$ is estimated by maximum likelihood.

In certain situations the Laplace approximation $L^*(\boldsymbol{\theta})$ is inaccurate, and as a consequence, the resulting estimator of $\boldsymbol{\theta}$ may differ from the true maximum likelihood estimator (Breslow and Lin, 1995). To diagnose such instances, and to improve the approximation, we use the Laplace importance sampling method (Kuk, 1999; Skaug, 2002). To this end we consider the following generalization of the objective function (3):

$$L^{\dagger}(\boldsymbol{\theta}) = F\left\{\widehat{\boldsymbol{u}}(\boldsymbol{\theta}), \boldsymbol{H}(\boldsymbol{\theta}), \boldsymbol{\theta}\right\}, \tag{4}$$

where $F(\boldsymbol{u}, \boldsymbol{H}, \boldsymbol{\theta})$ is a smooth function. Clearly, the Laplace approximation is a special case of this obtained by putting $F(\boldsymbol{u}, \boldsymbol{H}, \boldsymbol{\theta}) = |\det(\boldsymbol{H})|^{-1/2} \exp\{g(\boldsymbol{u}, \boldsymbol{\theta})\}$.

The methods developed in the present paper have been implemented as a module for the software package AD Model Builder (Fournier, 2001). Given as input C++ code for $g(\boldsymbol{u}, \boldsymbol{\theta})$, the system automatically maximizes $L^*(\boldsymbol{\theta})$. In addition to explaining the computational methodology, our goal is to make comparison to existing mixed-model software with respect to computational speed and numerical accuracy. The rest of the paper is organized as follows: Section 2 sets up the framework for hierarchical models, Section 3 outlines the computational methods, while examples are studied in Section 4. Section 5 provides some concluding remarks.

## 2. Hierarchical models

Let $\boldsymbol{y} = (y_1, \ldots, y_n)$ be a vector of observations, and let $\boldsymbol{u} = \left(u_1, \ldots, u_q\right)$ be a vector of latent random variables (random effects) influencing the value of $\boldsymbol{y}$. The conditional density of $\boldsymbol{y}$ given $\boldsymbol{u}$ is denoted by $f(\boldsymbol{y}|\boldsymbol{u})$, and the marginal