Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Econometrics 128 (2005) 301–323
JOURNAL OF Econometrics
www.elsevier.com/locate/econbase

Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects ☆

Sophia Rabe-Hesketh[a,*], Anders Skrondal[b], Andrew Pickles[c]

[a]*Graduate School of Education, University of California, 3659 Tolman Hall, Berkeley, CA 94720, USA*
[b]*Biostatistics Group, Division of Epidemiology, Norwegian Institute of Public Health, Oslo, Norway*
[c]*School of Epidemiology and Health Sciences & CCSR, The University of Manchester, UK*

Received 28 March 2002; received in revised form 15 July 2003
Available online 19 October 2004

Abstract

Gauss–Hermite quadrature is often used to evaluate and maximize the likelihood for random component probit models. Unfortunately, the estimates are biased for large cluster sizes and/or intraclass correlations. We show that adaptive quadrature largely overcomes these problems. We then extend the adaptive quadrature approach to general random coefficient models with limited and discrete dependent variables. The models can include several nested random effects (intercepts and coefficients) representing unobserved heterogeneity at different levels of a hierarchical dataset. The required multivariate integrals are evaluated efficiently using spherical quadrature rules. Simulations show that adaptive quadrature performs well in a wide range of situations.

JEL classification: C1; C8

Keywords: Random effects; Random coefficients; Multilevel models; Hierarchical models; Numerical integration; Adaptive quadrature; Spherical quadrature rules; GLLAMM

☆This paper was completed while the first author was employed by the Institute of Psychiatry, King's College London.

*Corresponding author.
E-mail addresses: sophiarh@berkeley.edu (S. Rabe-Hesketh), anders.skrondal@fhi.no (A. Skrondal), andrew.pickles@man.ac.uk (A. Pickles).

0304-4076/$ - see front matter © 2004 Published by Elsevier B.V.
doi:10.1016/j.jeconom.2004.08.017

1. Introduction

We consider novel approaches to maximum likelihood estimation of random effects models for limited and discrete dependent variables based on numerical integration. The simplest model includes a single random component or intercept that varies between clusters of observations and induces dependence within these clusters. Random effects models are useful for modeling panel data or grouped cross-sectional data where the responses for the same person or group cannot be assumed to be independent after conditioning on exogenous variables. In the grouped cross-sectional case the groups or clusters could be for instance households, firms or geographical entities. Multilevel or hierarchical models accommodate more than one level of clustering, an example being panel data with time-points (level 1) nested in individuals (level 2) who are nested in firms (level 3). Nested random intercepts at the firm and individual levels can then be used to model unobserved heterogeneity between firms and between individuals within firms. The firm-level random intercept induces dependence among individuals in the same firm and the individual-level random intercept induces additional dependence among observations on the same individual. Random coefficients can be included to model unobserved heterogeneity in the effects of variables between firms and/or individuals. Recent publications on random effects and multilevel models in economics and econometrics include Antweiler (2001), Baltagi et al. (2001), Beron et al. (1999), Blundell and Windmeijer (1997), Cardoso (2000), Carey (2000), Davis (2002) and Rice and Jones (1997). We also refer to Baltagi (2001) and Hsiao (2003) for discussions of multilevel models.

In limited and discrete dependent variable models with normally distributed random effects, the marginal likelihood generally does not have a closed form. A standard approach to parameter estimation is therefore to evaluate the marginal likelihood numerically using Gauss–Hermite quadrature. For two-level random component (also called random intercept) binary probit models, this approach is often attributed to Butler and Moffitt (1982) although it was introduced earlier for closely related models by Bock and Lieberman (1970).

Gaussian quadrature tends to work well with moderate cluster sizes as typically found in panel data. However with large cluster sizes, which are common in grouped cross-sectional data, the estimates become biased. This problem was pointed out recently by Borjas and Sueyoshi (1994) and Lee (2000) for probit models, by Albert and Follmann (2000) for Poisson models and by Lesaffre and Spiessens (2001) for logit models. Lee (2000) attributes the poor performance of quadrature to numerical underflow and develops an algorithm to overcome this problem. For probit models his algorithm works well in simulations with clusters as large as 100 when the intraclass correlation is 0.3 but produces biased estimates when the correlation is increased to 0.6. A likely reason for this is that for large clusters and high intraclass correlations, the integrands of the cluster contributions to the likelihood have very sharp peaks that may be located between adjacent quadrature points. Albert and Follmann (2000) and Lesaffre and Spiessens (2001) illustrate this problem for Poisson and logit models,

respectively. Naylor and Smith (1982) suggest a solution to a similar problem encountered in Bayesian statistics where numerical integration is often used to compute posterior densities. Essentially, the solution consists of scaling and translating the quadrature locations to place them under the peak of the integrand. A slightly different version of this *adaptive* quadrature approach has been suggested by Liu and Pierce (1994).

In this paper we initially describe and implement Naylor and Smith's version of adaptive quadrature for random component probit models. In a simulation study we show that, in contrast to the method suggested by Lee (2000), adaptive quadrature provides unbiased estimates for random component probit models with clusters as large as 500 and intraclass correlations as high as 0.9. Even for smaller cluster sizes and intraclass correlations, where ordinary quadrature is adequate, adaptive quadrature is superior since it requires fewer quadrature points. We extend the estimation method to models including (1) nested random effects and (2) random coefficients in addition to random intercepts. Although adaptive quadrature has previously been implemented for generalized linear mixed models with a single level of clustering (Pinheiro and Bates, 1995) and for multidimensional probit item factor analysis (Bock and Schilling, 1997), this is to our knowledge the first generalization for multilevel models. We carry out simulations to assess the performance of adaptive quadrature in the multilevel setting.

For models including random coefficients, the likelihood involves multidimensional integrals which are usually evaluated using cartesian product quadrature (e.g. Bock and Aitkin, 1981; Lillard, 1993). We suggest using spherical quadrature rules specifically designed for integrating over multivariate normal densities (Stroud, 1971) since these rules require fewer quadrature points to achieve a given accuracy. Simulations are carried out to assess the performance of adaptive quadrature using spherical rules.

2. Estimation using adaptive and spherical quadrature

In Section 2.1 we describe adaptive quadrature for random component binary probit models. In Section 2.2 we extend adaptive quadrature to multilevel random coefficient models. Here cartesian product quadrature is used to evaluate multivariate integrals. Section 2.3 describes spherical quadrature rules as a more efficient alternative to cartesian quadrature. Section 4 shows how the methods are applied to models with other types of discrete and limited dependent variables.

2.1. Adaptive quadrature for random component probit models

The random component binary probit model can be written as

$$y_{ij}^* = x_{ij}'\beta + u_j + \varepsilon_{ij},$$
$$y_{ij} = \mathrm{I}(y_{ij}^* > 0),$$

where $i = 1, \ldots, n_j$ indexes the individual observations, $j = 1, \ldots, N$ indexes clusters of observations, x_{ij} is a vector of explanatory variables, β is a vector of corresponding regression coefficients, u_j is the random intercept for cluster j and ε_{ij} is an error term. In a panel data setting, i is a time-point, j an individual and u_j represents time constant unobserved heterogeneity in the behavior of the individual which renders his or her n_j observations correlated. The random terms u_j and ε_{ij} are mutually independent, $u_j \sim N(0, \sigma^2)$ and $\varepsilon_{ij} \sim N(0, 1)$ and independent of the explanatory variables x_{ij}. The residual intraclass correlation for the underlying responses is

$$\rho \equiv \mathrm{Cor}(y^*_{ij}, y^*_{i'j} | x_{ij}, x_{i'j}) = \frac{\sigma^2}{1 + \sigma^2}.$$

The likelihood contribution of the jth cluster is a multivariate integral over the correlated total error terms $u_j + \varepsilon_{ij}$, $i = 1, \ldots, n_j$. Using an idea at least known since Dunnett and Sobel (1955), Bock and Lieberman (1970) and Butler and Moffitt (1982) simplify this integral to a univariate integral by exploiting the fact that the error terms are conditionally independent given the random effect. For a given cluster j, the likelihood contribution therefore is

$$f^{(2)}_j(\theta) = \int g(u_j; 0, \sigma^2) \prod_{i=1}^{n_j} f^{(1)}_{ij}(\theta | u_j) \mathrm{d}u_j, \tag{1}$$

where θ is the vector of all parameters, $g(\cdot; \mu, \sigma^2)$ is the normal density with mean μ and variance σ^2 and $f^{(1)}_{ij}(\theta | u_j)$ is the conditional likelihood contribution of unit ij given the random effect,

$$f^{(1)}_{ij}(\theta | u_j) = y_{ij} \Phi(\eta_{ij}) + (1 - y_{ij}) \Phi(-\eta_{ij}), \tag{2}$$

where Φ is the standard normal cumulative distribution function and η_{ij} is the linear predictor

$$\eta_{ij} = x'_{ij}\beta + u_j.$$

The integral, which cannot be solved analytically, can instead be evaluated numerically using Gauss–Hermite quadrature (see e.g., Stroud and Secrest, 1966). Instead of integrating over u_j, we will integrate over $v_j = u_j/\sigma$ with standard normal density $\phi(v_j)$. The approximation then is

$$f^{(2)}_j(\theta) = \int \phi(v_j) \prod_{i=1}^{n_j} f^{(1)}_{ij}(\theta | v_j) \mathrm{d}v_j \approx \sum_{r=1}^{R} p_r \prod_{i=1}^{n_j} f^{(1)}_{ij}(\theta | a_r), \tag{3}$$

where $\sqrt{\pi} p_r$ and $a_r/\sqrt{2}$ are the weights and locations of R point Gaussian quadrature for integrals of the form $\int \exp(-x^2) f(x) \mathrm{d}x$. The method is exact if $f(x)$ is a polynomial of degree up to $2R - 1$.

In the context of Bayesian inference, Naylor and Smith (1982) suggest an improved integration method that is *adaptive* in the sense that it takes into account the properties of the integrand $\phi(v_j) \prod_{i=1}^{n_j} f^{(1)}_{ij}(\theta | v_j)$. Note that the integrand is the product of the 'prior' density of v_j and the joint probability of the responses given v_j

which, after normalization with respect to v_j, is just the 'posterior' density of v_j given the observed responses. According to the Bayesian central limit theorem (e.g., Carlin and Louis, 2000, p. 122–124), posterior densities are approximately normal for large sample sizes, corresponding to large cluster sizes n_j in this application. If μ_j and τ_j^2 are the mean and variance of the posterior density, we would therefore expect the ratio $\phi(v_j)\prod_{i=1}^{n_j} f_{ij}^{(1)}(\theta|v_j)/g(v_j;\mu_j,\tau_j^2)$ to be well approximated by a low-degree polynomial. Writing the integral as

$$f_j^{(2)}(\theta) = \int g(v_j;\mu_j,\tau_j^2)\left(\frac{\phi(v_j)\prod_{i=1}^{n_j} f_{ij}^{(1)}(\theta|v_j)}{g(v_j;\mu_j,\tau_j^2)}\right)\mathrm{d}v_j,$$

changing the variable of integration from v_j to $z_j = (v_j - \mu_j)/\tau_j$ and applying the standard quadrature rule yields

$$f_j^{(2)}(\theta) \approx \sum_{r=1}^{R} \pi_{jr} \prod_{i=1}^{n_j} f_{ij}^{(1)}(\theta|\alpha_{jr}), \tag{4}$$

where

$$\alpha_{jr} = \mu_j + \tau_j a_r, \tag{5}$$

$$\pi_{jr} = \sqrt{2\pi}\tau_j \exp(a_r^2/2)\phi(\mu_j + \tau_j a_r)p_r. \tag{6}$$

Pinheiro and Bates (1995) point out that this approach is essentially a deterministic version of importance sampling with $g(v_j;\mu_j,\tau_j^2)$ as importance density. The advantage of adaptive quadrature can be seen in Fig. 1 which illustrates for $R = 5$ how adaptive quadrature translates and scales the locations so that they lie directly under the integrand.

The posterior means and standard deviations required for adaptive quadrature are themselves obtained using adaptive quadrature so that the integration is iterative. Using starting values $\mu_j^0 = 0$ and $\tau_j^0 = 1$ to define α_{jr}^0 and π_{jr}^0, the posterior means and

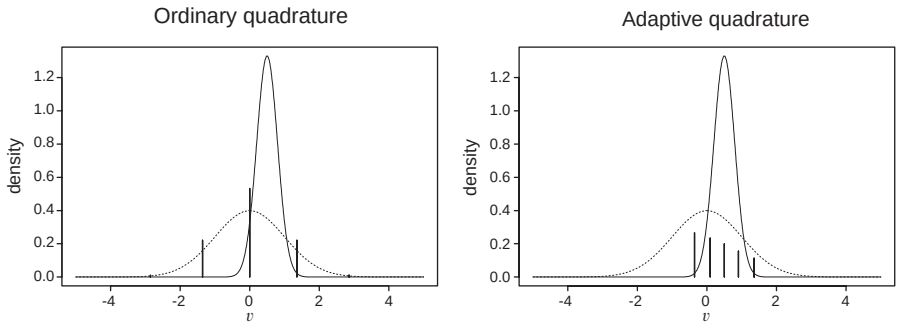

Fig. 1. Prior (dotted curve) and posterior (solid curve) densities and quadrature weights (bars) for ordinary and adaptive quadrature. The integrand is proportional to the posterior density.

standard deviations are updated in the kth iteration using

$$f_j^{(2)k}(\theta) = \sum_{r=1}^{R} \pi_{jr}^{k-1} \prod_{i=1}^{n_j} f_{ij}^{(1)}(\theta|\alpha_{jr}^{k-1}),$$

$$\mu_j^k = \frac{\sum_{r=1}^{R} (\alpha_{jr}^{k-1})\pi_{jr}^{k-1} \prod_{i=1}^{n_j} f_{ij}^{(1)}(\theta|\alpha_{jr}^{k-1})}{f_j^{(2)k}(\theta)},$$

$$\tau_j^k = \sqrt{\frac{\sum_{r=1}^{R} (\alpha_{jr}^{k-1})^2 \pi_{jr}^{k-1} \prod_{i=1}^{n_j} f_{ij}^{(1)}(\theta|\alpha_{jr}^{k-1})}{f_j^{(2)k}(\theta)} - (\mu_j^k)^2}, \tag{7}$$

followed by evaluation of α_{jr}^k and π_{jr}^k using (5) and (6). This sequence is repeated until convergence. A similar iterative algorithm is described in another context by Naylor and Smith (1988). The algorithm can converge very slowly or fail to converge if insufficient quadrature points are used to evaluate the posterior moments accurately, giving a useful warning that the approximation is poor.

Liu and Pierce (1994) describe an integration method based on a first order Laplace approximation (Tierney and Kadane, 1986) where μ_j is the mode of the integrand and τ_j is the standard deviation of the normal density approximating the integrand at the mode. Pinheiro and Bates (1995) use this method in the context of two-level random coefficient models. An advantage of their approach is that μ_j and τ_j do not themselves rely on the quadrature approximation so that an iterative process of the kind described above is not required. However, the method is also computationally demanding since numerical optimization and differentiation are required to determine μ_j and τ_j for each cluster. In addition, the posterior mean and standard deviation may better reflect the shape of the integrand when its tails are heavier than that of a normal density. Most importantly for our purposes, the first order Laplace approximation cannot be readily extended to multilevel problems as we will see in the next section. Both methods are of course equivalent if the posterior distribution is normal.

So far we have only addressed the problem of evaluating the marginal likelihood for given parameter values θ. The next problem is to maximize this marginal likelihood with respect to θ. Bock and Aitkin (1981) and others use Gaussian quadrature within an EM algorithm. We use a Newton–Raphson algorithm where the Hessian is obtained by numerical differentiation. Interestingly, numerical derivatives may be more accurate than numerically integrated analytical derivatives since the integrals for the derivatives are often very poorly approximated by quadrature or adaptive quadrature (Lesaffre and Spiessens, 2001). Numerical differentiation requires repeated evaluation of the marginal likelihood in the neighborhood of the 'current' parameter values. We do not update the quadrature locations and weights for each of these evaluations but keep them fixed for a full iteration of the Newton–Raphson procedure. The algorithm alternates between a step of Newton–Raphson to update the parameter values and the set of iterations in (7) to update the quadrature locations and weights. The reasons for not updating the quadrature locations and weights during numerical differentiation are that it would

be computationally demanding and that large changes in these quantities could make the likelihood surface appear discontinuous.

2.2. Adaptive quadrature for multilevel random coefficient models

A general three-level random coefficient model can be written as

$$\eta_{ijk} = x'_{ijk}\beta + x^{(2)\prime}_{ijk}u^{(2)}_{jk} + x^{(3)\prime}_{ijk}u^{(3)}_{k}, \tag{8}$$

where i, j and k index the units at levels 1, 2 and 3, respectively (e.g. time-points in individuals in firms), $x'_{ijk}\beta$ is the fixed effects part, $x^{(2)}_{ijk}$ is a vector of explanatory variables with random effects $u^{(2)}_{jk}$ at level 2 and $x^{(3)}_{ijk}$ is a vector of explanatory variables with random effects $u^{(3)}_{k}$ at level 3. The random effects at a given level have a multivariate normal distribution and the random effects at different levels are mutually independent and independent of the residual error term ε_{ijk} and explanatory variables. The general L level version of this model can be written as

$$\eta = x'\beta + \sum_{l=2}^{L} x^{(l)\prime}u^{(l)},$$

where subscripts are omitted to simplify notation. The marginal log-likelihood is

$$L(\theta) = \sum \ln f^{(L)}(\theta),$$

where $f^{(L)}(\theta)$ is the likelihood contribution of a unit at the highest level L. Let $U^{(l)} = (u^{(l)\prime}, \ldots, u^{(L)\prime})'$ for $l \leqslant L$. Exploiting conditional independence among level-$(l-1)$ units given the random effects $U^{(l)}$ at levels l and above, the likelihood contribution of a given level-l unit can be obtained recursively as

$$\begin{aligned} f^{(l)}(\theta|U^{(l+1)}) &= \int g(u^{(l)}; 0, \Sigma^{(l)}) \prod f^{(l-1)}(\theta|U^{(l)}) \mathrm{d}u^{(l)}, \qquad l = 2, \ldots, L-1 \\ f^{(L)}(\theta) &= \int g(u^{(L)}; 0, \Sigma^{(L)}) \prod f^{(L-1)}(\theta|u^{(L)}) \mathrm{d}u^{(L)}, \end{aligned} \tag{9}$$

where $f^{(1)}(\theta|U^{(2)})$ is the conditional level-1 likelihood contribution given in (2) for binary probit models (see Section 4 for other response models), $g(u^{(l)}; 0, \Sigma^{(l)})$ is the multivariate normal density of $u^{(l)}$ with covariance matrix $\Sigma^{(l)}$ and the product is over all level-$(l-1)$ units within the level-l unit as shown explicitly for a two-level model in (1).

Instead of integrating over the correlated random effects $u^{(l)}$, we will integrate over independent standard normal variables $v^{(l)}$ with

$$u^{(l)} = Q^{(l)}v^{(l)}, \tag{10}$$

where $Q^{(l)}$ is the Cholesky decomposition of $\Sigma^{(l)}$. Letting $V^{(l)} = (v^{(l)\prime}, \ldots, v^{(L)\prime})'$, the integral over the $M^{(l)}$ random effects at level l can then be approximated by cartesian

product quadrature,

$$f^{(l)}(\theta|V^{(l+1)}) = \int \phi(v_M)\ldots\int \phi(v_1)\prod f^{(l-1)}(\theta|v_1,\ldots,v_M,V^{(l+1)})\mathrm{d}v_1\cdots\mathrm{d}v_M$$
$$\approx \sum_{r_M} p_{r_M}\cdots\sum_{r_1} p_{r_1}\prod f^{(l-1)}(\theta|a_{r1},\ldots,a_{rM},V^{(l+1)}), \quad (11)$$

where we have omitted the (l) superscript for M and the variables being integrated over and will continue to do so in the remainder of this section.

We can improve the approximation by using adaptive quadrature. Although the multivariate integrals in (11) are evaluated as nested sets of univariate integrals, first over v_1, then over v_2, up to v_M, we cannot simply apply the adaptive quadrature rule in (5) and (6) to each univariate integral. This is because when integrating over a given v_m, the integrand is proportional to the posterior density of v_m conditional on all random effects not yet integrated over, i.e. v_{m+1} to v_M and all higher level random effects. Since the random effects will generally have non-zero posterior correlations, we would therefore require the *conditional* posterior moments of v_m given all random effects not yet integrated over. We can simplify the problem considerably by transforming to a new set of random effects with zero posterior correlations so that the *marginal* moments can be used. Naylor and Smith (1988) discuss this problem in a Bayesian context and suggest the orthogonalizing transformation

$$w_1 = v_1,$$
$$w_s = v_s + \sum_{t=1}^{s-1}\gamma_{st}w_t, \quad s = 2,\ldots,S \quad (12)$$

with

$$\gamma_{st} = -\mathrm{cov}(v_s,w_t)/\mathrm{var}(w_t),$$

where we have omitted the (l) superscript and let v_s denote the sth of all random effects (in some order) with $s = 1,\ldots,S$ and $S = \sum_l M^{(l)}$. The transformation has unit Jacobian.

The sequence of transformations therefore starts with random effects z_s with zero posterior means and covariances and unit posterior variances which are evaluated at the Gauss–Hermite quadrature locations a_r, $r = 1,\ldots,R$. These random effects are rescaled to $w_s = \mu_s + \tau_s z_s$, giving the adaptive quadrature locations for univariate integration, α_{sr} in (5), and transformed to v_s via (12). The adaptive quadrature locations for multivariate integration are therefore given by

$$A_{sr} = \alpha_{sr} - \sum_{t=1}^{s-1}\gamma_{st}\,\alpha_{tr}$$

with corresponding weights

$$P_{sr} = \sqrt{2\pi}\,\tau_s\,\exp(a_r^2/2)\phi(A_{sr})p_r.$$

The weights P_{sr} for the sth random effect depend on A_{sr} and hence on the locations α_{tr} of all preceding random effects $t<s$. In order to keep the weights of higher level

effects constant when integrating over the lower level effects, the v_s should be ordered from the highest to lowest level, the ordering within a level being arbitrary. For two random effects, the transformation from (z_1, z_2) to (v_1, v_2) and hence from (a_{r_1}, a_{r_2}) to (A_{1r_1}, A_{2r_2}) is illustrated in the first row of Fig. 2. It is clear that, for given posterior means and standard deviations, adaptive quadrature will be particularly superior to ordinary quadrature when the variables v_s have marked posterior correlations. Note that we would expect substantial negative posterior correlations between random intercepts at different levels since the effect (on the posterior distribution) of increasing the higher level random intercept can to some degree be counteracted by decreasing the lower level one and vice versa.

The γ_{st} required for the transformations in (12) as well as the posterior moments μ and τ of w can be obtained from the posterior means, variances and covariances of v. For given adaptive quadrature locations and weights, the algorithm computes the marginal likelihood and posterior moments of v recursively from level 2 to L. The terms evaluated at a given level l are displayed in Table 1.

After evaluating all terms up to level L, the posterior variances and covariances are found using

$$\mathrm{cov}(v_m^{(k)} v_n^{(l)}) = E[v_m^{(k)} v_n^{(l)}] - E[v_m^{(k)}]E[v_n^{(l)}].$$

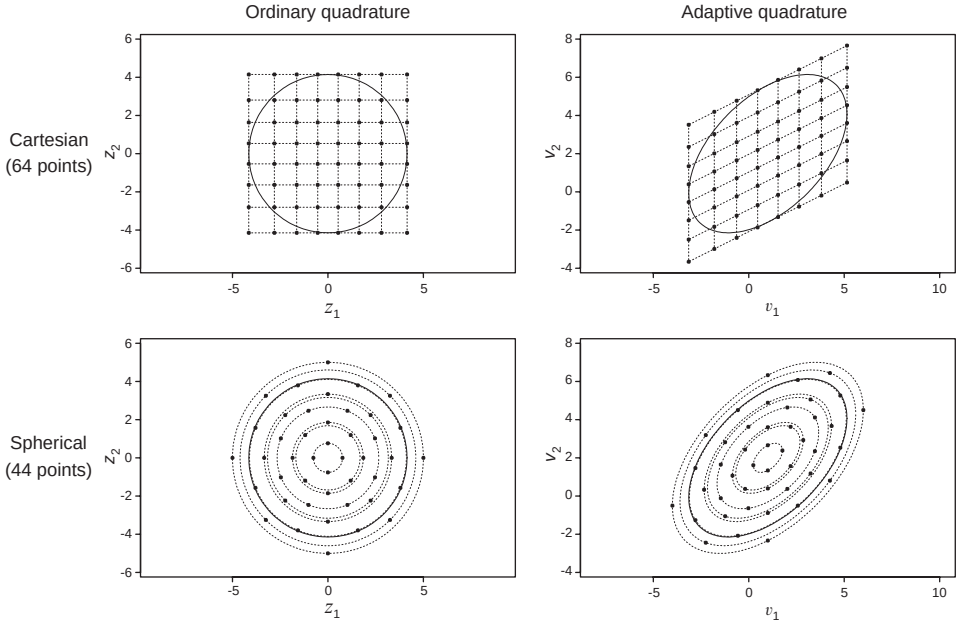

Fig. 2. Locations for ordinary and adaptive integration in two dimensions using cartesian and spherical quadrature rules with $d = 7$, where $\mu_1 = 1$, $\mu_2 = 2$, $\tau_1 = \tau_2 = 1$ and the posterior correlation is 0.5.

Table 1
Quantities evaluated at level l to obtain the likelihood by adaptive quadrature

Likelihood:

$$f^{(l)}(\theta|V^{(l+1)}) = \sum_{r_M} P_{Mr_M} \cdots \sum_{r_1} P_{1r_1} \prod f^{(l-1)}(\theta|A_{1r_1}, \ldots, A_{Mr_M}, V^{(l+1)}).$$

First order moments:

$$E[v_m^{(l)}|V^{(l)}] = A_m^{(l)},$$

$$m = 1, \ldots, M^{(l)},$$

$$E[v_m^{(k)}|V^{(l+1)}] = \frac{\sum_{r_M} P_{Mr_M} \cdots \sum_{r_1} P_{1r_1} E[v_m^{(k)}|V^{(l)}] \prod f^{(l-1)}(\theta|A_{1r_1}, \ldots, A_{Mr_M}, V^{(l+1)})}{f^{(l)}(\theta|V^{(l+1)})},$$

$$k = 1, \ldots, l;\ m = 1, \ldots, M^{(k)}.$$

Second order moments:

$$E[v_m^{(l)} v_n^{(l+i)}|V^{(l)}] = A_m^{(l)} A_n^{(l+i)},$$

$$m = 1, \ldots, M^{(l)};\ i = 0, \ldots, L-l;\ n = \begin{cases} m, \ldots, M^{(l)} & i = 0 \\ 1, \ldots, M^{(l+i)} & i > 0 \end{cases}$$

$$E[v_m^{(k)} v_n^{(k+i)}|V^{(l+1)}] = \frac{\sum_{r_M} P_{Mr_M} \cdots \sum_{r_1} P_{1r_1} E[v_m^{(k)} v_n^{(k+i)}|V^{(l)}] \prod f^{(l-1)}(\theta|A_{1r_1}, \ldots, A_{Mr_M}, V^{(l+1)})}{f^{(l)}(\theta|V^{(l+1)})},$$

$$k = 1, \ldots, l;\ m = 1, \ldots, M^{(k)};\ i = 0, \ldots, L-k;\ n = \begin{cases} m, \ldots, M^{(k)} & i = 0 \\ 1, \ldots, M^{(k+i)} & i > 0. \end{cases}$$

These moments can be used to update the quadrature locations and weights and we can iterate as in the univariate case until convergence. This set of iterations is then alternated with single steps of a Newton–Raphson procedure as described in the univariate case.

Note that adaptive quadrature as described here, based on the posterior moments, can be applied as easily to multilevel models as to two-level models. This is in contrast to the first order Laplace approximation suggested by Liu and Pierce (1994). Applying their method to two-level models is straightforward—the mode with respect to all the random effects is found and the covariance matrix of the

approximating multivariate normal density is found from the inverse Hessian matrix of the log of the integrand. However, in multilevel models, finding the mode with respect to $v^{(l)}$ would require integrating out all lower level random effects $v^{(2)},\ldots,v^{(l-1)}$ for each value of $v^{(l)}$ during numerical optimization and differentiation with respect to $v^{(l)}$.

2.3. Multivariate integration using spherical quadrature rules

Cartesian product quadrature in (11) is a straightforward application of Gauss–Hermite quadrature to multidimensional integration. However, as pointed out in Naylor and Smith (1988), integrals of the form

$$\int\cdots\int \exp(-x_1^2-\cdots-x_M^2)f(x_1,\ldots,x_M)\mathrm{d}x_1\ldots\mathrm{d}x_M$$

can often be integrated more efficiently using spherical quadrature rules. These rules are located on concentric hyperspheres as illustrated for two dimensions in the bottom left panel of Fig. 2.

A rule of degree d is exact if $f(x_1,\ldots,x_M)$ is a linear combination of monomials of the form $x_1^{k_1}\ldots x_M^{k_M}$ with $k_1+\cdots+k_M\leqslant d$. Cartesian product quadrature with R points per dimension is exact for monomials with degree $d=2R-1$. In addition, cartesian product quadrature is exact for monomials with $k_1+\cdots+k_M>d$ as long as $k_1\leqslant d,\ldots,k_M\leqslant d$.

A compilation of quadrature rules for multidimensional integrals is given in Stroud (1971) and has been updated by Cools and Rabinowitz (1993) and Cools (1999). The most efficient quadrature rules for a certain dimension M and degree d are those that require the fewest number of points; rules with positive weights are generally more accurate than rules with some negative weights. For integrals of the form above, the most efficient published degree 7 rules with positive weights that we are aware of use 2^M+2M^2+1 points for $M=3,4,6$ and $2^{M+1}+4M^2$ for $M\geqslant 3$ and are given in Stroud (1971). For example, in six dimensions, the rule requires 137 points compared with 4096 $(=4^M)$ for cartesian product quadrature. Unfortunately, we are aware of published higher degree rules with positive weights only for $M=2,3$.

We can use these spherical rules to evaluate the $M^{(l)}$ dimensional integrals at each level $l=2,\ldots,L$ using ordinary or adaptive quadrature. However, we cannot use a single $S=\sum_l M^{(l)}$ spherical rule for integrating over the random effects at all levels. This is because, as shown in (9), integration with respect to $u^{(l)}$ to compute $f^{(l)}(\theta|U^{(l+1)})$ requires $f^{(l-1)}(\theta|U^{(l)})$ to be evaluated by complete integration with respect to $u^{(l-1)}$ for each value of $u^{(l)}$. Cartesian quadrature provides such nested integration as seen in (11) and by considering the case $S=2$ illustrated in the first row of Fig. 2 where the sum along a given column of quadrature points corresponds to complete integration with respect to v_2 for a given value of v_1. In contrast, as remarked by Naylor and Smith (1988), spherical quadrature does not permit such 'marginalization'.

3. Simulation study

3.1. Simple random component probit model

We first investigate the bias in parameter estimates using both ordinary and adaptive quadrature for the random component or random intercept binary probit model. The following model was simulated:

$$y_{ij}^{*} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2j} + u_j + \varepsilon_{ij}, \quad \text{var}(\varepsilon_{ij}) = 1,$$

where x_{1ij} varies between level-1 units ij and takes on the values 0 and 1 with probabilities equal to 0.5, whereas x_{2j} varies between clusters j also taking on values 0 and 1 with probabilities 0.5 independently of x_{1ij}.

The fixed parameters were set to $\beta_0 = 0$, $\beta_1 = 1$, $\beta_2 = 1$, σ was varied so that $\rho = 0.30, 0.45, 0.60, 0.75, 0.90$ and combined with cluster sizes $n_j = 10, 100, 500$. We used 1000 clusters to obtain precise estimates of the biases with 50 replications. For each simulated dataset, the parameters were estimated by ordinary quadrature with 10, 20 and 40 points and by adaptive quadrature with 5, 10 and 20 points. If the relative change in mean log-likelihood with increasing numbers of quadrature points was no more than 5×10^{-5}, the smaller number of quadrature points was considered adequate; otherwise the maximum number of quadrature points was used even if it appeared inadequate. Table 2 shows the number of quadrature points and the means and standard deviations of $\widehat{\sigma}$. The corresponding results for the regression coefficients are given in Table 3. Fig. 3 shows boxplots of the relative bias of $\widehat{\sigma}$ defined as $(\widehat{\sigma} - \sigma)/\sigma$.

Table 2
Estimates of σ using R-point ordinary and adaptive quadrature

			Ordinary quadrature			Adaptive quadrature		
n_j	ρ	σ	Mean $\widehat{\sigma}$	(sd)	R	Mean $\widehat{\sigma}$	(sd)	R
10	0.30	0.655	0.658	(0.025)	10	0.659	(0.025)	5
10	0.45	0.905	0.903	(0.037)	20	0.904	(0.038)	10
10	0.60	1.225	1.224	(0.041)	20	1.225	(0.042)	10
10	0.75	1.732	1.739	(0.067)	40	1.740	(0.067)	20
10	0.90	3.000	2.812	(0.106)*	40	2.986	(0.133)	20
100	0.30	0.655	0.649	(0.017)*	40	0.653	(0.018)	5
100	0.45	0.905	0.878	(0.028)*	40	0.910	(0.024)	5
100	0.60	1.225	1.073	(0.030)*	40	1.229	(0.030)	5
100	0.75	1.732	1.332	(0.044)*	40	1.713	(0.067)	20
100	0.90	3.000	1.768	(0.062)*	40	2.935	(0.090)*	20
500	0.30	0.655	0.543	(0.023)*	40	0.654	(0.019)	5
500	0.45	0.905	0.661	(0.023)*	40	0.910	(0.021)	5
500	0.60	1.225	0.782	(0.030)*	40	1.240	(0.034)*	5
500	0.75	1.732	0.951	(0.043)*	40	1.732	(0.050)	20
500	0.90	3.000	1.224	(0.056)*	40	2.991	(0.081)	20

*True value outside approximate 95% confidence interval.

Table 3
Estimates of β_0, β_1 and β_2 (true values 0,1,1) using ordinary and adaptive quadrature with the same number of quadrature points R as in Table 2

		Ordinary quadrature						Adaptive quadrature					
		$\widehat{\beta}_0$		$\widehat{\beta}_1$		$\widehat{\beta}_2$		$\widehat{\beta}_0$		$\widehat{\beta}_1$		$\widehat{\beta}_2$	
n_j	ρ	Mean	(sd)	Mean	(sd)	Mean	(sd)	Mean	(sd)	Mean	(sd)	Mean	(sd)
10	0.30	−0.01	(0.04)	0.99	(0.03)	1.02	(0.06)	−0.01	(0.04)	0.99	(0.03)	1.02	(0.06)
10	0.45	0.01	(0.05)	1.00	(0.04)	1.00	(0.07)	0.01	(0.05)	1.00	(0.04)	1.01	(0.07)
10	0.60	−0.01	(0.07)	1.00	(0.04)	1.01	(0.09)	−0.01	(0.07)	1.00	(0.04)	1.01	(0.09)
10	0.75	0.00	(0.08)	0.99	(0.04)	0.99	(0.11)	0.00	(0.08)	0.99	(0.04)	0.99	(0.11)
10	0.90	−0.05	(0.23)	1.00	(0.05)	1.02	(0.30)	−0.02	(0.15)	1.01	(0.05)	0.96	(0.22)
100	0.30	−0.01	(0.03)	1.00	(0.01)	1.01	(0.04)	0.00	(0.03)	1.00	(0.01)	1.01	(0.04)
100	0.45	0.00	(0.08)	1.00	(0.01)	1.00	(0.11)	0.00	(0.04)	1.00	(0.01)	1.02	(0.06)
100	0.60	0.03	(0.15)	1.00	(0.01)	0.95	(0.21)	0.01	(0.06)	1.00	(0.01)	1.02	(0.06)
100	0.75	−0.05	(0.25)	0.99	(0.01)	1.02	(0.33)	−0.01	(0.07)	1.00	(0.01)	1.04	(0.08)
100	0.90	−0.11	(0.36)	0.98	(0.02)	0.90	(0.47)	−0.06	(0.14)	1.00	(0.02)	0.97	(0.20)
500	0.30	0.00	(0.08)	1.00	(0.01)	0.99	(0.08)	0.00	(0.03)	1.00	(0.01)	1.01	(0.04)
500	0.45	0.00	(0.13)	1.00	(0.00)	0.98	(0.16)	−0.01	(0.04)	1.00	(0.00)	1.01	(0.06)
500	0.60	−0.01	(0.18)	1.00	(0.01)	0.88	(0.26)	0.01	(0.05)	1.00	(0.01)	1.02	(0.07)
500	0.75	−0.03	(0.39)	0.99	(0.01)	0.75	(0.41)	0.01	(0.08)	1.00	(0.01)	0.99	(0.13)
500	0.90	−0.08	(0.62)	0.99	(0.01)	0.56	(0.89)	−0.06	(0.15)	1.00	(0.01)	0.96	(0.23)

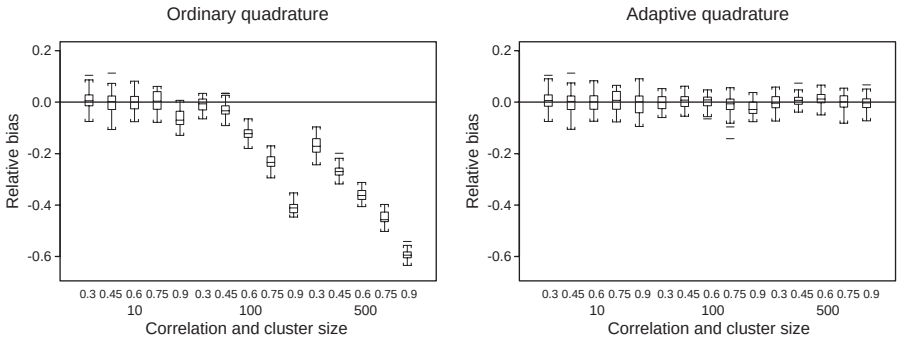

Fig. 3. Relative bias of $\widehat{\sigma}$ for ordinary and adaptive quadrature.

Adaptive quadrature requires a considerably smaller number of quadrature points than ordinary quadrature to achieve a stable log-likelihood. Using ordinary quadrature, the standard deviation estimates become increasingly biased as the cluster size and intraclass correlation increase, 40 points being clearly inadequate for correlations above 0.45 when $n_j = 100$ and above 0.3 when $n_j = 500$. Adaptive quadrature performs very well for all combinations of n_j and ρ with no more than 20 quadrature points and fewer for lower intraclass correlations. As expected, adaptive

quadrature appears to work better for larger cluster sizes where the posterior distribution is closer to normal. Somewhat surprisingly, the estimates of the intercept β_0 and of the regression coefficient β_1 of the within-cluster covariate are fairly unbiased even where the estimates of the standard deviation σ are biased using ordinary quadrature. However, using ordinary quadrature, the estimates of the regression coefficient β_2 of the between-cluster covariate have severe downward bias for large clusters and high intraclass correlation. Moreover, in many cases the standard deviations of the estimates of β_0 and β_2 are substantially larger than for adaptive quadrature, meaning that the estimates for a particular dataset can be very poor.

3.2. Three-level probit model

We now consider three-level binary probit models of the form

$$y^*_{ijk} = \beta_0 + u^{(2)}_{jk} + u^{(3)}_k + \varepsilon_{ijk}, \quad \text{var}(\varepsilon_{ijk}) = 1,$$

where the level-2 random intercept $u^{(2)}_{jk}$ has variance σ_2^2 and the level-3 random intercept $u^{(3)}_k$ has variance σ_3^2. In particular, we will assess the performance of adaptive quadrature for different cluster sizes and intraclass correlations. There are two cluster sizes for the three-level model, the number of level-1 units in each level-2 unit, n_2, and the number of level-2 units in each level-3 unit, n_3. The posterior density of $u^{(2)}_{jk}$, conditional on $u^{(3)}_k$ becomes increasingly normal as n_2 increases. Therefore fewer quadrature points should be required at level 2 for larger n_2. The posterior density of $u^{(3)}_k$ is the product of the prior density and a product of n_3 level-2 likelihood contributions. This density will become increasingly normal as n_3 increases but also as n_2 increases, since the level-2 likelihood contributions themselves then become closer to normal. Therefore, generally, fewer quadrature points may be required at level 3 than at level 2. In addition to estimating the parameters with 5 and 10 point quadrature per dimension, we will therefore also try using a larger number of quadrature points at level 2 (10 points) than level 3 (5 points).

There are several ways of defining intraclass correlations. The marginal correlation between units in the same level-2 and level-3 units is

$$\rho_{23} \equiv \text{cor}(y^*_{ijk}, y^*_{i'jk}) = \frac{\sigma_2^2 + \sigma_3^2}{\sigma_2^2 + \sigma_3^2 + 1},$$

whereas the conditional correlation, conditioning on the level-3 random effect, is

$$\rho_{2|3} \equiv \text{cor}(y^*_{ijk}, y^*_{i'jk} | u^{(3)}_k) = \frac{\sigma_2^2}{\sigma_2^2 + 1}.$$

The correlation between units in the same level-3 unit but different level-2 units is

$$\rho_3 \equiv \text{cor}(y^*_{ijk}, y^*_{i'j'k}) = \frac{\sigma_3^2}{\sigma_2^2 + \sigma_3^2 + 1}.$$

The mean parameter estimates over 50 simulations for different combinations of n_2, n_3, $\rho_{2|3}$ and ρ_3 are given in Table 4.

Consistent with the results of the previous section, 5-point adaptive quadrature at level 2 is inadequate when the level-2 cluster size, n_2, is 10 and the intraclass correlation $\rho_{2|3}$ is 0.6. These biases are greater when ρ_3 is large, but, surprisingly, lower when n_3 is small. When both intraclass correlations are high, σ_3 is poorly estimated with 5-point quadrature even when n_2 is large. A striking result is that in all simulations where 10-point quadrature per dimension performed better than 5-point quadrature, the combination of 10 points at level 2 and 5 points at level 3 worked nearly as well.

3.3. Random coefficient probit models

We simulated data for 1000 clusters j each with 10 level-1 units ij from the two-level binary probit model

$$y_{ij}^* = \beta_0 + \beta_1 x_{ij} + u_{0j} + u_{1j} x_{ij} + \varepsilon_{ij}, \quad \mathrm{var}(\varepsilon_{ij}) = 1,$$

where x_{ij} varies between level-1 units and equals 0 or 1 with equal probabilities, $\beta_0 = -0.5$, $\beta_1 = 1$ and the random intercept u_{0j} and slope u_{1j} have unit standard deviations $\sigma_0 = \sigma_1 = 1$ and covariance $\sigma_{01} = 0.5$. Conditional on x_{ij}, the correlation between y_{ij}^* and $y_{i'j}^*$ is then 0.5 if $x_{ij} = x_{i'j} = 0$, 0.53 if $x_{ij} \neq x_{i'j}$ and 0.75 if $x_{ij} = x_{i'j} = 1$.

We estimated the parameters using adaptive quadrature with both cartesian and spherical rules of degrees 7, 11 and 15, requiring 16, 36 and 64 points for cartesian quadrature and 12, 28 and 44 for spherical quadrature. The results are shown in Table 5.

The spherical rules give nearly identical results as the same degree cartesian rules and the eleven degree rule appears to be adequate. We repeated the simulations with $\sigma_0 = \sigma_1 = 1.5$ and $\sigma_{01} = 1.25$ so that the intraclass correlations ranged between 0.67 and 0.87. As expected due to the higher intraclass correlations, the eleven degree rule no longer appears adequate and a fifteen degree rule is required. The spherical rules of a given degree now appear a little inferior to the cartesian rules of the same degree.

To illustrate the usefulness of spherical rules for estimating models with many random coefficients, we simulated a dataset with six correlated random effects from the model

$$y_{ij}^* = x_{ij}'\beta + x_{ij}'u_j + \varepsilon_{ij}, \quad \mathrm{var}(\varepsilon_{ij}) = 1,$$

where $x_{1ij} = 1$ and x_{2ij} to x_{6ij} are mutually independent, equal to 0 and 1 with probabilities 0.5. We simulated 100 clusters of size 100 and estimated the parameters using adaptive quadrature with a 137-point degree 7 spherical rule (rule 7-1 in Stroud, 1971). The true and estimated parameters are given in Table 6. Out of the 27 parameters, 20 were within a standard error of the true value and only 3 were more than two standard errors away from the true value.

Table 4
Mean parameter estimates (standard deviations) using adaptive quadrature for some three-level models

n_2,n_3	$\rho_{2\|3}$	ρ_3	ρ_{23}	Param.	True value	Mean estimate 5, 5 points	Mean estimate 10, 10 points	Mean estimate 10, 5 points
10,100	0.6	0.3	0.72	σ_2	1.225	1.217 (0.014)	1.222 (0.015)	1.222 (0.014)
				σ_3	1.035	2.542 (1.125)	1.076 (0.117)	1.065 (0.092)
				β_0	0.000	−0.026 (0.876)	−0.017 (0.110)	−0.020 (0.104)
				Log-lik.	–	−46399.57	−46345.76	−46345.25
100,10	0.6	0.3	0.72	σ_2	1.225	1.246 (0.025)	1.237 (0.040)	1.232 (0.031)
				σ_3	1.035	1.231 (0.111)	1.102 (0.092)	1.032 (0.068)
				β_0	0.000	−0.041 (0.158)	−0.016 (0.133)	0.004 (0.111)
				Log-lik.	–	−39690.86	−39686.62	−39685.72
10,100	0.6	0.6	0.84	σ_2	1.225	*	1.226 (0.021)	1.221 (0.019)
				σ_3	1.936	*	2.021 (0.166)	2.033 (0.187)
				β_0	0.000	*	−0.072(0.221)	−0.066 (0.187)
				Log-lik.	–	*	−35623.16	−35623.19
100,10	0.6	0.6	0.84	σ_2	1.225	1.249 (0.050)	1.235 (0.037)	1.235 (0.037)
				σ_3	1.936	2.919 (0.422)	1.965 (0.138)	1.979 (0.140)
				β_0	0.000	0.021 (0.407)	−0.005 (0.245)	−0.011 (0.244)
				Log-lik.	–	−30317.49	−30302.50	−30302.57
10,10	0.6	0.6	0.84	σ_2	1.225	1.222 (0.050)	1.231 (0.051)	1.231 (0.051)
				σ_3	1.936	2.110 (0.238)	1.967 (0.192)	1.969 (0.201)
				β_0	0.000	0.002 (0.293)	−0.003 (0.187)	−0.004 (0.190)
				Log-lik.	–	−3665.78	−3664.53	−3664.61

*Converged only for 40 datasets and gave very many large estimates of σ_3.

Table 5
Mean estimates (standard deviations) using cartesian and spherical adaptive quadrature of different degrees for random coefficient probit models

		Degree 7		Degree 11		Degree 15	
True param.		Cartesian	Spherical	Cartesian	Spherical	Cartesian	Spherical
β_0	(−0.5)	−0.500 (0.041)	−0.499 (0.041)	−0.500 (0.041)	−0.503 (0.042)	−0.501 (0.042)	−0.501 (0.041)
β_1	(1.0)	1.006 (0.048)	1.001 (0.049)	1.003 (0.048)	1.000 (0.048)	1.001 (0.048)	1.002 (0.048)
σ_0	(1.0)	1.007 (0.045)	1.011 (0.045)	0.993 (0.044)	0.993 (0.044)	0.994 (0.044)	0.994 (0.044)
σ_1	(1.0)	1.001 (0.070)	1.008 (0.070)	0.999 (0.068)	0.995 (0.066)	1.001 (0.068)	0.997 (0.068)
σ_{01}	(0.5)	0.539 (0.078)	0.552 (0.076)	0.508 (0.071)	0.501 (0.071)	0.507 (0.071)	0.507 (0.071)
Log-lik.		−5197.2 (64.75)	−5196.7 (64.82)	−5198.0 (64.64)	−5198.3 (64.62)	−5198.0 (64.64)	−5198.0 (64.63)
β_0	(−0.5)	−0.479 (0.055)	−0.465 (0.054)	−0.477 (0.065)	−0.492 (0.057)	−0.497 (0.059)	−0.486 (0.059)
β_1	(1.0)	1.019 (0.071)	1.040 (0.072)	1.007 (0.070)	1.002 (0.070)	0.997 (0.071)	1.003 (0.072)
σ_0	(1.5)	1.534 (0.064)	1.554 (0.066)	1.479 (0.074)	1.486 (0.062)	1.489(0.062)	1.493 (0.063)
σ_1	(1.5)	1.512 (0.099)	1.538 (0.101)	1.476 (0.089)	1.436 (0.078)	1.498 (0.091)	1.463 (0.085)
σ_{01}	(1.25)	1.184 (0.203)	1.323 (0.205)	1.090 (0.175)	1.036 (0.184)	1.098 (0.176)	1.097 (0.184)
Log-lik.		−4496.5 (73.84)	−4494.8 (74.17)	−4501.1 (73.70)	−4502.2 (73.45)	−4500.4 (73.38)	−4501.3 (73.32)

Table 6
True parameters in bold and estimates (standard errors) using adaptive spherical quadrature for probit model with six random effects (log-likelihood $=-5103.82$)

$$\beta = \begin{bmatrix} \mathbf{.0} & .03\ (.05) \\ \mathbf{.5} & .50\ (.06) \\ \mathbf{-1.0} & -1.00\ (.06) \\ \mathbf{1.0} & .97\ (.07) \\ \mathbf{-.5} & -.43\ (.07) \\ \mathbf{.0} & -.07\ (.06) \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} \mathbf{.25} & .14\ (.04) & & & & & & & & & & \\ \mathbf{.12} & .13\ (.04) & \mathbf{.25} & .20\ (.05) & & & & & & & & \\ \mathbf{-.12} & -.07\ (.03) & \mathbf{-.16} & -.15\ (.04) & \mathbf{.25} & .23\ (.06) & & & & & & \\ \mathbf{.09} & .12\ (.03) & \mathbf{.12} & .15\ (.04) & \mathbf{-.12} & -.15\ (.04) & \mathbf{.25} & .30\ (.06) & & & & \\ \mathbf{-.09} & -.10\ (.04) & \mathbf{-.12} & -.18\ (.04) & \mathbf{.12} & .17\ (.08) & \mathbf{-.09} & -.22\ (.05) & \mathbf{.25} & .37\ (.09) & & \\ \mathbf{.12} & .14\ (.03) & \mathbf{.16} & .19\ (.05) & \mathbf{-.16} & -.20\ (.05) & \mathbf{.12} & .12\ (.04) & \mathbf{-.12} & -.23\ (.05) & \mathbf{.25} & .28\ (.06) \end{bmatrix}$$

4. Other types of dependent variables

The same adaptive quadrature method can be used for counts, durations, continuous, censored and ordinal dependent variables, discrete choices and rankings. The likelihoods have the same form except for the level-1 contribution $f^{(1)}(\theta|U^{(2)})$, given for binary dependent variables in (2).

For counts, the likelihood contribution of a Poisson model is

$$f^{(1)}(\theta|U^{(2)}) = \frac{[\exp(\eta)]^s}{s!} \quad \text{if } y = s, \quad s = 0, 1, \ldots. \tag{13}$$

Random effects models for counts are discussed in Cameron and Trivedi (1998). If a piecewise exponential proportional hazards model is assumed for durations with hazards remaining constant for intervals of time, Holford (1980) and Clayton (1988) show that each observed duration contributes a product of terms of the form of (13) to the likelihood, namely one term for each interval it exceeds. For continuous dependent variables, we can specify for instance a normal, gamma or inverse Gaussian density depending on the shape of the distribution.

For limited dependent variables, we assume that the underlying continuous variable can be modeled as

$$y^* = \eta + \varepsilon,$$

where ε is normally distributed with standard deviation v. For continuous responses subject to left-censoring at b_l (Tobin, 1958), right-censoring at b_r, or both (Rosett and Nelson, 1975), or for grouped (or interval censored) dependent variables with boundaries b_l and b_r (Stewart, 1983), the likelihood contribution is

$$f^{(1)}(\theta|U^{(2)}) = \begin{cases} \phi(y/v)/v & \text{if uncensored} \\ \Phi([\eta - b_r]/v) & \text{if right-censored} \\ \Phi([b_l - \eta]/v) & \text{if left-censored} \\ \Phi([b_r - \eta]/v) - \Phi([b_l - \eta]/v) & \text{if grouped,} \end{cases} \tag{14}$$

where b_l and b_r are usually constant but can vary across units. For ordinal responses with categories s, $s = 1, \ldots, S$, the likelihood is as for grouped dependent variables with unknown thresholds κ_{s-1} and κ_s in place of fixed censoring limits b_l and b_r when $y = s$, where $-\infty = \kappa_0 < \kappa_1 < \cdots < \kappa_S = \infty$ (Aitchison and Silvey, 1957). A number of other random effects models suitable for ordered responses and discrete time durations are described in Rabe-Hesketh et al. (2001c).

For discrete choices, we can model the utility for alternative s, $s = 1, 2, .., S$ as

$$y_s^* = \eta_s + \varepsilon_s,$$

so that

$$y = s \quad \text{if } y_s^* > y_\ell^*, \quad \forall \ell, \ \ell \neq s.$$

If ε_s is Gumbel (extreme value of Type I), with density function $\exp(-\varepsilon_s - \exp(-\varepsilon_s))$, the likelihood contribution is a multinomial logit

$$f^{(1)}(\theta|U^{(2)}) = \frac{\exp(\eta_s)}{\sum_\ell^S \exp(\eta_\ell)} \quad \text{if } y = s. \tag{15}$$

An 'exploded logit' is obtained for rankings (e.g. Beggs et al., 1981; Hausman and Ruud, 1987). See Skrondal and Rabe-Hesketh (2003) for a treatment of multilevel random effects models for discrete choices and rankings. Skrondal and Rabe-Hesketh (2004) discuss models with many different types of dependent variables including mixed types.

5. Discussion

As far as we are aware, this is the first generalization of adaptive quadrature for multilevel modeling. Our simulations show that the method performs well in a wide variety of situations including large cluster sizes and high intraclass correlations where ordinary quadrature often fails. Adaptive quadrature requires lower degree integration rules than ordinary quadrature, particularly for the higher level random effects. Further gains in efficiency can be achieved by using spherical quadrature rules. Unfortunately, however, there are to our knowledge no published higher degree spherical rules for integrals in four or more dimensions to be used for problems where degree 7 rules are insufficient. Another advantage of adaptive quadrature is that it gives empirical Bayes predictions of cluster or individual-specific random effects and their standard errors as a by-product. These are often of both substantive interest and of importance for checking model specification.

Adaptive quadrature is slower than alternative estimation methods such as penalized quasilikelihood (PQL) (Breslow and Clayton, 1993), for example as implemented in the iterative generalized least squares algorithm (Goldstein, 1991). Unfortunately, the parameter estimates from PQL tend to be biased for binary dependent variables with small cluster sizes and high intraclass correlations (e.g. Rodriguez and Goldman, 1995, 2001). Moreover, PQL does not involve a likelihood which prohibits the use of likelihood based inference such as likelihood ratio tests and likelihood based confidence intervals. Improved results can be achieved using a sixth order Laplace approximation for the marginal likelihood, LaPlace6, (Raudenbush et al., 2000) which worked as well as 7-point adaptive quadrature in simulations of a two-level binary dependent variable model. However, an advantage of adaptive quadrature is that the precision can be increased by simply using more quadrature points whereas increasing the degree of the Taylor expansion for the Laplace method would require more work (Raudenbush et al., 2000).

Computer intensive alternatives to adaptive quadrature include simulation based approaches such as Markov Chain Monte Carlo (MCMC) (e.g. Gelman et al., 2003) and maximum simulated likelihood (MSL) (Hajivassiliou and Ruud, 1994). The hierarchical structure of multilevel models lends itself naturally to MCMC using for instance Gibbs sampling. If vague priors are specified, the method essentially yields

maximum likelihood estimates. Unfortunately, a problem with this approach is how to ensure that a truly stationary distribution has been obtained. Another important shortcoming is that there is no diagnostic for assessing empirical identification (e.g., Keane, 1992). Regarding simulated maximum likelihood, a merit is that conditional independence specifications implicit in standard multilevel models may be relaxed. This can be useful in panel data models where ARMA(p,q) processes and their special cases are sometimes specified for the level-1 errors ε_{ij}. Furthermore, unlike methods based on quadrature, simulation methods allow statistical analysis of the approximation error.

We have confined the simulations to multilevel random effects probit models for binary dependent variables, although the estimation method can be used for many other types of dependent variable as outlined in Section 4. In comparison to binary responses, these other response types tend to yield more concentrated posterior densities where ordinary quadrature can be expected to perform poorly (see e.g. Albert and Follmann, 2000). An example with count data is given in Rabe-Hesketh et al. (2002) where adaptive quadrature recovers previous estimates and standard errors whereas ordinary quadrature fails.

The adaptive quadrature method can also be used for the more general class of multilevel factor and structural equation models (Rabe-Hesketh et al., 2004) since they have the same conditional independence structure as random coefficient models: variables (at level 1) are conditionally independent given the factors which in turn are conditionally independent given higher level factors, etc. The marginal likelihood has the same form as that of random coefficient models, the only difference being the form of the linear predictor η. Factor models are useful for generating flexible covariance structures using only a small number of latent variables, see for example Rabe-Hesketh and Skrondal (2001). They are also useful for inducing dependence between multiple processes as required for selection and endogenous treatment models and their multilevel extensions (e.g. Skrondal and Rabe-Hesketh, 2004, Chapter 14).

Maximum likelihood estimation and empirical Bayes prediction for all of these models using adaptive quadrature is implemented in gllamm (Rabe-Hesketh et al., 2000, 2001a,b, 2002) which runs in Stata (StataCorp, 2003). The program can also handle discrete random effects including nonparametric maximum likelihood (Heckman and Singer, 1984; Rabe-Hesketh et al., 2003) and is available from `http://www.gllamm.org`.

References

Aitchison, J., Silvey, S., 1957. The generalization of probit analysis to the case of multiple responses. Biometrika 44, 131–140.

Albert, P.S., Follmann, D.A., 2000. Modeling repeated count data subject to informative dropout. Biometrics 56, 667–677.

Antweiler, W., 2001. Nested random effects estimation in unbalanced panel data. Journal of Econometrics 101, 295–313.

Baltagi, B.H., 2001. Econometric Analysis of Panel Data, 2nd Edition. Wiley, London.

Baltagi, B.H., Song, S., Jung, B., 2001. The unbalanced nested error component regression model. Journal of Econometrics 101, 357–381.

Beggs, S., Cardell, S., Hausman, J., 1981. Assessing the potential demand for electric cars. Journal of Econometrics 16, 1–19.

Beron, K., Murdoch, J., Thayer, M., 1999. Hierarchical linear models with application to air pollution in the South Coast air basin. American Journal of Agricultural Economics 81, 1123–1127.

Blundell, R., Windmeijer, F., 1997. Cluster effects and simultaneity in multilevel models. Health Economics 1, 6–13.

Bock, R.D., Aitkin, M., 1981. Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. Psychometrika 46, 443–459.

Bock, R.D., Lieberman, M., 1970. Fitting a response model for n dichotomously scored items. Psychometrika 33, 179–197.

Bock, R.D., Schilling, S., 1997. High-dimensional full-information item factor analysis. In: Berkane, M. (Ed.), Latent Variable Modelling and Applications to Causality. Springer, New York, NY, pp. 164–176.

Borjas, G.J., Sueyoshi, G.T., 1994. A two-stage estimator for probit models with structural group effects. Journal of Econometrics 64, 165–182.

Breslow, N.E., Clayton, D.G., 1993. Approximate inference in generalized linear mixed models. Journal of the American Statistical Association 88, 9–25.

Butler, J.S., Moffitt, R., 1982. A computationally efficient quadrature procedure for the one-factor multinomial probit model. Econometrica 50, 761–764.

Cameron, A.C., Trivedi, P.K., 1998. Regression Analysis of Count Data. Cambridge University Press, Cambridge.

Cardoso, A.R., 2000. Wage differentials across firms: an application of multilevel modelling. Journal of Applied Econometrics 15, 343–354.

Carey, K.A., 2000. Multilevel modelling approach to analysis of patient costs under managed care. Health Economics 9, 435–446.

Carlin, B.P., Louis, T.A., 2000. Bayes and Empirical Bayes Methods for Data Analysis, 2nd Edition. Chapman & Hall/CRC, Boca Raton, FL.

Clayton, D., 1988. The analysis of event history data: a review of progress and outstanding problems. Statistics in Medicine 7, 819–841.

Cools, R., 1999. Monomial cubature rules since "Stroud": a compilation—part 2. Journal of Computational and Applied Mathematics 112, 21–27.

Cools, R., Rabinowitz, P., 1993. Monomial cubature rules since "Stroud": a compilation. Journal of Computational and Applied Mathematics 48, 309–326.

Davis, P., 2002. Estimating multi-way error components models with unbalanced data structures. Journal of Econometrics 106, 67–95.

Dunnett, C.W., Sobel, M., 1955. Approximations to the probability integral and certain percentage points of a multivariate analogue of Student's *t*-distribution. Biometrika 42, 258–260.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 2003. Bayesian Data Analysis, 2nd Edition. Chapman and Hall/CRC, Boca Raton, FL.

Goldstein, H., 1991. Nonlinear multilevel models with an application to discrete response data. Biometrika 78, 45–51.

Hajivassiliou, V.A., Ruud, P.A., 1994. Classical estimation methods for LDV models using simulation. In: Engle, R.F., McFadden, D.L. (Eds.), Handbook of Econometrics, Vol. IV. Elsevier, New York, NY, pp. 2383–2441.

Hausman, J.A., Ruud, P.A., 1987. Specifying and testing econometric models for rank-ordered data. Journal of Econometrics 34, 83–103.

Heckman, J.J., Singer, B., 1984. A method of minimizing the impact of distributional assumptions in econometric models for duration data. Econometrica 52, 271–320.

Holford, T.R., 1980. The analysis of rates and survivorship using log-linear models. Biometrics 36, 299–305.

Hsiao, C., 2003. Analysis of Panel Data, 2nd Edition. Cambridge University Press, Cambridge.

Keane, M.P., 1992. A note on identification in the multinomial probit model. Journal of Business and Economic Statistics 10, 193–200.

Lee, L.-F., 2000. A numerically stable quadrature procedure for the one-factor random-component discrete choice model. Journal of Econometrics 95, 117–129.
Lesaffre, E., Spiessens, B., 2001. On the effect of the number of quadrature points in a logistic random-effects model: an example. Applied Statistics 50, 325–335.
Lillard, L.A., 1993. Simultaneous-equations for hazards—marriage duration and fertility timing. Journal of Econometrics 56, 189–217.
Liu, Q., Pierce, D.A., 1994. A note on Gauss–Hermite quadrature. Biometrika 81, 624–629.
Naylor, J.C., Smith, A.F.M., 1982. Applications of a method for the efficient computation of posterior distributions. Applied Statistics 31, 214–225.
Naylor, J.C., Smith, A.F.M., 1988. Econometric illustrations of novel numerical integration strategies for Bayesian inference. Journal of Econometrics 38, 103–125.
Pinheiro, J.C., Bates, D.M., 1995. Approximations to the log-likelihood function in the nonlinear mixed-effects model. Journal of Computational Graphics and Statistics 4, 12–35.
Rabe-Hesketh, S., Skrondal, A., 2001. Parameterization of multivariate random effects models for categorical data. Biometrics 57, 1256–1264.
Rabe-Hesketh, S., Pickles, A., Taylor, C., 2000. sg129: Generalized linear latent and mixed models. Stata Technical Bulletin 53, 47–57.
Rabe-Hesketh, S., Pickles, A., Skrondal, A., 2001a. GLLAMM: A general class of multilevel models and a Stata program. Multilevel Modelling Newsletter 13, 17–23.
Rabe-Hesketh, S., Pickles, A., Skrondal, A., 2001b. GLLAMM Manual. Technical Report 2001/01. Department of Biostatistics and Computing, Institute of Psychiatry, King's College, University of London. Downloadable from http://www.gllamm.org.
Rabe-Hesketh, S., Yang, S., Pickles, A., 2001c. Multilevel models for censored and latent responses. Statistical Methods in Medical Research 10, 409–427.
Rabe-Hesketh, S., Skrondal, A., Pickles, A., 2002. Reliable estimation of generalized linear mixed models using adaptive quadrature. The Stata Journal 2, 1–21.
Rabe-Hesketh, S., Pickles, A., Skrondal, A., 2003. Correcting for covariate measurement error in logistic regression using nonparametric maximum likelihood estimation. Statistical Modelling 3, 215–232.
Rabe-Hesketh, S., Skrondal, A., Pickles, A., 2004. Generalized multilevel structural equation modeling. Psychometrika 69, 167–190.
Raudenbush, S.W., Yang, M.L., Yosef, M., 2000. Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. Journal of Computational and Graphical Statistics 9, 141–157.
Rice, N., Jones, A., 1997. Multilevel models and health economics. Health Economics 6, 561–575.
Rodriguez, G., Goldman, N., 1995. An assessment of estimation procedures for multilevel models with binary responses. Journal of the Royal Statistical Society, A 158, 73–89.
Rodriguez, G., Goldman, N., 2001. Improved estimation procedures for multilevel models with binary response: a case study. Journal of the Royal Statistical Society, A 164, 339–355.
Rosett, R.N., Nelson, F.D., 1975. Estimation of a two-limit probit regression model. Econometrica 43, 141–146.
Skrondal, A., Rabe-Hesketh, S., 2003. Multilevel logistic regression for polytomous data and rankings. Psychometrika 68, 267–287.
Skrondal, A., Rabe-Hesketh, S., 2004. Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models. Chapman & Hall/CRC, Boca Raton, FL.
StataCorp., 2003. Stata Statistical Software: Release 8. Stata Press, College Station, TX.
Stewart, M.B., 1983. On least-squares estimation when the dependent variable is grouped. Review of Economic Studies 50, 737–753.
Stroud, A.H., 1971. Approximate Calculation of Multiple Integrals. Prentice-Hall, Englewood Cliffs, NJ.
Stroud, A.H., Secrest, D., 1966. Gaussian Quadrature Formulas. Prentice-Hall, Englewood Cliffs, NJ.
Tierney, L., Kadane, J.B., 1986. Accurate approximations for posterior moments and marginal densities. Journal of the American Statistical Association 81, 82–86.
Tobin, J., 1958. Estimation of relationships for limited dependent variables. Econometrica 26, 24–36.