Research Report

Adaptive Quadrature for Item Response Models

Shelby J. Haberman

Research & Development

September 2006
RR-06-29

Adaptive Quadrature for Item Response Models

Shelby J. Haberman

ETS, Princeton, NJ

October 2006

Abstract

Adaptive quadrature is applied to marginal maximum likelihood estimation for item response models with normal ability distributions. Even in one dimension, significant gains in speed and accuracy of computation may be achieved.

Key words: Normal distribution, Gauss-Hermite integration, Rasch model, 2PL model

Acknowledgments

This paper has benefited from conversations with Matthias von Davier and Sandip Sinharay. Boaz Haberman provided programming assistance.

In marginal maximum likelihood estimation for models for item responses in which the ability distribution is normal, evaluation of the log likelihood and its partial derivatives requires quadrature. Gauss-Hermite quadrature is attractive given the normal ability distribution, but this method of integration may be less efficient than adaptive Gauss-Hermite quadrature for relatively long tests.

To study this issue, Gauss-Hermite and adaptive Gauss-Hermite quadrature are described in sections 1 and 2. Application to the Rasch and 2PL models is made using data from the Praxis™ series.

Implications of results for psychometric practice are considered in section 3. On the whole, adaptive quadrature appears attractive even for one-dimensional cases.

1 Gauss-Hermite Quadrature

Gauss-Hermite quadrature is a classical numerical integration technique based on Hermite polynomials (Ralston, 1965, pp. 93–97). It has been applied to marginal estimation for a long period of time (Bock & Lieberman, 1970). In general, the Gauss-Hermite approach is applied to an integral of the form

$$I(f) = \int_{-\infty}^{\infty} f(x)\exp(-x^2)dx.$$

The r-point approximation

$$I_r(f) = \sum_{k=1}^{r} w_k f(x_k)\exp(-x_k^2)$$

is designed so that $I_r(f) = I(f)$ for any real function f such that, for an integer $k \leq 2r-1$, $f(x) = x^k$ for all real x. Tables of the x_k and w_k are readily available both in general works on mathematical functions (Davis & Polonsky, 1965, p. 924) and in specialized works on Gaussian quadrature (Stroud & Secrest, 1966). An important feature of Gauss-Hermite quadrature is that the difference $e_r(f) = I(f) - I_r(f)$ satisfies

$$e_r(f) = \frac{r!(2\pi)^{1/2} f_{2r}(\eta)}{2^r(2r)!}$$

for some real η if f has continuous derivatives f_k for $1 \leq k \leq 2r$.

In marginal estimation, it is more convenient to consider expressions based on the normal density function. The standard formula in calculus for change of variables implies that, for the standard normal density function ϕ with value $\phi(x) = \exp(-x^2/2)/(2\pi)^{1/2}$ for x in R and for a real function g, the integral

$$J(g) = \int_{-\infty}^{\infty} g(y)\phi(y)dy$$

is equal to $I(f)$ if $f(x) = g(2^{1/2}x)/\pi^{1/2}$ for real x. If $y_k = 2^{1/2}x_k$ and $v_k = w_k/\pi^{1/2}$ for $1 \leq k \leq r$, then

$$J_r(g) = \sum_{k=1}^{r} g(y_k)v_k = I_r(f),$$

and $J(g) - J_r(g)$ equals

$$e_r(f) = \frac{r!(2)^{1/2}g_{2r}(2^{1/2}\eta)}{(2r)!},$$

where g_{2r} is the $2r$th derivative of g. It is important in the discussion in this paper to note that each v_k is positive, $v_k = v_{r-k+1}$, $y_k = -y_{r-k+1}$, and $\sum_{k=1}^{r} v_k = 1$, so that $J_r(g)$ is symmetric about 0 in the sense that $J_r(g) = J_r(h)$ if $h(x) = g(-x)$ for all real x. If f is a bounded and continuous function on the real line, then $e_r(f)$ approaches 0 as r approaches ∞ (Breiman, 1968, p. 181). More generally, if f is continuous and $|f(x)|/[1 + |x|^k]$ is bounded for real x for some real $k \geq 0$, then $e_r(f)$ approaches 0 as r approaches ∞ (Breiman, 1968, p. 164). In addition, if Φ is the distribution function of a standard normal random variable and if Φ_r is the distribution function such that $\Phi_r(x)$, x real, is the sum of the v_k for integers k, $1 \leq k \leq r$, such that $y_k \leq x$, then $\Phi_r(x)$ approaches $\Phi(x)$ as r approaches ∞ (Breiman, 1968, p. 159).

To illustrate use of Gauss-Hermite quadrature, two simple examples of one-dimensional item response models will be considered. Both will be applied to data from the Praxis series in which 45 items are right-scored for 8,686 examinees.

In both applications, n examinees and q items are present. For examinee i from 1 to n, X_{ij} is the response code for item j, $1 \leq j \leq q$. It is assumed that X_{ij} is 1 for a correct response and 0 for an incorrect or missing response. The vectors $\mathbf{X}_i$ with coordinates X_{ij}, $1 \leq j \leq q$, are assumed to be independent and identically distributed random vectors. Let Γ be the set of q-dimensional vectors with coordinates that are 0 or 1, and let S be the set of arrays $\mathbf{s}$ with nonnegative coordinates $s(\mathbf{x})$, $\mathbf{x}$ in Γ, such that the sum $\sum_{\mathbf{x}\in\Gamma} s((x) = 1$. Let $\mathbf{p}$ in S be the array of probabilities $p(\mathbf{x}) = P(\mathbf{X}_i = \mathbf{x})$ for $\mathbf{x}$ in Γ, so that $\mathbf{p}$ defines the distribution of $\mathbf{X}$, and the log likelihood function ℓ is

$$\ell(\mathbf{p}) = \sum_{i=1}^{n} \log p(\mathbf{X}_i).$$

For a nonempty subset T of S, consider a model M that $\mathbf{p}$ is in T. The log likelihood $\ell(\mathbf{p})$ has maximum $\ell(T)$ for $\mathbf{p}$ in T, and a maximum-likelihood estimate $\hat{\mathbf{p}}$ in T satisfies $\ell(\hat{\mathbf{p}}) = \ell(T)$. If

g is a function on T, then $\hat{g} = g(\hat{\mathbf{p}})$ is a maximum-likelihood estimate of $g(\mathbf{p})$. The estimated minimum logarithmic penalty per item for prediction of $\mathbf{X}_i$ by use of a probability vector in T is

$$\hat{H} = -(nq)^{-1}\ell(T)$$

(Gilula & Haberman, 1994, 1995). If $\mathbf{p}$ is in T, then $\hat{H}$ provides an estimate of the entropy per item of the responses $\mathbf{X}_i$.

Associated with each examinee i is an ability variable θ_i. The θ_i, $1 \le i \le n$, are assumed to be independent and identically distributed random variables with common distribution function D, and, for each i, $1 \le i \le n$, the X_{ij}, $1 \le j \le q$, satisfy the local independence requirement that they are conditionally independent given θ_i. It is further assumed that the pairs $(\mathbf{X}_i, \theta_i)$ are mutually independent. For each examinee i, $1 \le i \le n$, the conditional probability that $X_{ij} = 1$ given $\theta_i = \theta$ is $P_j(\theta) > 0$, and the conditional probability that $X_{ij} = 0$ is $Q_j(\theta) > 0$. The item logit function λ_j of item j, $1 \le j \le q$, is then $\log(P_j/Q_j)$ (Holland, 1990). The item logit vector $\boldsymbol{\lambda}$ is then the q-dimensional function with coordinate j, $1 \le j \le q$, equal to λ_j.

For q-dimensional vectors $\mathbf{a}$ and $\mathbf{b}$ with respective coordinates a_j and b_j for $1 \le j \le q$, let

$$\mathbf{a}'\mathbf{b} = \sum_{j=1}^{q} a_j b_j.$$

and let

$$V = \prod_{j=1}^{q} Q_j = \prod_{j=1}^{q} [1 + \exp(\lambda_j)]^{-1}. \tag{1}$$

Then

$$p(\mathbf{x}) = \int V \exp(\mathbf{x}'\boldsymbol{\lambda}) dD \tag{2}$$

for all $\mathbf{x}$ in Γ, so that

$$\ell(\mathbf{p}) = \sum_{i=1}^{n} \log \int V \exp(\mathbf{X}_i'\boldsymbol{\lambda}) dD.$$

In this report, one-parameter (1PL, Rasch) and two-parameter (2PL) models are considered. In the one-parameter case, let $\mathbf{1}$ be the q-dimensional vector with each coordinate 1. Let Λ_1 be the set of functions $\boldsymbol{\lambda}$ such that

$$\boldsymbol{\lambda}(\theta) = a\theta\mathbf{1} - \boldsymbol{\gamma} \tag{3}$$

for some common item discrimination a and some q-dimensional vector $\boldsymbol{\gamma}$ with coordinates γ_j, $1 \le j \le q$, so that γ_j/a is the difficulty parameter for item j, $1 \le j \le q$. Then $\boldsymbol{\lambda}$ is assumed to be in Λ_1.

In the two-parameter case, let Λ_2 be the set of functions $\boldsymbol{\lambda}$ such that

$$\boldsymbol{\lambda}(\theta) = \theta\mathbf{a} - \boldsymbol{\gamma} \tag{4}$$

for some q-dimensional vector $\mathbf{a}$ with coordinates $a_j > 0$ for $1 \leq j \leq q$ and some q-dimensional vector $\boldsymbol{\gamma}$ with coordinates γ_j, $1 \leq j \leq q$, so that a_j is the item discrimination and γ_j/a_j is the difficulty parameter for item j, $1 \leq j \leq q$. Then $\boldsymbol{\lambda}$ is assumed to be in Λ_2.

It is quite common to assume that the common distribution function D of the θ_i is the standard normal distribution function Φ, so that

$$p(\mathbf{x}) = J(V\exp(\mathbf{x}'\boldsymbol{\lambda})) \tag{5}$$

for all $\mathbf{x}$ in Γ, and

$$\ell(\mathbf{p}) = \sum_{i=1}^{n} \log J(V\exp(\mathbf{X}_i'\boldsymbol{\lambda})). \tag{6}$$

For example, in the normal 1PL (normal Rasch) model M_1, it is assumed that $\mathbf{p}$ is in T_1, where T_1 consists of arrays $\mathbf{p}$ in S such that (1) and (2) hold for some $\boldsymbol{\lambda}$ in Λ_1 and $D = \Phi$. Similarly, in the normal two-parameter model M_2, it is assumed that $\mathbf{p}$ is in T_2, where T_2 consists of arrays $\mathbf{p}$ in S such that (1) and (2) hold for some $\boldsymbol{\lambda}$ in Λ_2 and $D = \Phi$.

If Gauss-Hermite quadrature with r points is used to approximate the log likelihood for model M_k, where k is 1 or 2, then the practical effect is to replace M_k with the model M_{kr} that $\mathbf{p}$ is in T_{kr}, where T_{kr} is the set of arrays $\mathbf{p}$ in S such that (1) and (2) hold for some $\boldsymbol{\lambda}$ in Λ_k and $D = \Phi_r$. Thus

$$p(\mathbf{x}) = J_r(V\exp(\mathbf{x}'\boldsymbol{\lambda})) \tag{7}$$

for all $\mathbf{x}$ in Γ, and

$$\ell(\mathbf{p}) = \sum_{i=1}^{n} \log J_r(V\exp(\mathbf{X}_i'\boldsymbol{\lambda})). \tag{8}$$

For model M_{kr}, maximization of the log likelihood corresponds to maximization of a log likelihood function for a log-linear model of a frequency table in which not all cells are directly observed, so that algorithms developed for such models can be applied (Haberman, 1988).

In practice, results for model M_{kr} are quite similar to results for the model M_k even for integers r of moderate size. Consider the example from the Praxis program. In the case of M_1, the normal Rasch model, $r = 20$ points suffice for quite accurate results. Differences between corresponding maximum-likelihood estimates of the parameters a and γ_j under models M_{1r} and

M_1 do not exceed 0.00003 in magnitude, a very satisfactory result, especially given that the estimated asymptotic standard deviations for these parameters are at least 0.02. Results for $r = 9$ are a bit less accurate but still relatively satisfactory. In this case, the magnitude of differences does not exceed 0.007 in the case of a and 0.001 for the γ_j. To compare estimated minimum logarithmic penalties per item, let $\hat{H}_k$ be $-(nq)^{-1}\ell(T_k)$, and let $\hat{H}_{kr}$ be $-(nq)^{-1}\ell(T_{kr})$. For the normal Rasch model, $\hat{H}_1 = 0.59639$, while $\hat{H}_{1r}$ is 0.59641 for $r = 9$ and $\hat{H}_{1r}$ is 0.59639 for $r = 20$. Indeed, the difference $\hat{H}_{1r} - \hat{H}_1$ is about 4×10^{-8} for $r = 20$.

In the case of the 2PL model, results are somewhat similar, but not quite as accurate for the approximations by Gaussian quadrature. For $r = 32$, differences in maximum-likelihood estimates of the a_j and γ_j for models M_{2r} and M_2 do not exceed 0.00003 in magnitude. For $r = 20$, differences are within 0.0006 in magnitude, and for $r = 9$, differences are within 0.04 in magnitude. In this case, $\hat{H}_2 = 0.59157$, $\hat{H}_{2r} = 0.59165$ for $r = 9$, and $\hat{H}_{2r} = 0.59157$ for $r = 20$. The difference $\hat{H}_{2r} - \hat{H}_2$ is about 4×10^{-7} for $r = 20$ and 4×10^{-8} for $r = 32$.

These differences should be kept in perspective. Use of a standard likelihood-ratio chi-square test for M_1 versus M_2 yields a statistic of about 3,770 on 44 degrees of freedom, so that model M_1 can hardly hold in any case, and the difference $\hat{H}_1 - \hat{H}_2$ of 0.00474 is much larger than any differences $\hat{H}_{kr} - \hat{H}_k$ that are encountered here.

Even in the case of the normal 2PL model, an interaction model (Haberman, 2004) yields a value of $\hat{H}$ of 0.59112. Here the model used assumes that $\mathbf{p}$ is in S and

$$\log p(\mathbf{x}) = \tau_s - \sum_{j=1}^{q}(\beta_j + s\gamma_j)$$

for $\mathbf{x}$ in Γ and $\sum_{j=1}^{q} x_j = s$ for some real τ_s, $0 \leq s \leq q$, β_j, $1 \leq j \leq q$, and γ_j, $1 \leq j \leq q$. Thus effects of approximation of integrals by Gaussian quadrature are relatively small compared to basic differences between models.

The results for Gauss-Hermite quadrature are certainly adequate for $r = 32$ for both examples, and they are rather satisfactory for $r = 20$. The question arises whether one can manage to achieve higher accuracy with a smaller number of points by modification of the procedures used for quadrature. This change can be especially important in cases such as the 2PL model in which simple sufficient statistics are not available to simplify computations.

One alternative approach to quadrature is to use the approach found in the version of Parscale used in the National Assessment of Educational Progress (NAEP). In this method, integration points are $z_k = (k-21)/5$ for $1 \le k \le 41$, so that z_k ranges from -4 to 4, and weights are $u_k = d\exp(-z_k^2/2)$ for $1 \le k \le 41$, where

$$d^{-1} = \sum_{k=1}^{41} \exp(-z_k^2/2).$$

Let D_N be the distribution function such that, for x real, $D_N(x)$ is the sum of the u_k for all k such that $z_k \le x$. Then maximum-likelihood estimates for model M_k are computed in effect for the model M_{kN} that $\boldsymbol{\lambda}$ is in Λ_k and $D = D_N$. This approach is a bit less accurate than is Gauss-Hermite integration with $r = 20$. For example, for M_{1N}, the corresponding estimated log penalty per item is $\hat{H}_{1N} = 0.59639$, but the differences between estimates for a and γ_j from model M_{1N} and from model M_1 have absolute values as large as 0.0002. Thus the NAEP approach does not appear attractive given that better accuracy can be obtained by Gauss-Hermite integration with fewer quadrature points. A more promising alternative is adaptive Gauss-Hermite quadrature.

2 Adaptive Gauss-Hermite Quadrature

Adaptive Gauss-Hermite quadrature has been used in statistical analysis for some time (Naylor & Smith, 1982). The key feature involves careful use of distinct linear transformations designed for each integral $J(V\exp(\mathbf{X}_i'\boldsymbol{\lambda}))$. The linear transformations vary for each iteration of an iterative algorithm. For iteration $t \ge 0$, let the maximum-likelihood estimate $\hat{\boldsymbol{\lambda}}$ of $\boldsymbol{\lambda}$ have an approximation $\boldsymbol{\lambda}_t$ with coordinates λ_{jt} for $1 \le j \le q$, let

$$V_t = \prod_{j=1}^{q} [1 + \exp(\lambda_{jt})]$$

be the approximation of V that corresponds to $\boldsymbol{\lambda}_t$, let

$$P_{jt} = \frac{\exp(\lambda_{jt})}{1 + \exp(\lambda_{jt})}$$

be the approximation of P_j that corresponds to λ_{jt}, and let $Q_{jt} = 1 - P_{jt}$. In the case of M_1,

$$\boldsymbol{\lambda}_t(\theta) = a_t\theta\mathbf{1} - \boldsymbol{\gamma}_t$$

for an approximation a_t to the maximum-likelihood estimate of a and an approximation $\boldsymbol{\gamma}$ to the maximum-likelihood estimate of $\boldsymbol{\gamma}$. Similarly, in the 2PL case,

$$\boldsymbol{\lambda}_t(\theta) = \theta\mathbf{a}_t - \boldsymbol{\gamma}_t$$

for an approximation $\mathbf{a}_t$ to the maximum-likelihood estimate of $\mathbf{a}$ with coordinates a_{jt} for $1 \le j \le q$ and an approximation $\boldsymbol{\gamma}$ to the maximum-likelihood estimate of $\boldsymbol{\gamma}$.

Let

$$L_{it} = \log V_t + \mathbf{X}_i' \boldsymbol{\lambda}_t + \log \phi$$

for each examinee i, so that L_{it} is the logarithm of the probability density of θ_i that corresponds to the item logit function $\boldsymbol{\lambda}_t$. Observe that for both models M_1 and M_2, L_{it} is a strictly concave function such that $L_{it}(\theta)$ approaches $-\infty$ as $|\theta|$ approaches ∞. Thus L_{it} has a unique maximum μ_{it}. Under the 1PL model, the derivative L_{it}' of L_{it} satisfies

$$L_{it}'(\mu_{it}) = a_t X_{i+} - a_t \sum_{j=1}^{q} P_{jt}(\mu_{it}) - \mu_{it} = 0,$$

and the second derivative of L_{it} at μ_{it} is

$$L_{it}''(\mu_{it}) = -1 - a_t^2 \sum_{j=1}^{q} P_{jt}(\mu_{it}) Q_{jt}(\mu_{it}) < 0.$$

Under the 2PL model,

$$L_{it}'(\mu_{it}) = \sum_{j=1}^{q} a_{jt}[X_{ij} - P_{jt}(\mu_{it})] - \mu_{it} = 0,$$

and the second derivative of L_i at t_i is

$$L_{it}''(\mu_{it}) = -1 - \sum_{j=1}^{q} a_{jt}^2 P_{jt}(\mu_{it}) Q_{jt}(\mu_{it}) < 0.$$

Adaptive quadrature uses the location μ_{it} of the maximum of L_{it} and on $\sigma_{it} = 1/[-L_{it}''(\mu_{it})]^{1/2}$.

Consider use of a scale change based on the linear function z_{it} such that

$$z_{it}(\theta) = (\theta - \mu_{it})/\sigma_{it}$$

for real θ. Then z_{it} has inverse

$$z_{it}^{-1}(z) = \mu_{it} + \sigma_{it} z$$

for real z. Let

$$\Delta_{it} = L_{it}(z_{it}^{-1}) - L_{it}(\mu_{it}) - \log \phi,$$

so that Δ_{it} is a rescaled version of L_{it} that has first and second derivative 0 at 0. Consider $\mathbf{p}$ in S such that (1) and (5) hold. Let

$$g_{it} = [\boldsymbol{\lambda} - \boldsymbol{\lambda}_t]' \mathbf{X}_i - \log V + \log V_t,$$

so that

$$p(\mathbf{X}_i) = J(\exp(g_{it} + L_{it})/\phi)$$

and

$$\ell(\mathbf{p}) = \sum_{i=1}^{n} \log J(\exp(g_{it} + L_{it})/\phi).$$

Let

$$h_{it} = g(z_{it}^{-1}).$$

Then use of standard formulas from calculus for change of variables shows that

$$\ell(\mathbf{p}) = \sum_{i=1}^{n} \log(\sigma_{it}) + \sum_{i=1}^{n} L_{it}(\mu_{it}) + \sum_{i=1}^{n} \log J(\exp(\Delta_{it} + h_{it})). \tag{9}$$

Iteration t proceeds as if $\ell(\mathbf{p})$ in (9) is replaced by

$$\ell_r(\mathbf{p}) = \sum_{i=1}^{n} \log(\sigma_{it}) + \sum_{i=1}^{n} L_{it}(\mu_{it}) + \sum_{i=1}^{n} \log J_r(\exp(\Delta_{it} + h_{it})).$$

The potential advantage of the adaptive approximation is that Δ_{it} is normally much less variable than is $V_t \exp(\mathbf{X}_i' \boldsymbol{\lambda}_t)$, as is evident from consideration of derivatives at 0. In addition, iteration t is based on behavior of $\boldsymbol{\lambda}$ for $\boldsymbol{\lambda}$ close to $\boldsymbol{\lambda}_t$. The potential complication is that μ_{it} must be found by an iterative computation essentially the same as that used to find the posterior mean of an ability distribution.

For both models M_1 and M_2 for the Praxis example, it is quite adequate to let $r = 9$. In both cases, maximum-likelihood estimates of parameters are accurate to within 0.00001. The adaptive procedure was substantially faster than the ordinary Gauss-Hermite quadrature with 32 points. In the 2PL case, on the particular personal computer on which the calculations were made, the adaptive quadrature calculations required about 30 seconds, and the ordinary routine required about 90 seconds. Even use of $r = 5$ only results in a modest decrease in accuracy, for parameter estimates in the 2PL case remain accurate to about 0.0001. Thus adaptive Gauss-Hermite quadrature is quite attractive. It is reasonable to expect that adaptive Gauss-Hermite quadrature is increasingly attractive as the number of items increases, so that the σ_{it}, $1 \le i \le n$, become increasingly small.

3 Conclusions

Results in this report support the conclusion that adaptive Gauss-Hermite quadrature is quite attractive in marginal estimation when a normal ability distribution is assumed. Generalization to multivariate normal ability distributions should be quite feasible. Such generalizations should be important in NAEP models that employ such ability distributions.

References

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, *35*, 179–197.

Breiman, L. (1968). *Probability.* Reading, MA: Addison-Wesley.

Davis, P. J., & Polonsky, I. (1965). Numerical interpolation, differentiation, and integration. In M. Abramowitz & I. A. Stegun (Eds.), *Handbook of mathematical functions* (pp. 875–924). New York: Dover.

Gilula, Z., & Haberman, S. J. (1994). Models for analyzing categorical panel data. *Journal of the American Statistical Association*, *89*, 645–656.

Gilula, Z., & Haberman, S. J. (1995). Prediction functions for categorical panel data. *The Annals of Statistics*, *23*, 1130–1142.

Haberman, S. J. (1988). A stabilized Newton-Raphson algorithm for log-linear models for frequency tables derived by indirect observation. *Sociological Methodology*, *18*, 193–211.

Haberman, S. J. (2004). *Joint and conditional maximum likelihood estimation for the Rasch model for binary responses* (ETS RR-04-20). Princeton, NJ: ETS.

Holland, P. W. (1990). The Dutch identity: A new tool for the study of item response models. *Psychometrika*, *55*, 5–18.

Naylor, J. C., & Smith, A. F. M. (1982). Applications of a method for the efficient computation of posterior distributions. *Applied Statistics*, *31*, 214.

Ralston, A. (1965). *A first course in numerical analysis.* New York: McGraw-Hill.

Stroud, A., & Secrest, D. (1966). *Gaussian quadrature formulas.* Englewood Cliffs, NJ: Prentice-Hall.