Gauss-Hermite Quadrature: Numerical or Statistical Method?

Vahid Partovi Nia^{*} Institute of Mathematics, Ecole Polytechnique Fédéral de Lausanne

June 9, 2006

Abstract

Gauss-Hermite Quadrature (GHQ) is often used for numerical approximation of integrals with Gaussian kernels. In generalized linear mixed models random effects are assumed to have Gaussian distributions, but often the marginal likelihood, which has the key role in parameter estimation and inference, is analytically intractable. In addition to Monte Carlo methods, first or second order Taylor expansion, Laplace approximation or GHQ are feasible tools for numerical evaluation of the integrals. In this paper we review the key ideas of GHQ. Nonparametric Maximum Likelihood (NPML) estimation is shown to be a flexible version of GHQ. A binary nested random effects model is fitted to a real data set using GHQ.

Keywords: Function Interpolation; Generalized Linear Mixed Model; Hermite Polynomial; Mixing Distribution; Nonparametric Maximum Likelihood.

^{*}This work has been done at the Ferdowsi University of Mashhad.

1 Introduction

Generalized linear models are powerful tools for data analyses, specially for discrete response data. Overdispersion is a common problem in many data sets which are fitted by such models. This problem makes inference on parameters unreliable. Adding a scale parameter is suggested in these cases but, this ignores the source of overdispersion. Generalized linear mixed models are often used to model overdispersion, but then the marginal likelihood is usually analytically intractable. Different approximation methods are available, of which GHQ is one of the most used. Poor approximation of marginal likelihood may occur using GHQ, and knowing its basics helps the user to recognize these cases.

1.1 Interpolation

The fundamental idea of numerical approximation of integrals is approximation of integrable functions. Hence, function interpolation plays the key role. Assume that n observations of an unknown function f are available: $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ and good approximation to f is desired.

By Taylor's theorem, the function f which is contiguously arbitrary order differentiable on the closed interval [a, b] can be approximated as a polynomial function. Approximation can be improved by adding higher orders.

1.2 Ordinary Interpolation

We look for a polynomial of degree at most n - 1 that passes through all n observations $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$. This can be found analytically by solving a system of n polynomial equations, which is often hard to solve, especially for large n. Least squares fit of a saturated polynomial model is another solution. In this case the residual sum of squares is exactly equal to zero which is not statistically interesting. To find the polynomial which passes through all n observations we may use the auxiliary polynomial functions as well:

$$\pi(x) = (x - x_1) (x - x_2) \cdots (x - x_n),$$

$$l_i(x) = \frac{(x - x_1) (x - x_2) \cdots (x - x_{i-1}) (x - x_{i+1}) \cdots (x - x_n)}{(x_i - x_1) (x_i - x_2) \cdots (x_i - x_{i-1}) (x_i - x_{i+1}) \cdots (x_i - x_n)}.$$
(1)

The function $l_i(x)$ is a polynomial function of degree at most n-1. By definition $l_i(x_j) = \delta_{ij}$ where δ_{ij} is the Kronecker delta. Hence,

$$f(x) = \sum_{i=1}^{n} l_i(x) y_i,$$
 (2)

that is a polynomial function of degree n-1 passes through n data points, so is the desired polynomial.

1.3 Hermite Interpolation

Assume that we have observations in triples (x_i, y_i, y'_i) , i = 1, ..., n, where y'_i is the first derivative (gradient) of reference unknown function f at the point x_i . Hermite interpolation is a generalization of ordinary interpolation when gradients are available in addition to the values. Fort (1948) generalized it to derivatives of arbitrary order.

We look for a polynomial function of order at most 2n - 1 that passes through all values (x_i, y_i) , with first derivative equal to y'_i at x_i , i = 1, ..., n. Similar to (2) we write

$$f(x) = \sum_{i=1}^{n} h_i(x) y_i + \sum_{i=1}^{n} \overline{h}_i(x) y'_i,$$

where $h_i(x)$ and $\overline{h}_i(x)$, i = 1, ..., n, are polynomials of order at most 2n - 1. They satisfy

$$h_i(x_j) = \delta_{ij}, \qquad h'_i(x_j) = 0, \qquad \overline{h}'_i(x_j) = \delta_{ij}.$$
(3)

From (1) we know $l_i(x_j) = \delta_{ij}$ is a polynomial of order n - 1, we conclude $l_i^2(x)$ that is a polynomial of order 2n - 2 seems to be a good candidate for h_i and $\overline{h_i}$. To make h_i and $\overline{h_i}$ polynomials of order 2n - 1 we assume:

$$h_i(x) = r_i(x) l_i^2(x), \qquad \overline{h}_i(x) = s_i(x) l_i^2(x),$$

where $r_i(x)$ and $s_i(x)$ are both at most linear functions of x. We find $r_i(x)$ and $s_i(x)$ such that h_i and \overline{h}_i satisfy (3). Hence,

$$r_i(x_i) = 1, \quad s_i(x_i) = 0, \quad r'_i(x_i) + 2l'_i(x_i) = 0, \quad s'_i(x_i) = 1.$$
 (4)

The functions $s_i(x) = x - x_i$ and $r_i(x) = 1 - 2l'_i(x_i)(x - x_i)$ satisfy (4), so we have the Hermite interpolation formulae

$$h_i(x) = [1 - 2l'_i(x_i)(x - x_i)] l_i^2(x), \qquad \overline{h}_i(x) = (x - x_i) l_i^2(x).$$

Figures 1 and 2 compare ordinary interpolation and Hermite interpolation for two functions with 5 nodes established at points (-2.02,-0.96,0,0.96,2.02). Ordinary interpolation gives a polynomial of degree 4 and Hermite interpolation gives a polynomial of degree 9.



2 Integral Evaluation

A way to do numerical approximation of integrals is to divide integral bound [a, b] into equally lengthed subintervals and evaluate Riemann upper and lower bounds. This method is feasible but often computationally time consuming. We look for a method to approximate the integral using fewer summations and greater accuracy.



Assume an integral of the form $\int_{a}^{b} f(x) w(x) dx$. The functions f(x), w(x) and g(x) = f(x) w(x) are called the integrable, weighting and product functions respectively. Numerical approximation using quadrature method is based on situations that the form of f is too complex to calculate the integral using antiderivative function or even analytically intractable. However, integration of $\int_{a}^{b} P_n(x) w(x) dx$ is easy. The function $P_n(x)$ is an *n*th order polynomial approximation of f(x) and hence we expect two integrals to be almost equal if we could approximate it well in [a, b].

2.1 Hermite Quadrature

To approximate an integral of the form $\int_{a}^{b} f(x) w(x) dx$ using Hermite Quadrature (HQ), function f is substituted by its Hermite interpolation:

$$\int_{a}^{b} f(x) w(x) dx \doteq \int_{a}^{b} P_{2n-1}(x) w(x) dx = \sum_{i=1}^{n} H_{i} y_{i} + \sum_{i=1}^{n} \overline{H}_{i} y_{i}',$$

where

$$H_{i} = \int_{a}^{b} h_{i}(x) w(x) dx = \int_{a}^{b} [1 - 2l'_{i}(x) (x - x_{i})] l_{i}^{2}(x) w(x) dx, \quad (5)$$

$$\overline{H}_{i} = \int_{a}^{b} \overline{h}_{i}(x) w(x) dx = \int_{a}^{b} (x - x_{i}) l_{i}^{2}(x) w(x) dx.$$

The integrals H_i and \overline{H}_i are known as the quadrature weights. As an example figures 3 and 4 show function f(x) w(x) versus P(x) w(x) both with $w(x) = e^{-x^2}$.



Fig. 3: $f(x) = 0.5e^{-(x-1)^2} + e^{-(x+1)^2}$.

As the figures show, the choice of nodes in [a, b] plays the key role in a good approximation of integrals. Thompson (2000) suggests using a random process to choose nodes named the random quadrature (RQ) method.

2.2 Gaussian Quadrature

Gaussian quadrature (GQ) maybe regarded as another reformulation of (5) in situations where \overline{H}_i vanishes:

$$\overline{H}_{i} = \int_{a}^{b} \overline{h}_{i}(x) w(x) dx = \int_{a}^{b} \pi(x) l_{i}(x) w(x) dx,$$

$$H_{i} = \int_{a}^{b} h_{i}(x) w(x) dx = \int_{a}^{b} l_{i}^{2}(x) w(x) dx - 2 \int_{a}^{b} l_{i}'(x) \pi(x) l_{i}(x) w(x) dx.$$

The polynomial $\pi(x)$ is an arbitrary polynomial function of degree n. If $\pi(x)$ is a polynomial perpendicular to $l_i(x)$ with respect to the weighting function w(x), that is

$$\int_{a}^{b} \pi(x) l_{i}(x) w(x) dx = 0, \qquad (6)$$

the second term of HQ formulae vanishes and simplifies as follows:

$$\int_{a}^{b} f(x) w(x) dx \doteq \sum H_{i} y_{i}.$$

Quadrature nodes (x_i) are roots of polynomial $\pi(x)$, (1). Note that we fit a polynomial function of degree 2n - 1 to the integrable function f using just n pairs (x_i, y_i) !



2.3 Gauss–Hermite Quadrature

GQ method for the Gaussian kernel as the weighting function $(w(x) = e^{-x^2})$ with integral bounds $(-\infty, +\infty)$ often is called Gauss-Hermite quadrature (GHQ). Fortunately function π that satisfies (6) is the well-known Hermite polynomial. The first five Hermite polynomials are:

$$\begin{array}{rcl} H_0\left(x\right) &=& 1, & H_1\left(x\right) = 2x, \\ H_2\left(x\right) &=& 4x^2 - 2, & H_3\left(x\right) = 8x^3 - 12x, \\ H_4\left(x\right) &=& 16x^4 - 48x^2 + 12, & H_5\left(x\right) = 32x^5 - 160x^3 + 120x. \end{array}$$

After some algebraic calculations we find quadrature weights as follows:

$$H_i = \frac{2^{n+1} n! \sqrt{\pi}}{H_{n+1}(x_i)},$$

where $x_1, ..., x_n$ are quadrature nodes, the roots of $H_n(x)$ the Hermite polynomial of order n. The main computational problem in GHQ method is finding the form of $H_n(x)$. Fortunately the iterative Hermite polynomial unity helps to find analytical form of $H_n(x)$ from $H_{n-1}(x)$ and $H_{n-2}(x)$:

$$H_{n}(x) = 2xH_{n-1}(x) - 2(n-1)H_{n-2}(x)$$

The Hermite polynomials can be calculated using algebraic software such as MAPLE or MATHEMATICA up to any arbitrary order. However, quadrature nodes and weights up to 20 nodes are available in Abramowitz and Stegun (1965).

Figure 5 represents quadrature nodes of GHQ method (horizontal axis) versus quadrature weights (vertical axis) that confirms quadrature nodes are symmetric about zero.

2.4 Adaptive Gauss–Hermite Quadrature

Quadrature nodes in both GQ and GHQ methods are not arbitrary unlike the HQ method, so appropriate choice of number of quadrature points to approximate the integrable and/or product functions accurately has an important role. Sometimes many quadrature points may be needed to approximate the functions adequately, depending on form of curvature and the nonvanishing domain of the integrable and product functions. Centralization of

the product function about zero or standardization is recommended in Liu and Pierce (1994) under the name of adaptive Gauss-Hermite quadrature (AGHQ). AGHQ often decreases the number of required quadrature points, especially for product functions with maxima far from zero, but it inflates the computational complexity.

Fast vanishing tails of the Gaussian kernel (weighting function) help GHQ to approximate functions accurately with a moderate number of quadrature points. Equivalence of GHQ using many quadrature points with AGHQ using few has been reported for a binary random effects model by Lessaffre and Spiessens (2000). Rabe-Hesketh et al. (2002) have described AGHQ and have implemented it in STATA. SAS PROC NLMIXED also uses AGHQ to fit generalized linear mixed models. Often use of different numbers of quadrature points is suggested to investigate stability of the approximated marginal likelihood.



Fig. 5: Gauss-Hermite quadrature nodes and weights.

3 Application to a Binary Random Effects Model

Assume the following three level binary random effects model that often is used for analysis of repeated measurements, Anderson and Aitkin (1985).

$$y_{ijk} \mid \mu, O_i, R_{ij} \stackrel{i.i.d}{\sim} Bin(n_P, p_{ij}), \qquad (7)$$

$$\log\left(\frac{p_{ij}}{1 - p_{ij}}\right) = \mu + O_i + R_{ij},$$

 $i = 1, ..., n_O, \quad j = 1, ..., n_R, \quad k = 1, ..., n_P,$
 $O_i \sim N\left(0, \sigma_O^2\right), \qquad R_{ij} \sim N\left(0, \sigma_P^2\right).$

Writing $y_{ij.} = \sum_{k=1}^{n_P} y_{ijk}$, after some algebraic calculations the marginal log-likelihood $\ell(\mu, \sigma_O, \sigma_P)$ can be written

$$\sum_{i=1}^{n_O} \log \left\{ \int_{-\infty}^{+\infty} \left(\prod_{j=1}^{n_R} \int_{-\infty}^{+\infty} \left[\frac{\exp\{\left(\mu + \sqrt{2}\sigma_O O_i + \sqrt{2}\sigma_R R_{ij}\right) y_{ij.}\}}{1 + \exp\{\left(\mu + \sqrt{2}\sigma_O O_i + \sqrt{2}\sigma_R R_{ij}\right) y_{ij.}\}^{n_P}} \right] e^{-R_{ij}^2} dR_{ij} \right\} e^{-O_i^2} dO_i \right\}$$

Unfortunately the integrals above are analytically intractable. The approximated marginal log-likelihood using GHQ is

$$\ell\left(\mu,\sigma_{O},\sigma_{P}\right)\stackrel{\approx}{\propto} \sum_{i=1}^{n_{O}}\log\left[\sum_{v=1}^{n_{v}}w_{v}\left(v\right)\left\{\prod_{j=1}^{n_{R}}\sum_{u=1}^{n_{u}}w_{u}\left(u\right)\left(\frac{\exp\left[\left\{\mu+\sqrt{2}\sigma_{O}x_{v}\left(v\right)+\sqrt{2}\sigma_{R}x_{u}\left(u\right)\right\}y_{ij.}\right]}{1+\exp\left[\left\{\mu+\sqrt{2}\sigma_{O}x_{v}\left(v\right)+\sqrt{2}\sigma_{R}x_{u}\left(u\right)\right\}y_{ij.}\right]^{n_{P}}}\right)\right\}\right]$$

The triples (n_u, x_u, w_u) and (n_v, x_v, w_v) are the number of quadrature points, quadrature nodes and quadrature weights with respect to the level 2 and level 1 of the data respectively.

Approximated marginal log-likelihood can be maximized by the standard nlminb function of S-PLUS using Quasi-Newton algorithm. Roots of Hermite polynomials also are achievable using polyroot.

The model (7) has been used to investigate effect of appraisers and trial repeats on ability of distinguishing correct parts from failed parts. The data is reported in Automotive Industry Action Group (2002) p. 127 and analyzed

by another method there using Kappa measure of agreement. Here we fit binary mixed effect model.

Parameter estimations and 95% confidence intervals found using parametric bootstrap is reported in the following table.

	ML Estimates	95% Confidence Interval
μ	2.91	(2.47, 3.37)
σ_O	0.00	(0.00, 0.48)
σ_R	0.71	(0.00, 1.10)

Insignificant variance components agrees with analysis of Automotive Industry Action Group (2002) that is insignificant appraiser and trial repeats effect. To investigate validity of estimated parameters slice approximated log-marginal likelihood with respect to the parameters has been shown in Figures 6.

4 Nonparametric Maximum Likelihood

GHQ and AGHQ are useful tools for parameter estimation with Gaussian mixing distributions. For other parametric mixing distributions, McCulloch (1997) recommends MCEM and MCNR algorithms for ML estimation of parameters, which works well especially in high-dimensional integrals. However, Heckman and Singer (1984) showed that parameter estimates in mixed models are sensitive to the choice of mixing distribution, Davies (1987) also reported similar results. Nonparametric Maximum Likelihood (NPML) of mixing distribution is introduced by Laird (1978). Aitkin (1999) made a general framework for NPML estimation of parameters in generalized linear mixed models and has implemented it in GLIM.

The form of approximated likelihood using GHQ helps to achieve NPML estimators:

$$l\left(\boldsymbol{\theta}\right) = \sum_{i} \log\left(\int f_{\mathbf{y}|\tau_{i}} f_{\tau_{i}} d\tau_{i}\right).$$

The conditional distribution $f_{\mathbf{y}|\tau_i}$ also may have similar integrations in itself that can be handled similarly, so we can approximate $l(\boldsymbol{\theta})$ as

$$l(\boldsymbol{\theta}) \doteq \sum_{i} \log \left(\sum_{v} w_{v}(v) f_{\mathbf{y}|\tau_{i}}(x_{v}(v)) \right).$$



Fig. 6: Slice marginal likelihoods.

For the case of Gaussian mixing distribution (f_{τ} is Gaussian), w_v and x_v are known as quadrature nodes and weights in GHQ as we discussed. However, in general case it is not easy to find. After observing the data, maximum likelihood is used to estimate w_v and x_v in addition to the parameter vector $\boldsymbol{\theta}$. NPML approach may be regarded as a method which discretizes unknown mixing distribution and estimates position of discretization and probability mass at that position both, in addition to the model parameters. Fotouhi (2003) compared MCMC, GHQ and NPML methods and found NPML is preferable in most cases. Unfortunately computational complexity of NPML increases exponentially as the number of sample size increases. Vermunt (2003) introduced a modified EM algorithm to solve the exponential increase of the computational complexity for NPML estimates in three-level models.

Acknowledgement

I am grateful to Prof. Anthony C. Davison for reviewing the paper and his helpful suggestions which improved the draft manuscript and to my master thesis professors Mohammad M. Tabatabey and Gholam H. Shahkar for their encouragement.

References

Abramowitz, **M. and Stegun**, **J.** (Eds.) (1965) *Handbook of Mathematical Functions* Dover, New York.

Aitkin, M. (1999) A General Maximum Likelihood Analysis of Variance Components in Generalized Linear Models Biometrics, 55, 117–128.

Anderson, D. A. and Aitkin, M. (1985) Variance Component Models with Binary Response: Interviewer Variability Journal of Royal Statististical Society, B, 47, 203–210.

Automotive Industry Action Group (2002) Measurement Systems Analysis Reference Manual 3rd Ed. Detroit, MI.

Davies, R. B. (1987) Masspoint Method for Dealing with Nuisance Parameters in Longitudinal Studies in Longitudinal Data Analysis, 88–109.

Fort, D. (1948) Finite Differences and Difference Equations in Real Domain Oxford University Press, Fair Lawn, W.E.

Fotouhi, A. R. (2003) Comparison of Estimation Procedures for Nonlinear Multilevel Models Joural of Statistical Software, 10, 8–28.

Thompson, J. (2000) Simulation, A Modeler's Approach John Wiley and Sons, New York.

Heckman, J. and Singer, B. (1984) A Method for Minimizing Impact of Distributional Assumptions in Econometric Models for Duration Data Econometrica, 52, 2, 271-320.

Hildebrand, F. B. (1988) Introduction to Numerical Analysis, McGraw Hill, India.

Laird, N. (1978) Nonparametric Maximum Likelihood Estimation of a Mixing Distribution Journal of the American Statistical Association, 73, 805–811. Lessaffre, E. and Spiessens, B. (2000) On the Effect of Number of Quadrature Points in a Logistic Random Effects Model: An Example, Applied Statistics, 50, 3, 325–335.

Liu, Q. and Pierce, D.A. (1994) A Note on Gauss-Hermite Quadrature Biometrika, 81, 3, 624–629.

McCulloch, C. E. (1997) Maximum Likelihood Algorithms for Generalized Linear Mixed Models Journal of the American Statistical Association, 92, 162–170.

Rabe-Hesketh, S., Skrondal, A. and Pickles, A. (2002) Reliable Estimation in Generalized Linear Mixed Models using Adaptive Quadrature The Stata Journal, 2, 1–21.

Vermunt, J. K. (2003) An Expectation-Maximization Algorithm for Generalized Linear Three-Level Models Multilevel Newletter, 14, 2, 3–12.