Distribution Theory and Inference for Polynomial-Normal Densities

M. Evans
Department of Statistics
University of Toronto
Toronto, Ontario M5S 1A1

T. Swartz
Department of Mathematics and Statistics
Simon Fraser University
Burnaby, British Columbia V5A 1S6

Key Words and Phrases : polynomial-normal densities; distribution theory; conditional inference; Gram-Charlier approximations; importance sampling.

ABSTRACT

This paper considers a class of densities formed by taking the product of nonnegative polynomials and normal densities. These densities provide a rich class of distributions that can be used in modelling when faced with non-normal characteristics such as skewness and multimodality. In this paper we address inferential and computational issues arising in the practical implementation of this parametric family in the context of the linear model. Exact results are recorded for the conditional analysis of location-scale models and an importance sampling algorithm is developed for the implementation of a conditional analysis for the general linear model when using polynomial-normal distributions for the error.

1. INTRODUCTION

Parametric statistical inference often relies on an assumption of normality for an error distribution. This assumption, although supported by the Central

Limit Theorem, is widely recognized as artificial. It would therefore be useful if there were a family of distributions that could be used in applied problems to model a variety of non-normal shapes and for which exact (as opposed to asymptotic) inferences were available. In general, the difficulty with exact inference is that it often leads to intractable expressions; expressions involving difficult and possibly high-dimensional integrals.

This paper discusses the family of polynomial-normal densities. These densities are defined as nonnegative polynomials times normal densities. They can be used in regression analysis to model a wide variety of non-normal shapes. Also the form of these densities allows for the development of algorithms for exact conditional inference methods. These two points suggest that this family can make an important contribution to practical data analysis.

In Section 2 a closure property for this family under linear combinations is proved that does not appear to exist in the literature. In Section 3 exact results are obtained for the implementation of a conditional analysis, as discussed in Fisher(1934) or Fraser(1979), of a location-scale model when the distribution form is a known polynomial-normal density. In Section 4 a fitting algorithm, for selecting a polynomial-normal density when the distribution form is not known, is developed. In Section 5 the relation of the polynomial-normal family to truncated Gram-Charlier expansions is discussed; see for example Cramer(1946), Kendall and Stuart(1958), Johnson and Kotz(1970) for a discussion of these. Also in Section 5 the relation of the contents of this paper to the work of Bartlett(1935), Gayen(1949, 1950a, 1950b, 1951), Davis(1976) and Geweke(1989) is discussed.

In Section 6 an importance sampling algorithm, based on the multivariate t-distribution, is developed for the implementation of a conditional analysis of a general linear model when the error distribution is polynomial-normal. This permits the practical analysis of complicated data sets allowing for highly non-normal shapes for the error distribution. The development of this algorithm relies intrinsically on the form of the polynomial-normal densities to produce finite variance Monte Carlo estimates for various quantities of inferential interest. This is the most important contribution of the paper and it is the content of Proposition 5. A numerical example of such an analysis is also presented. Copies of Fortran programs for all computer algorithms used in the paper can be obtained by writing the authors.

2. THE FAMILY AND SOME DISTRIBUTION THEORY

A polynomial-normal density function takes the form

$$f(x) = kp\left(\frac{x-\tau_1}{\tau_2}\right)\frac{1}{\tau_2}\phi\left(\frac{x-\tau_1}{\tau_2}\right) \tag{1}$$

where $\tau_1 \in \Re$, $\tau_2 > 0$, p is a nonnegative polynomial, ϕ is the $N(0,1)$ density and the norming constant k is specified by

$$k^{-1} = \int_{-\infty}^{\infty} p(x)\phi(x)dx. \tag{2}$$

We will study this class of densities with no other restrictions on p but note the possibility of examining interesting sub-classes by adding further constraints on the polynomial.

Since the moments of the standard normal distribution are all known, the integral (2) can be easily evaluated. Alternatively, as p is a polynomial of degree $2m$, the integral (2) can be exactly evaluated by a Gauss-Hermite rule of order $m+1$. That is to say

$$k^{-1} = \sum_{i=1}^{m+1} w_i p(x_i)$$

where the w_i and x_i are Gauss-Hermite weights and points respectively and this is perhaps the preferred method of evaluation as it avoids the need to calculate high-order moments of the normal. For more information on Gauss-Hermite quadrature, the theory of orthogonal polynomials and quadrature in general, the reader is referred to Davis and Rabinowitz(1984).

Similarly, if the random variable X has a polynomial-normal density as specified by (1) with $\deg(p) = 2m$, then

$$E(X^n) = \int_{-\infty}^{\infty} (\tau_1 + \tau_2 x)^n kp(x)\phi(x)dx$$

which can be evaluated exactly by a Gauss-Hermite rule of order $\lceil (n+2m+1)/2 \rceil$. We note that the moment calculations can also be obtained by using a lower order Gauss rule corresponding to the particular polynomial-normal density. The weights and points for these quadrature rules can be derived using the algorithm of Golub and Welsch(1969) and Hermite integration or via an algorithm of Kautsky and Golub(1983).

In addition to the availability of the moments of X, an application of Rodrigue's formula (see Davis and Rabinowitz(1984)) yields a convenient expression for the distribution function $F(x)$. It is given by

$$F(x) = \Phi\left(\frac{x-\tau_1}{\tau_2}\right) - \sum_{j=1}^{2m} \frac{c_j}{\sqrt{j}}\, h_{j-1}\left(\frac{x-\tau_1}{\tau_2}\right)\phi\left(\frac{x-\tau_1}{\tau_2}\right) \tag{3}$$

where Φ is the distribution function for the $N(0,1)$, h_j is the j^{th} orthonormal Hermite polynomial and

$$c_j = \int_{-\infty}^{\infty} h_j(x) kp(x) \phi(x) dx$$

which can be evaluated exactly via Gauss-Hermite integration. Therefore given expression (3) and a uniform(0,1) variate u, a random polynomial-normal variate x can be easily generated by solving the equation $F(x) = u$ through bisection.

An interesting result concerns the closure of the polynomial-normal family under linear operations. This is developed more fully below and stated precisely in Proposition 2.

Proposition 1. Consider a density function $f(x)$ with corresponding moment generating function $m(t)$. The density $f(x)$ is polynomial-normal of degree $2m$ if and only if $m(t) = e^{\tau_1 t + \tau_2^2 t^2/2} q(t)$ where $q(t)$ is a nonnegative polynomial of degree $2m$.

Proof : Suppose that $f(x)$ is a polynomial-normal density of degree $2m$. Then it's moment generating function is

$$m(t) = \int_{-\infty}^{\infty} e^{xt} kp\left(\frac{x-\tau_1}{\tau_2}\right) \frac{1}{\tau_2} \phi\left(\frac{x-\tau_1}{\tau_2}\right) dx = e^{\tau_1 t + \tau_2^2 t^2/2} q(t)$$

where

$$q(t) = \int_{-\infty}^{\infty} kp(x + \tau_2 t) \phi(x) dx \tag{4}$$

is a nonnegative polynomial of degee $2m$ in t.

Conversely suppose that $m(t) = e^{\tau_1 t + \tau_2^2 t^2/2} q(t)$ for some nonnegative polynomial q of degree $2m$. Then the characteristic function corresponding to $f(x)$ is $m(it)$ and by the Inversion Theorem and some straight-forward manipulations we obtain

$$\begin{aligned} f(x) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} q(it) exp\left[-\frac{\tau_2^2}{2}(t^2 + 2it(x-\tau_1)/\tau_2^2)\right] dt \\ &= \phi\left(\frac{x-\tau_1}{\tau_2}\right) \frac{1}{\tau_2} \int_{-\infty}^{\infty} q\left(\frac{iz}{\tau_2} + (x-\tau_1)/\tau_2^2\right) \phi(z) dz. \end{aligned}$$

Now $\int_{-\infty}^{\infty} q\left(\frac{iz}{\tau_2} + (x-\tau_1)/\tau_2^2\right) \phi(z) dz$ is a polynomial of degree $2m$ in x and since f is a density it must be nonnegative.

Note that (4) provides a convenient formula for computing the moment generating function of a polynomial-normal density. However, not all nonnegative polynomials q give rise to valid moment generating functions. The following corollary characterizes those that do.

Corollary 1. Consider a function $m(t) = e^{\tau_1 t + \tau_2^2 t^2/2} q(t)$ where q is a nonnegative polynomial. Then m is a moment generating function if and only if $q(0) = 1$ and

$$p(x) = \int_{-\infty}^{\infty} q(\frac{it}{\tau_2} + \frac{x}{\tau_2})\phi(t)dt$$

is a nonnegative polynomial. Further when m is a moment generating function it is associated with density (1) where p is given as above and $k = 1$.

Proof : Suppose that m is a moment generating function. Then $1 = m(0) = q(0)$ and the nonnegativity of p follows from the if part of Proposition 1.

Conversely put $f(x) = p(\frac{x-\tau_1}{\tau_2})\phi(\frac{x-\tau_1}{\tau_2})\frac{1}{\tau_2}$ and note that f is nonnegative everywhere. Applying the proof of Proposition 1, which does not require f be a probability density, only that $m(t)$ exist in an open interval about 0, then $q(t)$ is given by (4) and this implies $1 = q(0) = \int_{-\infty}^{\infty} p(x)\phi(x)dx$ which implies f is a density.

Proposition 2. If $X_1, \ldots, X_n$ are mutually statistically independent random variables and X_i has moment generating function $e^{\tau_{1i} t + \tau_{2i}^2 t^2/2} q_i(t)$ where q_i is a nonnegative polynomial of degree $2m_i$, then $Y = a_1 X_1 + \ldots + a_n X_n + c$ has density $p\left(\frac{y-\mu}{\sigma}\right)\frac{1}{\sigma}\phi\left(\frac{y-\mu}{\sigma}\right)$ where $\mu = c + \sum_{i=1}^n a_i \tau_{1i}$, $\sigma^2 = \sum_{i=1}^n a_i^2 \tau_{2i}^2$ and p is the nonnegative polynomial of degree $2(m_1 + \ldots + m_n)$ given by

$$p(x) = \int_{-\infty}^{\infty} \prod_{j=1}^{n} q_j\left(\frac{i a_j t}{\sigma} + \frac{a_j x}{\sigma}\right)\phi(t)dt. \tag{5}$$

Proof : The moment generating function of Y is $e^{(c + \sum_{j=1}^n a_j \tau_{1j})t + \sum_{j=1}^n a_j^2 \tau_{2j}^2 t^2/2} \times \prod_{j=1}^n q_j(a_j t)$ and thus by Proposition 1, Y has a polynomial-normal density and by Corollary 1, $p(x)$ is given by (5).

The Hermite expansion of p in (5) can be evaluated by Hermite integration and thus we have an exact expression for the density of Y.

To deal effectively with the polynomial-normal family of densities we require a convenient parametrization of the polynomials. We use the fact that

a polynomial is nonnegative if and only if its roots are conjugate pairs. Hence we write

$$p(x) = \prod_{i=1}^{m} (\lambda_i x - 1)(\overline{\lambda}_i x - 1) \tag{6}$$

where $\lambda_1, \ldots, \lambda_m \in \mathcal{C}$ and hence $1/\lambda_i$ is a root of p. Note that, by letting some of the $\lambda_i = 0$, with this parametrization we get all nonnegative polynomials of degree less than or equal to $2m$ which do not have 0 as a root. In particular the normal distribution is obtained as a polynomial-normal density when $\lambda_1 = \ldots = \lambda_m = 0$. Alternative parameterizations can be chosen which do not exclude the polynomials having 0 as a root; e.g. use the roots rather than the reciprocals of the roots as we have done. Our choice of parameterization is motivated by the fitting algorithm developed for this family in section 4. For note that if $\lambda_i = 0$ then the i-th factor is constant and this is an important feature for that algorithm. Little is lost by excluding the polynomials with 0 as a root as the corresponding polynomial-normal density can be approximated by a polynomial-normal density for which $p(0) \neq 0$. To see this suppose f is any density with first and second moments μ_1 and μ_2. Then for $\lambda \in \mathcal{C}$, $g_\lambda(x) = (\mid \lambda \mid^2 \mu_2 - (\lambda + \bar{\lambda})\mu_1 + 1)^{-1} \ (\lambda x - 1)(\bar{\lambda} x - 1) f(x)$ is the density obtained by adding a quadratic factor corresponding to the root $1/\lambda$. As $\mid \lambda \mid \to \infty$ we have that $g_\lambda(x) \to g(x) = \mu_2^{-1} x^2 f(x)$ which is the density obtained by adding the quadratic factor corresponding to the root 0.

We denote by $\mathcal{F}_m$ the class of polynomial-normal densities of the form (1) where p is a nonnegative polynomial having parametrization (6), and τ_1 and τ_2 are defined by $\tau_1 = -\tau_{1*}/\tau_{2*}$ and $\tau_2 = 1/\tau_{2*}$ where

$$\tau_{1*} = \int_{-\infty}^{\infty} x k p(x) \phi(x) dx$$

and

$$\tau_{2*}^2 + \tau_{1*}^2 = \int_{-\infty}^{\infty} x^2 k p(x) \phi(x) dx.$$

Therefore the class of standardized polynomial-normal densities $\mathcal{F}_m$ consists of polynomial-normal densities of the form (1) having zero mean, unit variance, $\deg(p) \leq 2m$ and $p(0) \neq 0$. The correspondence between $f \in \mathcal{F}_m$ and $\lambda_1, \ldots, \lambda_m$ is not one-to-one as permutations of $\lambda_1, \ldots, \lambda_m$ do not alter the density. However, the correspondence is one-to-one between $f \in \mathcal{F}_m$ and sets $\{\lambda_1, \ldots, \lambda_k\}$ where $k \leq m$ and $\lambda_i \in \mathcal{C}$.

Since the family $\mathcal{F}_m$ is completely determined by $\lambda_1, \ldots, \lambda_m$, there are $2m+2$ free parameters in the location-scale model

$$Y = \mu + \sigma Z \tag{7}$$

where $Z \sim f \in \mathcal{F}_m$. Similarly, in a regression problem with k predictors $X_1, \ldots, X_k$ there are $2m+k+1$ free parameters in the model

$$Y = \beta_1 X_1 + \ldots + \beta_k X_k + \sigma Z \tag{8}$$

where $Z \sim f \in \mathcal{F}_m$.

To illustrate the previous theory we consider some examples.

Example 1. Suppose that $X_1, \ldots, X_n$ is a sample from the distribution with polynomial-normal density

$$f(x) = kp(x)\phi(x) = (1+|\lambda|^2)^{-1}(\lambda x - 1)(\overline{\lambda} x - 1)\phi(x)$$

for some $\lambda \in \mathcal{C}$. It is easy to verify that the mean and variance of X_i are given by $\mu = -(\lambda + \overline{\lambda})/(1+|\lambda|^2)$ and $\sigma^2 = (3|\lambda|^2+1)/(1+|\lambda|^2) - \mu^2$ respectively.

Using Proposition 2 we have that the density of $Y = X_1 + \ldots + X_n$ is $p_n\left(\frac{y}{\sqrt{n}}\right)\frac{1}{\sqrt{n}}\phi\left(\frac{y}{\sqrt{n}}\right)$ where

$$\begin{aligned} p_n(x) &= \int_{-\infty}^{\infty} q^n\left(\frac{it}{\sqrt{n}} + \frac{x}{\sqrt{n}}\right)\phi(t)dt \\ &= \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} kp\left(\frac{it}{\sqrt{n}} + \frac{x}{\sqrt{n}} + z\right)\phi(z)dz\right]^n \phi(t)dt. \end{aligned}$$

The polynomial p_n can be evaluated exactly using Gauss-Hermite rules for both the inner and outer integrals. However, the best way to do this, when repeated calculations of $p_n(x)$ are required, is to first evaluate the Hermite expansion of q and then of p_n, as this cuts down on the amount of computation.

To consider a specific case we choose the complex number $\lambda = (1,1)$ and plot the density of the standardized variate $Z = \sqrt{n}(\overline{X} - \mu)/\sigma$ for $n = 1$ (Figure 1) and for $n = 5$ (Figure 2). The densities are plotted alongside the $N(0,1)$ density for comparison purposes. From these plots we see that although the underlying distribution is strikingly different from the normal, the Central Limit Theorem is providing a good approximation by the time $n = 5$.

Example 2. In this example we consider the underlying density $f(x) = k\prod_{i=1}^{4}(\lambda_i x - 1)(\overline{\lambda}_i x - 1)\phi(x)$ where $\lambda_1 = (0,1)$, $\lambda_2 = (1/2, 1/2)$, $\lambda_3 = (1,0)$ and

$\lambda_4 = (2, 1/2)$. In Figures 3 and 4 we give the plots of the standardized densities for $n = 1$ and $n = 5$ respectively together with plots of the $N(0,1)$ density. We see that in this example the normal distribution is not providing a good approximation. As n increases the normal will inevitably be adequate by the Central Limit Theorem. The point of this example is to show that the normal approximation cannot always be relied upon with this family. Thus the exact distribution must be computed or some alternative approximation technique used. In the following sections we develop the conditional approach to inference with these models and show how to use Monte Carlo to obtain accurate approximations to the conditional distributions of statistics of interest.

3. INFERENCE

In this section we consider inference for the parameters μ and σ in the location-scale model (7) when we assume that the underlying density $f \in \mathcal{F}_m$ is known. In Section 5 we discuss the selection of the nonnegative integer m and the polynomial-normal density $f \in \mathcal{F}_m$.

Consider then a sample $y_1, \ldots, y_n$ from the location-scale model (7) where we define $s_y^2 = \|\underline{y} - \overline{y}\underline{1}\|^2$, $s^2 = \|\underline{z} - \overline{z}\underline{1}\|^2$ and $\underline{d} = (\underline{y} - \overline{y}\underline{1})/s_y = (\underline{z} - \overline{z}\underline{1})/s$. Since $\underline{z}$ is a sample from the polynomial-normal distribution $f \in \mathcal{F}_m$, any functional of $\underline{z}$ has a distribution which is independent of the location parameter μ and the scale parameter σ. Therefore quantities such as $\overline{z}$, s and $\underline{d}$ all have distributions that do not depend on (μ, σ).

In the case where σ is known, inferences concerning μ can be based on the pivotal $\overline{z} = (\overline{y} - \mu)/\sigma$. For example, the distribution of $\overline{z}$ can be obtained from Proposition 2 and using the formula for the cdf given in (3), constants a_1 and a_2 can be found such that $P(a_1 \leq \overline{z} \leq a_2) = .95$ It follows that a 95% confidence interval for μ is given by $(\overline{y} - a_2\sigma, \overline{y} - a_1\sigma)$.

In the case where σ is unknown, marginal inference concerning the parameters μ and σ can be based on the pivotal quantities $t = \overline{z}/s = (\overline{y} - \mu)/s_y$ and $s = s_y/\sigma$ respectively. In general, the distributions of the random variables s and t are computationally difficult. However, for the class of standardized polynomial-normal densities (ie. $z \sim f \in \mathcal{F}_m$), the distributions are manageable, at least for small to moderate n, and are given in Corollaries 2 and 3 respectively. This approach is referred to as the unconditional approach.

It has been argued (see Fisher(1934) or Fraser(1979)) that the conditional approach is the preferred approach for the analysis of the location-scale model (7) and more generally for the regression model (8). In the conditional analysis we proceed as above except that we base our inferences on the pivotals s and t conditioned on the observed value of the statistic $\underline{d}$. This simplifies

calculations and is partly justified by the fact that $\underline{d}$ is an ancillary statistic. The conditional distributions of s and t are given in Propositions 3 and 4 respectively. We note that in the case of a normal error the conditional and unconditional approaches are exactly the same since $\underline{d}$ is independent of both s and t. Thus the polynomial-normal family offers an opportunity to compare the conditional and unconditional analyses for a non-normal family.

First we consider the distribution of s where we denote the density of a Chi(n) variable by $g_n(s) = s^{n-1}e^{-s^2/2}/[2^{(n-2)/2}\Gamma(\frac{n}{2})]$.

Proposition 3. Consider the location-scale model (7) based on a sample of size n. The conditional density of s given $\underline{d}$ is given by

$$g_{s|\underline{d}}(s|\underline{d}) = r(\underline{d})p_1(s,\underline{d})g_{n-1}\left(\frac{s}{\tau_2}\right)\frac{1}{\tau_2} \qquad s>0$$

where $r^{-1}(\underline{d}) = \int_0^\infty p_1(\tau_2 s,\underline{d})g_{n-1}(s)ds$ is the inverse of the norming constant and

$$p_1(s,\underline{d}) = \int_{-\infty}^{\infty}\prod_{i=1}^{n} kp(u/\sqrt{n} + sd_i/\tau_2)\phi(u)du$$

is a positive polynomial in s.

Proof : Since $z_1,\ldots,z_n$ is a sample from the polynomial-normal distribution, the joint density of $\underline{z}$ is $\prod_{i=1}^n kp\left(\frac{z_i-\tau_1}{\tau_2}\right)\frac{1}{\tau_2}\phi\left(\frac{z_i-\tau_1}{\tau_2}\right)$. Letting $P = (\underline{p}_1,\ldots,\underline{p}_{n-1})$ be an orthonormal basis of $\mathcal{L}^-(\underline{1}) \subseteq \Re^n$, we write $\underline{v} = P'\underline{d}$ and note that $\|\underline{v}\|^2 = \|\underline{d}\|^2 = 1$. Following Theorem 2.1.3 of Muirhead(1982) we can express $\underline{v} \in \Re^{n-1}$ in polar coordinates in terms of $\theta_1,\ldots,\theta_{n-2}$, say as $\underline{v} = \underline{u}(\underline{\theta})$. Making the change of variables $\underline{z} \to (\overline{z},s,\underline{\theta})$, the Jacobian of the transformation is $n^{1/2}s^{n-2}j(\underline{\theta})$ where $j(\underline{\theta}) = \sin^{n-3}\theta_1\sin^{n-4}\theta_2\cdot\cdots\sin\theta_{n-3}$ is derived in Muirhead(1982). Therefore the joint density of $(\overline{z},s,\underline{\theta})$ is

$$\prod_{i=1}^{n} kp\left(\frac{\overline{z}+sd_i(\underline{\theta})-\tau_1}{\tau_2}\right)\phi\left(\frac{\overline{z}-\tau_1}{\tau_2/\sqrt{n}}\right)\frac{\sqrt{n}}{\tau_2}g_{n-1}\left(\frac{s}{\tau_2}\right)\frac{1}{\tau_2}A_{n-1}^{-1}j(\underline{\theta}) \qquad (9)$$

where $A_{n-1} = 2\pi^{(n-1)/2}/\Gamma\left(\frac{n-1}{2}\right)$ is the surface area of the unit sphere in $\Re^{n-1}$ and $\underline{d}(\underline{\theta}) = P\underline{u}(\underline{\theta})$. The result then follows immediately from (9) as $\underline{d}$ is a 1-1 function of $\underline{\theta}$.

Note that the quantities $r(\underline{d})$ and $p_1(s,\underline{d})$ can be evaluated exactly using the moment formulae for the Chi and normal distributions respectively. Alternatively, these quantities can be calculated exactly using Gauss rules for the

Chi(n-1) distributon and Hermite rules respectively. From the conditional density for s we get the form of the marginal density for s.

Corollary 2. Consider the location-scale model (7) based on a sample of size n. The (unconditional) density of s is given by

$$g_s(s) = p_1(s) g_{n-1}\left(\frac{s}{\tau_2}\right)\frac{1}{\tau_2} \qquad s > 0$$

where $p_1(s) = \int_0^{2\pi}\int_0^{\pi}\cdots\int_0^{\pi} p_1(s, \underline{d}(\underline{\theta}))A_{n-1}^{-1} j(\underline{\theta}) d\underline{\theta}$ is a positive polynomial in s.

Proof : The result follows immediately from (9).

Note that the evaluation of $p_1(s)$ involves the integration of powers of sines and cosines. This can be carried out exactly using spherical quadrature rules. We see that the conditional and unconditional densities of s take the form of a positive polynomial times a Chi(n-1) density, a natural generalization of normal theory. Further these polynomials are related in a simple way and demonstrate the difference between a conditional and unconditional analysis for σ.

Now we consider the distribution of t.

Proposition 4. Consider the location-scale model (7) based on a sample of size n. The conditional density of $t = \overline{z}/s$ given $\underline{d}$ is given by

$$g_{t|\underline{d}}(t|\underline{d}) = r(\underline{d}) \exp\{-n\tau_1^2/2\tau_2^2\} p_2(t, \underline{d}) h_{n-1}(\sqrt{n(n-1)}t)\sqrt{n(n-1)}$$

where

$$p_2(t, \underline{d}) = \int_0^{\infty} exp\left(\frac{n\tau_1 tv}{\tau_2\sqrt{1+nt^2}}\right)\prod_{i=1}^{n} kp\left(\frac{v(t+d_i)}{\sqrt{1+nt^2}} - \frac{\tau_1}{\tau_2}\right) g_n(v) dv,$$

$h_{n-1}(t)$ is the density of the Student$(n-1)$ distribution and $r(\underline{d})$ is given in Proposition 3.

Proof : From expression (9) we obtain the conditional density of $(\overline{z}, s)$ given $\underline{d}$ by dividing by $r^{-1}(\underline{d}(\underline{\theta}))A_{n-1}^{-1} j(\underline{\theta})$. Then make the transformation $(\overline{z}, s) \rightarrow (t, v)$ where $t = \overline{z}/s$ and $v = s\sqrt{1+nt^2}/\tau_2$. This transformation has Jacobian $v\tau_2^2/(1+nt^2)$. Integrating out the variable v gives $p_2(t, \underline{d})$ from which follows the conditional density of t given $\underline{d}$.

Note that the troublesome term in the evaluation of the conditional density of t given $\underline{d}$ is $p_2(t, \underline{d})$. This can be approximated using the Gauss quadrature rule based on $g_n(v)$. Alternatively an exact calculation can be obtained by observing that the essential integration in the evaluation of $p_2(t, \underline{d})$ takes the form $\int_0^\infty v^k e^{av} e^{-v^2/2} dv = e^{a^2/2} \int_{-a}^\infty (u+a)^k e^{-u^2/2} du = e^{a^2/2} \sum_{i=0}^k \begin{pmatrix} k \\ i \end{pmatrix} a^{k-i} \times \int_{-a}^\infty u^i e^{-u^2/2} du$ for some positive integer k and some constant a. A single application of integration by parts gives $\int_{-a}^\infty u^i e^{-u^2/2} du = (-a)^{i-1} e^{-a^2/2} + (i-1) \int_{-a}^\infty u^{i-2} e^{-u^2/2} du$. Repeated applications then lead to either $\int_{-a}^\infty u e^{-u^2/2} du = e^{-a^2/2}$ or $\int_{-a}^\infty e^{-u^2/2} du = \sqrt{2\pi}(1 - \Phi(-a))$ depending on whether i is odd or even. This establishes that $p_2(t, \underline{d})$ is a bounded function of t. We shall see in section 6 that this fact proves useful in our development of a more general computational approach. The form of the marginal density of t now follows easily.

Corollary 3. Consider the location-scale model (7) based on a sample of size n. The unconditional density of $t = \overline{z}/s$ is given by

$$g_t(t) = \exp\{-n\tau_1^2/2\tau_2^2\} p_2(t) h_{n-1}(\sqrt{n(n-1)}t)\sqrt{n(n-1)}$$

where $p_2(t) = \int_0^{2\pi} \int_0^\pi \cdots \int_0^\pi p_2(t, \underline{d}(\underline{\theta})) A_{n-1}^{-1} j(\underline{\theta}) d\underline{\theta}$.

Proof : Take the product of $g_{t|\underline{\theta}} = g_{t|\underline{d}}$ from Proposition 4 and the marginal density of θ which is $r^{-1}(\underline{d}(\underline{\theta})) A_{n-1}^{-1} j(\underline{\theta})$ and can be derived from (9). This gives the joint density of $(t, \underline{\theta})$. The result follows after integrating out the variable $\underline{\theta}$.

Hence the conditional and unconditional densities of t generalize the the normal theory result for t.

Thus closed form expressions have been obtained for both the conditional and unconditional analyses of the location-scale model (7) when the error distribution is given by a polynomial-normal density. Although the formulae are at times complex, they are well within the capabilities of modern computing, particularily for the conditional analyses. In fact, for small n exact expressions for the various densities may be derived using symbolic languages such as Mathematica and Maple. Practically speaking, however, the formulae are more interesting for the way they show how inference changes when we move away from normal errors and the contrast between the conditional and unconditional analyses.

Similar formulas can also be obtained for the regression model (8) based on polynomial-normal errors. However the densities have even more complicated expressions. As such, the general approach to the distribution theory, as discussed in this section, is not practical. Consequently, an alternative approach is developed in Section 6 and this is also the preferred approach in the location-scale context when actually doing calculations.

4. FITTING THE FAMILY

In the previous section we developed the inferential theory for the parameters μ and σ in the location-scale model (7) based on polynomial-normal errors. In doing so we assumed that the underlying polynomial-normal error density $f \in \mathcal{F}_m$ is known. In this section we discuss the selection of the nonnegative integer m and the polynomial-normal density $f \in \mathcal{F}_m$ for the location-scale model. The discussion generalizes easily to the regression model (8).

One approach is to select m rather large, as $\mathcal{F}_m$ contains all polynomial-normal densities of degree $2m$ or less, and then select $f \in \mathcal{F}_m$ using only $\underline{d}$. One possibility is to maximize the marginal likelihood $r^{-1}(\underline{d})$ as a function of $f \in \mathcal{F}_m$, see for example Fraser(1979). In this context this is unwieldy as we are forced to carry out a maximization of a complicated function of $2m$ variables. Note that this difficulty also prevents us from using more elaborate model selection criteria such as AIC.

Instead we adopt an approach that is computationally feasible in this context and that works well in the examples we have tried. For this we view $\sqrt{n-1}\underline{d}$ as an estimate of the error vector $\underline{z}$. This is supported by the following result.

Lemma 1. If $z_1, \ldots, z_n$ is a sample from a continuous distribution P having mean 0 and variance 1, then for any interval A

$$\frac{1}{n}\sum_{i=1}^{n} I_A(\sqrt{n-1}d_i) \longrightarrow P(A)$$

almost surely as $n \to \infty$.

Proof : See the appendix.

In the case of the regression model (8) a result similar to Lemma 1 can be obtained. For the regression model we treat $\sqrt{n-k}\underline{d}$ as a sample from $f \in \mathcal{F}_m$ where $\underline{d} = (\underline{y} - X\underline{\hat{\beta}})/s_y$, $s_y = \|\underline{y} - X\underline{\hat{\beta}}\|$, $\underline{\hat{\beta}} = (X'X)^{-1}X'\underline{y}$ and $X = (\underline{x}_1, \ldots, \underline{x}_k)$.

Now viewing $\sqrt{n-1}\underline{d}$ as a sample from some $f \in \mathcal{F}_m$, a first thought in fitting a member of the family $\mathcal{F}_m$ to the data, might be to choose $\hat{f} \in \mathcal{F}_m$ whose distribution function $\hat{F}$ is "closest" to the empirical distribution of $\sqrt{n-1}\underline{d}$. This could be done by first specifying l points $x_1, \ldots, x_l$ and then minimizing

$$\sum_{j=1}^{l} (\hat{F}(x_j) - p_j)^2 \tag{10}$$

where $p_j = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty, x_j]}(\sqrt{n-1}d_i)$. There does not appear to be an optimal way to select l and the x_i. We note the reasonable conjecture, however, that if $l \geq 2m$ then $f \in \mathcal{F}_m$ is completely determined by the values of $F(x_i)$ independent of specific values of the x_i. This conjecture is discussed further in section 5 where it is proven to be true for an important subclass of $\mathcal{F}_m$. Irrespective of this conjecture it makes sense to choose l relatively high to avoid over-fitting at just a few points and to choose the x_i to span the region where probability will be concentrated for a distribution with mean 0 and variance 1. Note that Chebyshev's inequality implies that 8/9 of the probability lies in (-3,3) for such a distribution.

For all but the smallest values of the parameter m, the $2m$-dimensional minimization of (10) is again computationally difficult. Therefore we consider a compromise; a fitting algorithm which tackles the minimization iteratively in manageable steps and also provides a method of determining m. Letting $\hat{f}_0 \in \mathcal{F}_0$ denote the standard normal density, in step i of the algorithm, we require that $\hat{f}_i \in \mathcal{F}_i$ take the form

$$\hat{f}_i(x) = k_i \left[\lambda_i \left(\frac{x - \tau_{1i}}{\tau_{2i}}\right) - 1\right] \left[\overline{\lambda}_i \left(\frac{x - \tau_{1i}}{\tau_{2i}}\right) - 1\right] \hat{f}_{i-1} \left(\frac{x - \tau_{1i}}{\tau_{2i}}\right)$$

and minimize (10) over all such choices.

Hence we are fitting by quadratic factors and the determination of $\hat{f}_i$ from $\hat{f}_{i-1}$ involves a 2-dimensional minimization only. This can be carried out by a simple search algorithm which we describe below. Recall from Section 1 that the parameters k_i, τ_{1i} and τ_{2i} are prescribed given the quadratic polynomial $(\lambda_i x - 1)(\overline{\lambda}_i x - 1)$. The algorithm is terminated and m is specified when a subsequent step of the algorithm provides a quadratic factor with λ_i sufficiently close to $(0, 0)$. As the search is over a finite grid which includes (0,0) the algorithm typically terminates with this choice.

To implement the minimization step of the algorithm, we consider an alternative parametrization of the polynomial. This allows the search for the minimum to be carried out over a 2-dimensional compact set. We put $h_{0i}(x) = 1$,

$h_{1i}(x) = x$ and $h_{2i}(x) = D_i^{-1}(x^2 - \mu_{3i}x - 1)$ where $\mu_{3i} = \int_{-\infty}^{\infty} x^3 \hat{f}_{i-1}(x)dx$ and $D_i^2 = \int_{-\infty}^{\infty}(x^2 - \mu_{3i}x - 1)^2 \hat{f}_{i-1}(x)dx$. These are the first 3 orthonormal polynomials with respect to $\hat{f}_{i-1}(x)$. We can then write a quadratic polynomial q_i uniquely as $q_i(x) = a_i h_{0i}(x) + b_i h_{1i}(x) + c_i h_{2i}(x)$. Now given $\hat{f}_{i-1}(x)$, Lemma 2 shows that the only possible values for the coefficients (a_i, b_i, c_i) lie in a 2-dimensional compact set.

Lemma 2. The function $g(x) = q_i(x)\hat{f}_{i-1}(x)$ is a density function if and only if $a_i = 1$, $0 \leq c_i \leq D_i$ and $(b_i D_i - c_i \mu_{3i})^2 \leq 4c_i(D_i - c_i)$.

Proof : See the appendix.

Example 3. Consider the density $f \in \mathcal{F}_2$ specified by $\lambda_1 = (1/2, \sqrt{3}/2)$ and $\lambda_2 = (1, \sqrt{3})$. Using $l = 11$ points given by $x_i \in \{0, \pm.25, \pm.5, \pm1, \pm2, \pm3\}$ we use the algorithm of fitting by quadratic factors to determine $\hat{f}_1 \in \mathcal{F}_1$ and $\hat{f}_2 \in \mathcal{F}_2$. Here we use the exact values of $F(x_i)$ instead of estimates and so there is no sampling error involved. The search algorithm is based on 100 equispaced points in each of the c_i and b_i coordinates. Figures 5 and 6 give plots of the fitted densities $\hat{f}_1$ and $\hat{f}_2$ respectively together with the underlying density f. Fitting another factor produced no further improvement. The results indicate that $\hat{f}_2$ provides an excellent approximation to f. Of course, a finer grid for the search would produce a better approximation.

Example 4. In this example we fit a polynomial-normal density from the location-scale model to the logarithms of 100 stress-rupture lifetimes of Kevlar pressure vessels. The dataset can be found in Andrews and Herzberg(1985), page 183. Using the fitting algorithm as described in example 3 we obtained $m = 1$ and $\lambda_1 = (-1.565, 0)$. The fitted density together with the standard normal density is plotted in Figure 7. Fitting another factor produced no further improvement. It is clear that this is a highly non-normal dataset. This is confirmed in a histogram plot of $\sqrt{n-1}\underline{d}$ which indicates that the secondary mode is not a spurious aspect of the fitting algorithm.

The examples above and other examples, see example 5 for a context where its merits can be assessed when sampling error is present, support the use of the proposed fitting algorithm. In addition to the empirical evidence, the algorithm is convenient and intuitive. It begins by finding the "best" quadratic-normal density according to criterion (10), and continues with each subsequent factor improving on the fit in the sense of (10).

As previously mentioned, provided that the number l of points x_i chosen is greater than $2m$, it is reasonable to conjecture that $f \in \mathcal{F}_m$ is completely determined by the values $F(x_i)$. If this conjecture holds, and we were to fit $f \in \mathcal{F}_m$ by minimizing (10) over all such f, then Lemma 1 would establish the consistency of this estimate. The examples indicate that proceeding by quadratic factors does not produce serious problems for the accuracy of our inferences. Further research is needed, however, to assess the effect of sampling errors on this algorithm and also to see if there are better approaches. At this point the evidence indicates that the algorithm is working well and we note the lack of any practical alternatives.

5. COMPARISON WITH GRAM-CHARLIER APPROXIMATION

Following Kendall and Stuart(1958), page 148, with appropriate changes of notation, the Gram-Charlier Type A series for the density $g(x)$, where $g(x)$ has all its moments and in particular mean τ_1 and variance τ_2^2, is given by

$$\sum_{i=0}^{\infty} b_i h_i \left(\frac{x-\tau_1}{\tau_2}\right) \phi\left(\frac{x-\tau_1}{\tau_2}\right) \frac{1}{\tau_2}$$

where $b_i = \int_{-\infty}^{\infty} h_i\left(\frac{x-\tau_1}{\tau_2}\right) g(x)dx$ and h_i is the i^{th} orthonormal Hermite polynomial. This is a formal series and is not necessarily convergent. Using the fact that $h_0(x) = 1$, $h_1(x) = x$ and $h_2(x) = (x^2-1)/\sqrt{2}$, we have that $b_0 = 1$, $b_1 = 0$ and $b_2 = 0$.

The corresponding Gram-Charlier approximation of degree $2m$ is then given by

$$g_m(x) = p\left(\frac{x-\tau_1}{\tau_2}\right) \phi\left(\frac{x-\tau_1}{\tau_2}\right) \frac{1}{\tau_2} \tag{11}$$

where $p(x) = \sum_{i=0}^{2m} b_i h_i(x)$. Since odd degree polynomials cannot be nonnegative, we consider only the even degree case. It is easy to verify that the mean and variance of g_m are also τ_1 and τ_2^2.

The traditional approach to fitting (11) to data is based on the method of moments as τ_1, τ_2 and the b_i are all simple functions of the moments of g. Hence if $x_1, \ldots, x_n$ is a sample we put $\tau_1 = \bar{x}, \tau_2 = s_x/\sqrt{n-1}$ and $b_i = \frac{1}{n}\sum_{i=1}^{n} h_i(\sqrt{n-1}d_i)$.

There are several differences between a Gram-Charlier approximation (11) and a polynomial-normal density function (1). We point these out below and also discuss some of their relationships.

First, τ_1 and τ_2 in (11) are constrained to be the mean and standard deviation of the distribution. The polynomial-normal densities do not have

this restriction. Hence there are two additional degrees of freedom with a polynomial-normal density of the same degree. This is reflected in the fact that the weight function $\phi(\frac{x-\tau_1}{\tau_2})\frac{1}{\tau_2}$ can take on a wider range of possibilities.

Second, there is no guarantee that (11) is a legitimate probability density. The polynomial may take on negative values. We could restrict, as is done in Geweke(1989), our attention to Gram-Charlier approximations $g_m(x)$ where the polynomial is nonnegative. We refer to these as Gram-Charlier densities and let $\mathcal{G}_m$ denote the class of Gram-Charlier densities of the form $p(x)\phi(x)$ where $p(x)$ is a nonnegative polynomial of degree $\leq 2m$. Thus the class $\mathcal{G}_m$ consists of all Gram-Charlier densities having zero mean and unit variance. Since $b_0 = 1$, $b_1 = 0$ and $b_2 = 0$, $\mathcal{G}_m$ is a $(2m-2)$-dimensional subfamily of $\mathcal{F}_m$. A practical problem with $\mathcal{G}_m$ is its lack of a convenient parametrization. In other words, it is difficult to find appropriate constraints on the b_i, $i > 2$ to ensure nonnegativity.

We now show that the method given in Section 4 for fitting is consistent when we have a location-scale model with $Z \sim g \in \mathcal{G}_m$. The distribution function for such a g takes the form, following (3), $G(x) = \Phi(x) + r(x)\phi(x)$ where $r(x)$ is a polynomial of degree less than $2m$. Hence if we specify G at $l \geq 2m$ points x_i then this specifies the value of $r(x) = (G(x) - \Phi(x))/\phi(x)$ at l points. Since r is a polynomial of degree less than $2m$ these values completely prescribe r and hence g. This together with Lemma 1 implies that if we fit $g \in \mathcal{G}_m$ by minimizing (10) over such g then this gives a consistent estimate as well. This result supports our approach to fitting elements of $\mathcal{F}_m$ and our conjecture that $f \in \mathcal{F}_m$ is prescribed by the values of its distribution function at $2m$ points.

The traditional approach to fitting Gram-Charlier densities cannot be applied to fit a general element of $\mathcal{F}_m$ since τ_1 and τ_2 in (1) are complicated functions of the moments of f which we have not been able to obtain in closed form. Hence we developed the alternative approach of section 4. Also, the Gram-Charlier series for $f \in \mathcal{F}_m \setminus \mathcal{G}_m$ need not converge; for a discussion of the convergence of such series see Kendall and Stuart(1958).

Geweke(1989) has used the method of maximum likelihood to fit Gram-Charlier densities. We have also experimented with maximum likelihood in the fitting of polynomial-normal densities. Like Geweke(1989), we have experienced the difficulties associated with multimodal likelihoods. For this reason, and some of the reasons stated above, we prefer our algorithm of fitting by quadratic factors.

Previous authors have also developed distribution theory results when the basic density is a polynomial times a normal density. Bartlett(1935) derives an

asymptotic expression for the unconditional density of the t-statistic when the polynomial is of degree 4, not necessarily nonnegative and with retrictions on the coefficients. Under similar assumptions Gayen(1949, 1950a, 1950b, 1951) derives asymptotic expressions for the unconditional densiities of many statistics of inferential interest. Under the assumption that the basic density has a Gram-Charlier expansion, Davis(1976) derives asymptotic series expressions for the unconditional densities of many statistics. In section 6 we develop an algorithm which permits the computation of conditional probabilities to practically meaningful accuracies for any statistic when the model is a linear model with polynomial-normal error.

6. COMPUTATIONS FOR REGRESSION

Whereas exact expressions can often be obtained for the distributions of quantities of inferential interest, these expressions do not necessarily provide the best approach to implementing the computations. This is particularly true in the regression context, where instead, we consider a version of adaptive importance sampling. Adaptive importance sampling ideas have been used by various authors including Smith et. al.(1987), Evans, Gilula and Guttman(1989) and Oh and Berger(1992).

In the regression model (8) we define $s_y = \|\underline{y} - X\hat{\underline{\beta}}\|$, $s = \|\underline{z} - X\underline{b}\|$ and $\underline{d} = (\underline{y} - X\hat{\underline{\beta}})/s_y = (\underline{z} - X\underline{b})/s$ where $\hat{\underline{\beta}} = (X'X)^{-1}X'\underline{y}$, $\underline{b} = (X'X)^{-1}X'\underline{z}$ and $X = (\underline{x}_1, \ldots, \underline{x}_k)$. Inference concerning σ is based on the conditional distribution of $s = s_y/\sigma$ given $\underline{d}$ and inference concerning $\underline{\beta}$ is based on the conditional distribution of $\underline{t} = \underline{b}/s = (\hat{\underline{\beta}} - \underline{\beta})/s_y$ given $\underline{d}$.

To implement a conditional analysis we need to be able to calculate conditional expectations of various functions of $(\underline{b}, s)$; e.g. to tabulate the conditional distribution function of t_i which then leads to conditional confidence intervals for β_i. While we don't need the explicit form of the joint density of $(\underline{b}, s)$ to describe our algorithm, an unnormalized version is given in (15). It turns out to be convenient for the necessary integrations to make the transformations $(\underline{b}, s) \to (\underline{t}, s)$ and $(\underline{t}, s) \to \underline{a}$ where

$$\underline{a} = \begin{pmatrix} \sqrt{n-k}R\underline{t} \\ \sqrt{2(n-k)}log(s/\sqrt{n-k}) \end{pmatrix}$$

$X = QR$, R is an upper triangular matrix with positive diagonal elements, and Q is a column-orthonormal matrix. With normal errors, the vector $\underline{a}$ is asymptotically normal with mean vector $\underline{0}$ and variance-covariance matrix I.

Denote the unnormalized conditional density of $\underline{a}$ given $\underline{d}$, obtained from

(15) and the transformations, by $L(\underline{a}|\underline{d})$. Thus the general problem is to calculate

$$E[u|\underline{d}] = \int_{-\infty}^{\infty} u(\underline{a})L(\underline{a}|\underline{d})d\underline{a} \ / \ \int_{-\infty}^{\infty} L(\underline{a}|\underline{d})d\underline{a} \tag{12}$$

for various functions u. Note that the denominator is the norming constant.

If we generate values $\underline{a}_1, \ldots, \underline{a}_N$ from a density w on $\Re^{k+1}$, we get an importance sampling estimator of (12) as given by

$$\hat{E}[u|\underline{d}] = \sum_{i=1}^{N} u(\underline{a}_i)\frac{L(\underline{a}_i|\underline{d})}{w(\underline{a}_i)} \ / \ \sum_{i=1}^{N} \frac{L(\underline{a}_i|\underline{d})}{w(\underline{a}_i)}. \tag{13}$$

The Strong Law of Large Numbers gives almost sure convergence of $\hat{E}[u|\underline{d}]$ to (12). In addition, provided that the numerator and denominator in (13) both have finite variance, then using Theorem A, page 122 of Serfling(1980), (13) is asymptotically normal. The following proposition gives conditions under which the numerator and denominator in (13) have finite variance for certain functions u when w is a multivariate Student density.

Proposition 5. Suppose that u is bounded in absolute value by a polynomial in $\underline{a}$ of degree r. Further let w be the density of $\underline{\mu} + \Sigma_{\Delta}\underline{\epsilon}$, where $\underline{\mu} \in \Re^{k+1}$, $\Sigma_{\Delta} \in \Re^{(k+1)\times(k+1)}$ is a lower triangular matrix with positive diagonal elements and the random vector $\underline{\epsilon}$ is distributed as a $(k+1)$-dimensional Student(λ) with $\lambda \leq n - 2k - 4 - 2r$. Then (13) has finite variance.

Proof : The variance of the numerator of (13) is finite if and only if

$$\int_{-\infty}^{\infty} u^2(\underline{a})\frac{L^2(\underline{a}|\underline{d})}{w(\underline{a})}d\underline{a} < \infty. \tag{14}$$

The result follows when we have established the finiteness of (14) for $u(\underline{a}) = a_i^r$ for $i = 1, \ldots, k+1$.

The proof proceeds as follows. We first show that all of the conditional moments of s exist and determine what conditional moments of $\underline{t}$ will exist when sampling from a polynomial-normal density. The next step is to note that $L^2(\underline{a}|\underline{d})$, when considered as a function of $(\underline{b}, s)$, takes the form of an unnormalized conditional density of $(\underline{b}, s)$, when sampling from a polynomial-normal density, times a power of s. Then using $w(\underline{a}) \propto (1 + (\underline{a} - \underline{\mu})'\Sigma^{-1}(\underline{a} - \underline{\mu}))^{-\frac{\lambda+k+1}{2}}$, and combining this with the above steps, the result will be obtained.

Generalizing the proof of Proposition 3 it can be shown that the conditional density of $(\underline{b}, s)$ given $\underline{d}$ is proportional to

$$\prod_{i=1}^{n} p\left(\frac{1}{\tau_2}\sum_{j=1}^{k}(x_{ij}b_j + sd_i) - \frac{\tau_1}{\tau_2}\right) \exp\{-(s^2 + \underline{b}'X'X\underline{b})/2\tau_2^2\} \times \exp\{\tau_1(\underline{b}'X'\underline{1} + s\underline{d}'\underline{1})/\tau_2^2\} s^{n-k-1}. \quad (15)$$

From (15) it is immediate that all of the conditional moments of s exist as $e^{-s^2/2\tau_2^2}$ times e^{cs}, for any constant c, times a polynomial has a finite integral. In (15) make the transormation $(\underline{b}, s) \to (\underline{t}, s)$, which has Jacobian s^k, and then integrate out s to get that the conditional density of $\underline{t}$ is proportional to $(1 + \underline{t}'X'X\underline{t})^{-n/2}$ times

$$\int_0^\infty \prod_{i=1}^{n} p\left(\frac{u}{\tau_2}[\frac{\sum_{j=1}^{k} x_{ij}t_j + d_i}{\sqrt{1 + \underline{t}'X'X\underline{t}}}] - \frac{\tau_1}{\tau_2}\right) \exp\{\frac{\tau_1}{\tau_2}u\frac{(\underline{t}'X'\underline{1} + \underline{d}'\underline{1})}{\sqrt{1 + \underline{t}'X'X\underline{t}}}\} g_n(u)du. \quad (16)$$

Now observe that (16) is of the same form as $p_2(\cdot, \underline{d})$ in Proposition 3 and the argument after the proof of that result shows that (16) is bounded as a function of $\underline{t}$. Hence the factor $(1 + \underline{t}'X'X\underline{t})^{-n/2}$ determines that all the moments of $\underline{t}$ exist of order less than or equal to $n - k - 3$. This completes the first step of the proof.

Now make the transformation $\underline{a} \to (\underline{b}, s)$ in (14). The Jacobian of the transformation is $1/s^{k+1}$. It follows that $L^2(\underline{a}|\underline{d})/s^{k+1}$, as a function of $(\underline{b}, s)$, is proportional to (15) squared times s^{k+1}. Therefore, looking at the form of (15), we have that $L^2(\underline{a}|\underline{d})/s^{k+1}$, as a function of $(\underline{b}, s)$, is proportional to s^n times the conditional density of $(\underline{b}, s)$ given $\underline{d}$ when $\underline{z}' = (z_1, \ldots, z_n)$ is a sample from $p_* \left(\frac{z-\tau_1}{\tau_2/\sqrt{2}}\right) \phi \left(\frac{z-\tau_1}{\tau_2/\sqrt{2}}\right) \frac{1}{\tau_2/\sqrt{2}}$ with $p_*(x) = p^2(x/\sqrt{2})$. This completes the second step.

Now as a function of a_i, $w^{-1}(\underline{a})$ behaves like $|a_i|^{\lambda+k+1}$ for large $|a_i|$. If $u(\underline{a}) = a_i^r$ for $i \leq k$, then $u^2(\underline{a}) = (n-k)^{2r}(\underline{e}_i'R\underline{t})^{2r}$ where $\underline{e}_i$ is the i-th standard basis vector. Combining this with steps one and two we get that (14) holds for such a function u whenever $2r + \lambda + k + 1 \leq n - k - 3$ or $\lambda \leq n - 2k - 4 - 2r$. If $u(\underline{a}) = a_{k+1}^r = (2(n-k))^{r/2}(logs/\sqrt{n-k})^r$, then $u^2(\underline{a}) = O(s^{2r})$ for large s. Since all of the conditional moments of s exist, it follows that (14) also holds for this coordinate. This completes the proof.

We note that Proposition 5 provides a rare instance where one can actually establish the appropriate degrees of freedom for the multivariate Student to

guarantee the finite variance of the importance sampling estimator. Hence we will use importance samplers w that are affine transformations of multivariate Student distributions with appropriate degrees of freedom.

The question arises as to which affine transformation should be chosen. A general principle of importance sampling is to choose w to mimic $L(\cdot|\underline{d})$. In that way the variance of the estimator $\hat{E}[u|\underline{d}]$ is typically small. To that end, we attempt to match the mean vector and variance matrix of w to the conditional mean and conditional variance matrix of $\underline{a}$. As the conditional mean and variance matrix of $\underline{a}$ are unknown, we proceed adaptively. We begin with $\mu_0 = \underline{0}$ and $\Sigma_0 = I$. Then we generate $\underline{a}_1, \ldots, \underline{a}_N$ from w_0 where w_0 is the density corresponding to $\underline{\mu}_0 + \sqrt{\frac{\lambda-2}{\lambda}}\Sigma_0\underline{\epsilon}$ with $\underline{\epsilon}$ distributed as a Student(λ) distribution and the factor $\sqrt{\frac{\lambda-2}{\lambda}}$ standardizes the Student distribution to have variance matrix equal to the identity. From this sample, we obtain the estimates $\underline{\mu}_1$ and Σ_1 of the conditional mean vector and variance matrix of $\underline{a}$ using (13) and appropriate choices of the u functions. Next we generate $\underline{a}_1, \ldots, \underline{a}_N$ from $w_1(\underline{a})$, the density corresponding to $\underline{\mu}_1 + \sqrt{\frac{\lambda-2}{\lambda}}\Sigma_1\underline{\epsilon}$. We then combine this sample and the previous sample, weighting each sample point by the appropriate multivariate Student density, to produce new estimates $\underline{\mu}_2$ and Σ_2. We continue this for finitely many steps and obtain a final importance sampler $w_*(\underline{a})$. We then use $w_*(\underline{a})$ to estimate expectations of interest via straight importance sampling.

Example 5. We consider a regression model where we observe 50 y-values, 10 from each of $Y = \beta_i + \sigma Z$, $i = 1, \ldots, 5$. Setting $\beta_i = i$ and $\sigma = 1$, we simulate the data by generating 50 independent and identically distributed variates from $f \in \mathcal{F}_3$ specified by $\lambda_1 = (1/2, \sqrt{3}/2)$, $\lambda_2 = (-1/2, \sqrt{3}/2)$ and $\lambda_3 = (1/2, 0)$. Applying the fitting algorithm to the residual vector $\sqrt{n-k}\underline{d}$ we obtain $m = 1$ with $\lambda = (.7686, .5771)$. The error density f together with the fitted density are plotted in Figure 8. While the estimate of m is not very good we notice that the fitted density is doing reasonably well at picking up the main features of the error distribution.

We now use the fitted error distribution and the adaptive importance sampling algorithm to calculate conditional confidence intervals for the β_i. Applying Proposition 5 we require a 6-dimensional Student(34) distribution to ensure that the importance sampling estimators have finite variance when we are only concerned with estimating 1st and 2nd moments and probability contents. To be conservative we use a 6-dimensional Student(10) distribution. With $N = 1000$ we carry out 10 iterations of adaptive importance sampling,

where we begin adapting to the variance matrix after the third iteration. We obtain the estimates

$$\hat{E}[\underline{a}|\underline{d}] = (-.031, -.015, -.162, -.483, 1.065, .083)'$$

and

$$\hat{V}[\underline{a}|\underline{d}] = \begin{pmatrix} .993 & & & & & \\ .025 & .599 & & & & \\ .070 & .026 & .844 & & & \\ .049 & .017 & .016 & .652 & & \\ -.002 & -.083 & -.072 & -.073 & .703 & \\ .376 & .106 & .251 & .154 & -.036 & 1.330 \end{pmatrix}$$

Based on a much larger simulation we conclude that these estimates are accurate to within $\pm.02$. Under normality, $E[a_i|\underline{d}] = 0$, $Var[a_i|\underline{d}] = 1$ and $Cov[a_i, a_j|\underline{d}] = 0$ whenever $1 \leq i, j \leq 5$. Clearly the non-normal analysis is making a big difference. These computations require only 2 seconds of CPU time on an IRIS 40/220 computer.

At the end of the 10 iterations of adaptive importance sampling, a final importance sampler w_* is obtained. Additional simulations are then carried out, from which the quantiles of the conditional distribution of $\underline{a}$ given $\underline{d}$ are estimated. Transforming back, we obtain the following 95% confidence intervals for the β_i based on a further sample of 10,000.

$$\begin{array}{lll} \beta_1: & (0.38, 1.58) & \text{where } \hat{\beta}_1 = 0.95 \\ \beta_2: & (0.91, 2.05) & \text{where } \hat{\beta}_2 = 1.50 \\ \beta_3: & (2.61, 3.72) & \text{where } \hat{\beta}_3 = 3.10 \\ \beta_4: & (3.22, 4.18) & \text{where } \hat{\beta}_4 = 3.55 \\ \beta_5: & (4.34, 5.35) & \text{where } \hat{\beta}_5 = 5.17 \end{array}$$

Note that the confidence intervals may be skewed. In particular, the interval for β_4 is highly asymmetric.

APPENDIX

Proof of Lemma 1 : Let $\epsilon > 0$ and $z_1, z_2, \ldots$ be a sequence such that $\frac{1}{n}\sum_{i=1}^{n} I_A(z_i) \to P(A)$ whenever A is an interval and $\bar{z} \to 0, s/\sqrt{n-1} \to 1$ as $n \to \infty$. This holds, with probability 1, for a sequence of i.i.d. values from P by the Glivenko-Cantelli Theorem and the Strong Law of Large Numbers.

Let $A = (a, b)$ be an interval in $\Re$. By the continuity of the distribution P, there exists $a_1 < a < a_2 < b_1 < b < b_2$ such that $B = (a_1, a_2) \cup (b_1, b_2)$ satisfies $P(B) \leq \epsilon/4$. Now $|I_A(\sqrt{n-1}d_i) - I_A(z_i)| = |I_A(\sqrt{n-1}(z_i - \overline{z})/s) - I_A(z_i)| = |I_{A \triangle (\overline{z} + (s/\sqrt{n-1})A)}(z_i)|$ where $\triangle$ denotes symmetric set difference. Since $\overline{z} \to 0$ and $s/\sqrt{n-1} \to 1$ we can find N_1 such that $n > N_1$ implies $A \triangle (\overline{z} + (s/\sqrt{n-1})A) \subseteq B$. We can also find N_2 and N_3 such that $n > N_2$ implies $|\frac{1}{n}\sum_{i=1}^n I_A(z_i) - P(A)| < \epsilon/2$ and $n > N_3$ implies $|\frac{1}{n}\sum_{i=1}^n I_B(z_i) - P(B)| < \epsilon/4$. Putting the above pieces together, we have that $n > \max\{N_1, N_2, N_3\}$ implies

$$\begin{aligned}
& |\frac{1}{n}\sum_{i=1}^n I_A(\sqrt{n-1}d_i) - P(A)| \\
\leq\ & |\frac{1}{n}\sum_{i=1}^n I_A(\sqrt{n-1}d_i) - \frac{1}{n}\sum_{i=1}^n I_A(z_i)| + |\frac{1}{n}\sum_{i=1}^n I_A(z_i) - P(A)| \\
\leq\ & \frac{1}{n}\sum_{i=1}^n I_B(z_i) + \epsilon/2 \leq |\frac{1}{n}\sum_{i=1}^n I_B(z_i) - P(B)| + P(B) + \epsilon/2 \\
\leq\ & \epsilon/4 + \epsilon/4 + \epsilon/2 = \epsilon
\end{aligned}$$

and this establishes the result for finite intervals. A straightforward modification gives the result whenever $|a|$ or $|b|$ is infinite.

Proof of Lemma 2 : We have that $g(x)$ is a density if and only if $1 = \int_{-\infty}^{\infty} g(x)dx = a_i$ and $q_i(x) \geq 0$. Given that $a_i = 1$, $q_i(x) = c_i D_i^{-1} x^2 + (b_i - c_i D_i^{-1}\mu_{3i})x + 1 - c_i D_i^{-1}$, for $q_i(x) \geq 0$, the coefficient of x^2 must be nonnegative. That is, $c_i D_i^{-1} \geq 0$. In addition for $q_i(x) \geq 0$, the discriminant corresponding to the equation $q_i(x) = 0$ must be nonpositive. That is, $(b_i - c_i D_i^{-1}\mu_{3i})^2 - 4c_i D_i^{-1}(1 - c_i D_i^{-1}) \leq 0$. Together these last two inequalities imply $0 \leq c_i \leq D_i$ and the result follows.

ACKNOWLEDGEMENTS

Both authors were supported in part by grants from the Natural Sciences and Engineering Research Council of Canada. The authors thank two referees for a number of constructive and useful comments.

BIBLIOGRAPHY

Andrews, D.F. and Herzberg, A.M. (1985). Data. Springer-Verlag.

Bartlett, M.S. (1935). The effect of non-normality on the t distribution. Proc. Cam. Phil. Soc., 223-231.

Cramer, H. (1946). Mathematical Methods of Statistics. Princeton University Press.

Davis, A.W. (1976). Statistical distributions in univariate and multivariate Edgeworth populations. Biometrika, 63, 3, 661-670.

Davis, P.J. and Rabinowitz, P. (1984). Methods of Numerical Integration, Second Edition. Academic Press.

Evans, M., Gilula, Z. and Guttman, I. (1989). Latent class analysis of two-way contingency tables. Biometrika, 76, 557-562.

Evans, M. and Swartz, T. (1988). Sampling from Gauss rules. SIAM J. Sci. Stat. Comput., Vol. 9, No. 5, 950-961.

Fisher, R.A. (1934) . Two new properties of mathematical likelihood. JRSSA, 144, 285-307.

Fraser, D.A.S. (1979). Inference and Linear Models. McGraw-Hill Inc. London.

Gayen, A.K. (1949). The distribution of Student's t in random samples of any size drawn from non-normal universes. Biometrika, 36, 353-369.

Gayen, A.K. (1950a). Significance of difference between the means of two non-normal samples. Biometrika, 37, 399-408.

Gayen, A.K. (1950b). The distribution of the variance ratio in random samples of any size drawn from non-normal universes. Biometrika, 37, 236-255.

Gayen, A.K. (1951). The frequency distribution of the product-moment correlation coefficient in random samples of any size drawn from non-normal universes. Biometrika, 38, 219-247.

Geweke, J. (1989). Modelling with normal polynomial expansions. In Economic Complexity. Chaos, Sunspots and Nonlinearity. Edited by W.A. Bennett, J. Geweke and K. Shell. Cambridge Univ. Press.

Golub, G.H. and Welsch, J.H. (1969). Calculation of Gauss quadrature rules. Math. Comp. 23 221-230.

Johnson, N. and Kotz, S. (1970). Continuous Univariate Distributions. Houghton Mifflin Company.

Kautsky, J. and Golub, G.H. (1983). On the calculation of Jacobi matrices. Linear Algebra Appl. 52/53 439-455.

Kendall, M. and Stuart, A. (1958). The Advanced Theory of Statistics, Vol. 1, Distribution Theory. Charles Griffin and Company.

Muirhead, R.J. (1982). Aspects of Multivariate Statistical Theory. John Wiley & Sons, Inc. New York.

Oh, M-S. and Berger, J.O. (1992). Adaptive importance sampling in Monte Carlo integration. To appear in J. Stat. Comp. and Simul.

Serfling, R.J. (1980). Approximation Theorems of Mathematical Statistics. John Wiley & Sons, Inc. New York.

Smith, A.F.M., Skene, A.M., Shaw, J.E.H. and Naylor, J.C. (1987). Progress with numerical and graphical methods for practical Bayesian statistics. The Statistician, 36, 75-82.