KATHOLIEKE UNIVERSITEIT
LEUVEN

Faculty of Business and Economics

Goodness-of-fit tests in mixed models

G. Claeskens and J.D. Hart

DEPARTMENT OF DECISION SCIENCES AND INFORMATION MANAGEMENT (KBI)

KBI 0902A

Goodness-of-fit tests in mixed models

Gerda Claeskens
ORSTAT and Leuven Statistics Research Center
Katholieke Universiteit Leuven, Naamsestraat 69, 3000 Leuven, Belgium
e-mail: gerda.claeskens@econ.kuleuven.be

Jeffrey D. Hart
Department of Statistics
Texas A&M University, College Station, TX77843, USA
e-mail: hart@stat.tamu.edu

February 2009

Abstract

Mixed models, with both random and fixed effects, are most often estimated on the assumption that the random effects are normally distributed. In this paper we propose several formal tests of the hypothesis that the random effects and/or errors are normally distributed. Most of the proposed methods can be extended to generalized linear models where tests for non-normal distributions are of interest. Our tests are nonparametric in the sense that they are designed to detect virtually any alternative to normality. In case of rejection of the null hypothesis, the nonparametric estimation method that is used to construct a test provides an estimator of the alternative distribution.

Keywords: mixed model, hypothesis test, nonparametric test, minimum distance, order selection.

AMS Subject classification codes: Primary 62G10

1 Introduction

Availability of large sets of data, some with many variables but only a few replicates, and others with many repeated observations per subject, asks for advanced models. Often, one uses a mixture of random and fixed effects for describing these data. For example, in microarray experiments one typically has information on thousands of genes, with only a few replicates. This is a situation where a "classical" model with only fixed effects would fail, since the number p of variables (genes) greatly exceeds n, the number of observations (replicates). Using a random effects model and estimating effects distributions is often preferable to trying to estimate all the individual effects. Often, normality is assumed for effects distributions. In this paper we address ways in which we can test the assumption of normality.

As another example, consider small area estimation where one usually has only a few observations per area. Typically, the areas are modelled as random effects. Their distribution is then important in constructing prediction intervals for specific areas. Using an incorrect distribution can lead to incorrect prediction bounds, with possibly important consequences for the conclusions drawn from such an analysis.

Mixed linear models, for example for longitudinal studies (with $p < n$), might ask that random effects distributions be more complex than the classical Gaussian. Approaches based on normal mixtures Komárek and Lesaffre (2008), penalized model fitting (Ghidey et al., 2004), or Hermite expansions (Zhang and Davidian, 2001; Chen et al., 2002) provide more flexible alternatives. Again, the question arises whether such approaches are justified by the data, or whether the simpler normal random effect distribution would suffice.

Rejecting the null hypothesis of normality in a linear mixed model and looking at the estimated alternative distribution might suggest missing variables in the model. For example, a missing fixed binary covariate might lead to a mixture of two distributions.

In this paper we provide several strategies for constructing formal statistical tests of the hypothesis of normality of random effects and/or error distributions. The tests will be nonparametric in the sense that we do not assume a single parametric form for the alternative model. The omnibus nature of the tests leads to good power for a wide range of alternative distributions.

In Sections 3 and 4 we give a (non-exhaustive) overview of simple diagnostic measures (mainly based on plots) for checking the distribution of random effects and error terms. Section 5 proposes series-based tests that have a close connection to order selection and Neyman smooth type tests developed for testing hypotheses in (fixed effects) regression models and testing the fit of an error density (without covariates present), respectively. Section 7 explains a minimum distance testing approach that could be used in mixed effects models for which each random effect has at least two replicates. These tests are applied to some data examples in Sections 6 and 8. A discussion follows in Section 9.

2 Notation

The main part of this paper will work with linear mixed models, even though several of the proposed methods can be applied to more general mixed models. A linear mixed model takes the form

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{Y}$ is the vector of length N of response values, $\boldsymbol{X}$ is the $N \times K$ design matrix of fixed effect covariates, $\boldsymbol{\beta}$ is the vector of length K of fixed effect parameters, $\boldsymbol{Z}$ is the $N \times d$ design matrix of random effect covariates, $\boldsymbol{\gamma}$ is the vector of length d of random effects and $\boldsymbol{\varepsilon}$ is the vector of random errors, assumed to be independent of the random effects $\boldsymbol{\gamma}$. Standard assumptions include independence between random effects $\boldsymbol{\gamma}$ and errors $\boldsymbol{\varepsilon}$, as well as normality for both random effects and error distributions. Individual components of the vector $\boldsymbol{Y}$ are often denoted using multiple indices. For example, in longitudinal studies Y_{jk} denotes the kth observation for the jth subject. The number of subjects is given by n, while n_j denotes the number of replicated observations for subject j.

The covariance matrix of $\boldsymbol{Y}$ is denoted by $\boldsymbol{V}$, while those of $\boldsymbol{\gamma}$ and $\boldsymbol{\varepsilon}$ are denoted by $\boldsymbol{\Sigma}_g$ and $\boldsymbol{\Sigma}_\varepsilon$, respectively. The likelihood of the data $\boldsymbol{Y}$ will be denoted by $L(\cdot)$, possibly stressing dependence on parameter values in the function argument, when this would not be clear from the context. We use $\boldsymbol{\theta}$ as a notation for the combined parameters in the mixed model (coming from both fixed and random effects and error distributions).

3 Graphical diagnostics in mixed models

Calvin and Sedransk (1991) describe two methods to construct residuals for graphically checking the normality assumption on the error terms; however, their methods cannot be used to check assumptions on random effects distributions. Their first method consists of premultiplying the response vector by the inverse of the square root of the estimated variance matrix $\boldsymbol{V}$ of the response variables. This leads to residuals that are approximately standard normally distributed. Disadvantages of this approach include the smoothing effect that averaging of residuals has (which might mask effects of outlying observations), and the effect of using estimated variance components rather than the true values in the standardization. A similar transformation of residuals has recently been investigated by Jacqmin-Gadda et al. (2007), who multiply the residuals $\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ by the Cholesky square root of the covariance matrix to obtain residuals that are uncorrelated. These are then used in a QQ-plot to check normality. Like the first Calvin-Sedransk approach, this method does not provide a means of testing assumptions concerning the random effects distribution. The second approach described in Calvin and Sedransk (1991) uses best linear unbiased predictors (BLUP) of the random effects, predicts the response values, and computes BLUP residuals of the form $\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{Z}\hat{\boldsymbol{\gamma}}$. While this does not introduce averaging of residuals, the resulting residuals are still correlated.

The diagnostic plots of Lange and Ryan (1989) use standardized empirical Bayes estimates of the random effects in a weighted normal QQ-plot. This method works in particular for graphically checking the distribution of random effects. The choice of weights allows one to test normality of multiple random effects by computing a linear combination of effects. As in the first Calvin-Sedransk approach, this method can be adversely affected by having to estimate unknown parameters, namely fixed effects and variance components.

The QQ-plots of Park and Lee (2004) for longitudinal data are based on the fact that, under normality, a quadratic form in the residuals $\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ is approximately chi-squared distributed when estimated variances are inserted in the covariance matrix.

4 Traditional tests for normality adapted to mixed models

Formal tests have not been studied extensively. Most approaches try to transform the correlated residuals of a linear mixed model to uncorrelated residuals, in order to apply a classical test for normality. Hwang and Wei (2006) apply such a method to a two-stage cluster sampling design corresponding to a mixed model of the form

$$Y_{jk} = \mu_j + \gamma_j + \varepsilon_{jk},\ j = 1, \dots, n;\ k = 1, \dots, n_j,$$

with the ε_{jk} and γ_j independent mean-zero random variables with variance σ_e^2 for the errors and σ_γ^2 for the random effect. Under normality of both error and random effects, Hwang and Wei (2006) construct a transformation of the response values Y_{jk} that results in uncorrelated transformed variables. These are then used for testing univariate normality using classical test statistics (such as a Shapiro-Wilk test or tests based on skewness). When the null hypothesis of normality is rejected, one cannot say whether this is due to a misspecified random effects distribution, or to a wrong error distribution.

Pearson χ^2-type tests for mixed models have been studied by Jiang (2001), who assumes a linear random effects model with independent additive random effects of the form

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}_1\boldsymbol{\gamma}_1 + \ldots + \boldsymbol{Z}_s\boldsymbol{\gamma}_s + \boldsymbol{\varepsilon},$$

and performs hypothesis tests about the random effects and error distributions. The Pearson "χ^2" statistic (which in this case, however, does not have a χ^2 distribution), is based on a partitioning of the range of response values into disjoint intervals. One computes the observed "cell counts," indicating how many Y_{jk} are within each cell, and compares that to the estimated expected cell counts under the hypothesized distributions (by inserting estimators of fixed effects and variance components). The test statistic is a multiple of the sum of squared differences of observed and expected cell counts. Since the response values are correlated by construction of the mixed model, the observed cell counts are not a sum of independent and identically distributed values, causing the limiting distribution to be different from χ^2, and making the choice of the normalizing factor difficult. For the case of a single random effect where $Y_{jk} = X_{jk}\beta + \gamma_j + \varepsilon_{jk}$ with $j = 1, \ldots, n$ and $k = 1, \ldots, n_j$, the normalizing factor is taken to be n, while the choice is less clear in the case of multiple random effects. Moreover, this test requires the same difficult choices as does the classical Pearson χ^2 test, namely what should the number of cells be and should the cells be of equal length or of equal probability. Since the null hypothesis simultaneously specifies the random effects distributions and the error distribution, it is not clear in case of rejection what has been the cause nor what is a good form for the distributions. The tests that we will construct in Sections 5 and 7 explicitly suggest alternative distributions.

A test that does not not specifically address testing the distribution of random effects, but does test a parametric mixed effect model against a semiparametric mixed effect model, is studied by Lombardía and Sperlich (2008).

5 Order selection-type goodness-of-fit tests

5.1 The concept of order selection tests

Until further notice we assume that model (1) holds. We first address the problem of testing normality of the random effects distribution assuming that the error distribution is known up to finitely many parameters.

Order selection tests were introduced by Eubank and Hart (1992) to test the fit of a regression mean function. This testing approach is based on an (orthogonal) series expansion of the function of interest about the hypothesized null model. To apply an order selection test to our problem, the random effects density is expanded in a series about a normal density; see for example equation (2). For an overview of such estimation methods for random effects densities, see Section 5.2. The series expansion is truncated after M terms, yielding an approximation to the underlying density that improves as the truncation point, or *order*, M increases. One may fit several models, each with a different value of M, and then use a suitable model selection criterion to determine an appropriate M. Eubank and Hart (1992) used a modified Mallows' C_p to determine the order. An intuitively appealing test can now be constructed as follows. If the model selection criterion selects a model with more parameters than the null model, then the null hypothesis is rejected. Otherwise, normality of the random effects distribution is not rejected. In Section 5.3.1 we apply the theoretical results of Aerts et al. (1999) to obtain the limiting distribution of this test.

5.2 Density estimation methods

In recent years, nonparametric estimation approaches have been developed to estimate the random effects densities as smooth density functions. Zhang and Davidian (2001) use the so-called 'semi-nonparametric' representation of a density function as studied by Gallant and Nychka (1987). This estimator takes the form of a Hermite series where the normal density function is multiplied by the square of a polynomial and suitably normalized to arrive at a proper density function. Clearly, the hypothesized normal density function is obtained when setting the polynomial equal to 1, while adding more terms to the series expansion allows one to obtain density functions with more features than a normal density. This expansion, and in particular the necessary number of terms in a truncated series approximation, will form the basis of our formal testing approach. See Section 5.3.1 for more details.

A finite mixture of normal density functions is fit by Verbeke and Lesaffre (1996). To test for the presence of such heterogeneity, they discourage the use of a likelihood ratio test, but rather suggest transforming the response vector $\boldsymbol{Y}_j$ for the jth individual by taking a linear combination $\boldsymbol{a}_j^t \boldsymbol{Y}_j$, where the vector $\boldsymbol{a}_j$ corresponds to the eigenvector belonging to the largest eigenvalue of $\boldsymbol{Z}_j \widehat{\mathrm{Var}}(\boldsymbol{\gamma}_j) \boldsymbol{Z}_j^t$, and then using a Kolmogorov-Smirnov or Shapiro-Wilk test. Again, an alternative could be to base a test on the data-driven selection of the number of components in the mixture distribution; see Section 5.3.3.

A further extension of the finite mixture model is studied by Ghidey et al. (2004) who, based on the idea of penalized spline estimators (P-splines), fit a large number of mixture components and introduce a penalty on the finite differences of coefficients related to the mixture proportions.

A comparison of the three estimation methods mentioned above, together with the 'smoothing by roughening' method of Shen and Louis (1999), which uses empirical Bayes estimators, is given in Ghidey et al. (2008).

5.3 Tests for normality of random effects assuming that the error distribution is normal

Here we consider the mixed model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Z\gamma} + \boldsymbol{\varepsilon}$, where we assume that $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma_\varepsilon^2 \boldsymbol{I}_N)$ and pose no assumption on the distribution of $\boldsymbol{\gamma}$. Our interest lies in testing the null hypothesis

$$H_0 : \boldsymbol{\gamma} \sim N_d(\boldsymbol{\mu}_\gamma, \boldsymbol{\Sigma}_\gamma),$$

where the covariance structure is not specified. By the reparametrization $\boldsymbol{\gamma} = \boldsymbol{\mu}_\gamma + \boldsymbol{GU}$, with $\boldsymbol{\mu}_\gamma$ the mean of $\boldsymbol{\gamma}$ and $\boldsymbol{\Sigma}_\gamma = \boldsymbol{GG}^t$, it is sufficient to test the hypothesis that $\boldsymbol{U} \sim N_d(\boldsymbol{0}, \boldsymbol{I})$.

5.3.1 Semi-nonparametric Hermite expansions

As a first approach we follow Zhang and Davidian (2001), who use a so-called semi-nonparametric (SNP) estimator for the distribution of the random effects $\boldsymbol{\gamma}$, as developed by Gallant and Nychka (1987). This estimator is based on a Hermite expansion of the unknown density of $\boldsymbol{\gamma}$ about the normal density. More specifically, we approximate the density f_U of the standardized variable $\boldsymbol{U}$ in $\boldsymbol{\gamma} = \boldsymbol{\mu}_g + \boldsymbol{GU}$. By an Edgeworth expansion of the density of $\boldsymbol{U}$ around the normal density ϕ (here given for the one-dimensional case; see e.g. Severini (2000), Section 2.3, which also contains an expression for the multi-dimensional case),

$$f_U(u) = \phi(u)\{1 + k_3 H_3(u) + k_4 H_4(u) + \ldots\},$$

where k_3, k_4 are related to the cumulants of U and the Hermite polynomials satisfy $H_j(u)\phi(u) = (-1)^j \frac{d^j \phi(u)}{du^j}$, and are hence polynomials in u. For example, $H_3(u) = u^3 - 3u$, $H_4(u) = u^4 - 6u^2 + 3$.

By a reordering of the terms in the expansion we may approximate the infinite series by the seminonparametric density, here given for the d-dimensional case,

$$\widehat{f}_{U,M}(\boldsymbol{u}) = P_M^2(\boldsymbol{u})\phi(\boldsymbol{u}), \tag{2}$$

where ϕ is the d-dimensional standard normal density and $P_M(\cdot)$ is a d-variable polynomial, defined as

$$P_M(\boldsymbol{u}) = \sum_{|\boldsymbol{\lambda}|\le M} a_{\boldsymbol{\lambda}} \boldsymbol{u}^{\boldsymbol{\lambda}}.$$

Here, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d)$, $|\boldsymbol{\lambda}| = \sum_{l=1}^{d} \lambda_l$, $\boldsymbol{u}^{\boldsymbol{\lambda}} = u_1^{\lambda_1} \ldots u_d^{\lambda_d}$, and the coefficients $a_{\boldsymbol{\lambda}}$ for each $\boldsymbol{\lambda}$ in the sum satisfy conditions to ensure that the integral of $\hat{f}_{U,M}$ is equal to one. The integer M is called the *order* of the polynomial. For example, in the case $M = 2$ and $d = 2$, the integers λ_1 and λ_2 satisfy $|\lambda_1 + \lambda_2| \le 2$ and the polynomial is

$$P_2(\boldsymbol{u}) = a_{00} + a_{10}u_1 + a_{01}u_2 + a_{20}u_1^2 + a_{11}u_1u_2 + a_{02}u_2^2.$$

This leads immediately to an approximation of the density function f_γ of γ. For linear mixed models, this constraint on the coefficients $\boldsymbol{a}$, which is the vector containing all $a_{\boldsymbol{\lambda}}$, can be given explicitly by requesting that $\boldsymbol{a}^t\boldsymbol{A}\boldsymbol{a} = 1$, where the matrix $\boldsymbol{A}$ is defined in terms of moments of d-dimensional uniform random variables (see Zhang and Davidian, 2001, for details). For generalized linear mixed effects models (see for example Chen et al., 2002) no such explicit constraint can be given, and instead one introduces a normalizing constant which, together with setting the first coefficient $a_{0\ldots0} = 1$, ensures that $\widehat{f}_{U,M}$ integrates to one. The obtained densities can take various shapes, with tails ranging from lighter than normal to t-like tails. From a computational point of view, this method is attractive since it avoids the construction of residuals as needed for ordinary density estimation. It phrases the model again as a closed-form likelihood and performs maximum likelihood estimation of all parameters directly, namely fixed effects, variance components and polynomial coefficients $\boldsymbol{a}$.

Note that the integer M plays the role of a smoothing parameter, with larger values of M corresponding to less smooth distributions, i.e., ones with more features. Chen et al. (2002) construct an informal test for normality by letting M take the values 0, 1 and 2, and choose one of those three possibilities by application of the information criteria AIC (Akaike, 1973), BIC (Schwarz, 1978; Akaike, 1978) or Hannan and Quinn's criterion (1979). This idea can be formalized by making the connection to order selection and Neyman smooth type tests; see Eubank and Hart (1992) and Ledwina (1994).

To estimate a one-dimensional distribution, only one term at a time is added to the truncated series, leading to the limit distribution as described in Eubank and Hart (1992), Hart (1997) and Aerts et al. (1999).

For a two-dimensional distribution of random effects, the number of terms added in the construction of the Hermite series in each step (that is, going from $M-1$ to M) is equal to $M+1$. This corresponds precisely to model sequence (b) of Aerts et al. (2000). For the summation limit in (2) equal to M, a total of $N_{2,M} = \binom{M+2}{M} = 0.5M(3+M)+1$ terms are contained in the sum. Upon invoking the constraint that $\hat{f}_{U,M}$ integrates to 1, the model $\hat{f}_{U,M}$ contains $N_{2,M} - 1$ more estimated parameters than the null model. Note that the fixed effect parameters and the error covariance matrix $\boldsymbol{\Sigma}_\varepsilon$ enter in the same way in all models. Hence, for the penalized criteria it suffices to consider in the penalty only the number of terms added in the series expansion. Indeed, adding a constant to the penalty that is the same for all models will not change the selected order. The test that rejects whenever the order $\widehat{M}_{\text{aic}} > 0$, is equivalent to rejecting H_0 when (see Aerts et al., 1999)

$$T_{\text{OS},n} = \max_{1 \le m \le M} \frac{2\{\log L(\hat{\boldsymbol{a}}_m) - \log L_{H_0}\}}{m(m+3)/2} > 2, \tag{3}$$

where $L(\hat{\boldsymbol{a}}_m)$ denotes the maximized likelihood under the alternative $f_U \equiv \hat{f}_{U,m}$, and L_{H_0} denotes the maximized likelihood under the null hypothesis of normality. Under the regularity conditions on the likelihood function as stated in Aerts et al. (1999), and using the extension to more than one dimension as in Aerts et al. (2000), we obtain that $T_{\text{OS,n}} \xrightarrow{\mathcal{D}} T_{\text{OS}}$ as n tends to infinity, with

$$T_{\text{OS}} = \max_{r \ge 1} \frac{V_r}{r(r+3)/2}, \text{ where } V_r = \sum_{j=1}^{r} \chi^2_{j+1}, \tag{4}$$

$\chi^2_2, \chi^2_3, \ldots$ are independent random variables, and χ^2_j is distributed chi-square with j degrees of freedom, $j \ge 2$. Using 100,000 simulated values of this distribution, it turns out that the level of test (3) which rejects for values bigger than 2, is about 0.18. This is the situation of the informal testing approach when using the standard AIC.

Let C_α be the $1-\alpha$ quantile of T_{OS}, and define a modified version of AIC by $\text{AIC}_{C_\alpha}(M) = -2$ log-likelihood$_M + C_\alpha(N_{2,M} - 1)$. If $\widehat{M}_{\text{AIC}}$ is the value of M that minimizes AIC_{C_α}, then a test that rejects the hypothesis of normality if and only if $\widehat{M}_{\text{AIC}} > 0$ has limiting level α. This test is equivalent to working directly with $T_{\text{OS},n}$ and rejecting H_0 if and only if $T_{\text{OS},n} > C_\alpha$. The latter approach would allow calculation of an approximate P-value by comparison of the observed value of $T_{\text{OS},n}$ with the distribution of T_{OS}.

For a d-dimensional density estimator, at step M, there are $\binom{M+d-1}{d-1}$ terms added to the series. In total, after M steps, this leads to

$$N_{d,M} = \sum_{m=0}^{M} \binom{m+d-1}{d-1} = \binom{M+d}{M}$$

terms. Thus, the traditional AIC for this model with the series truncated at value M takes the form $\text{AIC}(M) = -2$ log-likelihood$_M + 2(N_{d,M} - 1)$. Rejecting the null hypothesis of normality at level α is equivalent to rejecting when

$$T_{\text{OS},d,n} = \max_{1 \le m \le M} \frac{2\{\log L(\hat{a}_m) - \log L_{H_0}\}}{(N_{d,m} - 1)} > C_n, \tag{5}$$

with C_n appropriately chosen as the $(1-\alpha)$ quantile of the corresponding distribution. The expression in (3) is the special case with $d = 2$. Under the same set of regularity conditions, it can be shown that $T_{\text{OS},d,n}$ has limiting distribution

$$T_{\text{OS},d} = \max_{r \ge 1} \frac{V_r}{(N_{d,r} - 1)}, \tag{6}$$

with $V_r = \sum_{j=1}^{r} \chi^2_{n(d,j)}$, $n(d,j) = \binom{j+d-1}{d-1}$ (the number of terms added in step j for the d-dimensional density estimator) and $\chi^2_2, \chi^2_3, \ldots$ defined as before. Again, the critical value C_n or a P-value is easily simulated for any d. As an alternative, one may apply a bootstrap procedure, which might be advantageous, especially for small data sets.

We wish to stress that this particular way of testing may not be very powerful for large dimensions (i.e., large d) because of the curse of dimensionality. For alternative testing procedures and different schemes for entering terms in a series expansion, see Aerts et al. (2000). The so-called frequentist-Bayes tests of Hart (2008) may also be adapted to the setting of the current paper and it would be worthwhile comparing them to order selection tests in high dimensional cases.

Both $\mathrm{BIC}(M) = -2$ log-likelihood$_M + \log(n)(N_{d,M} - 1)$ and the Hannan-Quinn criterion $\mathrm{HQ}(M) = -2$ log-likelihood$_M + \log\log(n)(N_{d,M} - 1)$ are consistent model selection criteria (see, for example, Claeskens and Hjort, 2008, Ch. 4). This implies that if the null hypothesis holds, then the null model will be selected with probability tending to 1 as $n \to \infty$. This has important consequences for the construction of a test statistic. In order to construct a valid test (with a non-trivial distribution under the null hypothesis), we have to omit $M = 0$ from the model choice list. In other words, we do not allow that the null model is chosen by the BIC or HQ criterion. This construction is used in the goodness of fit testing setting by Ledwina (1994). As a test statistic, we can take the value of the likelihood ratio statistic at the model with the series truncated at the BIC or HQ selected model order. Note that, originally, Ledwina (1994) used a score test. In the construction of a nested sequence of models, where terms are added to the series expansion one by one, as is the case in the one-dimensional density estimation setting, this approach results in a limiting χ^2_1 distribution under the null hypothesis, and a non-central χ^2_1 under local alternatives. (See, however, Claeskens and Hjort (2004) for alternative schemes with better power properties.) In general, for estimation of a d-dimensional density using the Hermite series approach, the simplest model (excluding the normal model) contains d more estimated parameters than the null model. This implies that the limiting distribution of a test based on BIC or HQ order selection has a limiting χ^2_d distribution when the null model is excluded from the model search.

5.3.2 Log-linear expansions

An alternative to the Hermite expansion is the following log-linear expansion of the density function

$$f_M(\boldsymbol{u};\boldsymbol{a}) = \phi(\boldsymbol{u})c_M(\boldsymbol{a})^{-1}\exp\left\{\sum_{j=1}^{M} a_j\psi_j(\boldsymbol{u})\right\}, \tag{7}$$

where the basis functions ψ_j are orthogonal with respect to the null density ϕ in the sense that $\int \phi(\boldsymbol{u})\psi_j(\boldsymbol{u})\psi_k(\boldsymbol{u})d\boldsymbol{u} = I(j = k)$, $\boldsymbol{a} = (a_1,\ldots,a_M)$ and the normalizing constant is given by $c_M(\boldsymbol{a}) = \int \phi(\boldsymbol{u})\exp\{\sum_{j=1}^{M} a_j\psi_j(\boldsymbol{u})\}d\boldsymbol{u}$.

To estimate the unknown parameters in the linear mixed model with this type of log-linear expression for the random effect distribution, we can proceed as follows. The marginal density of the response vector $\boldsymbol{Y}_i$ for subject i ($i = 1,\ldots,n$) can be written in terms of the conditional density of $\boldsymbol{Y}_i$ given the (standardized) random effects $\boldsymbol{U}_i = \boldsymbol{u}$, and the marginal density of the random effects $f_M(\boldsymbol{u};\boldsymbol{a})$ in the following way

$$f(\boldsymbol{Y}_i;\theta) = \int f(\boldsymbol{Y}_i|\boldsymbol{u};\theta)\phi(\boldsymbol{u})c_M(\boldsymbol{a})^{-1}\exp\left\{\sum_{j=1}^{M} a_j\psi_j(\boldsymbol{u})\right\}d\boldsymbol{u}.$$

If the random effects have a d-variate standard normal distribution, then $f(\boldsymbol{Y}_i|\boldsymbol{u};\theta)\phi(\boldsymbol{u}) = g(\boldsymbol{u}|\boldsymbol{Y}_i)g(\boldsymbol{Y}_i;\theta)$, where $g(\boldsymbol{Y}_i;\theta)$ denotes the marginal density of $\boldsymbol{Y}_i$ under the null model and $g(\boldsymbol{u}|\boldsymbol{Y}_i)$ the conditional density of the random effects, given $\boldsymbol{Y}_i$. Hence, the log likelihood of the data can be written as

$$\sum_{i=1}^{n} \log g(\boldsymbol{Y_i}; \theta) + \sum_{i=1}^{n} \log \left(c_M^{-1}(\boldsymbol{a}) E_{U_i|Y_i,\theta} \left[\exp \left\{ \sum_{j=1}^{M} a_j \psi_j(\boldsymbol{U}_i) \right\} \right] \right).$$

This expression is to be maximized for the unknown parameter values. The first sum is simply the log-likelihood of normal data, while the second sum needs the evaluation of conditional means of $\exp\{\psi_j(\boldsymbol{U})\}$ with $\boldsymbol{U}$ standard normal. A test of the hypothesis that $\boldsymbol{\gamma}$ has a d-variate normal distribution can now proceed in the same way as with the Hermite expansions. Models with several values of approximation level M are fit to the data (but with otherwise the same random and fixed effects) and a model selection criterion, or equivalently an order selection test statistic, is applied. Asymptotic distribution theory as in Aerts et al. (1999) justifies the approach. By expanding around another distribution than the normal one in (7), other null hypotheses can be tested, which makes this type of test interesting for use in, for example, generalized linear mixed models.

5.3.3 Mixtures of normal distributions

An interesting alternative to a series expansion to model a more flexible random effect distribution is through the use of a mixture of normal distributions. Verbeke and Lesaffre (1996) used this approach in mixed linear models. In their heterogeneity model, the random effects are assumed to be sampled from a mixture of G normal distributions with different (unknown) means and identical (unknown) covariance matrix,

$$\boldsymbol{\gamma}_i \sim \sum_{g=1}^{G} \pi_g N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_\gamma),$$

with the constraints that $\sum_{g=1}^{G} \pi_g = 1$ and $E(\boldsymbol{\gamma}_i) = \sum_{g=1}^{G} \pi_g \boldsymbol{\mu}_g = 0$.

This implies that the response $\boldsymbol{Y}_i$ also follows a mixture of normal distributions, namely

$$\boldsymbol{Y}_i \sim \sum_{g=1}^{G} \pi_g N(\boldsymbol{X}_i + \boldsymbol{Z}_i \boldsymbol{\mu}_j, \boldsymbol{V}_i),$$

where the covariance matrix is $\boldsymbol{V}_i = \boldsymbol{Z}_i \boldsymbol{\Sigma}_\gamma \boldsymbol{Z}_i^t + \sigma_\varepsilon^2 \boldsymbol{I}_{n_i}$. While Verbeke and Lesaffre (1996) use the EM algorithm to estimate the unknown parameters, Proust and Jacqmin-Gadda (2005) use a Marquardt algorithm. An extension of the above model is provided by Ghidey et al. (2004) who use a large number of mixture components and add a penalty to deal with the resultant possibility of overfitting, similarly to the penalized spline fitting idea.

The model under the null hypothesis results when $G = 1$, while values of $G > 1$ allow for more flexible shapes of the distribution. Testing the null hypothesis of normality can proceed by means of tests on the value of G, in a similar way as for order selection tests. Proust and Jacqmin-Gadda (2005) suggested the use of AIC and BIC to determine the number of components. They constructed an AIC-type model selection criterion by penalizing twice the value of the attained log likelihood by twice the number of parameters, the latter composed by adding the number of fixed effects parameters, the number of parameters in the covariance matrix and the number of components G in the mixture. This is in the spirit of AIC for linear models, without random effects and mixture distributions. It turns out (see for example Naik et al., 2007) that for mixture models, this is not a good course to follow. In Naik et al. (2007) the mixture regression criterion is developed and it is shown to be an efficient selection criterion to determine jointly the number of components in the mixture and the regression parameters in the (fixed effects) linear regression model.

A value of AIC for models with random effects (with a single normal distribution) is studied by Vaida and Blanchard (2005). A 'marginal AIC' is there formed by considering the mixed model as a linear model though with a correlation structure for the errors that is determined by the random effects. The marginal AIC is defined as twice a penalized log-likelihood value, where the penalty is the total number of parameters (fixed effects and parameters in the covariance matrix). For conditional models where the random effects are themselves of main interest, a different formula is needed (see Vaida and Blanchard, 2005). A criterion in the spirit of the AIC for mixture models with random effects can be formed by combining these methods. Instead of the constant 2 in the penalty part of the AIC, we could use a value C that is found via simulation, or bootstrap, to yield the desired level of the test. In contrast to orthogonal series tests, in the context of mixture models, the orthogonality is no longer present, which complicates the asymptotic distribution. Rather than developing such theory here, we suggest using the bootstrap for practical application.

5.4 Simultaneous tests on error and random effects distributions

The way of testing as described in the previous section can be extended to simultaneously testing normality hypotheses of both random effects and error distributions. We here describe the approach based on Hermite series expansions; a similar method results for log-linear expansions.

Consider approximation (2) for the random effects density, and similarly, write for the error density

$$\widehat{f}_{\varepsilon,M_\varepsilon}(\boldsymbol{v}) = \left\{ \sum_{|\boldsymbol{\delta}|\le M_\varepsilon} b_{\boldsymbol{\delta}} \boldsymbol{v}^{\boldsymbol{\delta}} \right\}^2 \phi(\boldsymbol{v}).$$

This leads to modeling the marginal density of $\boldsymbol{Y}$ in the following way,

$$\begin{aligned} f_{M,M_\varepsilon}(\boldsymbol{y};\theta) &= \int \widehat{f}_\varepsilon(\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Zu}|\boldsymbol{u},\theta)\widehat{f}_{U,M}(\boldsymbol{u})d\boldsymbol{u} \\ &= \int \phi(\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Zu};\theta)\phi(\boldsymbol{u}) \left\{ \sum_{|\boldsymbol{\delta}|\le M_\varepsilon} b_{\boldsymbol{\delta}} (\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Zu})^{\boldsymbol{\delta}} \right\}^2 \\ &\quad \times \left\{ \sum_{|\boldsymbol{\lambda}|\le M} a_{\boldsymbol{\lambda}} \boldsymbol{u}^{\boldsymbol{\lambda}} \right\}^2 d\boldsymbol{u}. \end{aligned}$$

Maximizing the likelihood of the data yields maximum likelihood estimates of the model parameters θ, a_λ and b_δ. The null model of normality for both random effects and error distribution is obtained when $M = M_\varepsilon = 0$. The order selection idea now uses a model selection method such as AIC, with appropriately chosen penalty constant, to determine data-driven values for M and M_ε. A particular advantage of this test is that, in case of rejection, it can indicate where the discrepancy is; for example, when only one of the two orders exceeds zero.

6 Examples and simulation results

6.1 Framingham cholesterol data

As a first example we consider the Framingham cholesterol data as used by Zhang and Davidian (2001). This dataset consists of information on 200 individuals, with cholesterol levels measured at the start of the study and further every two years for 10 years. Other information given is the age at the start of the study and the individual's gender. Not all measurements for all subjects were recorded. The following mixed model, with a random intercept and random slope for the time effect, is fit to the data:

$$Y_{jk} = \beta_0 + \beta_1 \text{age}_j + \beta_2 \text{gender}_j + \beta_3 \text{time}_{jk} + \gamma_{0j} + \gamma_{1j} \text{time}_{jk} + \varepsilon_{ij}.$$

Zhang and Davidian (2001) used the SNP estimation method on this dataset and tried models with series truncation point $M = 0, 1, 2$. They then applied AIC with penalty term twice the number of parameters in the model, BIC with penalty term the log of the total number of observations N (although n could have been another choice here) multiplied by the number of parameters, and the Hannan and Quinn criterion with penalty term $\log\log(N)$ times the number of parameters. They report that all three criteria prefer the model with $M = 1$ over the models with $M = 0$ or $M = 2$.

We will test the null hypothesis of bivariate normality of the random effects

$$H_0 : (\gamma_0, \gamma_1) \sim N_2(\mathbf{0}, \mathbf{\Sigma}_\gamma)$$

using the order selection test based on the semi-nonparametric Hermite expansion. Models were fit with truncation points $M = 0$, corresponding to the null model, and with $M = 1, 2, \ldots, 5$. Using the asymptotic distribution, we obtain the following simulated critical values of the test statistic: at nominal level 10% $C_n = 2.474$, at 5% $C_n = 3.084$ and at level 1% $C_n = 4.584$. The observed value of the test statistic T_{OS} is equal to 12.39, with the corresponding P-value equal to 10^{-5}, which is clearly evidence that the null hypothesis of bivariate normality of the random effects should be rejected. The chosen value of the series truncation point is $M = 1$, also indicating that a more complex model is needed than just bivariate normality for the random effects. These values are computed using log likelihood values for the models with $M = 0, \ldots, 5$, as in the following table:

M	0	1	2	3	4	5
log lik.	-160.986	-148.597	-146.891	-145.470	-142.917	-142.606
$N_{2,M} - 1$	0	2	5	9	14	20

6.2 The sleep study

The data for the sleep study are obtained from R's library lme4, using the syntax data(sleepstudy). This considers the reaction times in a sleep deprivation study for 18 individuals on 10 consecutive days. On day zero, the subjects had their normal amount of sleep, but starting that night they were restricted to only three hours of sleep per night. The response variable is the average reaction time on a series of tests given each day to each person in the study. This dataset is balanced, with all observations recorded. The fitted model contains a random subject specific intercept and slope

$$Y_{jk} = \beta_0 + \beta_1 \text{Day}_k + \gamma_{0j} + \gamma_{1j} \text{Day}_k + \varepsilon_{jk}.$$

We test the null hypothesis of bivariate normality of the random effects

$$H_0 : (\gamma_0, \gamma_1) \sim N_2(\mathbf{0}, \mathbf{\Sigma}_\gamma).$$

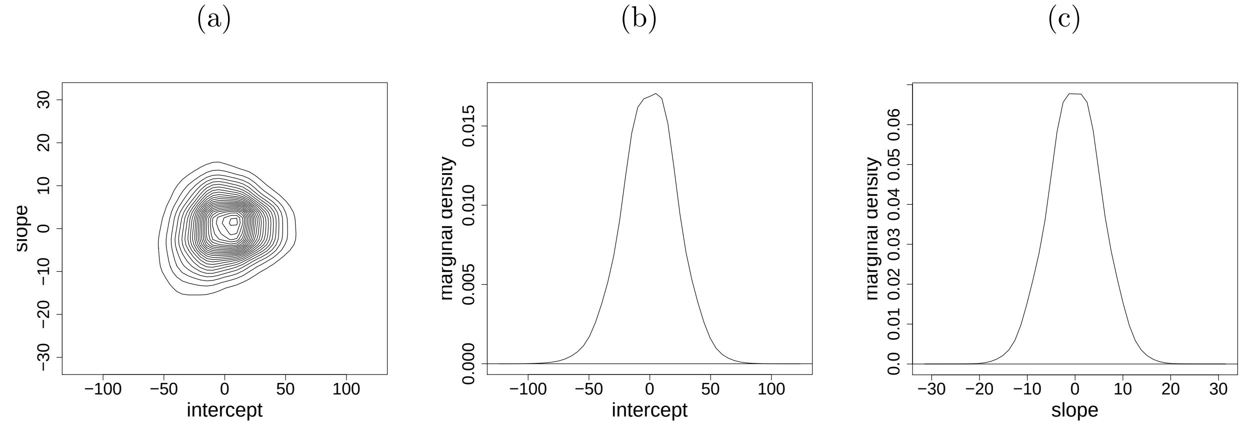

Figure 1: Sleep study. (a) Contour plot of the estimated density of the random slope and random intercept. (b) Marginal density estimate for the random intercept. (c) Marginal density estimate for the random slope. Estimates are obtained by using a penalized Gaussian mixture linear mixed model.

Table 1 contains the differences between AIC for models with $M = 1, \dots, 9$ and the AIC value under H_0, computed as

$$\text{aic}_C(M) = 2\{\log L(\hat{\boldsymbol{a}}_M) - \log L_{H_0}\} - C(N_{2,M} - 1).$$

It is worth noting that *large* values of aic_C indicate good models. In order to perform the test at, say, level 5%, we use the penalty constant $C = 3.084$, which is obtained by simulation from the asymptotic distribution of $T_{\text{OS},2}$ in (6). For these data, all $\text{aic}_{3.084}$ values are negative, which indicates that the model of the null hypothesis is chosen as the best one. So, for the sleep study data we do not have evidence that the random intercept and slope have a more complicated distribution than bivariate normality. This is also found by using the alternative version of the order selection test. Computing the test statistic $T_{\text{OS},2,n}$ as in (5) gives the value 2.63, with a corresponding simulated P value of 0.084. Interestingly, this is a case where the traditional AIC selects a nonnull model, but the penalty-modified AIC does not.

A graphical representation of these data is obtained using a penalized Gaussian mixture linear mixed model (Ghidey et al., 2004) with a grid of 10×10 density bases and a penalty term based on differences of adjacent coefficients. The AIC is used to obtain a data-driven value of the two penalty constants (one for each dimension). Figure 1 contains a contour plot of the bivariate density of the random intercept and slope, as well as marginal density plots. A visual inspection also shows no clear departure from normality.

6.3 Simulation results

We conducted a simulation study to compare the performance of the SNP-based order selection test $T_{\text{OS},1,n}$ to that of the Pearson test, as described by Jiang (2001). The data are generated according to the mixed model

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + \gamma_i + \varepsilon_{ij}, \tag{8}$$

with replicates $j = 1, 2, 3 = m$, $i = 1, \dots, n$, $\beta_0 = 1$, $\beta_1 = 2$, $x_{ij} \sim \text{Unif}(0, 10)$, $\varepsilon_{ij} \sim N(0, 0.3)$. As sample sizes we took $n = 35, 50$ and 100. We test for normality of the random effect, that is,

Table 1: Sleep study. Differences between AIC for models with various truncation points M in the Hermite series expansion and AIC for the null model. AIC$_2$ is the traditional AIC, with penalty twice the difference in numbers of parameters for the considered models. The AIC$_C$ uses $C = 3.084$ instead of 2 in the penalty term, corresponding to a 5% level for the order selection test.

M	1	2	3	4	5	6	7	8	9
$N_{2,M} - 1$	2	5	9	14	20	27	35	44	54
aic$_2$	1.25	-2.75	-4.46	-14.62	-24.33	-39.00	-51.08	-68.10	-87.55
aic$_C$	-0.91	-8.17	-14.21	-29.79	-46.01	-68.27	-89.02	-115.79	-146.08

$H_0 : \gamma \sim N(0, \sigma_\gamma^2)$. Under the null hypothesis we generate random effects with $\sigma_\gamma^2 = 0.1$.

The Pearson test divides the range of the response values into a number of bins, and compares the observed count O_k in each bin to the expected count E_k under the the hypothesized distribution for the random effect and error terms. We follow Jiang (2001) in the construction of this test. In particular, a range [0,22] is considered as likely values for the response values, and we construct for each sample size $M = \text{floor}\{(nm^2)^{0.2}\} = 3$ bins, resulting in the statistic

$$T_P = \frac{1}{nm^2} \sum_{k=1}^{3} (O_k - E_k)^2.$$

We tried to use the asymptotic distribution of this test, but did not obtain useful results for these small sample sizes, with all levels equal to zero. Instead we have used an empirical study of 10,000 simulated data sets to obtain quantiles of the null distribution of T_P. Subsequently, for each n, 1000 data sets were generated from the null model and each of two alternatives. In each setting, empirical quantiles were used for T_P and large sample quantiles for T_{OS}. To investigate the power of the tests, we generated data with true random effect $\gamma \sim t(1)$, which is a heavy tailed Cauchy distribution, and with γ coming from a mixture of normal distributions: with probability 0.1, a $N(-4, 0.1)$ distribution and with probability 0.9 a $N(4, 0.1)$ distribution. Table 2 shows that the order selection test is fairly conservative, but reaches good power under the alternative hypotheses. We reiterate that the critical values for the Pearson test are obtained from an empirical study, thus guaranteeing that they are nearly correct, while no corrections for the level were performed to obtain the power results for the order selection test. The order selection test has higher power for both alternatives. In particular for the mixture of normals case, it is able to detect the departure from normality, while with these sample sizes, the Pearson test is completely unable to do so.

7 Minimum distance methods

When replications are available, testing the fit of both error and random effects distributions becomes feasible. Initially we consider a fairly simple random effects model, and then discuss generalizations to more complex random effects and mixed models. Suppose the observations Y_{jk}, $k = 1, \ldots, n_j$, $j = 1, \ldots, n$, obey the model

$$Y_{jk} = \mu + \gamma_j + \varepsilon_{jk}, \quad k = 1, \ldots, n_j,\ j = 1, \ldots, n, \tag{9}$$

Table 2: Simulated data. Simulated rejection probabilities for testing the null hypothesis of normality of the random effect γ in model (8), for the SNP-based order selection test T_{OS} and the Pearson test T_P.

		H_0		$\gamma \sim t(1)$		$\gamma \sim$mixture	
Test	n	$\alpha = 0.10$	$\alpha = 0.05$	0.10	0.05	0.10	0.05
T_{OS}	35	0.024	0.006	0.163	0.150	0.239	0.206
T_P		0.098	0.045	0.101	0.081	0.002	0.000
T_{OS}	50	0.014	0.011	0.230	0.217	0.294	0.269
T_P		0.101	0.055	0.184	0.144	0.000	0.000
T_{OS}	100	0.020	0.010	0.336	0.329	0.362	0.355
T_P		0.103	0.052	0.220	0.196	0.000	0.000

where μ is a constant, $\gamma_1, \ldots, \gamma_n$ are i.i.d. mean 0 random variables having density g and ε_{jk}, $k = 1, \ldots, n_j$, $j = 1, \ldots, n$, are i.i.d. mean 0 random variables that are independent of $\gamma_1, \ldots, \gamma_n$ and have common density f. Of interest is testing the fit of parametric models for f and/or g.

Before proceeding to a discussion of our methodology, it is worthwhile to discuss the identifiability of model (9). Now, the model fails to be identifiable if and only if there exist distinct pairs (f_1, g_1) and (f_2, g_2) of densities that yield the same joint distribution for Y_{jk}, $k = 1, \ldots, n_j$, $j = 1, \ldots, n$. Since the Y_{jk} are i.i.d. for different j, it follows that identifiability is determined by the joint distribution of $Y_{m1}, \ldots, Y_{mn_m}$, where n_m is the largest of the n_js. In other words, the model is not identifiable if there exist distinct pairs (f_1, g_1) and (f_2, g_2) of densities such that the joint distribution of $Y_{m1}, \ldots, Y_{mn_m}$ is the same for both pairs.

Let ϕ_ε and ϕ_γ be the characteristic functions (cfs) of f and g, respectively. Reiersøl (1950) proved the remarkable result that *model (9) is identifiable when* $n_m = 2$ under the single condition that neither ϕ_ε nor ϕ_γ vanish throughout an interval. In essence this result implies that, under general conditions, both f and g can be consistently estimated in model (9) so long as the number of cases with $n_j \geq 2$ is unbounded as $n \to \infty$. This fact has been exploited in recent work by Li and Vuong (1998), Hall and Yao (2003), Delaigle et al. (2008) and Hart and Cañette (2008), all of whom propose methods for estimating f and g in model (9).

7.1 A test of fit for the error distribution

Consider the differences $\{\delta_{jkl} = Y_{jk} - Y_{jl} : 1 \leq k < l \leq n_j, j = 1, \ldots, n\}$. Obviously, $\delta_{jkl} = \varepsilon_{jk} - \varepsilon_{jl}$ for all j, k, l, and hence the differences are completely free of the random effect γ_j. Now, suppose one wishes to test the null hypothesis that f belongs to a parametric family $\mathcal{F}_0 = \{f(\cdot|\theta) : \theta \in \Theta\}$. One very straightforward way of doing so is to apply a standard goodness-of-fit test, such as the Kolmogorov-Smirnov (KS) or Cramér-von Mises (CVM), to the δ_{jkl}s to test the hypothesis that the distribution of $\varepsilon_{jk} - \varepsilon_{jl}$ is that induced by the assumption that $\varepsilon_{jk} \sim f(\cdot|\theta)$. There are at least two potential disadvantages of this approach. First of all, as argued by Rayner and Best (1989) and others, omnibus tests such as KS and CVM are often much less powerful than "directional" types of tests, such as smooth tests. A second disadvantage of basing a test on the marginal distribution of δ_{jkl} is that identifiability of f from the distribution of $\varepsilon_{jk} - \varepsilon_{jl}$ requires a fairly strong condition on the characteristic function of f. This entails that a test based on the marginal distribution of

δ_{jkl} will sometimes have very poor power. We will thus use a procedure that largely avoids the identifiability issue and also makes use of directional test statistics.

Let $\varepsilon_1, \varepsilon_2$, and ε_3 be independent and identically distributed as f, and let h be the joint density of $\varepsilon_1 - \varepsilon_2$ and $\varepsilon_1 - \varepsilon_3$. The previously mentioned result of Reiersøl (1950) implies that f is identifiable from h on the single condition that the characteristic function of f does not vanish throughout any interval. Assuming that $n_j \geq 3$ for a substantial proportion of the n_j, we may thus use information from all the pairs $(\delta_{jkl}, \delta_{jkm})$ such that $k \neq l$, $k \neq m$ and $l \neq m$ to estimate f. This is done using a variation of the minimum distance method of Wolfowitz (1957). Now, h has the form

$$h(x,y) = \int_{-\infty}^{\infty} f(z-x)f(z-y)f(z)\,dz,$$

where f is the density of ε_{jk}. Define

$$S_n = \{(j,k,l,m) : 1 \leq j \leq n, n_j \geq 3, 1 \leq k,l,m \leq n_j, l \neq k, m \neq k, l \neq m\}$$

and let $\hat{H}$ denote the empirical distribution of all pairs $(\delta_{jkl}, \delta_{jkm})$ such that $(j,k,l,m) \in S_n$. Then the parameters of a model $f(\cdot|\theta)$ may be estimated by choosing θ to maximize the Kullback-Leibler discrepancy

$$\begin{aligned} D(\theta) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \log h(x,y|\theta) d\hat{H}(x,y) \\ &= \frac{1}{\#S_n} \sum_{(j,k,l,m)\in S_n} \log h(\delta_{jkl}, \delta_{jkm}|\theta), \end{aligned}$$

where $\#S_n$ is the number of elements in S_n and

$$h(x,y|\theta) = \int_{-\infty}^{\infty} f(z-x|\theta)f(z-y|\theta)f(z|\theta)\,dz.$$

It is worth noting that $D(\theta)$ is not the likelihood of the pairs $(\delta_{jkl}, \delta_{jkm})$ since these pairs are not all independent. Rather, $D(\theta)$ measures the discrepancy of the parametric model $h(\cdot|\theta)$ from the true density h.

Our test of the null hypothesis that f is in $\mathcal{F}_0$ makes use of log-linear expansions as in Section 5.3.2. Define for $M = 1, 2, \ldots,$

$$f_M(x|\theta,\alpha) = f(x|\theta)c_M(\theta,\alpha)^{-1}\exp\left(\alpha\psi_M(x)\right),$$

and note that this model uses only one orthogonal function ψ_M and corresponding parameter α. For $M = 1, 2, \ldots$, define

$$D_M(\theta,\alpha) = \frac{1}{\#S_n} \sum_{(j,k,l,m)\in S} \log h_M(\delta_{jkl}, \delta_{jkm}|\theta,\alpha)$$

where

$$h_M(x,y|\theta,\alpha) = \int_{-\infty}^{\infty} f_M(z-x|\theta,\alpha)f_M(z-y|\theta,\alpha)f_M(z|\theta,\alpha)\,dz. \tag{10}$$

Our test procedure may be summarized as follows.

- For $M = 1, \ldots, q$, define
$$\Delta_M = \left.\frac{\partial D_M(\theta, \alpha)}{\partial \alpha}\right|_{(\theta,\alpha)=(\hat{\theta}_0,0)},$$
where $\hat{\theta}_0$ is the maximizer of $D_M(\theta, 0)$ with respect to θ. The statistic Δ_M is analogous to a score statistic. Let s_M be an estimator of the standard error of Δ_M. Then each of $\mathcal{S}_M = (\Delta_M/s_M)^2$, $M = 1, \ldots, q$, would serve as a test statistic for the null hypothesis that $f \in \mathcal{F}_0$.

- The statistics $\mathcal{S}_1, \ldots, \mathcal{S}_q$ are combined into an omnibus statistic as proposed in Hart (2008). This statistic has the form
$$T = \log\left(\sum_{j=1}^{q} j^{-2} \exp(\mathcal{S}_j/2)\right), \tag{11}$$
and the null hypothesis is rejected for large values of T.

- The null distribution of the statistic T is approximated by use of the parametric bootstrap. Independent and identically distributed random variables ε^*_{jk}, $k = 1, \ldots, n_j$, $j = 1, \ldots, n$, are generated from $f(\cdot|\hat{\theta}_0)$. Differences δ^*_{jkl} are computed from the ε^*_{jk}s, and a test statistic T^* is computed from these differences in exactly the same way T was computed from the δ_{jkl}s. This process is repeated a large number B of times, leading to bootstrap statistics $T^*_1, \ldots, T^*_B$. The null hypothesis is rejected at level α if T exceeds the $(1-\alpha)100$th percentile of the B bootstrap statistics.

7.2 Tests for the random effects distribution

If f were known, a method such as that in Section 5 could be used to test the fit of the random effects distribution. Here we do not assume that f is known. Our testing methodology requires estimates of f and g on the assumption that H_0 is true. There are at least two ways of estimating f. If $n_j \geq 3$ for most j, then, as described in Hart and Cañette (2008), we may compute a minimum distance estimate of f, an estimate that requires no parametric model. The advantage of this method is that the estimate of f is in no way influenced by fitting the parametric model $g(\cdot|\theta)$ for g. Alternatively, we may use a method in which f and $g(\cdot|\theta)$ are simultaneously estimated. An advantage of this method is that it only requires two replications for each j. The second method is the one described in this section. Specifically, we will use the minimum distance method of Hart and Cañette (2008) to estimate f and $g(\cdot|\theta)$. This method is similar to that of Beran and Millar (1994) for random coefficient regression models.

For ease of notation, let us assume that μ is known to be 0. The joint characteristic function (cf) of (Y_{jk}, Y_{jl}) $(k \neq l)$ is
$$\phi(s,t) = \phi_\gamma(s+t)\phi_\varepsilon(s)\phi_\varepsilon(t),$$
where ϕ_γ and ϕ_ε are the respective cfs of γ and ε. Defining the total number of pairs in the data as $n_{\text{total}} = \sum_{j=1}^{n} \binom{n_j}{2}$, a consistent estimator of $\phi(s,t)$ is the empirical cf
$$\hat{\phi}(s,t) = \frac{1}{n_{\text{total}}} \sum_{j=1}^{n} \sum_{k<l} \exp[isY_{jk} + itY_{jl}].$$

Now, let $\phi_\gamma(t|\theta)$ be the cf of the parametric model for g and $\hat{\phi}_\varepsilon$ be a candidate for ϕ_ε. Then we try to find θ and $\hat{\phi}_\varepsilon$ so that $\phi_\gamma(s+t|\theta)\hat{\phi}_\varepsilon(s)\hat{\phi}_\varepsilon(t)$ is a good match to $\hat{\phi}(s,t)$ for all (s,t). This, in essence, is the minimum distance method.

Let Q_f be the quantile function associated with f. The minimum distance method of Hart and Cañette (2008) produces estimates of $Q_f(u)$ at $u = (j-1/2)/q$, $j = 1,\ldots,q$. Let $\boldsymbol{Q} = (Q_1,\ldots,Q_q)$, where $Q_1 < \cdots < Q_q$ are estimates of $Q_f((j-1/2)/q)$, $j = 1,\ldots,q$. A corresponding estimate of ϕ_ε is

$$\hat{\phi}_\varepsilon(t) = \frac{1}{q}\sum_{j=1}^{q} e^{itQ_j}.$$

We propose that θ and $\boldsymbol{Q}$ be chosen to minimize

$$D(\theta,\boldsymbol{Q}) = \int\int \exp[-b^2(s^2+t^2)] \left|\hat{\phi}(s,t) - \tilde{\phi}(s,t)\right|^2 dsdt,$$

where

$$\tilde{\phi}(s,t) = \phi_\gamma(s+t|\theta)\hat{\phi}_\varepsilon(s)\hat{\phi}_\varepsilon(t).$$

Introducing the factor $\exp[-b^2(s^2+t^2)]$ into the discrepancy measure ensures integrability. The quantity b is a small positive number that plays the role of bandwidth. Indeed, $\exp[-b^2(s^2+t^2)/2]\hat{\phi}(s,t)$ is the cf of a kernel density estimate based on the observations $(\hat{Y}_{jk},\hat{Y}_{jl})$, $j = 1,\ldots,n$, $k < l$, and using bandwidth b and kernel equal to the product of Gaussian densities. A random search algorithm for determining the minimizer of D with respect to θ and $\boldsymbol{Q}$ is described in Hart and Cañette (2008).

As a test statistic, we propose $D(\hat{\theta},\widehat{\boldsymbol{Q}})$, where $\hat{\theta}$ and $\widehat{\boldsymbol{Q}}$ are the values determined to minimize D. The null distribution of the test statistic is approximated by use of the following bootstrap algorithm:

B1. Draw a random sample ε^*_{jk}, $k = 1,\ldots,n_j$, $j = 1,\ldots,n$, with replacement, from the set of quantiles $\widehat{\boldsymbol{Q}}$.

B2. Generate a random sample $\gamma^*_1,\ldots,\gamma^*_n$ from the density $g(\cdot|\hat{\theta})$.

B3. Construct bootstrap data $Y^*_{jk} = \gamma^*_j + \varepsilon^*_{jk}$, $k = 1,\ldots,n_j$, $j = 1,\ldots,n$.

B4. Compute the test statistic $D(\hat{\theta}^*,\widehat{\boldsymbol{Q}}^*)$ from the bootstrap data using all the steps used in computing $D(\hat{\theta},\widehat{\boldsymbol{Q}})$ from the original data.

B5. Repeat steps B1-B4 a large number of times and reject H_0 at level of significance α if $D(\hat{\theta},\widehat{\boldsymbol{Q}})$ exceeds the $(1-\alpha)$ percentile of all bootstrap statistics.

The choice of q, the number of quantiles, is worth some discussion. The test statistic depends on $\widehat{\boldsymbol{Q}}$ only through $\hat{\phi}_\epsilon$, and $\hat{\phi}_\epsilon$ is relatively insensitive to choice of q so long as q is sufficiently large. The only reason not to take q *very* large is computational, as the algorithm of Hart and Cañette (2008) is slower the larger q is. We have found $q = 100$ to be a good choice in practice.

7.3 Generalization to mixed models

Suppose now that we have a model of the form

$$Y_{jk} = \boldsymbol{x}_j^T\boldsymbol{\beta} + z_{jk}\gamma_j + \varepsilon_{jk}, \quad k = 1, \ldots, m,\ j = 1, \ldots, n, \tag{12}$$

where $\boldsymbol{\beta}$ is a p-vector of fixed effects and γ_j is a random effect. The index j denotes different main experimental units, while k denotes subunits within a main unit. We assume that each main unit has the same number m of subunits only to simply notation. Each main unit has a known covariate $\boldsymbol{x}_j$ and each subunit a known covariate z_{jk}. We make the following assumptions about the model:

A1. All covariate values $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ and z_{jk}, $k = 1, \ldots, n_j$, $j = 1, \ldots, n$, are fixed.

A2. The random variables ε_{jk}, $k = 1, \ldots, n_j$, $j = 1, \ldots, n$ are i.i.d. with $E(\varepsilon_{jk}) = 0$.

A3. The random effects $\gamma_1, \ldots, \gamma_n$ are i.i.d. with $E(\gamma_j) = 0$.

A4. The collections of random variables $\{\varepsilon_{jk}\}$ and $\{\gamma_j\}$ are independent of each other.

We wish to test the fit of models for the cumulative distribution functions G and/or F of γ_j and ε_{jk}, respectively. Because of the covariate z_{jk}, taking differences does not eliminate the random effect in this case, and hence we will estimate F and G simultaneously. If one wishes to test the fit of models for both F and G, we suggest that two separate tests be conducted, since then one will know which (if either) model exhibits lack of fit. Our methodology for the two cases is virtually the same and will be illustrated by testing the fit of a model for G.

We first note that the fixed effects add little difficulty to the inference of F and G. Let $\boldsymbol{X}$ be the $n \times p$ matrix with jth row equal to $\boldsymbol{x}_j^T$ and $\bar{\boldsymbol{Y}}$ be the column vector with jth element equal to $m^{-1}\sum_{k=1}^m Y_{jk}$, $j = 1, \ldots, n$. Under standard conditions, the least squares estimator $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\bar{\boldsymbol{Y}}$ is consistent for $\boldsymbol{\beta}$ as $n \to \infty$. We thus define the residuals

$$\hat{e}_{jk} = Y_{jk} - \boldsymbol{x}_j^T\widehat{\boldsymbol{\beta}}, \quad j = 1, \ldots, n,\ k = 1, \ldots, m,$$

and base our inference of F and G on these residuals.

Define $e_{jk} = Y_{jk} - \boldsymbol{x}_j^T\boldsymbol{\beta}$ and let ϕ_γ and ϕ_ε be the cfs of γ_j and ε_{jk}, respectively. Now consider the joint cf ϕ_{jkl} of (e_{jk}, e_{jl}) for $k < l$:

$$\begin{aligned}\phi_{jkl}(s,t) = E\left[\exp\left(ise_{jk} + ite_{jl}\right)\right] &= E\left[\exp\left(i\gamma_j(sz_{jk} + tz_{jl})\right)\right]\phi_\varepsilon(s)\phi_\varepsilon(t)\\ &= \phi_\gamma(sz_{jk} + tz_{jl})\phi_\varepsilon(s)\phi_\varepsilon(t).\end{aligned}$$

Averaging over all (j,k,l) yields

$$\begin{aligned}\phi_n(s,t) &= \frac{2}{nm(m-1)}\sum_{j=1}^n\sum_{k=1}^{m-1}\sum_{l=k+1}^m \phi_{jkl}(s,t)\\ &= \phi_\varepsilon(s)\phi_\varepsilon(t)\frac{2}{nm(m-1)}\sum_{j=1}^n\sum_{k=1}^{m-1}\sum_{l=k+1}^m \phi_\gamma(sz_{jk} + tz_{jl}).\end{aligned}$$

We may use the last expression as a basis for estimating ϕ_γ and ϕ_ε on the assumption that H_0 is true. Our methodology here closely parallels that in Section 7.2. First, we may estimate $\phi_n(s,t)$ by

$$\hat{\phi}_n(s,t) = \frac{2}{nm(m-1)}\sum_{j=1}^n\sum_{k<l}\exp(is\hat{e}_{jk} + it\hat{e}_{jl}).$$

Let $\phi_\gamma(t|\theta)$ be the cf of γ assuming that H_0 is true, and let $\hat{\phi}_\varepsilon$ be the cf corresponding to a finite set $\boldsymbol{Q}$ of candidate quantiles for F (as in Section 7.2). Then we may choose θ and $\boldsymbol{Q}$ to minimize

$$D(\theta, \boldsymbol{Q}) = \int\int \exp[-h^2(s^2+t^2)] \left|\hat{\phi}_n(s,t) - \tilde{\phi}_n(s,t)\right|^2 \, dsdt,$$

where

$$\tilde{\phi}_n(s,t) = \hat{\phi}_\varepsilon(s)\hat{\phi}_\varepsilon(t)\frac{2}{nm(m-1)}\sum_{j=1}^{n}\sum_{k<l}\hat{\phi}_\gamma(sz_{jk}+tz_{jl}|\theta).$$

Again, the algorithm of Hart and Cañette (2008) may be used to approximate the minimizer of D with respect to θ and $\boldsymbol{Q}$.

As a test statistic, we use $D(\hat{\theta}, \widehat{\boldsymbol{Q}})$, where $\hat{\theta}$ and $\widehat{\boldsymbol{Q}}$ are the values determined to minimize D. Virtually the same bootstrap algorithm as described in B1-B5 may be used to approximate the distribution of the test statistic. The bootstrap data take the form

$$Y_{jk}^* = \boldsymbol{x}_j^T\widehat{\boldsymbol{\beta}} + z_{jk}\gamma_j^* + \varepsilon_{jk}^*, \quad k = 1, \ldots, m, \ j = 1, \ldots, n,$$

and all the same steps used in calculating $D(\hat{\theta}, \widehat{\boldsymbol{Q}})$ from the Y_{jk}s are used in calculating $D(\hat{\theta}^*, \widehat{\boldsymbol{Q}}^*)$ from the Y_{jk}^*s.

As in the simpler model of Section 7.2, identifiability of the model is an important consideration. The results of Beran and Hall (1992) show that, under quite general conditions, both F and G can be estimated consistently in model (12). A sufficient condition for this result is that some sequence z_{jk_j}, $j = 1, \ldots, n$, represent i.i.d. draws from a distribution that has at least one of the points 0, $-\infty$ or ∞ in its support.

A generalization of model (12) is

$$Y_{jk} = \boldsymbol{x}_j^T\boldsymbol{\beta} + \boldsymbol{z}_{jk}^T\boldsymbol{\gamma}_j + \varepsilon_{jk}, \quad k = 1, \ldots, n_j, \ j = 1, \ldots, n,$$

where now $\boldsymbol{z}_{jk}$ is a column vector of r covariates and $\boldsymbol{\gamma}_j$ a column vector of r random effects. As before, ε_{jk}s are i.i.d. as f, $\boldsymbol{\gamma}_j$s are i.i.d with common r-variate density g, and ε_{jk}s are independent of $\boldsymbol{\gamma}_j$s. Tests analogous to those described for the case $r = 1$ may be constructed. Beran and Millar (1994) describe minimum distance methodology that could be used to estimate $(g(\boldsymbol{\gamma}|\theta), f)$ or $(g, f(\varepsilon|\theta))$, depending upon which goodness-of-fit hypothesis is of interest. They also provide conditions under which both g and f may be consistently estimated, a result that would ensure consistency of tests as proposed above.

8 A microarray example

Here we consider microarray data collected by Robert Chapkin and coworkers of his at Texas A&M University. The data we analyze are only part of a much larger data set, but provide a good example of methodology described in Sections 7.1 and 7.2. The data considered are Y_{jk}, $j = 1, \ldots, 8038$, $k = 1, \ldots, 5$, where j indexes genes, k indexes different rats, and Y_{jk} is the logarithm of the expression level for gene j and rat k. The five rats from which these data were collected were all subjected to the same treatment.

We assume the following model for the data:

$$Y_{jk} = R_k + \gamma_j + \varepsilon_{jk}, \quad j = 1, \ldots, 8038, \ k = 1, \ldots, 5,$$

Table 3: *Score statistics for testing the hypothesis that the rat data error distribution is normal. The statistic corresponding to M is based on the orthogonal function* $\cos(\pi M x)$. *This results in the value* $T = 469$ *for the statistic* T *in (11).*

M	1	2	3	4	5	6	7	8	9	10
Statistic	0.380	940	0.00119	99.3	1.23	0.0154	2.47	20.2	3.13	48.3

where R_k represents a rat effect, γ_j a gene effect, and ε_{jk} measurement error. Our assumptions about the γ_js and ε_{jk}s are the same as those we made for model (9).

We will test normality of each of g and f, the densities of γ_j and ε_{jk}, respectively. The first step in the analysis is to estimate rat effects by computing the mean of all data for each rat. Defining

$$Z_{jk} = Y_{jk} - \frac{1}{8038}\sum_{i=1}^{8038} Y_{ik}, \quad j = 1, \ldots, 8038, \ k = 1, \ldots, 5,$$

we may say that, to a good approximation, the Z_{jk}s follow model (9) since each rat effect is estimated by the mean of over 8000 observations. We will thus apply the methods of Sections 7.1 and 7.2 to test the two hypotheses of interest.

To test the null hypothesis that f is $N(0, \sigma^2)$, we compute differences of the form $\delta_{jkl} = Z_{jk} - Z_{jl}$. For computational expediency, we compute only one pair of differences for each gene. This is done by randomly selecting, for each gene j, three rats, k, l and m, say, and computing the differences $(\delta_{jkl}, \delta_{jkm})$. In applying a test as described in Section 7.1, we use $\psi_M(x) = \cos(\pi M \Phi(x/\sigma))$, where Φ is the standard normal cumulative distribution function. Defining h_M as in (10), we may then compute a test statistic based on the likelihood

$$L_M(\sigma, \alpha) = \prod_{j=1}^{8038} h(\delta_{jk(j)l(j)}, \delta_{jk(j)m(j)} | \sigma, \alpha),$$

where $(k(j), l(j), m(j))$ denote the three rats randomly selected for gene j. We then use standard methods to define a score statistic from the likelihood L_M for each of $M = 1, \ldots, 10$. These ten score statistics and the test statistic T (defined by (11)) are given in Table 3. The bootstrap algorithm of Section 7.1 was applied with $B = 500$. The largest bootstrap statistic was 3.95, indicating that the observed value of T is highly significant.

Examining a plot of the data reveals that the significance of T is not surprising. Figure 2 is a scatterplot of the pairs $(\delta_{jk(j)l(j)}, \delta_{jk(j)m(j)})$, which should follow a bivariate normal distribution if the errors are in fact normal. Instead the plot has an interesting pattern in which "arms" radiate from a central scatter. These arms are due to outlying differences, which in turn are due to outlying errors ε_{jk}. The minimum distance algorithm of Hart and Cañette (2008) was applied to the pairs $(\delta_{jk(j)l(j)}, \delta_{jk(j)m(j)})$ to obtain estimates $\widehat{Q}_1, \ldots, \widehat{Q}_{100}$ of the quantiles $Q_f((j - 1/2)/100)$, $j = 1, \ldots, 100$, respectively. A kernel estimate of f of the form $\hat{f}(x) = (100b)^{-1} \sum_{j=1}^{100} K((x - \widehat{Q}_j)/b)$, with K equal to a standard normal density, was then computed. The resulting estimate (scaled to have variance 1) is shown in Figure 3 along with a standard normal density. The error density is apparently leptokurtic with a longer right than left tail.

We next test for normality of the gene effect. We use the procedure described in Section 7.2 except that $\widehat{\boldsymbol{Q}}$ is taken to be the nonparametric estimate obtained from the analysis described

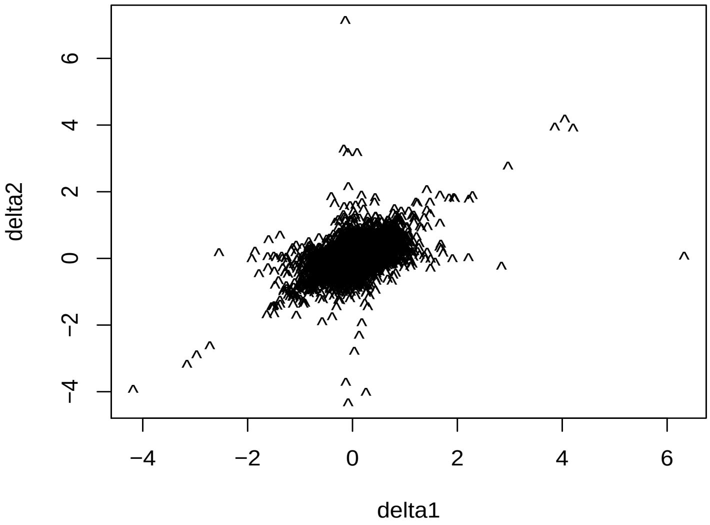

Figure 2: *Scatterplot of differences for the rat data.*

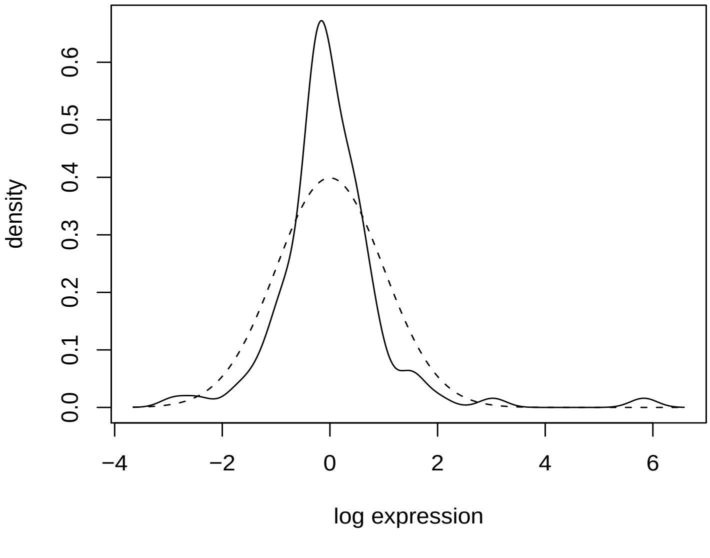

Figure 3: *Estimate of error density for the rat data. The solid and dashed lines are the density estimate and a standard normal density, respectively.*

immediately above and the estimate of the gene effect variance, $\sigma^2_{\gamma,0}$ is $\hat{\sigma}^2_\gamma$, the minimizer of $D(\sigma^2_\gamma, \widehat{\boldsymbol{Q}})$ with respect to σ^2_γ. Our test statistic is then $D(\hat{\sigma}^2_\gamma, \widehat{\boldsymbol{Q}})$. To speed up computations, only two of five

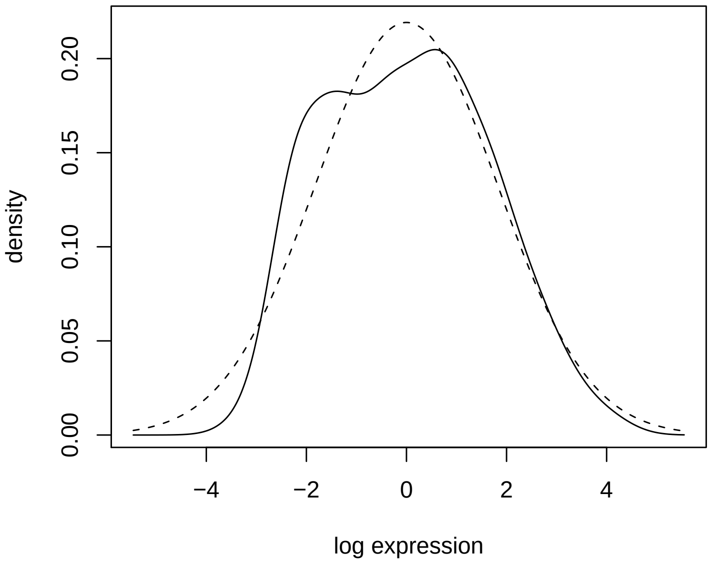

Figure 4: *Estimates of gene effect density for the rat data. The solid line is a nonparametric estimate of the density, and the dashed line is the normal density obtained on the assumption that gene effects are normally distributed.*

rats were used for a given gene, with the two rats being randomly selected for each gene. The value of the test statistic was 0.059, and the largest of five hundred bootstrap statistics was 0.000776, providing convincing evidence that the gene effects are not normally distributed. Applying the algorithm of Hart and Cañette (2008), led to the nonparametric estimate of the gene effect density shown in Figure 4.

9 Discussion

The use of flexible distributions in mixed effects models is relatively new. In addition to having good estimation methods, it is desirable to be able to test whether a more involved distributional model is really needed. The tests proposed in this paper are useful for that purpose. We construct smooth omnibus tests, which are well-studied and known to have good power properties in the context of linear models. Our proposed minimum distance tests are designed for mixed models with just a few replicates. A particular advantage of such tests is that they automatically provide an estimate of the underlying distribution. In case of rejection of the null hypothesis, they may suggest missing fixed effects in the model, for example in case of multimodality of the random effects distribution.

In generalized linear mixed models (GLMM), the conjugate distribution might be used as a random effect distribution instead of the normal distribution (Lee and Nelder, 1996). In such GLMM a closed form version of the likelihood is usually not available, which asks for alternative testing procedures that are not likelihood-based. An interesting direction for future research is a development of score-based goodness-of-fit tests that would work in combination with generalized estimating equations.

Another interesting direction is the development of Bayesian or frequentist-Bayesian tests to accompany Bayesian estimation methods in mixed models with flexible distributions for random effects.

Acknowledgements

Part of this research has been performed while G. Claeskens was visiting the Isaac Newton Institute at Cambridge University, U.K. The work of Professor Hart is partially supported by NSF Grant DM-0604801 and by Award No. KUS-C1-016-04, made by King Abdullah University of Science and Technology (KAUST). The authors wish to thank W. Ghidey and M. Davidian for providing some software. They also thank the reviewers for their constructive remarks.

References

Aerts, M., Claeskens, G., and Hart, J. D. (1999). Testing the fit of a parametric function. *Journal of the American Statistical Association*, 94:869–879.

Aerts, M., Claeskens, G., and Hart, J. D. (2000). Testing lack of fit in multiple regression. *Biometrika*, 87:405–424.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. and Csáki, F., editors, *Second International Symposium on Information Theory*, pages 267–281. Akadémiai Kiadó, Budapest.

Akaike, H. (1978). A new look at the Bayes procedure. *Biometrika*, 65:53–59.

Beran, R. and Hall, P. (1992). Estimating coefficient distributions in random coefficient regressions. *Ann. Statist.*, 20(4):1970–1984.

Beran, R. and Millar, P. (1994). Minimum distance estimation in random coefficient regression models. *Ann. Statist.*, 22(4):1976–1992.

Calvin, J. A. and Sedransk, J. (1991). Bayesian and frequentist predictive inference for the patterns of care studies. *Journal of the American Statistical Association*, 86(413):36–48.

Chen, J., Zhang, D., and Davidian, M. (2002). A Monte Carlo EM algorithm for generalized linear mixed models with flexible random effects distribution. *Biostatistics*, 3(3):347–360.

Claeskens, G. and Hjort, N. L. (2004). Goodness of fit via nonparametric likelihood ratios. *Scandinavian Journal of Statistics*, 31:487–513.

Claeskens, G. and Hjort, N. L. (2008). *Model Selection and Model Averaging.* Cambridge University Press, Cambridge.

Delaigle, A., Hall, P., and Meister, A. (2008). On deconvolution with repeated measurements. *Ann. Statist.*, 36(2):665–685.

Eubank, R. L. and Hart, J. D. (1992). Testing goodness-of-fit in regression via order selection criteria. *The Annals of Statistics*, 20:1412–1425.

Gallant, A. R. and Nychka, D. W. (1987). Semi-nonparametric maximum likelihood estimation. *Econometrica*, 55(2):363–390.

Ghidey, W., Lesaffre, E., and Eilers, P. (2004). Smooth random effects distribution in a linear mixed model. *Biometrics*, 60(4):945–953.

Ghidey, W., Lesaffre, E., and Verbeke, G. (2008). A comparison of methods for determining the random effects distribution of a linear mixed model. *Statistical Methods in Medical Research*, –(–):to appear.

Hall, P. and Yao, Q. (2003). Inference in components of variance models with low replication. *Ann. Statist.*, 31(2):414–441.

Hannan, E. J. and Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society, Series B*, 41:190–195.

Hart, J. (2008). Frequentist-Bayes lack-of-fit tests based on Laplace approximations. *Journal of Statistical Theory and Practice*, to appear.

Hart, J. and Cañette, I. (2008). Nonparametric estimation of distributions in a large-p, small-n setting. Technical report, Texas A&M University.

Hart, J. D. (1997). *Nonparametric Smoothing and Lack-of-fit Tests.* Springer-Verlag, New York.

Hwang, Y.-T. and Wei, P. F. (2006). A novel method for testing normality in a mixed model of a nested classification. *Comput. Statist. Data Anal.*, 51(2):1163–1183.

Jacqmin-Gadda, H., Sibillot, S., Proust, C., Molina, J.-M., and Thiébaut, R. (2007). Robustness of the linear mixed model to misspecified error distribution. *Comput. Stat. Data Anal.*, 51(10):5142–5154.

Jiang, J. (2001). Goodness-of-fit tests for mixed model diagnostics. *Ann. Statist.*, 29(4):1137–1164.

Komárek, A. and Lesaffre, E. (2008). Generalized linear mixed model with a penalized gaussian mixture as random effects distribution. *Computational Statistics and Data Analysis*, 52(7):3441–3458.

Lange, N. and Ryan, L. (1989). Assessing normality in random effects models. *Ann. Statist.*, 17(2):624–642.

Ledwina, T. (1994). Data-driven version of Neyman's smooth test of fit. *Journal of the American Statistical Association*, 89:1000–1005.

Lee, Y. and Nelder, J. A. (1996). Hierarchical generalized linear models. *J. Roy. Statist. Soc. Ser. B*, 58(4):619–678. With discussion.

Li, T. and Vuong, Q. (1998). Nonparametric estimation of the measurement error model using multiple indicators. *J. Multivariate Anal.*, 65:139–165.

Lombardía, M. J. and Sperlich, S. (2008). Semiparametric inference in generalized mixed effects models. *Journal of the Royal Statistical Society, Series B*, 70(5):913–930.

Naik, P. A., Shi, P., and Tsai, C.-L. (2007). Extending the Akaike information criterion to mixture regression models. *Journal of the American Statistical Association*, 102(477):244–254.

Park, T. and Lee, S.-Y. (2004). Model diagnostic plots for repeated measures data. *Biom. J.*, 46(4):441–452.

Proust, C. and Jacqmin-Gadda, H. (2005). Estimation of linear mixed models with a mixture of distribution for the random effects. *Computer Methods and Programs in Biomedicine*, 78:165–173.

Rayner, J. and Best, D. (1989). *Smooth Tests of Goodness of Fit.* Oxford University Press, New York.

Reiersøl, O. (1950). Identifiability of a linear relation between variables which are subject to error. *Econometrica*, 18:375–389.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464.

Severini, T. A. (2000). *Likelihood methods in statistics.* Oxford University Press.

Shen, W. and Louis, T. A. (1999). Empirical Bayes estimation via the smoothing by roughening approach. *J. Comput. Graph. Statist.*, 8(4):800–823.

Vaida, F. and Blanchard, S. (2005). Conditional akaike information for mixed-effects models. *Biometrika*, 92:351–370.

Verbeke, G. and Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*, 91(433):217–221.

Wolfowitz, J. (1957). The minimum distance method. *Ann. Math. Statist.*, 28:75–88.

Zhang, D. and Davidian, M. (2001). Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics*, 57(3):795–802.