Using the Posterior Predictive Distribution as a Diagnostic Tool for Mixed Models

Matthew Kramer[1]
[1]USDA/ARS/BCS, Bldg. 005, Room 130, 10300 Baltimore Ave., Beltsville, MD 20705, matt.kramer@ars.usda.gov

Abstract

The posterior predictive distribution (the distribution of data simulated from a model) has been used to flag model-data discrepancies in the Bayesian literature, and several approaches have been developed. The approach taken here differs from the others both conceptually and as realized. It works by comparing the "distance" between the data and a fitted model (as represented by pseudo-data simulated from this model) with "distance" within this model. The distance within the model is calculated by generating pseudo-data from it, using each set of these pseudo-data to re-estimate the model, and then generating pseudo-data from them, matching the way the original data are used to generate pseudo-data. "Distances" are calculated as the log of sums-of-squares between two sets of ranked observations (original data and a set of pseudo-data, or two sets of pseudo-data), and the test from comparing a mean distance (original data to pseudo-data) to a distribution of mean distances (pseudo-data to pseudo-data; this becomes the reference statistical distribution). The power of this method compares favorably with those of standard methods, e.g. *t*-tests, but it is more general since it can be used for most models in the GLMM framework, whether estimated using traditional or Bayesian methods. A new kind of plot (centipede plot), where the distribution of the ranked pseudo-data is compared to the original data at each ranked datum, is useful for determining the region of the data where the model fails, and should be a first step. For some model-data mismatches the centipede plot is more useful, others also require calculation of the test statistic. Neither method was helpful for diagnosing a missing interaction term in a complete block design with normally distributed data.

Key Words: mixed models, diagnostics, GLMM

1. Introduction

Linear mixed models (LMM's), and in recent years, generalized linear mixed models (GLMM's), are finding ever broader applications. In agriculture, they are used from modeling genomic data to data taken on an ecological scale, as well as their traditional role in designs using blocking (Gbur et al. 2012). Diagnostics for these models, especially for GLMM's, has lagged behind the developments in the software used to estimate them. Thus, currently we can relatively easily estimate model parameters and their standard errors and find the "best" model from a pool of candidate models, say using AIC; the first part of diagnostics (model comparisons). However, there is no guarantee that the best-fitting model is also the appropriate one, and with GLMM's there are many components (fixed and random independent variables, underlying data distribution, correlated errors, etc.), and thus many ways that the model can be 'wrong'. As an aside, we follow Gelman (2007) in assuming that all candidate models are wrong in some way and, had we sufficient data and the right diagnostics, this could be demonstrated. The goal is to find a reasonable model that produces data like those collected, appears to adequately represent the sort of processes we imagine to produce the observed data (i.e. is 'functionally' correct), and is usable for predictions and inference. There may not be such a model among the candidate models, thus the second part of diagnostics (model criticism) is to determine if the best candidate model fills these requirements. If not, new models need to be explored. For many statistical models, it is not easy to reliably determine whether the best candidate model matches the data collected, which requires some kind of goodness-of-fit check. Also, some kinds of mismatches are easier to detect than others, and all are easier to detect with larger sample sizes. In the Bayesian literature, diagnostics informing on data-model mismatches makes use of the posterior predictive distribution, essentially asking if discrepancies exist between the actual data and pseudo-data generated from the candidate model.

This has been implemented in various ways. Perhaps the simplest is to rank the data and individually rank the generated sets of pseudo-data, and see whether the lowest (highest) observation from the original data lies in the distribution of lowest (highest) observations of the ranked pseudo-data sets (Gelman, et al. 2013), an expanded graphical version (the 'centipede plot') is discussed below. This is an example of a broader technique, termed posterior predictive *p*-values (e.g. Meng 1994), where one compares a parameter set from fitting the model to the original data with parameter sets fit to pseudo-data generated from the model through a test statistic. The idea is that a quantity, similar to the *p*-value used in frequentist statistics, is calculated using $\Pr\{T(y^p) \geq (T(y) \mid y, H_0\}$, where $T(y)$ is a test statistic using the original data, y, y^p is the sampling distribution of the pseudo-data generated from the model represented as H_0; that is, y^p is conditioned on both the original data and the model selected (selected, in part, using the original data). For the model of fitting a normal distribution to data, the test statistic might use the mean and variance estimated using maximum likelihood. There can be computational problems implementing this method, in addition to deciding on the right test statistic(s); it becomes more difficult with increased model complexity, and there are also unresolved theoretical issues, e.g. those discussed by Bayarri and Berger (1998).

Other approaches have been advocated. Gelman, et al. (2013) suggest an omnibus χ^2 like measure, based on $\chi^2_i = (y_i - E(y^p_i))/\text{var}(y^p_i)$. Informal investigations of this measure by us, for example to check whether a mismatch was apparent if data generated from a Poisson distribution was fit with a normal distribution, were not promising. Note that generating data from a one-parameter distribution, then fitting them with a two-parameter distribution is a mismatch that is difficult to catch using any general single statistic, though the mismatch was obvious using a centipede plot, discussed below. The

minimum posterior predictive loss approach (Gelfand and Ghosh 1998), an extension of the L-criterion, is similar to the posterior predictive *p*-values in that estimated parameter values are used explicitly in the calculations. Gelfand and Ghosh (1998) discuss use of a squared error loss, the approach used in this paper, as a measure of fit, and specifically address GLMMs. Also mentioned is the importance of a penalty term for over-fit models. The method developed in this paper does not have an explicit penalty term, though over-fitting may be obvious because the fit is 'too' good, briefly mentioned below. The conditional predictive ordinate approach (Gelfand 1996) uses cross-validation (a leave-one-out) approach, so avoids the issues when re-using data. We do not have experience with this method.

The methods developed in this paper employ ranking. Ranking the data takes care of several problems. It avoids the need to look at groups of observations individually, rather the entire data set can be visualized at once. Also, visualizing data from different kinds of models (e.g. regression, ANOVA, GLMM, etc.) is essentially the same. Weaknesses of the model can be easily traced back to regions of the data, such as the model having difficulty with low or high values, this may not be obvious using other available diagnostic methods, especially for more complicated models that include random effects. Ranking also takes care of a more subtle issue, what the original data are compared to. When performing model criticism, we want to know if the model could have generated data similar to those collected. The most logical way is to compare the original data to data generated by the model, rather than by comparing original data to some parameter (or function of parameters) from the model. However, that requires that both data sets be organized in a similar way, so that like observations are being compared. While ranking the observations does not guarantee that like observations are being compared, if the model produces data similar to the original data, ranking should at least roughly align similar observations from the two data sets. Since the model may be missing predictor variables that could provide for a better alignment (i.e. within a group, generated pseudo-data may by i.i.d., so how should the generated data set be aligned with the original data?), ranking provides a viable method for determining how to pair observations from the two data sets without requiring more sophisticated statistical models.

The paper is organized as follows. An example of the posterior prediction check for lowest (highest) observations is given, then generalized to a graphic for all original data observations. A new statistic is described which allows one to calculate the probability the model is likely to have given rise to the actual data, based on calculating "distances" between pseudo-data (generated from a model) and the actual data; this can be used as a 'test', for example, if the probability is less than 0.05. Two examples, using published data from a field trial and laboratory mosquito data, analyzed within a GLMM framework, are given. Results from additional simulated data sets, from known models but fit with misspecified models are also shown. While this method (labeled PDT for "predictive distribution test") works both for models developed using frequentist and Bayesian approaches, it is perhaps better suited to the former because it relies on re-estimating models using pseudo-data, and model estimation takes longer using a Bayesian approach. The scope and number of data sets examined using this method is limited, and theoretical justification poorly developed. However, it is promising in that the PDT worked well for most data sets examined (the last example is one where it does not work), and would conceivably work as well on a much broader set of models than those used here. It has the distinct advantage of not requiring complicated coding, other than to simulate data from models and to store the results.

2. Example 1: posterior predictive distribution

A sample data set of 50 draws from a $N(0,1)$ distribution are taken. The model is a simple two parameter one, a mean, a variance, with the assumption that the parent population is normally distributed. Bayesian estimates of the two parameters, using rjags (Plummer 2013) in the R software (R Core Team 2014), yielded posterior density estimates of the two parameters given in Fig. 1, top right panels. One posterior distribution check advised by Gelman et al. (2013) is to determine how likely the extreme values (e.g. lowest) in the actual data would be generated by the model. For this, the data set was ranked and the lowest observation identified (Fig. 1, top left panel). Joint samples from the posterior distributions of means and standard deviations were taken, and from each 50 pseudo-data were generated. Each pseudo-data set was ranked, and the lowest value of each used to plot the histogram in Fig. 1, lower left panel, with the original lowest value indicated (blue dot). This original lowest value is squarely in the middle of the distribution of pseudo-data lowest values, indicating that the value of the lowest observation from the actual data is similar to those generated from this simple model, that is, the model and original data are consistent for lowest values. One could do this for every ranked actual observation, lowest to highest, producing Fig. 1, bottom right panel. In this panel, the distribution of the pseudo-data at each rank are depicted in green (mean and 95% credible interval), the blue dots are the superimposed ranked original data. This kind of plot is useful to see if the model misses the data at any of the rankings, and will be referred to as a "centipede plot", due to its obvious resemblance to that arthropod class. Since all the ranked original values lie well within the 95% credible interval of their respective rankings, based on this diagnostic the model used is consistent with these original data.

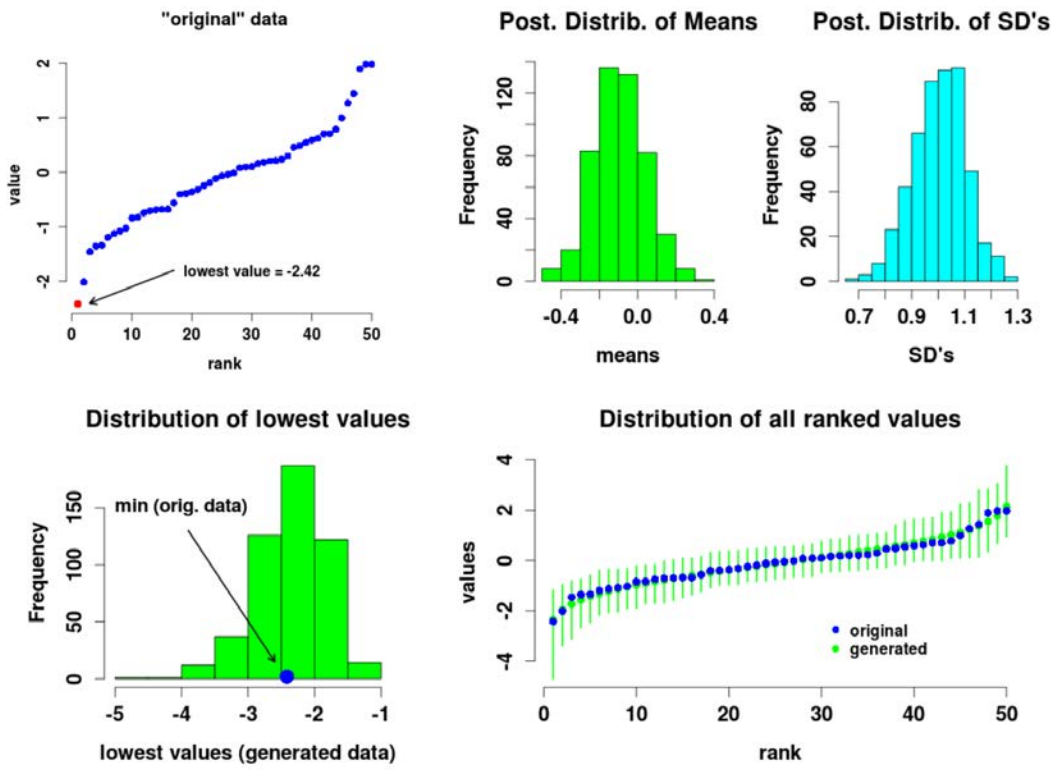

Figure 1: Top left panel: ranked "original" data generated from a $N(0, 1)$ distribution. Top right panels: distributions of posterior estimates for means and standard deviations (every 100th sample in the chain following the burn-in period). Bottom left panel: distribution of lowest values from pseudo-data generated from a normal distribution with mean and standard deviation sampled from the posterior distribution of parameter estimates. Bottom right panel: depiction of all rankings from the pseudo-data, from lowest to highest, with a 95% credible interval on each ranking (green vertical bar) and mean (green dot), original data superimposed with a blue dot.

3. Posterior Distribution Test (PDT)

While the visual inspection using a centipede plot is useful (especially for determining where in the data set the model fails, demonstrated below), a statistic that indicates how likely the candidate model would have generated the observed data is also needed. Toward that goal, a distance measure is proposed based on comparing pseudo-data generated by the model to the original data with pseudo-data generated by the model to other pseudo-data generated by the model; in short a model-data distance compared to a model-model distance, where pseudo-data is a proxy for the model. After investigating a number of ways to do this, the one described below was best in that it produces a test statistic that works in both the frequentist and Bayesian paradigms, works for all models examined (various GLMM's), produces the expected results when the true model is known (i.e. when a known model was used to generate the "original" data), that is, the nominal α value is preserved, and the method has reasonably good power to detect discrepancies between the model and the data, especially considering that the PDT can be used for most parametric models and statistical distributions. The key is that the within-model distances are computed the same way as the data-model distance.

Let $\mathbf{y}$ be the original data vector following ranking. Let $\mathbf{y}^p1, \ldots, \mathbf{y}^pn$ be vectors of ranked pseudo-data generated by the candidate model fit to $\mathbf{y}$, each the same length as $\mathbf{y}$, that is, for each observation in $\mathbf{y}$, predict a new observation. For $i = 1, \ldots, n$, calculate log $(\Sigma\ (\mathbf{y} - \mathbf{y}^pi)^2)$, this is the log of a sum-of-squares. Take the mean (m) of these values. The log transformation helps create a symmetric distribution (so the mean is a meaningful quantity), and using a sufficiently large number for n (examples in this paper were mostly done with $n = 300$) ensures that one has a good representation of possible values generated by the model. The value, m, represents the distance between the original data and the candidate model.

A similar procedure is needed for the within-model comparisons, the result will serve as a reference distribution for deciding whether m is excessively large. It is constructed in the following manner. For $i = 1, \ldots, n$, create a vector of pseudo-data (these can be the same vectors used in the comparison with the original data). For each one ($\mathbf{y}^pi$), fit the same model (i.e. the candidate model chosen based on the original data; the model form is the same but the parameters are re-estimated) and from it generate $j = 1, \ldots, n$ additional sets of pseudo-data, denoted as $\mathbf{y}^{pp}ij$, the double superscript indicating that this vector of pseudo-data is twice removed from the original data set. For each i, following ranking, calculate log $(\Sigma\ (\mathbf{y}^pi - \mathbf{y}^{pp}ij)^2)$, $j = 1, \ldots, n$. Take the mean of the j logs of sums-of-squares. Over all i's, this operation produces n means (a mean for each initial vector of pseudo-data, $\mathbf{y}^pi$). Locate the 95th quantile of the n means and check if m exceeds it. For $n = 300$, this requires 300 + 1 = 301 model estimations, which generally does not take long for frequentist models but can be time consuming in the Bayesian framework. Generating pseudo-data from a model and calculating log $(\Sigma\ (\mathbf{y}^pi - \mathbf{y}^{pp}ij)^2)$ takes little computer time.

4. Example 2: PDT example for normal distribution and power

To give a simple illustration of the PDT, a data set of 30 observations was generated from a $N(1.5, 1)$ and fit to a $N(1, 1)$ model, i.e. the data and model purposely do not match. Using the method described above and $n = 300$, m was calculated as 3.59. For the within model comparison, a histogram of the 300 means is given in Fig. 2, with the 95% quantile calculated as 2.47, less than $m = 3.59$. Thus the test statistic finds the probability that the model could have generated the data is extremely low (actually, none of the within model means was even close to m).

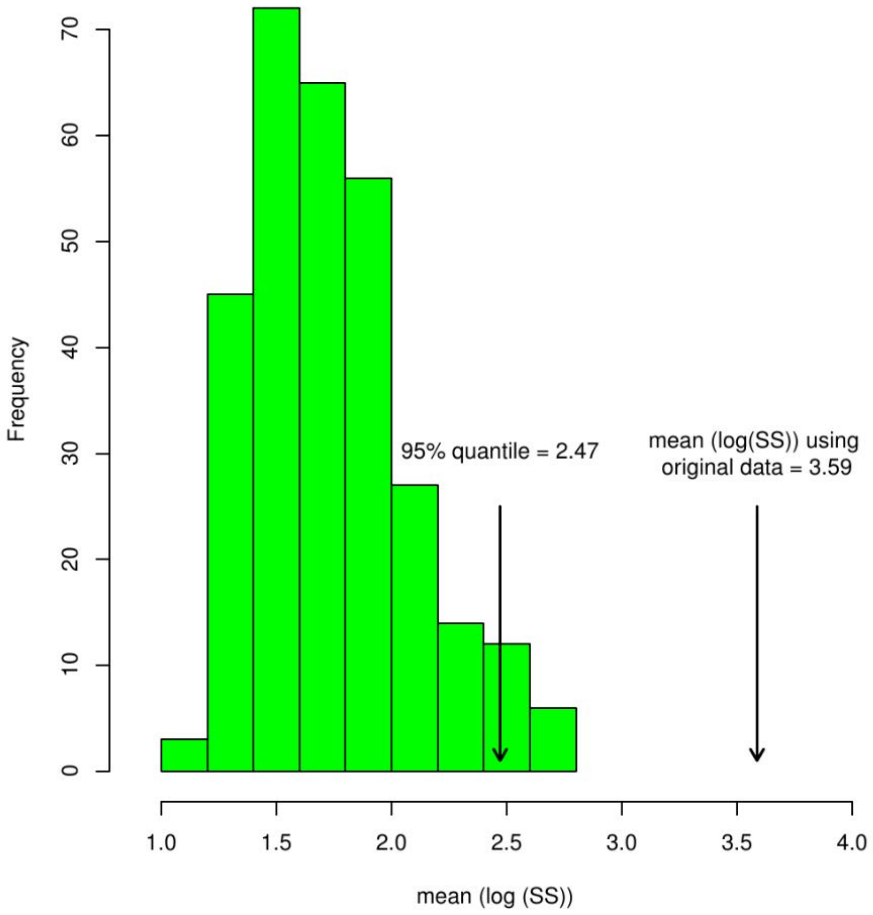

Figure 2: Distribution of the means of within-model comparisons where the model pseudo-data was generated from a $N(1, 1)$ distribution. The mean of the data-model comparisons is 3.59, indicated by an arrow on the right, and well above 2.47, the 95% quantile from the means of within-model comparisons, indicating that data and model do not match.

Power for the PDT for normally distributed data generated in a similar way to the example just described (data from $N(1.5, 1)$, model from $N(1, 1)$) was calculated empirically for sample sizes ranging from 3 to 30, using both frequentist and Bayesian methods. Results are given in Fig. 3, where the left panel gives the power under the Bayesian and the right under the frequentist paradigm. A reference power curve for a *t*-test is also plotted. The *t*-test is more powerful, not surprising because, unlike the method developed in this paper, it assumes the underlying distribution of the data is known to be normal. Note that the problems for very small samples sizes using the frequentist estimates, giving unrealistically high power, are not seen using Bayesian estimates. As the two parameters for this distribution are easy to calculate in both paradigms, there is little difference in the power curves. The curve for the frequentist estimates was calculated with fewer samples, $n = 300$ for frequentist versus 1000 for Bayesian, so are less smooth.

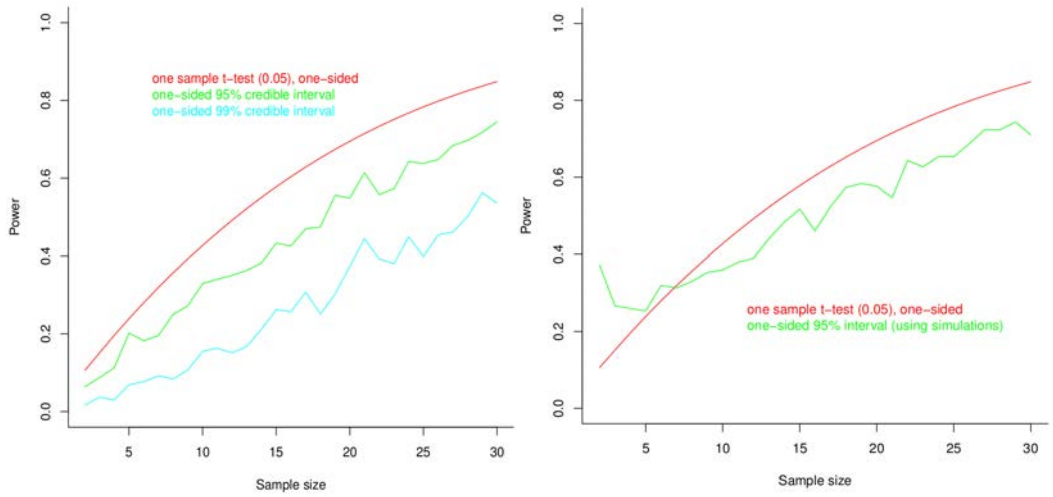

Figure 3: Empirical power curves for lack-of-fit for the PDT described in this paper, where data are generated from a $N(1.5, 1)$ distribution and fit assuming the distribution is $N(1, 1)$. Parameter estimates were made using Bayesian methods (left) and frequentist methods (right). The reference power curve for a *t*-test is given in red.

5. Example 3: linear mixed model

The data for this example were made available in the R agridat package (Wright 2013), and their analysis first described in Shafii and Price (1998). The dependent variable is rapeseed yield, the independent variables year (1987-1989), location (14 levels), genotype (6 levels), with three replications per 3-factor combination. Rather than use the Shafii and Price (1998) analysis, where the data were used to develop an AMMI model, two models were fit, a LM (linear model), and a LMM (linear mixed model). Zero values were removed, leaving 644 observations, and yield was transformed by raising it to the 0.35 power (this removed the positive relationship between the mean and variance). The LMM was fit using the R lme4 package (Bates et al. 2014), with random effects location, location by genotype, and location by year; other variables were considered fixed (including all two-way interactions of fixed effects).

First, a centipede plot was constructed for the LM (Fig. 4, left panel). Since there are many observations, only a sample of 50 are shown, and the distribution at each rank is depicted by box and whiskers symbols, rather than with a vertical line giving the 95% confidence interval. With the data spread out over such a large range, it is difficult to see where the model fails from this plot; a more informative plot can be constructed by subtracting out the data value at each rank, as in Fig. 4, right panel. It appears that the model underestimates values for low yields and some high yields, and overestimates them for average yields. The results from using the PDT are given in Fig. 5, the data-model mean is greater than the distribution of model-model means; it is unlikely that the LM could have generated this data set.

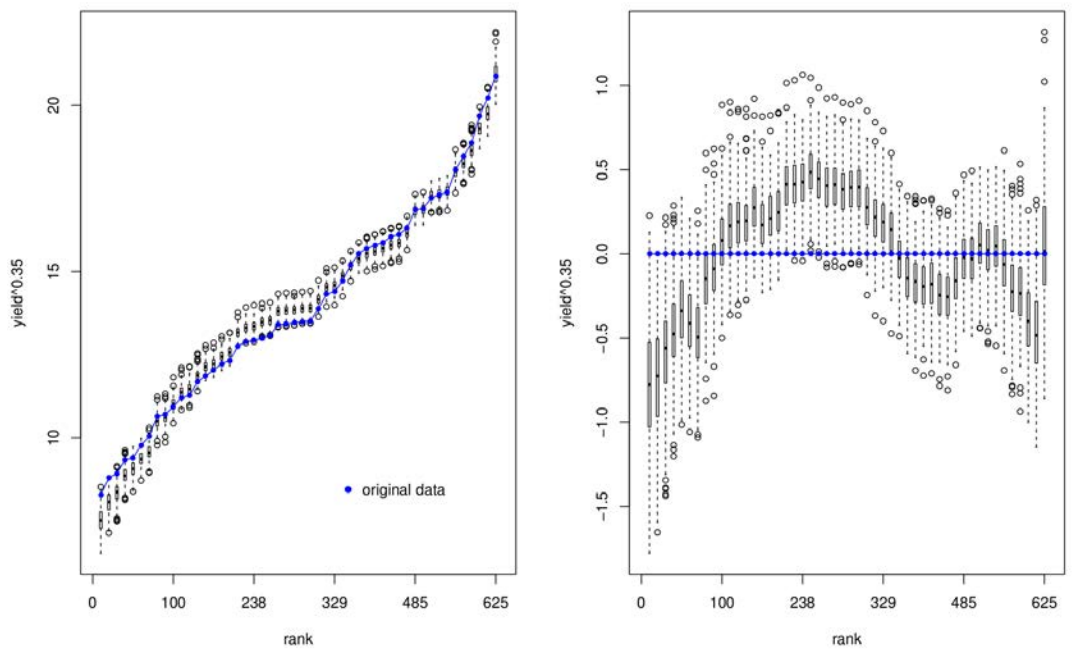

Figure 4: Centipede plots for the Shafii and Price data set, fit with a linear model. Left panel: random selection of 50 of the 644 observations used in modeling. Right panel: same sample of 50 observations, but with the observation value subtracted out (thus, the depicted "original data", connected blue dots, are all values of zero).

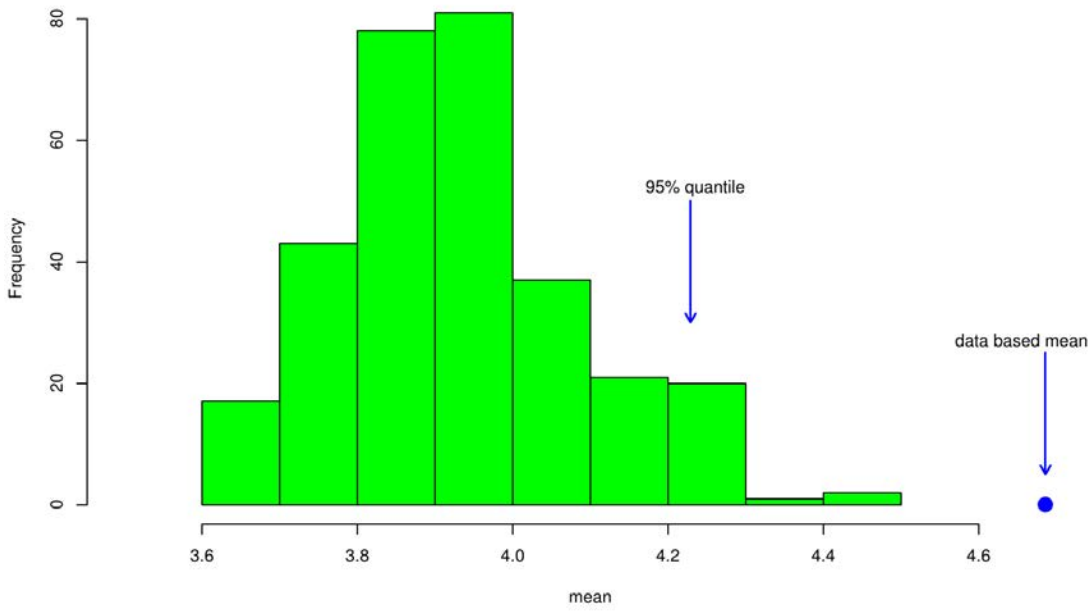

Figure 5: Results from using the PDT to assess the probability that LM could produce these data. Since the mean from the model-data comparison is far larger (right-most arrow) than the 95% quantile from the distribution of means of model-model comparisons, it is unlikely that this data set could have been produced by this model.

The LMM model appears to be more reasonable, based on the PDT (Figs. 6 and 7). The basic difference is that by relabeling location and its interactions with other independent variables as random effects, considerably more variability is introduced into the generated pseudo-data; the ranked data generally lie between the first and third quartiles of the distribution of pseudo-data at each rank. Note that levels of random effects are simulated as well, not held fixed, when data are generated. Models estimated using the lme4 package, when used to generate data (using the simulate function in R), have an option that allows one to hold the random effects fixed when generating data from the model; this would result in generated data with much lower variability, similar to that coming from a fixed model.

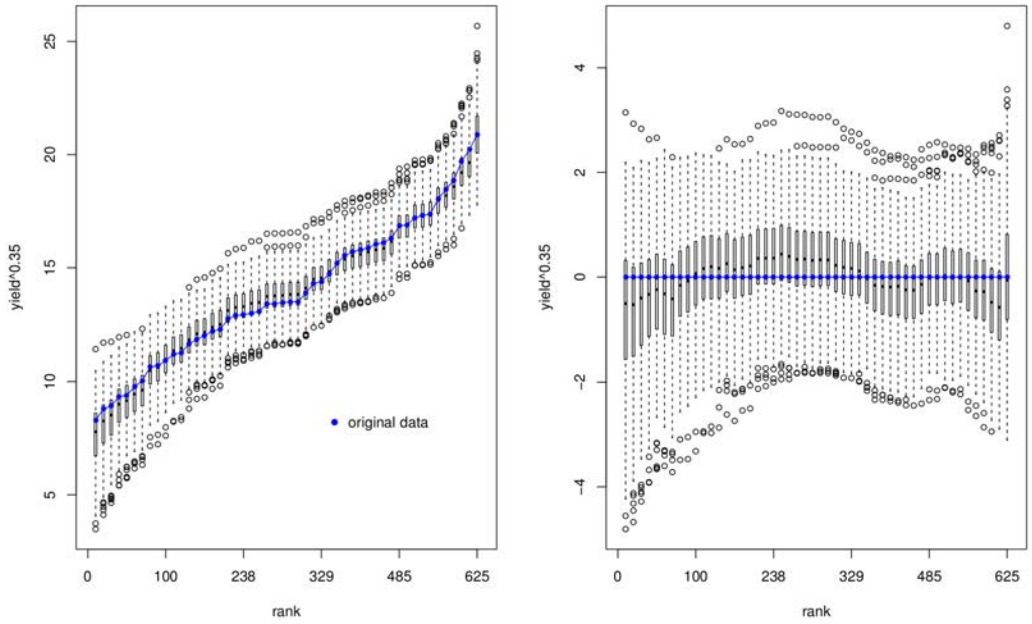

Figure 6: Centipede plots for the Shafii and Price data set, fit with a linear mixed model. Left panel: random selection of 50 of the 644 observations used in modeling. Right panel: same sample of 50 observations, but with the observation value subtracted out (thus, the depicted "original data" are all values of zero).

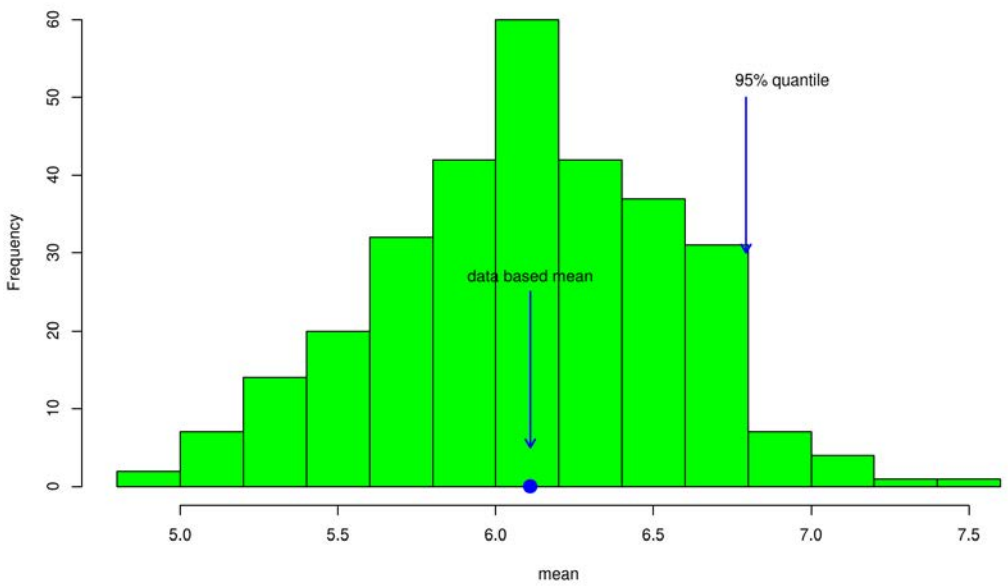

Figure 7: Results from using the PDT to assess the probability that the LMM could produce these data. The mean from the model-data comparison is well within (arrow in center of plot) the distribution of means of model-model comparisons indicating that this data set is similar to generated data sets produced by this model.

6. Example 4: mosquito data from wind tunnel experiments

A number of experiments have been conducted on mosquitoes in wind tunnels aimed at understanding their movement patterns to attractants and subsequent feeding behavior, in an effort to develop better repellents. The data in this example were analyzed in Klun et al. (2013). The set-up of the wind tunnel is given in Fig. 8. Data were taken on the number of mosquitoes (out of 20) that left the release cannister under three conditions; human breath, human breath scrubbed of CO_2, and nothing added to the incoming air. There were 10 replicates of each condition. For this analysis, the underlying distribution is assumed to be binomial, whose parameter changes with condition. Data were also taken on the number of probes recorded (aggregated over all mosquitoes on the membranes); for this an underlying Poisson distribution is assumed. Data were also taken on the number feeding (binomial), which will not be discussed here.

The distribution of the proportions leaving the cannister by condition is given in Fig. 9. Air mixed with human breath had the highest proportions. The data, especially for the scrubbed human breath condition, appears over-dispersed when compared to a binomial distribution. This is consistent with results from applying the PDT to two models, a GLM (generalized linear model) and a GLMM, the latter including a random trial effect. For the GLM, the mean of the distribution of log (sum-of-squares) for the data-model comparison is greater than the 95% quantile for the distribution representing model-model comparisons, i.e. the data and model do not match. For the GLMM, the data-model comparison is in the middle of the distribution of model-model comparisons, well below the 95% quantile, i.e. the data and model match. The centipede graph (not shown) shows no unusual patterns.

Figure 8: Left panel: wind tunnel. Air enters from the left and can be mixed with human breath or other gasses. Mosquitoes are released from the cannister at the right side of the tunnel (enlarged in the upper right panel). If they fly up-wind (to the left), they will encounter a feeding platform with warmed nutrient fluid under a membrane in several wells (one well is enlarged in the lower right panel). They may probe (pierce) the membrane and imbibe the nutrient liquid.

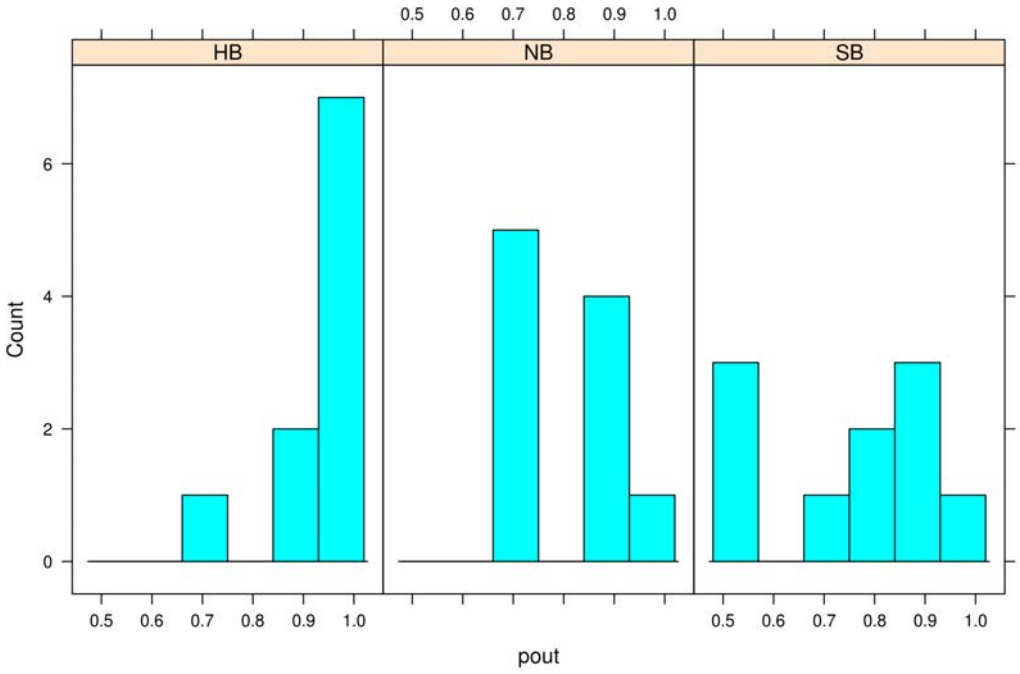

Figure 9: The proportion of mosquitoes that exited the cannister in each trial under the three breath conditions.

The distribution of the number of probes across the three breath conditions is given in Fig. 10. The number of probes was greatest for air mixed with human breath and uniformly low for the other two conditions. There is no indication of over-dispersion, in fact the data appear under-dispersed. A GLM with fixed effects alone appears adequate using the PDT, the data-model comparison mean is far smaller than the 95% quantile. In fact, only about 8% of the distribution of model-model comparisons was below the data-model comparison mean, suggesting that the data are under-dispersed (which may conceptually be a kind of over-fitting in the sense that the data more closely clustered around their means than is specified by the model). A centipede plot did not reveal unusual patterns, but clearly shows the increasing variability of the generated data with higher rankings, as expected for Poisson distributed data (Fig. 11).

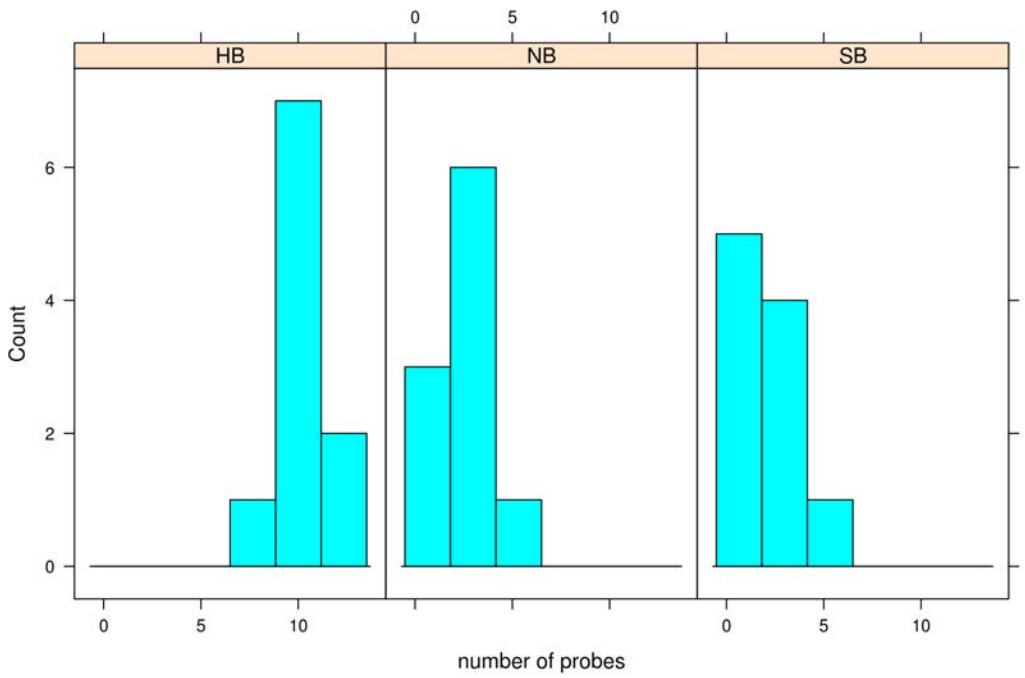

Figure 10: The total number of probes by mosquitoes on feeding stations in each trial for each condition.

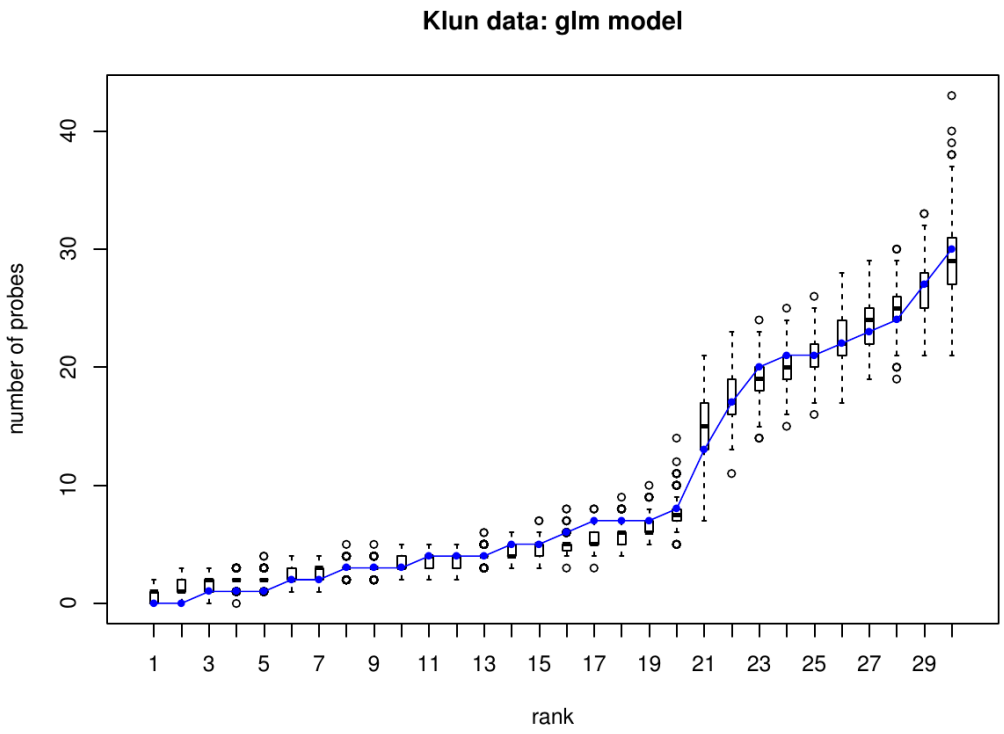

Figure 11: Centipede plot for the number of probes by mosquitoes. The data generated from the fitted GLM is represented by the box and whisker symbols, the original data superimposed with connected blue dots.

7. Example 5: simulated Poisson data from a block design fit with a normal distribution

Example data sets were simulated with the following R code.

```
n4 <- 5  # number of blocks
n5 <- 4  # number of treatments
n6 <- 1  # reps of treatments per block
mean2 <- rep (rpois (n5, 7), n6 * n4)
mean3 <- rep (as.character (1:n5), n6 * n4)
block2 <- rep (c(LETTERS,letters)[1:n4], each = n6 * n5)
block3 <- abs(rep (rnorm (n4,6,3), each = n6 * n5))
y4 <- rpois (n4*n5*n6, lambda = mean2 + block3)
D1 <- data.frame(y = y4, block = block2, trt = mean3)
```

This sets up a complete block design with 5 blocks, with a block effect taken from a normal (mean = 6, SD = 3) distribution, and four treatments per block, without replication. Treatment effects were samples from a Poisson distribution. Observations were samples from a Poisson distribution with λ = treatment effect + block effect. Such a data set satisfies the assumptions of a mixed model based on a Poisson process.

These data sets were then fit assuming they were generated from a normal distribution and centipede plots and the PDT statistic calculated. An example centipede plot is given in Fig. 12. The plot shows that the lowest values generated by the model can

lie below zero, a clear indication that this model is misspecified. The PDT statistic for this situation is not useful, *p*-values were greater than 0.05, though consistently slightly smaller than the same model fit assuming the data were Poisson (results not shown). Since the Poisson distribution has only one parameter and the normal distribution two, the latter distribution is more flexible, which partially compensates for not allowing the variance to increase with the mean. As can be seen in Fig. 12, the distributions around the higher ranked observations tend to miss the data more than they do around the lower ranked observations (compare with Fig. 11).

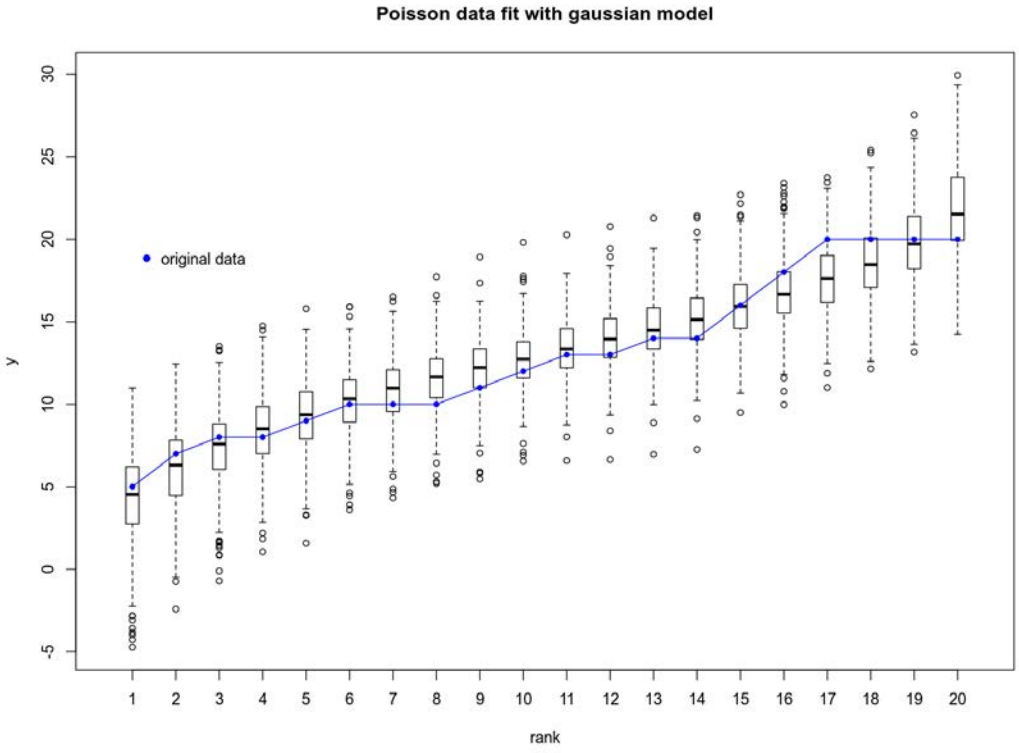

Figure 12: Centipede plot for Poisson data fit with a Gaussian model.

8. Example 6: simulated negative binomial data fit with a model assuming Poisson data

In this example of misspecification, 24 counts were generated with over-dispersion from a negative binomial distribution with $\varphi = 0.25$, block std. dev. = 0.25, and $\lambda = 10$, with a fixed treatment effect and a random block effect, similar to that described for Example 5. These simulated data were fit with a generalized linear mixed model assuming data were Poisson (so not over-dispersed), or Poisson but allowing for a random treatment-block interaction term (which does allow for over-dispersion, but different than that generated from a negative binomial distribution.

Twenty generated data sets were fit with the treatment and block (but no interaction) model. Fig. 13 gives examples of centipede plots from one that fails and one that didn't fail. In both cases, there is evidence from the centipede plots that the fit is not good. However, only five of the 20 failed the PDT at $\alpha = 0.05$. For this mismatch, the centipede plots provide more useful information than the PDT, perhaps not surprising for

such a small data set. In contrast, using the model with an interaction term, for 60 data sets, the centipede plots do not show evidence of poor fit (Fig. 14 gives one example) and none failed the PDT (despite a mismatch between the type of over-dispersion modeled and that created using the negative binomial distribution). Neither the graphical method nor PDT can discriminate between these two types of over-dispersion for this sized data set.

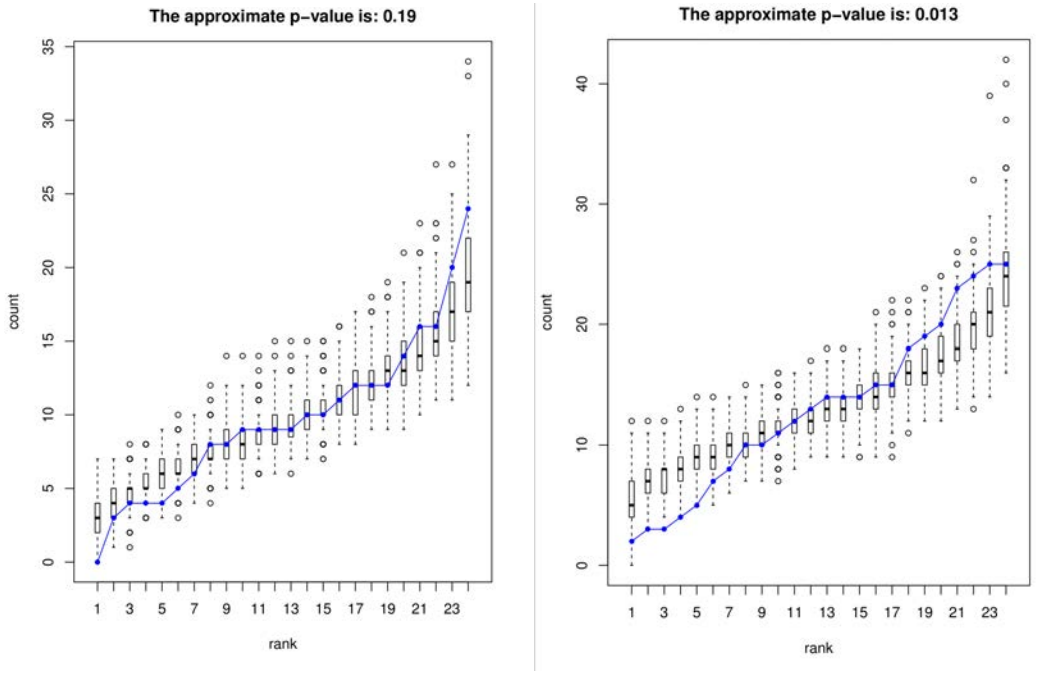

Figure 13: Two example centipede plots from counts generated from a negative binomial model and fit with a Poisson model.

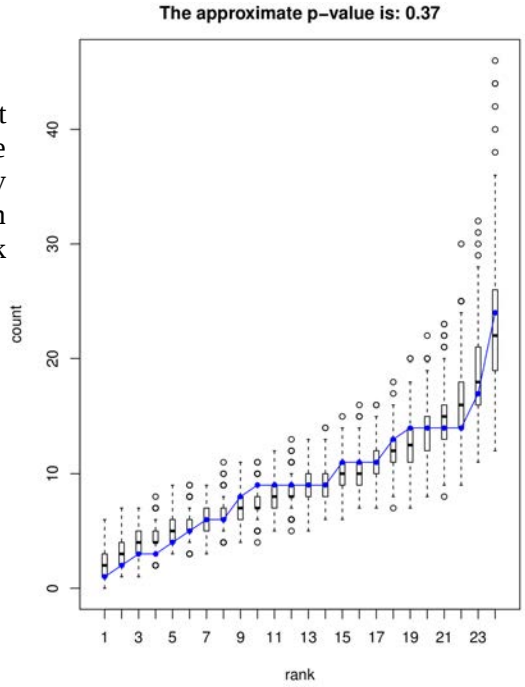

Figure 14: An example centipede plot from counts generated from a negative binomial model (with no treatment by block interaction) and fit with a Poisson model that includes a treatment by block interaction.

9. Example 7: simulated normal data with a block by treatment interaction fit without the interaction term

In mixed models, one common assumption is that there is no treatment by block (fixed effect by random effect) interaction. Indeed, many commonly used models are designed such that this interaction term is not estimable. For example, in a complete block design where there is one replicate of every treatment in each block, this interaction term cannot be estimated in the usual ANOVA way, although it could if there was some replications of treatments within a block. Another commonly used model, especially in diet and pharmaceutical trials, is a crossover design, where each subject (random effect) sequentially goes through various diets or treatments (fixed effect). Since subjects do not repeat treatments, the treatment-subject interaction is not easily estimable. There are various workarounds (e.g. Tukey's one d.f. test for interaction, AMMI models, see Kramer et al. 2012 for the use of AMMI models applied to crossover designs), but these do not correctly account for random effects.

In addition, if there is truly a treatment by block interaction that is not specified in the model, tests on the fixed effects using mixed models software are too liberal (*p*-values are too small). This can be readily seen by generating data from models with interaction and comparing tests on fixed effects from models with and without the interaction (Fig. 15 gives a scatterplot for *p*-values on the fixed effect from simulated data sets (with replicated treatments within blocks), modeled with and without the interaction term. Thus, it is worthwhile to ask if the methods described above might identify this type of misspecification.

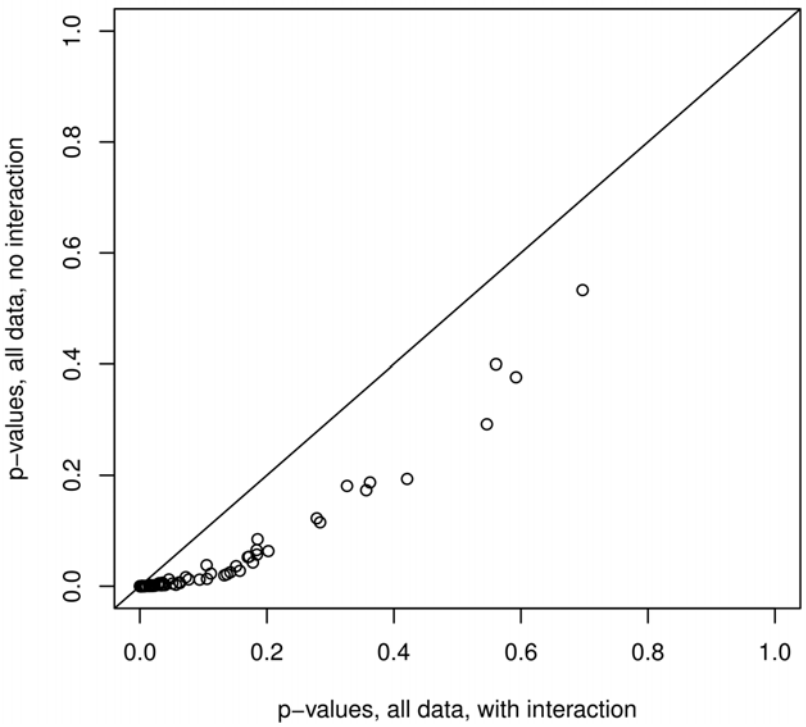

Figure 15: *P*-values for the fixed effect for data sets generated with a treatment by block interaction. Models were fit without this interaction term (y-axis) or with it (x-axis) using the R lmerTest package.

To examine whether this methodology was useful for this problem, 60 observations from a randomized complete block design (without replication but including a treatment by block interaction) were created. These data were fit by a model without accounting for this interaction. A centipede plot (Fig. 16) showed no evidence that the model was missing the interaction term. The PDT also did not find evidence of model misspecification (p = 0.62). Results were similar for other generated data sets. Thus, unfortunately, this methodology is not useful for identifying this data-model mismatch.

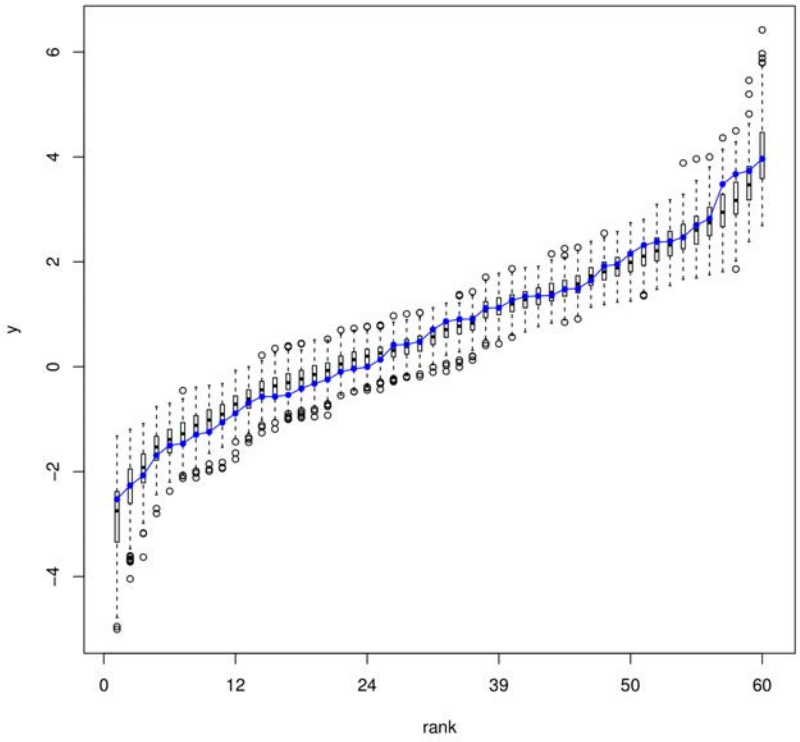

Figure 16: Centipede plot for data generated from a randomized complete block design with an interaction between the treatment and block, and fit with a model lacking this interaction term.

10. Conclusions

A diagnostic based on comparing two distances, a distance between the original data and that generated by a candidate model and a within model distance, comparing sets of generated data from the candidate model with each other, was developed to determine the probability the data could have arisen from the candidate model. It was developed from Bayesian ideas of using the posterior predictive distribution as a diagnostic. It is less powerful than a *t*-test, but is more broadly applicable, and was shown to be effective for several GLMM models. The method can probably be used for most parametric statistical models, whether in the frequentist or Bayesian paradigm. Since there is a paucity of diagnostics for GLMM models, especially to assess whether the process embodied by the candidate model could reasonably be believed to generate the data at hand, this method should help fill the gap. The GLMM examples included normal, binomial, and Poisson distributed data, with and without random effects. The PDT needs to be tested on a wider variety of model to better understand its strengths and weaknesses.

One weakness found was that the method does not flag Gaussian data from a randomized complete block desgin with a missing treatment-block interaction term. Testing can be done relatively easily and concisely using the R software, which contains a 'simulate' function, generating pseudo-data from models estimated by some of the widely used packages, like lm and glm, in the base package, and lme4. Using SAS to generate pseudo-data from these kinds of models requires writing more code, but is also not difficult, see Boykin et al. (2011) for tips. An R package for this method is planned.

References

Bayarri, M.J. and J.O. Berger. 1998. Quantifying surprise in the data and model verification. Bayesian Statistics 6, 53-82. (Includes discussion.)

Bates, D., M. Maechler, B. Bolker and S. Walker. 2014. lme4: Linear mixed-effects models using Eigen and S4. R package version 1.0-6. http://CRAN.R-project.org/package=lme4

Boykin, D., M.J. Camp, L. Johnson, M. Kramer, D. Meek, D. Palmquist, B. Vinyard, and M. West. 2011. Generalized linear mixed model estimation using Proc Glimmix: Results from simulations when the data and model match, and when the model is misspecified. In: Proceedings of the 22nd Annual Conference on Applied Statistics in Agriculture, Kansas State University, April 25-27, 2010. pp. 137-170.

Gelfand, A.E. 1996. Model determination using sampling-based methods. *In* W.R. Gilks, S. Richardson and D.J. Spiegelhalter (eds), Markov chain Monte Carlo in Practice. Chapman & Hall, pp. 145–161.

Gelfand, A.E. and S.K. Ghosh. 1998. Model choice: a minimum posterior predictive loss approach. Biometrika 85, 1-11.

Gelman, A. 2007. Comment: Bayesian checking of the second levels of hierarchical models. Statistical Science 22, 349-352.

Gelman, A., J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin. 2013. Bayesian Data Analysis, Third Ed. Chapman & Hall/CRC Press.

Gbur, E.E., W.W. Stroup, K.S. McCarter, S. Durham, L.J. Young , M. Christman, M. West, and M. Kramer. 2012. Analysis of Generalized Linear Mixed Models in the Agricultural and Natural Resources Sciences. American Society of Agronomy, Crop and Soil Science Society of America, Inc., Madison Wisconsin, USA.

Klun, J.A., M. Kramer and M. Debboun. 2013. Four simple stimuli that induce host-seeking and blood-feeding behaviors in two mosquito species, with a clue to DEET's mode of action. J. Vector Ecology 38, 143-153.

Kramer, M., S.C. Chen, S.K. Gebauer, and D.J. Baer. 2012. Estimating the subject by treatment interaction in non-replicated crossover diet studies. In: Proceedings of the 23rd annual Kansas State University Conference on Applied Statistics in Agriculture, Manhattan, Kansas, Weixin Yao (ed.). Pp. 96-110.

Meng, X-L. 1994. Posterior predictive *p*-values. Annals of Statistics 22, 1142-1160.

Plummer, M. 2013. rjags: Bayesian graphical models using MCMC. R package version 3-10. http://CRAN.R-project.org/package=rjags.

R Core Team. 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Shafii, B. and W.J. Price. 1998. Analysis of genotype-by-environment interaction using the additive main effects and multiplicative interaction model and stability estimates. JABES 3, 335–345.

Wright, K. 2013. agridat: Agricultural datasets. R package version 1.8. http://CRAN.R-project.org/package=agridat.