Are you Normal? The Problem of Confounded Residual Structures in Hierarchical Linear Models

Adam Loy and Heike Hofmann *

May 21, 2013

We encounter hierarchical data structures in a wide range of applications. Regular linear models are extended by random effects to address correlation between observations in the same group. Inference for random effects is sensitive to distributional mis-specifications of the model, making checks for (distributional) assumptions particularly important. The investigation of residual structures is complicated by the presence of different levels and corresponding dependencies. Ignoring these dependencies leads to erroneous conclusions using our familiar tools, such as Q-Q plots or normal tests. We first show the extent of the problem, then we introduce the *fraction of confounding* as a measure of the level of confounding in a model and finally introduce rotated random effects as a solution to assessing distributional model assumptions. This article has supplementary materials online.

Keywords: Diagnostic, Multilevel model, Q-Q plot, Random effects distribution

*Adam Loy is a PhD student (e-mail: loyad01@gmail.com) and Heike Hofmann is an Associate Professor in the Department of Statistics and Statistical Laboratory, Iowa State University, Ames, IA 50011-1210.

1. INTRODUCTION

There are a wide range of application areas—from the biological and physical sciences to the social sciences—in which we encounter nested data. Whether it is quality control in a manufacturing process that involves the monitoring of a set of components over time or students' performances in different schools across the country, analysts have to account for the correlation between observations in the same group. Hierarchical linear models, or multilevel models, allow us to do exactly that—but they also require us to make distributional assumptions on both the error terms and the random effects. These assumptions must hold to ensure the validity of the model and all of its resulting conclusions. Inference for the fixed effects in linear mixed models is fairly robust against model mis-specification (Butler and Louis, 1992; Verbeke and Lesaffre, 1997). This is different for random effects: they are sensitive to distributional mis-specifications and therefore have to be checked carefully, especially when they are central to the inferential goals, such as in the construction of a prediction interval for an unobserved group.

One approach to address this sensitivity is to avoid the assumptions made on the random effects distributions using semiparametric or nonparametric methods (Shen and Louis, 1999; Zhang and Davidian, 2001; Ghidey et al., 2004), or using a finite mixture of normal distributions for the random effects (Verbeke and Lesaffre, 1996). We refer the reader to Ghidey et al. (2010) for a recent review comparing these methods. The cost of increased robustness is increased computational complexity. These methods also have not been widely implemented in statistical software, making them less accessible. Another approach is to check this assumption using diagnostic tools. This is the approach on which we focus in this paper.

Quantile-quantile (Q-Q) plots (Wilk and Gnanadesikan, 1968) are our main graphical tool for visually evaluating a specific distributional assumption. For that, we plot the empirical distribution against the expected quantiles from the assumed distribution. In hierarchical linear models the investigation of residual structures is complicated by the presence of different levels. The nested structure of the data is reflected in the residual structure, and just as there is dependence between different levels in the data, we can expect dependencies between different levels in the residual

structure. Q-Q plots, in their weighted (Dempster and Ryan, 1985; Lange and Ryan, 1989) or unweighted form, do not account for this, which leads to erroneous conclusions in evaluating normality when there is a relatively high degree of shrinkage. Such situations are commonly encountered in practice, but are often overlooked in the literature. For example, Eberly and Thackeray (2005) explored properties of Lange and Ryan's weighted Q-Q plots for a balanced longitudinal data set and found that, for a properly specified mean structure, the weighted Q-Q plots can target the random effect distribution. This cannot be said for unbalanced data. Data imbalances lead to higher degrees of shrinkage, and in situations with high degrees of shrinkage weighted Q-Q plots cannot accurately target the random effect distribution, even if the mean structure is properly specified.

In this paper, we address the problem of distributional assessment due to confounding in residual structures. Section 2 illustrates the inadequacy of existing methods based on the predicted random effects. We introduce the concept of rotating the random effects into a reduced-dimensional subspace that is less confounded in Section 3, and illustrate how to obtain rotated random effects at all levels. In Section 4 we evaluate the sensitivity and specificity of tests of normality for the rotated random effects in a simulation study to investigate the behavior of Q-Q plots constructed from the rotated random effects. Finally, we demonstrate how this enables an appropriate graphical assessment of the distributional assumptions in Section 5.

2. MOTIVATING EXAMPLE

To illustrate the effect of confounding between different levels of residuals, we consider the data set discussed by Gelman and Pardoe (2006). This data set consists of a stratified random sample of 919 owner-occupied homes in 85 counties in Minnesota. Gelman and Pardoe (2006) suggest a hierarchical model of the form

$$\log(y_{ij}) = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2i} + b_{0i} + b_{1i} x_{1ij} + \varepsilon_{ij} \tag{1}$$

where $\log(y_{ij})$ denotes the radon measurement (in log pCi/L, i.e log picoCurie per liter) for house j within county i ($1 \leq j \leq n_i, 1 \leq i \leq 85$), x_{1ij} is a binary variable describing the level at which the measurement was taken (0 for the basement and 1 for a higher level), and x_{2i} denotes the average soil uranium content for county i. We assume i.i.d. normal errors $\varepsilon_{ij} \sim \mathcal{N}(0,\ \sigma_\varepsilon^2)$ and $\boldsymbol{b}_i \sim \mathcal{N}(\mathbf{0},\ \boldsymbol{D})$, where $\boldsymbol{D}$ allows for correlation between random effects within the same county i, b_{0i} and b_{1i}. Further, we assume independence between random effects and error terms.

Figure 1 shows a map of counties in Minnesota. The color shading represents average radon activity in a county. For two counties no data is available. Generally, more southern locations exhibit higher levels of radon activity. Figure 2 focuses on Hennepin (home to Minneapolis) and Winona (home to the city of the same name) counties, plotting radon level by floor level. Radon levels are usually the highest at the basement level of a house. The within-county sample sizes,

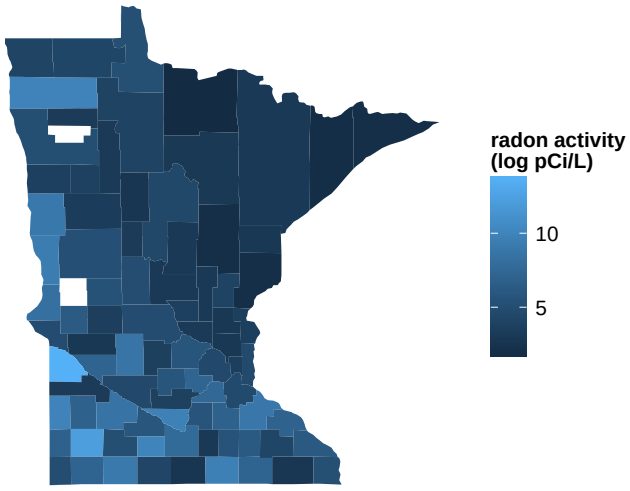


Figure 1: Map of the counties in Minnesota. The color shading represents average radon activity (in log pCi/L, i.e log picoCurie per liter).

n_i, are extremely unbalanced, ranging from one house to 116 houses, with 50% of the counties having between three and ten houses. Such unbalanced designs are common in applications, and result in a high degree of pooling in the predicted random effects, which results in quantities for many counties that are highly shrunken toward the global mean. It is this high degree of shrinkage that leads to dependence between predicted random effects and error terms (cf. eqns. 3 and 4), which in turn can lead us to draw erroneous conclusions for corresponding residual quantities.

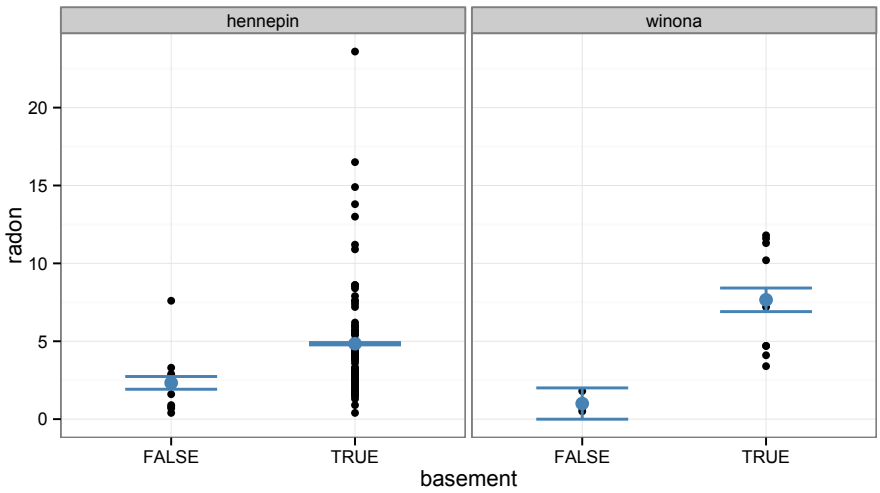
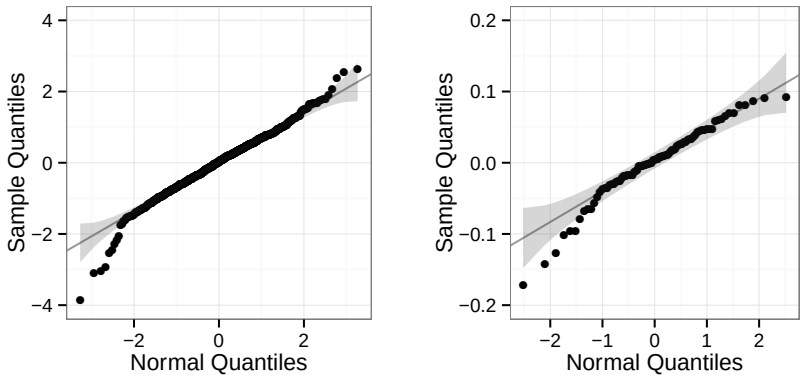


Figure 2: Activity of radon levels for Hennepin and Winona counties at basement (basement = TRUE) or higher in the residence. The bigger points indicate the sample means with 95% confidence intervals given by the error bars. Radon levels at the basement level are usually higher.



(a) Predicted error terms (b) Random intercepts

Figure 3: Q-Q plots of predicted residuals at different levels for model (1). Both plots suggest a deviation of residuals from a normal distribution. Note that random slopes (see figure 4) exhibit the largest deviation from normality.

In this example, Q-Q plots (Figure 3) for the residuals show that normality seems to be violated for the error terms and both random effects. But is this cause for concern? As there is little pooling at the observation level (level 1) we expect the distributional assessment of the error terms to be reliable, but the high degree of pooling for the random effects casts doubt on the reliability of their Q-Q plots. Next, we assess the reliability of the Q-Q plot for the random slope by utilizing the lineup protocol (Buja et al., 2009): Figure 4 shows a lineup of 20 Q-Q plots for the predicted

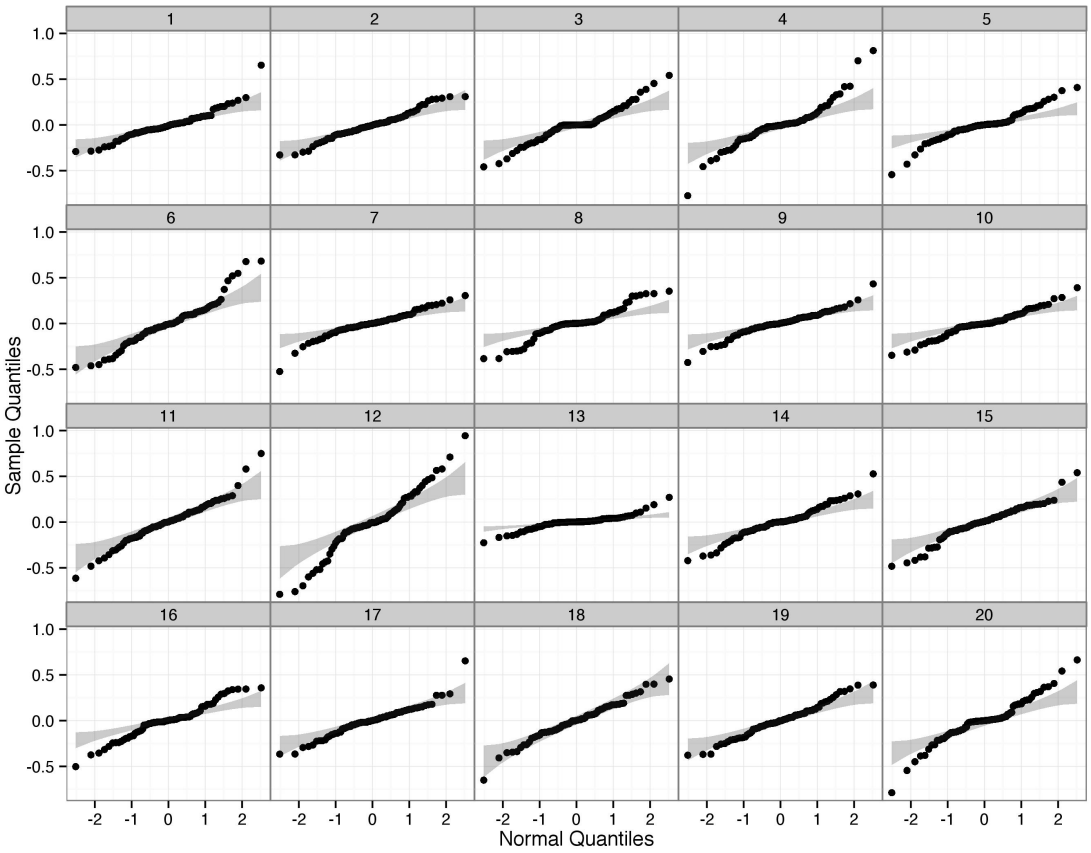


Figure 4: Lineup of normal Q-Q plots for the random slope term in model (1). The 19 null plots are obtained by simulation from the model. Can you identify the observed Q-Q plot?

random slope. The Q-Q plot of the observed random slopes is placed among 19 decoy plots of parametric bootstrap samples based on model (1) satisfying the normal distribution assumptions. The simulation parameters were set to the maximum likelihood estimates of model (1). If we can identify the real Q-Q plot in the lineup, this provides evidence that the distribution of the observed random slopes is not normal. However, the observed Q-Q plot (panel $12 + 2^2$) is virtually indistinguishable from the field of null plots. This suggests that the predicted random slopes from the data do not deviate significantly from model (1). Note that in practice we must blind ourselves from the true plot for proper use of lineups. In order to not violate this, we did not show the Q-Q plot of random slopes earlier.

What becomes apparent from the lineup, is that, astonishingly, *none* of the null plots conform to normality. To further investigate the apparent non-normal behavior of predicted random effects we

Table 1: Percentage of tests rejecting the null hypothesis of normality of the predicted random effects at the nominal 5% significance level when the error terms are normal, heavy tailed, and skewed. The percentages are hugely inflated under each setting compared to the nominal type I error rate.

(a) Random intercept

	Test		
	AD	CVM	KS
Normal	65.5	61.5	49.4
Heavy tailed	89.0	87.8	78.5
Skewed	84.1	83.0	71.5

(b) Random slope

	Test		
	AD	CVM	KS
Normal	87.4	86.9	81.5
Heavy tailed	96.5	96.7	92.7
Skewed	95.3	95.6	90.9

conduct a small simulation study: We generated 1000 parametric bootstrap samples of model (1) assuming normal random effects and level-1 residuals generated as normal ($\varepsilon^*_{ij} \overset{iid}{\sim} \mathcal{N}(\mathbf{0},\ \sigma^2_\varepsilon)$), heavy tailed ($\varepsilon^*_{ij} \overset{iid}{\sim} (\sigma_\varepsilon/\sqrt{3})\, t_3$), and skewed ($\varepsilon^*_{ij} \overset{iid}{\sim} \sigma_\varepsilon \{\text{Exp}(1) - 1\}$). For each simulated data set, we evaluated the assumption of normality for both the error terms and random effects using the Anderson-Darling (AD), Cramér von Mises (CVM) and Kolmogorov-Smirnov (KS) tests for normality. Table 1 shows the percentage of these tests rejecting the null hypothesis of normality at the 5% significance level.

The type I error rates are hugely inflated for both random effects, making an assessment of normality based on the empirical distribution impossible. For example, 84.1% of the AD tests of the random intercept rejected the null hypothesis of normality when the error terms were skewed. Use of standardized random effects and the weighted cumulative distribution function proposed by Lange and Ryan (1989) reduce the type I errors to the nominal level when the error terms are normal, but the type I errors remain inflated for non-normal error terms. Similarly, tests of non-normal random effects often failed to reject if the error terms were normally distributed. This inability to assess distributional assumptions correctly is a symptom typical of confounding between levels of residuals.

In situations with a large amount of pooling, confounding also affects the error terms, which in this particular example were the least affected and did not exhibit signs of deviation from normality.

In the remainder of this paper we investigate the root of concern that leads to the distributional

deviations, and derive residuals that address the issues introduced by pooling, allowing again for a familiar graphical assessment of these distributions.

3. ASSESSING THE DISTRIBUTION OF THE RANDOM EFFECTS

In this section we develop the rotated random effects and discuss computational and practical issues associated with their use. Before this discussion we present the model and notation used throughout the paper. Additionally, we review the problem of confounding which can be seen in the formulas for the residuals.

3.1 Model notation and residuals

The general stacked representation of the hierarchical linear model is given by

$$\begin{aligned} &\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\varepsilon}, \\ &\mathrm{E}\begin{bmatrix}\boldsymbol{b}\\ \boldsymbol{\varepsilon}\end{bmatrix} = \boldsymbol{0},\ \mathrm{Cov}\begin{bmatrix}\boldsymbol{b}\\ \boldsymbol{\varepsilon}\end{bmatrix} = \begin{bmatrix}\boldsymbol{D} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{R}\end{bmatrix} \end{aligned} \tag{2}$$

where $\boldsymbol{y}$ is an $n \times 1$ vector of observed responses, $\boldsymbol{X}$ ($n \times p$) and $\boldsymbol{Z}$ ($n \times q$) are design matrices, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown fixed effects, $\boldsymbol{b}$ is a $q \times 1$ vector of unobserved random effects, $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of unobserved errors, and $\boldsymbol{R}$ and $\boldsymbol{D}$ are positive definite covariance matrices.

Using this specification, the predicted error terms and random effects are given by

$$\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{R}\boldsymbol{P}\boldsymbol{y} = \boldsymbol{R}\boldsymbol{P}\boldsymbol{Z}\boldsymbol{b} + \boldsymbol{R}\boldsymbol{P}\boldsymbol{\varepsilon} \tag{3}$$

$$\widehat{\boldsymbol{b}} = \boldsymbol{D}\boldsymbol{Z}'\boldsymbol{P}\boldsymbol{y} = \boldsymbol{D}\boldsymbol{Z}'\boldsymbol{P}\boldsymbol{Z}\boldsymbol{b} + \boldsymbol{D}\boldsymbol{Z}'\boldsymbol{P}\boldsymbol{\varepsilon} \tag{4}$$

where $\boldsymbol{P} = \boldsymbol{V}^{-1}(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1})$. This set of equations reveals the inherent dependence between the residuals. Additionally, it is easily seen that both (3) and (4) lead to the analysis of correlated and potentially heteroscedastic disturbances as $\mathrm{Var}(\widehat{\boldsymbol{\varepsilon}}) = \boldsymbol{R}\boldsymbol{P}\boldsymbol{R}$ and $\mathrm{Var}(\widehat{\boldsymbol{b}}) = \boldsymbol{D}\boldsymbol{Z}'\boldsymbol{P}\boldsymbol{Z}\boldsymbol{D}$. The use of standardized residuals can correct the latter issue, but does

not address the fact that the residuals are interrelated. While problems may be expected at both levels of the model based on (3) and (4), we have found that the interpretation of Q-Q plots of the standardized predicted error terms

$$\boldsymbol{z}_{\varepsilon} = \text{diag}\left(\boldsymbol{RPR}\right)^{-1/2} \widehat{\boldsymbol{\varepsilon}}$$

is unaffected by this interrelationship. This is not the case with the standardized random effects. When the degree of pooling is high—as it is in the above radon example, and often is in practice—interpretation of the predicted random effects cannot be separated from the distribution of the error terms. Detailed simulation results documenting the utility of these residuals are available in the supplementary material.

3.2 Rotating the random effects

To combat confounding between different levels of residuals, we derive a reduced set of rotated residuals that are standardized, uncorrelated, and homoscedastic. We focus our discussion (and notation) on a two-level model with a single random effect in this section for ease of explanation, and describe how to extend this method at the end of this section.

First, we define the *fraction of confounding* for the random effects, which is minimized in the result below. This definition generalizes the fraction of confounding proposed by Hilden-Minton (1995).

Definition 1 (Fraction of confounding)**.** *Let* $\boldsymbol{b}$ *denote a vector of* q *random effects and* $\widehat{\boldsymbol{b}}$ *its predictions as defined in* (4)*. For a full rank matrix* $\boldsymbol{W} \in \mathbb{R}^{q \times s}$*, where* $s \leq q$*, the fraction of confounding in the* s*-dimensional space spanned by* $\boldsymbol{W}$ *is defined as*

$$FC(\boldsymbol{W}; \widehat{\boldsymbol{b}}) = \frac{1}{q} \text{Trace}\left(\left(\boldsymbol{W'BW}\right)^{-} \left(\boldsymbol{W'AW}\right) \right), \tag{5}$$

where $\boldsymbol{B}$ *is the covariance structure of* $\boldsymbol{b}$*,* $\boldsymbol{B} = \text{Var}(\widehat{\boldsymbol{b}})$*, and* $\boldsymbol{A}$ *is the conditional covariance structure of* $\widehat{\boldsymbol{b}}$ *given* $\boldsymbol{b}$*, i.e.* $\boldsymbol{A} = \text{Var}(\widehat{\boldsymbol{b}}|\boldsymbol{b})$*.*

Note that both $\boldsymbol{A}$ and $\boldsymbol{B}$ are positive semidefinite matrices by definition.

The fraction of confounding measures the contribution of the error terms to the variance of the random effects. FC $\in [0, 1]$, where 1 indicates that, due to confounding, the predicted random effects contain no information in addition to that found in the error terms, while 0 indicates no confounding. Notice that if $\boldsymbol{W}$ is the identity, then (5) measures the fraction of confounding in the original vector of predicted random effects.

In order to correct residuals for the impact of confounding, we propose using the linear transformation of the predicted random effects that substantially reduces the amount of confounding. To do this, we must determine the s-dimensional space in which confounding is substantially reduced and find the $\boldsymbol{W}$ that minimizes confounding within that space. To determine the dimension of the subspace spanned by the rotated residuals we propose the use of a visual tool similar to the scree plots used for selecting the number of principal components. The construction of such plots depends on the linear transformation $\boldsymbol{W}$, so we first discuss the selection of $\boldsymbol{W}$ given a fixed dimension s.

Selecting the optimal linear transformation. For a given s-dimensional space, the linear combination $\boldsymbol{W} \in \mathbb{R}^{q \times s}$ that minimizes (5) also minimizes

$$J_1(s) = \text{Trace}\left(\left(\boldsymbol{W'BW} \right)^{-1} \left(\boldsymbol{W'AW} \right) \right) \tag{6}$$

Mathematically, this problem is solved using the generalized eigenvalue decomposition

$$\boldsymbol{A}\boldsymbol{w}_k = \gamma_k \boldsymbol{B}\boldsymbol{w}_k \tag{7}$$

where γ_k and $\boldsymbol{w}_k$ are the k-th smallest eigenvalues and eigenvectors, respectively (Fukunaga, 1990). Computationally, we solve this problem by simultaneous diagonalization of $\boldsymbol{A}$ and $\boldsymbol{B}$ (McDonald et al., 1979; de Leeuw, 1982). Simultaneous diagonalization of $\boldsymbol{A}$ and $\boldsymbol{B}$ requires $\boldsymbol{W}$

to be $\boldsymbol{B}$-orthogonal, so the optimal $\boldsymbol{W}$ is found to be

$$\boldsymbol{W}^*(s) = \underset{\boldsymbol{W}\in\mathbb{R}^{q\times s},\ \boldsymbol{W'BW}=\boldsymbol{I}}{\arg\min} \operatorname{Trace}\left(\boldsymbol{W'AW}\right) \tag{8}$$

Below, we outline the procedure used to simultaneously diagonalize $\boldsymbol{A}$ and $\boldsymbol{B}$ for reference.

Algorithm 1 (Simultaneous diagonalization)**.** *Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be two positive semidefinite matrices. The transformation that simultaneously diagonalizes both matrices can be found through the following procedure:*

1. *Find a transformation that whitens $\boldsymbol{B}$. Such a transformation is given by $\boldsymbol{T_r}\boldsymbol{\Lambda_r}^{-1/2}$, where $\boldsymbol{T}_r$ and $\boldsymbol{\Lambda}_r$ are the first r eigenvectors and eigenvalues of $\boldsymbol{B}$, where $r = \operatorname{rank}(\boldsymbol{B})$.*

2. *Transform $\boldsymbol{A}$ and $\boldsymbol{B}$ to*

$$\boldsymbol{\Lambda_r}^{-1/2}\boldsymbol{T_r'}\boldsymbol{A}\boldsymbol{T_r}\boldsymbol{\Lambda_r}^{-1/2} = \boldsymbol{A}^* \tag{9}$$

$$\boldsymbol{\Lambda_r}^{-1/2}\boldsymbol{T_r'}\boldsymbol{B}\boldsymbol{T_r}\boldsymbol{\Lambda_r}^{-1/2} = \boldsymbol{I} \tag{10}$$

3. *Find an orthonormal transformation that diagonalizes $\boldsymbol{A}^*$. Such a transformation is given by the eigenvectors of $\boldsymbol{A}^*$, which we denote $\boldsymbol{U}$.*

Based on the above three steps, the transformation that simultaneously diagonalizes $\boldsymbol{A}$ and $\boldsymbol{B}$ is $\boldsymbol{T_r}\boldsymbol{\Lambda_r}^{-1/2}\boldsymbol{U}$.

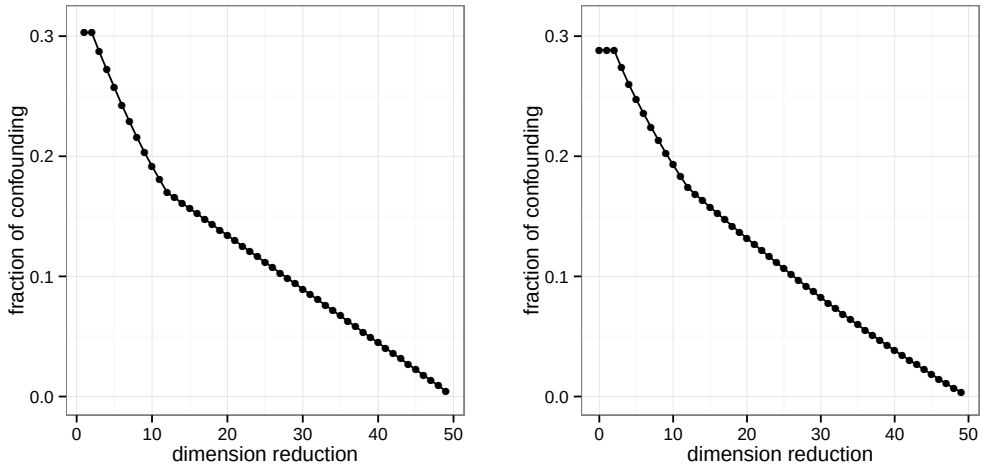
The above procedure can be used to find the general solution to (7). To find the more specific transformation defined by (8), we focus on the s eigenvectors associated with s the smallest eigenvalues of $\boldsymbol{A}^*$, $\boldsymbol{U}_s$, making

$$\boldsymbol{W}^* = \boldsymbol{T_r}\boldsymbol{\Lambda_r}^{-1/2}\boldsymbol{U}_s \tag{11}$$

The rotated random effects are then given by $\boldsymbol{W}^{*\prime}\widehat{\boldsymbol{b}}$, which are standardized, uncorrelated, and homoscedastic (see the appendix for a proof).

Selecting the dimension of the subspace. Selection of the dimension of the subspace spanned by the rotated residuals is central to our proposed method. Ideally, we would select the dimension such that the fraction of confounding is reduced to zero; however, this is not realistic in practice. Alternatively, we propose choosing the dimension that provides a substantial reduction in the fraction of confounding. Since our ultimate objective is distributional assessment, we must balance this reduction in the fraction of confounding with the loss in power of a test of the empirical distribution function (e.g., the Anderson-Darling test for normality) associated with dimension reduction. To guide this choice we suggest plotting the reduction in the fraction of confounding against the reduction in dimension, which is similar to the scree plot used to select the number of principal components. To illustrate the use of this plot we simulate two simple random intercept models with a group-level predictor: model $M1$ has 40 groups of 30 observations and 10 groups of 5 observations; model $M2$ also has 50 groups, with group sizes determined as random draws from either a Poisson(30) distribution (40 groups) or a Poisson(5) distribution (10 groups). Figure 5 shows two examples of such plots constructed for the simulated models. For both models, there is no decrease in the fraction of confounding for a one-dimensional reduction because a one-dimensional reduction simply adjusts for that rank deficiency of the covariance matrix, which can be thought of in terms of adjusting for the effective degrees of freedom. Both figures have an "elbow" in the plot corresponding to a reduction in the dimension of the subspace of 11; thus we would choose $s = 39$. Additionally, we see that the elbow in the plot for model $M2$ is less pronounced. This occurs because the groups sizes are more unbalanced so the difference between the large and small group sizes is reduced.

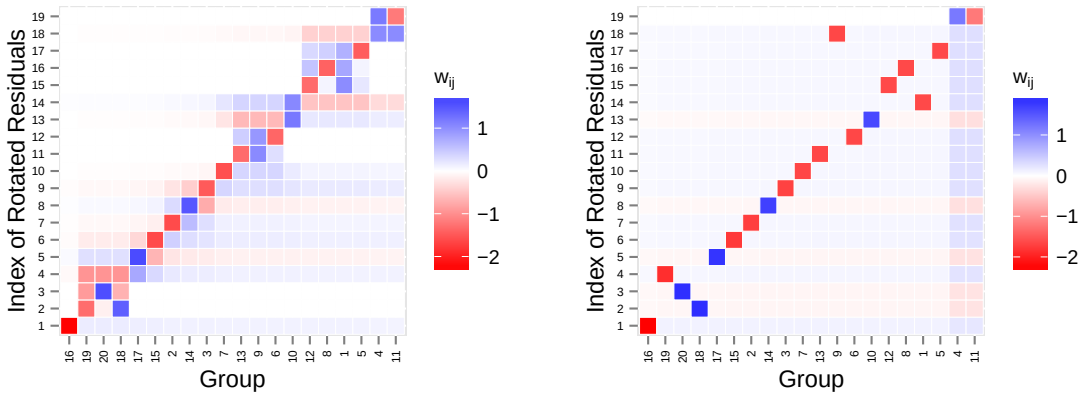
Correcting for supernormality. The transformation of the random effects results in a vector where each component is a linear combination of elements of $\widehat{\boldsymbol{b}}$. Consequently, the rotated residuals will appear more normal than the underlying distribution, if the underlying distribution



(a) Model $M1$ (b) Model $M2$

Figure 5: Plots of the fraction of confounding for each reduction in the dimension of subspace spanned by the rotated random intercepts from two simulation models. Model $M1$ has 50 groups: 40 groups of size 30, and 10 groups of size 5. Model $M2$ has 50 groups with group sizes determined as random draws from either a Poisson(30) distribution (40 groups) or a Poisson(5) distribution (10 groups). Based on these plots we choose $s = 39$, corresponding to a dimension reduction of 11.

is not normal. This issue is often referred to as supernormality (Atkinson, 1985). One approach to address supernormality in this context is to reduce the number of elements in the linear combinations, which should reduce the extent of the problem. To do this, we suggest using an orthogonal rotation of $\boldsymbol{W}^*$, which we denote $\boldsymbol{Q}$, just as we rotate the factor loadings in factor analysis. Using this approach, the rotated residuals are obtained by $\boldsymbol{Q}'\boldsymbol{W}^{*\prime}\widehat{\boldsymbol{b}}$. One rotation that will produce rotated residuals comprised of a small number of raw residuals is the raw varimax rotation (Johnson and Wichern, 2007). Figure 6 displays heat maps of $\boldsymbol{W}^{*\prime}$ (left) and $\boldsymbol{Q}'\boldsymbol{W}^{*\prime}$ (right) for a simulated random intercept model with 20 groups, and demonstrates that the raw varimax rotation reduces the number of groups loading highly on each rotated residual. Other orthogonal rotations could be used, but the varimax rotation is familiar to a wide range of analysts and is widely implemented in statistical software packages. A similar approach was used by Jensen and Ramirez (1999), who used the raw varimax rotation to produce recovered errors for distributional assessment in the ordinary regression model.



(a) $\boldsymbol{W}^{*\prime}$ (b) $\boldsymbol{Q}'\boldsymbol{W}^{*\prime}$

Figure 6: Heat map of $\boldsymbol{W}^{*\prime}$ and $\boldsymbol{Q}'\boldsymbol{W}^{*\prime}$ calculated where $\boldsymbol{Q}$ for a simulated random intercept model with 20 groups. Applying the raw varimax rotation, $\boldsymbol{Q}$, reduces the number of groups loading on a given rotated residual.

Extension to multiple random effects. Up to this point our discussion has ignored that a model may (and often will) contain numerous random effects. In this case, the assumptions made on each random effect should be checked; thus, we propose assessing each random effect separately. To this end we must define linear combinations $\boldsymbol{L}_k$ such that $\boldsymbol{L}_k'\widehat{\boldsymbol{b}}$ produces the kth marginal random effect. For example, in a model with a random intercept and random slope, if $\boldsymbol{Z}$ is organized as a block diagonal matrix—that is $\boldsymbol{Z} = \bigoplus_{i=1}^{m} \boldsymbol{Z}_i$ where $\bigoplus$ denotes the direct sum (Gentle, 2007, page 47)—then $\boldsymbol{L}_0 = \boldsymbol{I}_m \bigotimes(1,\ 0)$ produces the random intercepts and $\boldsymbol{L}_1 = \boldsymbol{I}_m \bigotimes(0,\ 1)$ produces the random slopes. The methodology presented in this section can be generalized to models with numerous random effects by substituting $\boldsymbol{L}_k'\widehat{\boldsymbol{b}}$ for $\widehat{\boldsymbol{b}}$.

4. SIMULATION STUDY

We conducted a simulation study to assess the specificity and sensitivity of tests of normality based on the two rotated residuals proposed in the previous section.

4.1 Design

We want to examine situations in which we correctly and incorrectly reject the null hypothesis of normality—that is, power and type I error, respectively. We compute the percentage of

Anderson-Darling (AD), Cramér von Mises (CVM), and Kolmogorov-Smirnov (KS) tests that rejected the null hypothesis of normality. These test statistics each measure the discrepancy between the empirical distribution of the rotated random effects and assumed distribution of the random effects, which sheds light on the behavior of Q-Q plots constructed from the rotated residuals.

The design matrices from model (1) were used as templates for realistic data generation; for simplicity of the simulation design, only the 60 counties with full rank $\boldsymbol{Z}$ matrices were included. Normal, heavy-tailed, and skewed distributions are used to generate the simulated errors and random effects. We use a rescaled t distribution with 3 degrees of freedom to generate heavy tailed residuals, and a centered and rescaled exponential distribution with a rate parameter of 1 to generate skewed residuals. For simplicity we require the distributions of the random slope and intercept to the be same and assume independence between the random effects. The nine distributional settings considered in the simulation study are summarized in Table 2.

Table 2: A summary of the nine distributional settings considered in the simulation study.

Distributions of		Random effects, F_2		
		$\mathcal{N}(0,\ \sigma_b^2)$	$(\sigma_b/\sqrt{3})\, t_3$	$\sigma_b \{\text{Exp}(1) - 1\}$
Error terms, F_1	$\mathcal{N}(0,\ \sigma_\varepsilon^2)$ $(\sigma_\varepsilon/\sqrt{3})\, t_3$ $\sigma_\varepsilon \{\text{Exp}(1) - 1\}$	$\varepsilon_{ij}^* \overset{iid}{\sim} F_1,\ \ b_{0i}^*, b_{1i}^* \overset{iid}{\sim} F_2$		

Additionally, the the fixed effects coefficients were set to the maximum likelihood estimates.

To investigate the effect that pooling has on the rotated random effects we considered three variance structures to represent different degrees of confounding for the random effects:

high: $\sigma_\varepsilon^2 = 4$ and $\sigma_{b_0}^2 = \sigma_{b_1}^2 = 1$

moderate: $\sigma_\varepsilon^2 = 1$ and $\sigma_{b_0}^2 = \sigma_{b_1}^2 = 1$

low: $\sigma_\varepsilon^2 = 1$ and $\sigma_{b_0}^2 = \sigma_{b_1}^2 = 4$

Under each simulation setting 1000 data sets were generated for each model and the rotated residuals were obtained using $s = \text{rank}(\boldsymbol{B})$ (which is 58 and 59 for the random intercept and slope, respectively) as well as $s = 55$, 50, 45, 40, 35, and 30.

4.2 Results

Figure 7 shows the average fraction of confounding for the rotated random intercept (left) and random slope (right) over the different values for s for each variance structure. As s is reduced, the fraction of confounding is reduced, which aligns with expectation as smaller choices of s reduce the contributions of more highly confounded groups.

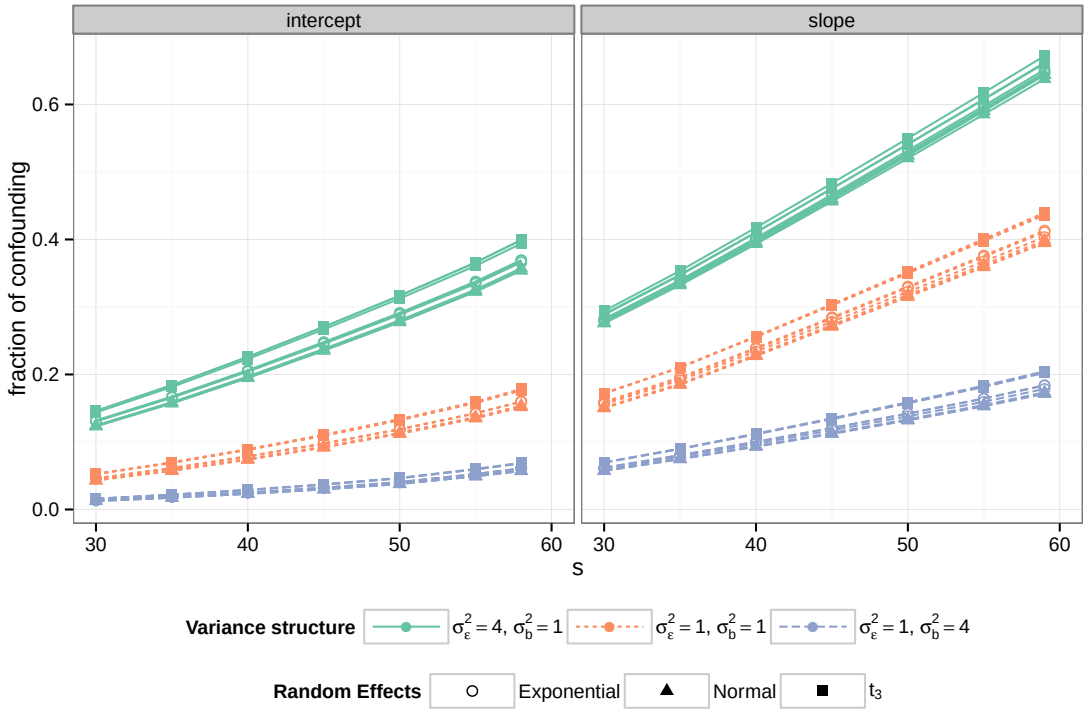


Figure 7: Change in the fraction of confounding (FC) as the dimension of the rotated random effects vector, s, is reduced for the three variance structures considered in the simulation study.

Table 3 and Figure 8 display the estimated type I error rates using the AD normality test ($\alpha = 0.05$) on the rotated and varimax rotated random intercepts and random slopes, respectively, when $\sigma^2_\varepsilon = 4$ and $\sigma^2_{b_0} = \sigma^2_{b_1} = 1$. The CVM and KS tests performed similarly and are omitted for brevity (full simulation results can be found in the supplementary material). Both figures show that the type I error rate is stabilized close to the nominal level with the appropriate choice of s. For the random intercept most choices of s perform reasonably well, with the type I error rate closest to the nominal level for all error distributions between 30 and 40. For the random slope, s must be

chosen to be 30 for type I error to be near the nominal level; however, s may need to be even smaller to achieve the nominal rate.

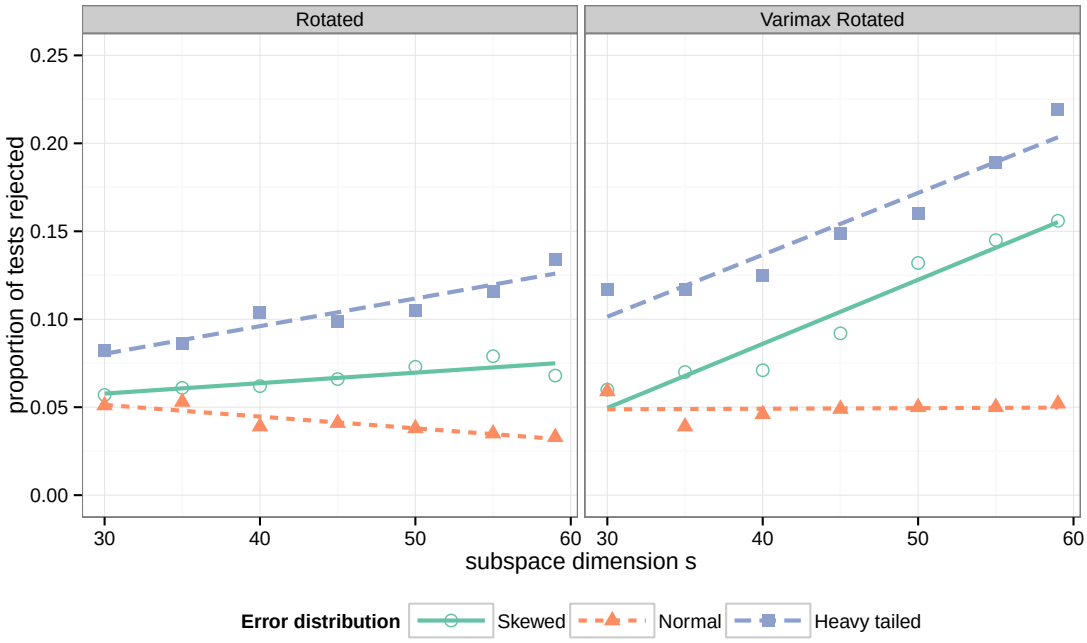


Figure 8: Estimated type I error rate using the Anderson-Darling normality test ($\alpha = 0.05$) on the rotated random slopes (left) and varimax rotated random slopes (right) by the distribution of the error terms and s.

Table 3 and Figure 9 show the estimated power of the AD test ($\alpha = 0.05$) on the rotated and varimax rotated random intercepts and random slopes, respectively, for the highly confounded variance structure. The estimated power to detect non-normal random effects distributions is amplified by the varimax rotation and larger choices of s. We also find that the estimated power is lower than would be expected from randomly sampled values from an exponential or t_3 distribution (what we will refer to as the "gold standard"). For example, when $s = 30$, simulations indicate the power of the AD test to detect a t_3 distribution to be approximately 0.4, whereas our simulations indicate nearly half the power, with the random slope generally having lower power than the random intercept. Interestingly, there is higher power to detect a heavy tailed distribution than a skewed distribution. Additional simulations (not shown) using a model with a continuous variable defining the random slope showed results similar to the random intercept (Table 3).

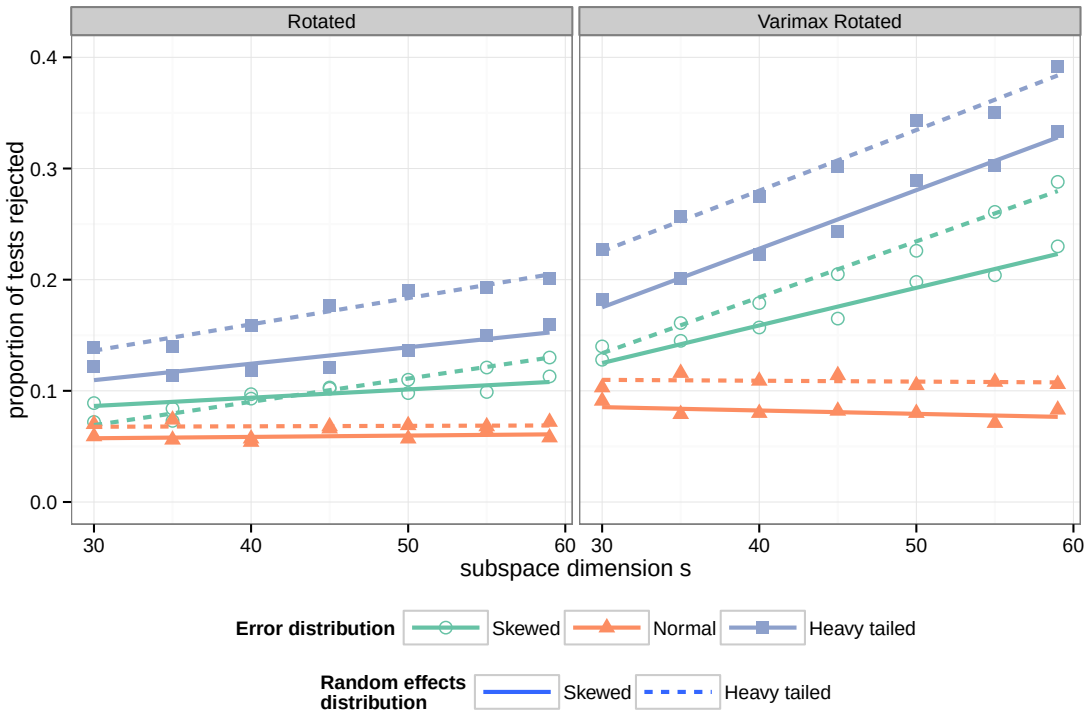


Figure 9: Linear smoother of the estimated power using the Anderson-Darling normality test ($\alpha = 0.05$) on the rotated random slopes (left) and varimax rotated random slopes (right) by s. The color denotes the distribution of the errors and the line type denotes the distribution of the random slope.

While the estimated power is lower than the gold standard, the fact that the type I error rate can be stabilized indicates that distributional problems detected using the rotated random effects will truly be problems; thus, providing more diagnostic information than the (unrotated) predicted random effects.

5. RADON DATA: REVISITED

Recall that in Section 2 we determined that the error terms were not normally distributed. Consequently, examination of Q-Q plots of the predicted random effects will likely lead to erroneous conclusions due to the high degree of shrinkage.

In order to construct Q-Q plots of the rotated random effects we first consider the choice of s. For model (1) the high degree of shrinkage leads to a large fraction of confounding for each random term: 0.72 for the random intercept and 0.70 for the random slope. In choosing s we

Table 3: Percentages of AD tests rejecting the null hypothesis of normality at the 5% significance level for the rotated (a) and varimax rotated (b) random intercepts by s. The gray shading indicates the situations where the random intercept is normal (i.e., type I error).

(a) Rotated random intercept.

		s						
Random intercept	Error term	58	55	50	45	40	35	30
Normal	Normal	4.4	4.0	4.6	4.3	4.4	5.0	5.7
	Heavy tailed	7.5	7.6	6.7	6.2	5.0	4.2	5.0
	Skewed	5.1	4.7	5.7	5.8	5.6	5.5	5.6
Heavy tailed	Normal	13.9	13.6	13.1	13.4	13.0	13.1	12.1
	Heavy tailed	19.0	18.6	16.7	16.1	16.0	14.8	13.9
	Skewed	15.5	15.1	14.2	13.6	13.2	12.7	11.9
Skewed	Normal	9.6	8.7	9.5	9.7	10.0	11.0	10.0
	Heavy tailed	12.6	12.5	12.0	11.3	10.1	11.3	11.0
	Skewed	13.4	13.4	12.2	12.2	11.0	11.3	10.8

(b) Varimax rotated random intercept.

		s						
Random intercept	Error term	58	55	50	45	40	35	30
Normal	Normal	4.9	5.3	5.2	5.3	5.3	5.2	5.5
	Heavy tailed	9.0	9.1	8.0	7.1	6.0	5.1	5.2
	Skewed	5.2	5.1	4.4	5.5	6.1	5.1	6.1
Heavy tailed	Normal	22.1	22.3	23.3	22.9	23.3	22.3	21.6
	Heavy tailed	34.4	33.3	32.1	31.6	30.1	27.0	26.6
	Skewed	27.8	26.7	25.6	27.0	24.4	23.1	21.8
Skewed	Normal	19.7	21.2	21.3	22.1	22.7	21.1	22.3
	Heavy tailed	29.7	28.4	27.1	25.0	25.5	25.0	23.8
	Skewed	22.2	23.5	21.7	23.1	21.1	22.9	21.1

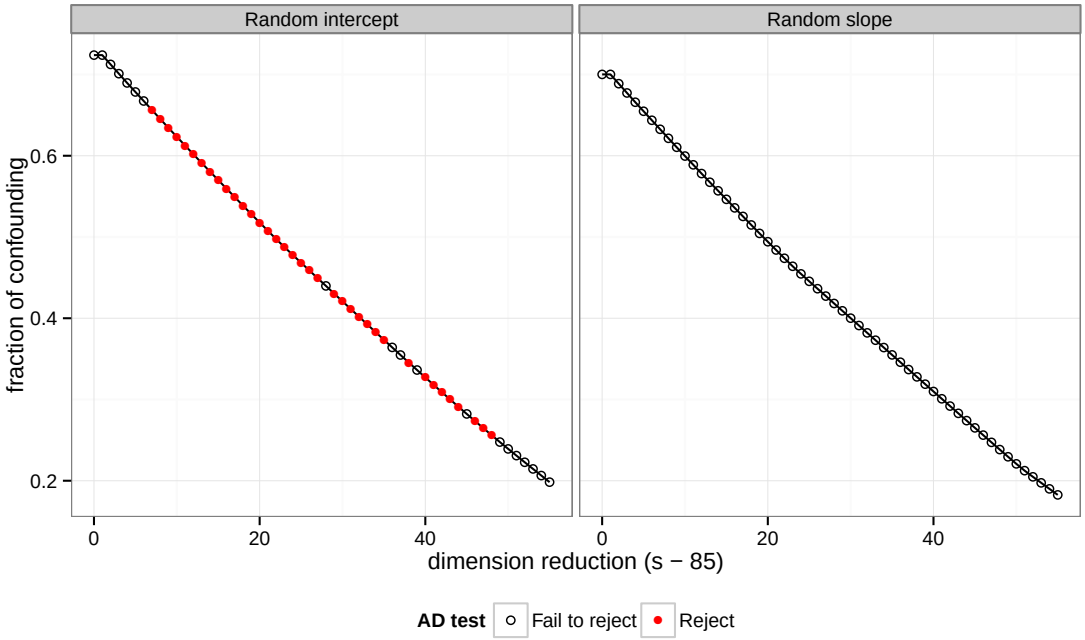
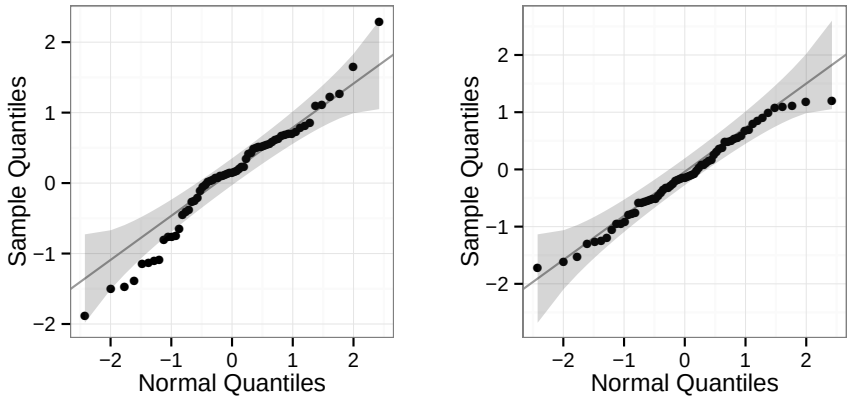


Figure 10: Plots of the fraction of confounding for each reduction in the dimension of subspace spanned by the rotated random intercept (left) and random slope (right) for the radon example. Filled points represent situations in which the AD test rejects normality of the varimax rotated random effects.

wish to substantially reduce the fraction of confounding, but restrict attention to $s \geq 30$ so as not to decrease the maximum possible power of a normality test too severely. Figure 10 shows the fraction confounding for $s \geq 30$.

We do not find elbows for either random effect in plots of the fraction of confounding against the reduction in dimensionality (Figure 10) making the choice of the dimension more difficult; however, because of the large number of counties with few observations, the smoothness of the plots is not unexpected. For a decision on the distribution, the exact choice of the subspace dimension s is also not particularly critical. The color (fill) of the points in figure 10 shows the results of the AD test at the 5% significance level for each dimension. The red (solid) points denote a rejection of the null hypothesis of normality. The results are stable until the dimensionality is reduced to the point where the AD test lacks power. To show an example of the Q-Q plots produced from the rotated random effects Figure 11 shows Q-Q plots of the rotated random effects for subspace

$s = 65$. We see that the random intercept significantly deviates from the assumption of normality while the random slope does not.



(a) Random intercept (b) Random slope

Figure 11: Normal Q-Q plots with point-wise 95% confidence bands of the marginals rotated random effects. The deviations from normality are much less pronounced than before resulting in the failure to reject the null hypothesis of normality.

6. DISCUSSION

In this paper we have discussed two graphical approaches to assess the distributional assumptions made on the random effects in hierarchical linear models. The first approach used the lineup protocol to compare the predicted random effects produced in estimating the observed model to those produced when estimating a properly specified model. This method only assumes a distributional specification for the random effects and does not directly compare the predicted random effects to their hypothesized distribution. Consequently, the conclusions that are drawn from this approach relate to evidence that the predicted random effects either are or are not consistent with what is expected under a correctly specified random effects distribution. The second approach rotates the predicted random effects so as to compare them directly to the hypothesized distribution using a Q-Q plot. We have shown that the rotated random effects are standardized, uncorrelated, and homoscedastic, and that the rotation addresses the confounding present allowing for the random effects to be targeted separately from the error terms.

Under either approach, a misspecified covariance structure may lead to erroneous rejection of the null hypothesis. Therefore, in practice we recommend an assessment of the structure of the within- and between-group covariance matrices prior to distributional assessment. An alternative approach would be the use of robust covariance estimation techniques to protect against such misspecification; however, it is not clear how this impacts the diagnostic tools. We will leave this investigation for future study.

It is important to note that formal tests have been proposed to detect mixture distributions in the random effects (Verbeke and Lesaffre, 1996) and for overall goodness-of-fit tests for both the error terms and random effects (Jiang, 2001); however, these methods do not lend themselves to graphical inspection and have not been implemented in statistical software. Our method, on the other hand, requires only byproducts of the model fitting procedure and the use of matrix decompositions for simultaneous diagonalization, which are widely accessible in standard software. All of the methods and graphics discussed in this paper are implemented in `R` (R Core Team, 2013). In particular, the rotated residuals are part of the package `HLMdiag` (Loy, 2013; Loy and Hofmann, in press).

Simulation has revealed that tests of normality using the rotated random effects achieve approximately nominal type I error rates with appropriate choice of the dimension, s. This indicates that assessment of the rotated residuals can target the distribution of the random effects in the presence of pooling, which the predicted random effects cannot. The power to detect non-normal random effects distributions is lower than the gold standard, which is to be expected as the rotated residuals consist of sums of predicted random effects, resulting in a total distribution that is closer to a normal distribution than its individuals. The varimax rotation reduces the impact of this supernormality effect. While we do think that the loss in power is troubling, the inflated type I error rates resulting from high levels of confounding is of a much bigger concern. Unlike before, any detection of a distributional deviation can now be trusted even in situations with high amounts of confounding between the levels of residuals.

7. APPENDIX: ADDITIONAL TECHNICAL DETAILS

We present the proof of the claim that the rotated residuals, $\boldsymbol{W}^{*\prime}\widehat{\boldsymbol{b}}$, are standardized, uncorrelated, and homoscedastic. Following the developments presented in Section 3.2, we present this discussion for the random effects assuming that there is only a random intercept. Generalization to the situation with multiple random effects follows as previously discussed.

Proof. Let $\boldsymbol{A} = \mathrm{Var}(\widehat{\boldsymbol{b}}|\boldsymbol{b})$, $\boldsymbol{B} = \mathrm{Var}(\widehat{\boldsymbol{b}})$, $r = \mathrm{rank}(\boldsymbol{B})$, and $q =$ the number of elements in $\widehat{\boldsymbol{b}}$. Note that by definition $\boldsymbol{A}$ and $\boldsymbol{B}$ are symmetric and positive semidefinite. Following from above, $\boldsymbol{T}_r$ and $\boldsymbol{\Lambda}_r$ follow from the spectral (or eigenvalue) decomposition of $\boldsymbol{B} = \boldsymbol{T_r}\boldsymbol{\Lambda_r}\boldsymbol{T_r'}$, and $\boldsymbol{U}$ follows from the spectral decomposition of $\boldsymbol{A^*} = \boldsymbol{\Lambda_r^{-1/2}}\boldsymbol{T_r'}\boldsymbol{A}\boldsymbol{T_r}\boldsymbol{\Lambda_r^{-1/2}} = \boldsymbol{U}\boldsymbol{\Gamma}\boldsymbol{U}'$. Then,

$$\begin{aligned}
\mathrm{Var}(\boldsymbol{W}^{*\prime}\widehat{\boldsymbol{b}}) &= \mathrm{Var}(\boldsymbol{U}'\boldsymbol{\Lambda}_{\boldsymbol{r}}^{-1/2}\boldsymbol{T}_{\boldsymbol{r}}'\widehat{\boldsymbol{b}}) \\
&= (\boldsymbol{U}'\boldsymbol{\Lambda}_{\boldsymbol{r}}^{-1/2}\boldsymbol{T}_{\boldsymbol{r}}')\mathrm{Var}(\widehat{\boldsymbol{b}})(\boldsymbol{T_r}\boldsymbol{\Lambda}_{\boldsymbol{r}}^{-1/2}\boldsymbol{U}) \\
&= (\boldsymbol{U}'\boldsymbol{\Lambda}_{\boldsymbol{r}}^{-1/2}\boldsymbol{T}_{\boldsymbol{r}}')\boldsymbol{B}(\boldsymbol{T_r}\boldsymbol{\Lambda}_{\boldsymbol{r}}^{-1/2}\boldsymbol{U}) \\
&= \boldsymbol{I}
\end{aligned}$$

proving that the rotated random effects are standardized, uncorrelated, and homoscedastic. □

SUPPLEMENTARY MATERIALS

The following supplemental materials can be obtained online:

Simulation results: The supplementary materials include the full simulation study discussed in Section 4. Additionally, the results of a simulation supporting the small simulation study discussed in Section 2 are presented and further show the need for alternative procedures to assess the distribution of the random effects.

`R` script for figures and simulations: The `R` code and data used to generate results discussed in this paper are available in the file `code_supplement.zip`.

R package `HLMdiag`: We have included the function to calculate the rotated random effects in the `R` package `HLMdiag`, which is available from the Comprehensive R Archive Network (CRAN, `http://cran.r-project.org/`).

References

Atkinson, A. C. (1985), *Plots, Transformations, and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis*, Oxford: Clarendon Press.

Buja, A., Cook, D., Hofmann, H., Lawrence, M., Lee, E.-K., Swayne, D. F., and Wickham, H. (2009), "Statistical Inference for Exploratory Data Analysis and Model Diagnostics," *Royal Society Philosophical Transactions A*, 367, 4361–4383.

Butler, S. and Louis, T. (1992), "Random effects models with non-parametric priors," *Statistics in Medicine*, 11, 1981–2000.

de Leeuw, J. (1982), "Generalized eigenvalue problems with positive semi-definite matrices," *Psychometrika*, 47, 87–93.

Dempster, A. P. and Ryan, L. M. (1985), "Weighted Normal Plots," *Journal of the American Statistical Association*, 80, 845–850.

Eberly, L. E. and Thackeray, L. M. (2005), "On Lange and Ryan's plotting technique for diagnosing non-normality of random effects," *Statistics & Probability Letters*, 75, 77–85.

Fukunaga, K. (1990), *Introduction to Statistical Pattern Recognition*, San Diego: Academic Press, 2nd ed.

Gelman, A. and Pardoe, I. (2006), "Bayesian Measures of Explained Variance and Pooling in Multilevel (Hierarchical) Models," *Technometrics*, 48, 241–251.

Gentle, J. E. (2007), *Matrix Algebra: Theory, Computations, and Applications in Statistics*, New York: Springer.

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

Ghidey, W., Lesaffre, E., and Verbeke, G. (2010), "A comparison of methods for estimating the random effects distribution of a linear mixed model," *Statistical Methods in Medical Research*, 19, 575–600.

Hilden-Minton, J. A. (1995), "Multilevel diagnostics for mixed and hierarchical linear models," Ph.D. thesis, University of California Los Angeles.

Jensen, D. R. and Ramirez, D. E. (1999), "Recovered errors and normal diagnostics in regression," *Metrika*, 49, 107–119.

Jiang, J. (2001), "Goodness-of-fit tests for mixed model diagnostics," *The Annals of Statistics*, 29, 1137–1164.

Johnson, R. A. and Wichern, D. W. (2007), *Applied Multivariate Statistical Analysis*, Upper Saddle River: Pearson, 6th ed.

Lange, N. and Ryan, L. (1989), "Assessing normality in random effects models," *The Annals of Statistics*, 17, 624–642.

Loy, A. (2013), *HLMdiag: Diagnostic tools for hierarchical (multilevel) linear models*, R package version 0.2.3.

Loy, A. and Hofmann, H. (in press), "HLMdiag: A Suite of Diagnostics for Hierarchical Linear Models in R," *Journal of Statistical Software*.

McDonald, R. P., Torii, Y., and Nishisato, S. (1979), "Some results on proper eigenvalues and eigenvectors with applications to scaling," *Psychometrika*, 44, 211–227.

R Core Team (2013), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Shen, W. and Louis, T. A. (1999), "Empirical Bayes Estimation via the Smoothing by Roughening Approach," *Journal of Computational and Graphical Statistics*, 8, 800–823.

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, 217–221.

— (1997), "The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data," *Computational Statistics & Data Analysis*, 23, 541–556.

Wilk, M. B. and Gnanadesikan, R. (1968), "Probability Plotting Methods for the Analysis of Data," *Biometrika*, 55, 1–17.

Zhang, D. and Davidian, M. (2001), "Linear Mixed Models with Flexible Distributions of Random Effects for Longitudinal Data," *Biometrics*, 57, 795–802.