Journal of Data Science **8**(2010), 579-595

The Effect of Sample Composition on Inference for Random Effects Using Normal and Dirichlet Process Models

Guofen Yan[1] and J. Sedransk[2]
[1] *University of Virginia and* [2] *Case Western Reserve University*

Abstract: Good inference for the random effects in a linear mixed-effects model is important because of their role in decision making. For example, estimates of the random effects may be used to make decisions about the quality of medical providers such as hospitals, surgeons, etc. Standard methods assume that the random effects are normally distributed, but this may be problematic because inferences are sensitive to this assumption and to the composition of the study sample. We investigate whether using a Dirichlet process prior instead of a normal prior for the random effects is effective in reducing the dependence of inferences on the study sample. Specifically, we compare the two models, normal and Dirichlet process, emphasizing inferences for extrema. Our main finding is that using the Dirichlet process prior provides inferences that are substantially more robust to the composition of the study sample.

Key words: Bayesian nonparametric method, extrema, heterogeneity, outlying clusters, robustness.

1. Introduction

Linear mixed-effects models are used extensively in practice when there are clustered observations. For example, in longitudinal studies repeated measurements are collected on the same subjects. In studies evaluating the performance of medical providers patients are clustered in hospitals.

Letting the response y_{ij} denote the jth observation in the ith cluster, the simplest model of this type is

$$y_{ij} \quad = \quad \mu + \alpha_i + \epsilon_{ij}\,, \quad i = 1, \ldots, m\,, j = 1, \ldots, n_i \tag{1.1}$$

where there are m clusters with n_i observations in cluster i. It is customarily assumed that the α_i and ϵ_{ij} are mutually independent, normally distributed random variables with $E(\alpha_i) = E(\epsilon_{ij}) = 0$, $\text{var}(\alpha_i) = \delta^2$ and $\text{var}(\epsilon_{ij}) = \sigma^2$. Extensions of (1.1) to include covariates are common (see, e.g., Verbeke and Molenberghs, 2000); a typical model is

$$Y_i \;=\; X_i\beta + Z_ib_i + \epsilon_i\,,\;\; i = 1, \ldots, m \tag{1.2}$$

where Y_i is a $(n_i \times 1)$ vector of response measurements for cluster i, β is a vector of fixed-effects parameters, b_i is a vector of random-effects parameters, and ϵ_i is a $(n_i \times 1)$ vector of random errors. In (1.2), X_i and Z_i are design matrices. Typically, b_i and ϵ_i are assumed to be normally distributed.

Analyzing (1.1) or (1.2) with an assumption of normality for the random effects, α_i in (1.1) or b_i in (1.2), may be problematic. It is well known (e.g., Box and Tiao, 1973; Verbeke and Lesaffre, 1996) that inferences are sensitive to the assumption of normality and also to the presence of outliers. For these reasons there has been extensive research to relax the assumption of normality of the random effects. The use of heavy-tailed distributions for the random effects is often more robust than the standard choice of a normal distribution (e.g., Pinheiro, Liu and Wu, 2001; Rosa, Gianola and Padovani, 2004). Verbeke and Lesaffre (1996) and Frühwirth-Schnatter, Tüchler and Otter (2004) assume that the distribution of the random effects is a mixture of normal distributions while Tao *et al.* (1999) use a predictive recursion method to obtain a nonparametric smooth density estimate.

A further problem is that the units being analyzed may not be a random sample from a single distribution, an important assumption for the α_i in (1.1) or the b_i in (1.2). For example, in a study of hospitals to identify outliers, the effects $\alpha_1, \ldots, \alpha_m$ corresponding to the m hospitals may not come from a single distribution. This situation often occurs because covariates associated with the hospitals cannot always be measured or controlled. In such situations a standard analysis of (1.1) or (1.2) may not recognize this heterogeneity of the random effects (Verbeke and Lesaffre, 1996).

To reduce the sensitivity of inferences to the assumption of normality and, especially, to take into account the possible heterogeneity of the random effects a potentially useful method is to replace the normal distribution assumption for the random effects with a nonparametric distribution using a Dirichlet process prior (DPP). Specifically, in a linear mixed-effects model, using a Dirichlet process (DP) as a prior distribution on the family of distributions for the random effects reflects our uncertainty about the distribution of the random effects. There is an extensive literature about the Dirichlet process: Dey, Müller and Sinha (1998), MacEachern and Müller (2000), and Müller and Quintana (2004) provide a good overview, additional references, applications and methodology. Applications where the usual parametric distribution for random effects has been replaced with a nonparametric distribution using a DPP include Kleinman and

Ibrahim (1988) and, more recently, Krnjajic, Kottas and Draper (2008), Müller, Quintana and Rosner (2007), Ohlssen, Sharples and Spiegelhalter (2007), van der Merwe and Pretorius (2003) and others. Implementation has been enhanced by the recent development of a package in R (Jara, 2007).

We believe, like many others, that using a DPP model will result in inferences that are more robust than those from a standard analysis, and investigate the extent to which this is true. Our specific focus is to examine the extent of changes in inference for *random effects* that may occur when additional, outlying units (i.e., outlying random effects) are included. That is, we start with a set of units, add outlying units and investigate the changes in inference about the original set of units (which may, before the additions, include both inlying and outlying units).

Although there has been published research that has compared parametric and semiparametric models with a DP prior, no one has studied the effect on inferences of the composition of the study sample as we do. Moreover, many of these comparisons have been concerned with inference for the fixed-effects (i.e., the population average effects) but not the random effects (i.e., subject-specific effects); see, e.g., Krnjajic, Kottas and Draper (2008). We believe that understanding how inference for the random effects changes is important as there are many applications where inference for the random-effects is desired and the results have a significant impact on decision-making. For example, in studies evaluating the performance of medical providers ("provider profiling") inference about the random effects is used to identify non-compliant hospitals, surgeons, etc. (Normand, Glickman and Gatsonis, 1997). In others such as child growth studies the individual growth trajectories $(b_1, \ldots, b_m)$ are of interest (Verbeke and Molenberghs, 2000; Hui and Berger, 1983; Diggle, Liang and Zeger, 1994).

Our study uses eight datasets. We start with the initial dataset, analyzed by Morris and Christiansen (1996), which consists of 23 kidney transplant hospitals where the outcome variable is graft failure. We use this dataset as a basis and then construct a second dataset by adding four hospitals whose data are simulated from populations different from those associated with the first 23. We then extend this study to other datasets with varying characteristics. We fit the customary one-way random-effects model, (1.1), and the associated DPP random-effects model, respectively, to these datasets. We contrast the results from the two models, focusing on two widely used summary measures, i.e., the posterior mean and the posterior probability that the outcome (failure rate) exceeds a threshold. The latter is chosen because it is of particular interest in quantifying the likelihood that the hospital is an outlier, an important aspect of "provider profiling." To make the results more transparent, we focus on the simpler model, (1.1), and clearly, the results can be applied to the general situation, (1.2).

In Section 2 we present the data and methods. We describe the remaining data sets and give all of our results in Section 3. There is additional discussion and a summary in Section 4.

Table 1: Posterior estimates in the normal and DPP random-effects models using the original sample (23 hospitals)

Hospital	# failures	n_i	y_i	$\sqrt{\phi_i}$	$E(\theta_i \mid y)$		$P(\theta_i > 0.25 \mid y)$	
ID					Normal	DPP	Normal	DPP
1	16	53	0.302	0.055	0.240	0.237	0.369	0.361
2	8	57	0.140	0.053	0.180	0.188	0.017	0.010
3	12	59	0.203	0.052	0.204	0.198	0.081	0.050
4	20	60	0.333	0.052	**0.255**	**0.261**	**0.528**	**0.558**
5	25	72	0.347	0.047	**0.266**	**0.274**	**0.649**	**0.662**
6	16	74	0.216	0.046	0.209	0.200	0.100	0.065
7	12	77	0.156	0.046	0.183	0.188	0.013	0.007
8	11	77	0.143	0.046	0.177	0.186	0.008	0.004
9	18	82	0.220	0.044	0.211	0.201	0.104	0.073
10	17	83	0.205	0.044	0.204	0.197	0.067	0.041
11	19	91	0.209	0.042	0.207	0.197	0.072	0.041
12	25	94	0.266	0.041	0.235	0.226	0.305	0.278
13	23	96	0.240	0.041	0.222	0.210	0.165	0.133
14	32	122	0.262	0.036	0.236	0.228	0.303	0.294
15	18	125	0.144	0.036	0.170	0.184	0.002	0.001
16	15	129	0.116	0.035	0.154	0.179	0.000	0.000
17	29	144	0.201	0.033	0.202	0.195	0.032	0.020
18	33	156	0.212	0.032	0.208	0.198	0.053	0.032
19	32	169	0.189	0.031	0.195	0.191	0.014	0.005
20	39	184	0.212	0.029	0.209	0.198	0.048	0.033
21	31	187	0.166	0.029	0.179	0.186	0.002	0.001
22	39	226	0.173	0.027	0.183	0.187	0.002	0.001
23	43	261	0.165	0.025	0.176	0.185	0.000	0.000

2. Data and Methods

Table 1 shows the original sample of the 23 kidney transplant hospitals (Morris and Christiansen 1996). The data, recorded during a 27-month period in the late 1980s, include the number of graft failures during this period (column 2), and the total number of kidney transplant operations performed (n_i) (column 3). Column 4 presents the proportions, y_i = # graft failures/n_i. Since the average of the observed failure rates is 20.5%, Morris and Christiansen (1996) use

$\sqrt{(0.2)(0.8)/n_i}$ to approximate the standard deviations $\sqrt{\phi_i}$ of the proportions (column 5).

Table 2: Posterior estimates in the normal and DPP random-effects models using the second sample (27 hospitals)

Hospital	# failures	n_i	y_i	$\sqrt{\phi_i}$	$E(\theta_i\|y)$		$P(\theta_i > 0.25\|y)$	
ID					Normal	DPP	Normal	DPP
Original sample								
1	16	53	0.302	0.055	0.298	0.239	**0.820**	0.376
2	8	57	0.140	0.053	0.155	0.189	0.029	0.006
3	12	59	0.203	0.052	0.210	0.195	0.209	0.038
4	20	60	0.333	0.052	0.326	0.276	**0.944**	**0.633**
5	25	72	0.347	0.047	0.340	0.298	**0.978**	**0.774**
6	16	74	0.216	0.046	0.221	0.197	0.254	0.052
7	12	77	0.156	0.046	0.165	0.189	0.026	0.005
8	11	77	0.143	0.046	0.153	0.188	0.012	0.003
9	18	82	0.220	0.044	0.223	0.198	0.269	0.052
10	17	83	0.205	0.044	0.209	0.194	0.166	0.030
11	19	91	0.209	0.042	0.213	0.195	0.182	0.034
12	25	94	0.266	0.041	0.266	0.223	**0.658**	0.263
13	23	96	0.240	0.041	0.241	0.204	0.410	0.106
14	32	122	0.262	0.036	0.263	0.223	**0.645**	0.267
15	18	125	0.144	0.036	0.151	0.187	0.003	0.000
16	15	129	0.116	0.035	0.125	0.183	0.000	0.000
17	29	144	0.201	0.033	0.204	0.193	0.078	0.011
18	33	156	0.212	0.032	0.214	0.195	0.131	0.026
19	32	169	0.189	0.031	0.192	0.191	0.029	0.004
20	39	184	0.212	0.029	0.214	0.195	0.106	0.019
21	31	187	0.166	0.029	0.170	0.188	0.003	0.000
22	39	226	0.173	0.027	0.175	0.188	0.002	0.000
23	43	261	0.165	0.025	0.168	0.187	0.000	0.000
Four additional hospitals								
24	82	120	0.683	0.037	0.660	0.632	1.000	1.000
25	77	130	0.592	0.035	0.575	0.590	1.000	1.000
26	74	140	0.529	0.034	0.516	0.567	1.000	1.000
27	89	150	0.593	0.033	0.578	0.591	1.000	1.000

The second dataset was constructed by adding four hospitals with characteristics different from those in the original sample. We generated their (failure) data independently from binomial distributions with n_i taken as 120, 130, 140, and 150, respectively, and a common failure rate, 0.6. These four simulated failure rates are 0.683, 0.592, 0.529, and 0.593, shown as IDs 24, 25, 26, and 27 in the lower panel of Table 2. Clearly, the second set (with 27 hospitals) is much less homogeneous than the first set. We have extended this study by modifying the characteristics of both the initial set of hospitals and the added set: see Section 3.3 for the description of the remaining data sets.

In practice, datasets containing outlying clusters will occur when the study population consists of two or more subpopulations. Alternatively, the clusters studied may be sampled from a single population of clusters but several were selected from the tail of the distribution.

Let y_i denote the observed graft failure rate in hospital i. Morris and Christiansen (1996) approximated the distribution of y_i as normal. While this assumption is simplistic we proceed in the same way because our conclusions are clearer than they would be if we used a more complex model for the y_i. Then the customary random-effects model is

$$\begin{aligned} & y_i = \mu + \alpha_i + \epsilon_i\,, \qquad \epsilon_i \sim N(0, \phi_i)\,, \\ & \{\alpha_i : i = 1, \ldots, m\} \mid \delta^2 \overset{iid}{\sim} N(0, \delta^2)\,, \\ & \pi(\mu, \delta^2) \;=\; \pi_1(\mu)\, \pi_2(\delta^2) \end{aligned} \tag{2.1}$$

where $m = 23$ for the first sample and $m = 27$ for the second sample. The $\{\phi_i : i = 1, \ldots, m\}$ are known and given in Tables 1 and 2. We took proper diffuse priors for μ and δ^2; i.e., a normal prior on μ with mean $\bar{y}$ and variance $100,000$, and an inverse gamma on δ^2 with parameters $(0.001, 0.001)$.

A DPP random-effects model replaces the normal prior on the α_i in (2.1) with a DP prior; i.e.,

$$\begin{aligned} & y_i = \mu + \alpha_i + \epsilon_i\,, \qquad \epsilon_i \sim N(0, \phi_i)\,, \\ & \{\alpha_i : i = 1, \ldots, m\} \mid G \sim G(\cdot)\,, \\ & G \mid \nu, \tau^2 \sim DP(\nu\, G_0)\,, \;\; G_0 = N(0, \tau^2)\,, \\ & \pi(\mu, \tau^2) = \pi_1(\mu)\, \pi_2(\tau^2) \end{aligned} \tag{2.2}$$

where DP denotes the Dirichlet Process prior for the cumulative distribution function, G. The prior for μ and τ^2 is the same as in (2.1).

There are two components in the DP: The base distribution G_0 defines the location of the DP prior, and the positive scalar ν is the precision parameter measuring the concentration of the prior for G around G_0. We take G_0 as the prior used in the normal random-effects model and consider ν to be a tuning constant. The additional stage of the DP prior for G expresses our uncertainty about the true distribution of the random effects relative to the prior, $G_0 = N(0, \tau^2)$, the distribution assumed for the random effects α_i in the normal model, (2.1). Therefore, the random-effects model with the DP prior generalizes the normal model

(2.1). The tuning parameter ν, which controls the number of distinct components in the distribution of the random effects, G, has been discussed extensively in the literature (see, e.g., Escobar and West, 1995, 1998). When ν is small, the random effects (corresponding to the hospital units) tend to form very few clusters, thus the resulting inference is close to that from a finite mixture model. In contrast, when ν is large, inferences under (2.1) and (2.2) are similar. Liu (1996) developed the relationship between ν and the number of distinct clusters k, i.e., $E(k) \approx \nu \log(1+m/\nu)$, where m is the number of random effects. We used small values, $\nu = 0.5$ and $\nu = 1.0$, corresponding to $E(k) \approx 2$ and 3, respectively. Subsequently, we found that using $\nu = 0.5$ and $\nu = 1$ gave very similar results.

Implementing this DPP random-effects model is straightforward: the *DPpackage* in R recently developed by Jara (2007) is available for applications. Following the computational strategies suggested in Escobar and West (1998), we analyzed the data using the Gibbs sampler and ran a burn-in of 10,000 draws with inference based on the next 10,000 draws. We examined the convergence by running several parallel chains with each of length 20,000 and using the second 10,000 draws to calculate $\widehat{R}$ (Gelman *et al.*, 2004, chap. 11). The $\widehat{R}$ is the suggested quantity that can be used to measure the extent to which the parallel chains are mixed. For the parameter of interest, say, τ^2 in (2.2), $\widehat{R} = \sqrt{\widehat{\text{var}}(\tau^2|y)/W}$ where $\widehat{\text{var}}(\tau^2|y) = (n-1)W/n + B/n$, the estimate of the marginal posterior variance, $\text{var}(\tau^2|y)$. Here, n is the length of chains after discarding the first half of the draws (n = 10,000), and W and B are the within-chain and between-chain sample variances calculated using the second half of the draws, respectively. When the chains have mixed, B and W should be similar. Thus, $\widehat{R}$ near 1 indicates that the second half of the draws are simulated from the target distribution, and can be used for inference. We calculated $\widehat{R}$ for parameters (τ^2, μ, α_i) which were all around 1. Also, the trace plots did not indicate any abnormal divergence.

We have fitted these two models to the two datasets as described above. Our inference about the hospital effects is based on summary statistics obtained from the posterior distributions of the θ_i where $\theta_i = (\mu + \alpha_i)$. The first one is the posterior mean of θ_i. The second is the posterior probability that the failure rate (i.e., θ_i) exceeds a given threshold; i.e.,

$$P(\theta_i > c|\, y)\,.$$

A large value of this probability suggests further investigation of this hospital as a possible outlier. Since it is acknowledged (Morris and Christiansen, 1996) that a graft failure rate that exceeds 25% is deemed unacceptable, we take $c = 0.25$.

3. Results

3.1 Comparing the two models

Analyzing the sample of 23 hospitals first, the posterior means and posterior probabilities, $P(\theta_i > 0.25|y)$, from the fit of the normal model, (2.1), are shown in columns 6 and 8, respectively, in Table 1. Among these 23 hospitals, hospitals 4 and 5 had the highest posterior means which also were greater than the threshold failure rate of 25% (25.5% and 26.6%, respectively). Their posterior probabilities of exceeding 25% were also highest, 52.8% and 64.9%, respectively. Results from the DPP model, (2.2), with $\nu = 1$ are shown in columns 7 and 9 of Table 1. (The results with $\nu = 0.5$ are similar.) Note that the same hospitals (4 and 5), regarded by Morris and Christiansen (1996) as "noncompliant", were identified in the DPP model. Based on these data we do not regard these two hospitals as "noncompliant," but ones where further investigation is indicated.

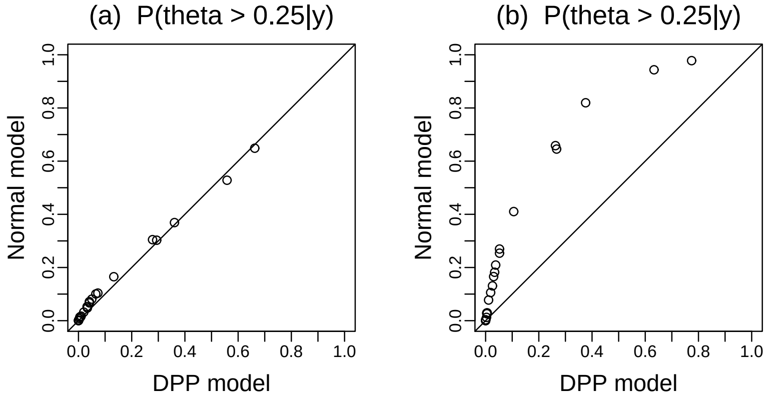

Figure 1: Scatterplots of the 23 paired posterior probabilities (corresponding to the 23 hospitals) that the failure rate exceeds a threshold of 0.25, one from the fit of the normal model and the other from the fit of the DPP model. Plot (a) uses the original sample of 23 hospitals, and plot (b) uses the second sample of 27 hospitals.

Figure 1(a) is a scatter plot of the 23 pairs of the posterior probabilities corresponding to the 23 hospitals, one from fitting the normal model and the other from fitting the DPP model. Since the points cluster along the 45 degree line the posterior probabilities from the two models are very similar. The posterior means from the two models are also similar (compare columns 6 and 7 in Table 1). In this case, where the data are homogeneous, inference about the individual hospital effects under these two approaches are very similar, i.e., the normal model, (2.1), fits the data appropriately.

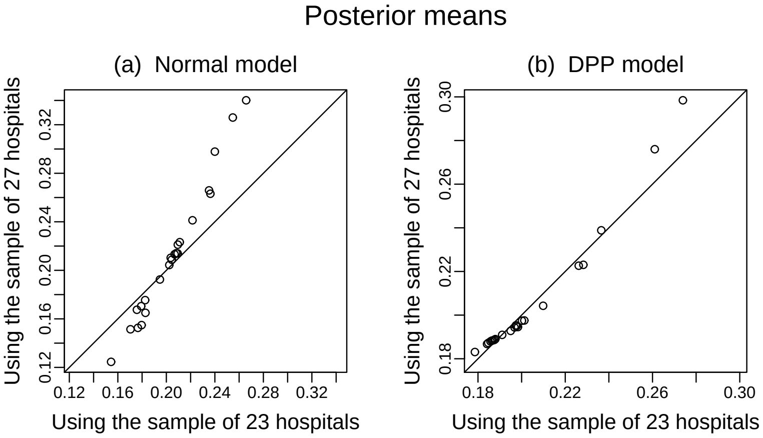

Figure 2: Scatterplots of the 23 paired posterior means corresponding to the original 23 hospitals, one from the analysis of the 23 hospitals and the other from the analysis of the 27 hospitals. Plot (a) is the fit of the normal model and plot (b) is the fit of the DPP model.

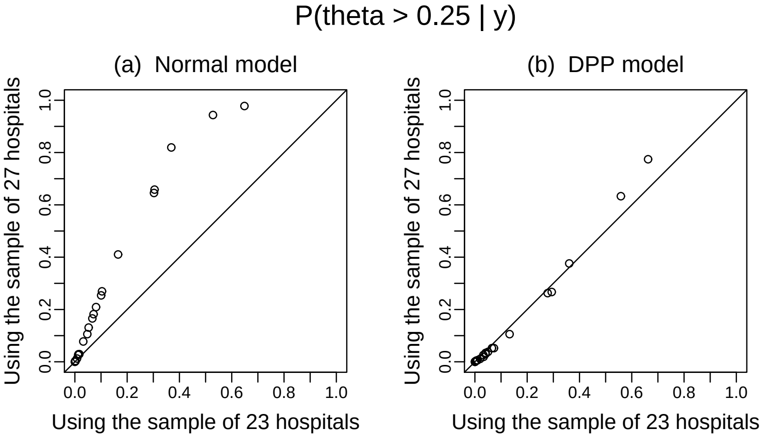

Figure 3: Scatterplots of the 23 paired posterior probabilities (corresponding to the 23 hospitals) that the failure rate exceeds a threshold of 0.25, one from the analysis of the 23 hospitals and the other from the analysis of the 27 hospitals. Plot (a) is the fit of the normal model and plot (b) is the fit of the DPP model.

Proceeding in the same manner, we fit the normal random-effects model, (2.1), and the DPP model, (2.2), to the second sample (27 hospitals). Do the compatible results from the two models seen in the original sample remain the same with the addition of these four hospitals? The results are summarized in Table 2. Unlike the compatible results seen in Table 1, there are substantial differences in both the posterior means and posterior probabilities; compare columns 6 and 7, and columns 8 and 9. Additionally, Figure 1(b) compares the posterior probabilities, one from the normal model and the other from the DPP model, for the *same set of*

23 hospitals. As we see, the posterior probabilities in Figure 1(b) are dramatically different for the two models. Since all of the posterior probabilities for the normal model are higher than those for the DPP model, the normal model apparently overstates the likelihood of the event. In this analysis (using 27 hospitals), the normal model yielded five hospitals (1, 4, 5, 12, 14) among the original 23 that had posterior probabilities larger than 0.5 whereas the DPP model had only two hospitals (4, 5) (see columns 8 and 9 in Table 2). Here, the conclusions from the normal random-effects model are very different from those from the DPP model. Again considering inference for the original 23 hospitals, there are larger discrepancies in the posterior means between the two models (normal vs. DPP) when the data from the 27 hospitals are used rather than the data from the 23 hospitals - compare columns 6 and 7 in Tables 1 and 2. Summarizing, it is clear that inferences about the 23 individual hospital effects are very different under these two models in the 27 hospital sample, but similar in the 23 hospital sample.

We next present the same results in a different way. We do this to highlight the difference between the two models regarding how inferences are affected by the composition of the study sample. Figure 2(a) compares the 23 pairs of posterior means in the normal model with one using data from the initial sample (23 hospitals) and the other using the second sample (27 hospitals), whereas Figure 2(b) provides the same comparison for the DPP model. Figures 3(a)-(b) present the results for the posterior probabilities. These Figures clearly show that for the normal model there are important changes in inference by using the two different samples. Conversely, by using a DPP model, there are only minor changes. This is so because in the normal model the addition of the four outlying hospitals shifted the posterior distributions for the random effects to the right, resulting in more hospitals having larger right tail-area probabilities. Since the DPP model automatically downweights outliers, the DPP estimates exhibit only small changes over the two samples. These results illustrate that conclusions from the normal model are not robust to different study samples and are vulnerable to the presence of outlying data.

The effects of including the outlying clusters can also be seen from the individual profiles of the hospital effects. Figure 4 displays the posterior distributions of the hospital effects for nine hospitals. Fitting the normal model (dotted lines), the posterior distributions exhibit nontrivial changes both in the scale and shape when the study sample changes (thin dotted line for the original sample and thick dotted line for the second sample). However, the posterior distributions from the DPP model (solid lines) are less influenced by the presence of the outlying hospitals.

Clearly, these results show that the DPP model is effective in reducing the dependence of inferences on the composition of the study sample.

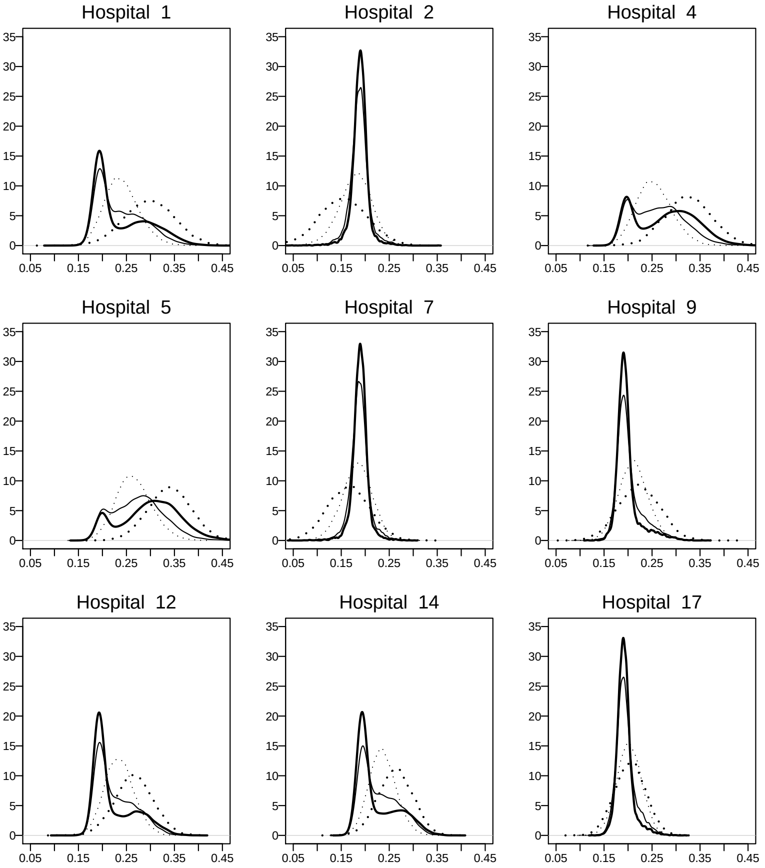

Figure 4: The posterior distributions of the failure rate for nine hospitals. Solid lines: the DPP model; Dotted lines: the normal model; Thin lines: using the original sample; Thick lines: using the second sample.

3.2 Shrinkage

Writing $\theta_i = \mu + \alpha_i$, it is well known that under the normal random effects model, (2.1),

$$E\{\theta_i|y, \delta^2\} = \lambda_i y_i + (1 - \lambda_i)\bar{y} \tag{3.1}$$

where $\lambda_i = \delta^2/(\delta^2 + \phi_i)$ and $\bar{y} = \sum_{k=1}^m \lambda_k y_k / \sum_{k=1}^m \lambda_k$. Thus, the relative weights on y_i and $\bar{y}$ in (3.1) depend on the ratio, ϕ_i/δ^2. Recall that δ^2 is the random effect variance whereas ϕ_i is the error variance.

It is well known that if (2.1) holds

$$E\{\theta_i|y\} = E_{\delta^2|y}\{\lambda_i y_i + (1 - \lambda_i)\bar{y}\} \tag{3.2}$$

provides improved inference for the θ_i and this is the estimator that Morris and Christiansen (1996) used to provide estimates similar to those in column 6 of Table 1. However, with the addition of the four outlying hospitals, the advantages from using (3.2) are lost because the between-hospital variance, δ^2, is very large relative to the within-hospital variance, ϕ_i. (The posterior median of δ^2 is 0.023 and the largest value of ϕ_i is 0.003.) Thus, for the 27 hospital sample $E(\theta_i|y) \approx y_i$, which can be seen in columns 4 and 6 of Table 2. However, for the 23 hospital sample (with no outliers), $E(\theta_i|y)$ is a weighted average of y_i and $\bar{y}$ with $E(\lambda_i|y)$ not near 1.

By contrast, using the DPP model the posterior means of the θ_i are approximately the same in Tables 1 and 2, as one may prefer because the data from hospitals 24-27 are outliers and should not affect the inferences about the original 23 hospitals. This contrast can be seen in Figure 3.

While there is no analytical expression for $E(\theta_i|y)$ in the DPP model, (2.2), it is clear (Escobar and West, 1998) that if the hospital effects are homogeneous, as in the first sample (23 hospitals), the posterior means from the normal and DPP models will be similar (as seen in Table 1). However, if there are subsets of the data that are quite different, as in the second sample (27 hospitals), inference about the hospital-specific effect θ_i will be based mostly on those hospitals whose values of y are close to y_i. That is, posterior inference for θ_i will be based on the data with values close to y_i. Thus, as we see from Figure 2(b), the posterior means of the θ_i using the DPP model are nearly the same with or without the outlying hospitals. This is so because the DPP properly ignores y_{24}, y_{25}, y_{26} and y_{27} when making inference about $\theta_1, \ldots, \theta_{23}$. Thus, DPP provides robust inference by downweighting the outlying hospitals.

3.3 Additional analyses

For further comparisons of (2.1) and (2.2) we have modified the original 23 hospital sample in different ways. We next summarize these results using typical examples. *First*, we added four hospitals to the original sample of 23 hospitals, but now the population failure rates for these hospitals are 0.4 rather than 0.6. Again, concerning the inference for the original 23 hospitals, the differences in the DPP posterior means and tail probabilities from fitting the original and new 27 hospital samples were smaller, as expected, than those from the normal model. This can be seen in Table 3 by comparing the results for "Case 0" and "Case 1." In Table 3, for a given fitted model (normal or DPP), $\triangle_i = E^*(\theta_i|y)$ - $E^{**}(\theta_i|y)$ was computed for each hospital where the first quantity is from using the original sample and the second one is from using the modified sample (here, the 23 and 27 hospital samples, respectively). The 25th, 50th and 75th sample percentiles of $|\triangle_i|$ are used to evaluate the effect of sample composition. The same summaries

Table 3: Distributions of the differences in posterior estimates from fitting the original and modified (addition/deletion of units) samples in normal and DPP models

	Posterior mean, $E(\theta_i\|y)$			Posterior probability, $P(\theta_i > 0.25\|y)$		
	25th$\|\triangle\|$†	Median $\|\triangle\|$	75th$\|\triangle\|$	25th$\|\triangle\|$‡	Median $\|\triangle\|$	75th$\|\triangle\|$
Case 0: Addition of four hospitals simulated from 0.6 true failure rate						
Normal model	0.007	0.012	0.026	0.008	0.078	0.205
DPP model	0.002	0.003	0.004	0.001	0.009	0.015
Additional cases						
Case 1: Addition of four hospitals simulated from 0.4 true failure rate						
Normal model	0.005	0.008	0.017	0.012	0.062	0.167
DPP model	0.001	0.004	0.005	0.002	0.017	0.028
Case 2: Addition of two hospitals simulated from 0.05 true failure rate						
Normal model	0.003	0.006	0.012	0.001	0.015	0.035
DPP model	0.000	0.001	0.004	0.000	0.002	0.005
Case 3: Deletion of hospital 16 (smallest rate)						
Normal model	0.001	0.002	0.003	0.001	0.003	0.008
DPP model	0.003	0.003	0.005	0.000	0.004	0.018
Case 4: Deletion of hospital 5 (largest rate)						
Normal model	0.001	0.003	0.005	0.004	0.014	0.037
DPP model	0.001	0.002	0.002	0.002	0.014	0.037
Case 5: Same as Case 0 but sample sizes n_i reduced to 5-20						
Normal model	0.021	0.031	0.039	0.241	0.290	0.330
DPP model	0.006	0.008	0.010	0.040	0.046	0.056

† Shown are the 25th, 50th and 75th sample percentiles of $|\triangle_i| = |E^*(\theta_i|y)$ - $E^{**}(\theta_i|y)|$ where the first quantity uses the original sample and the second one uses the modified sample.

‡ Shown are the 25th, 50th and 75th sample percentiles of $|\triangle_i(p)| = |P^*(\theta_i > 0.25|y)$ - $P^{**}(\theta_i > 0.25|y)|$ where the first quantity uses the original sample and the second one uses the modified sample.

are also provided in Table 3 for $\triangle_i(p) = P^*(\theta_i > 0.25|y)$ - $P^{**}(\theta_i > 0.25|y)$. For Case 0 (0.6 true failure rate for new outliers) the 75th percentile of $|\triangle_i(p)|$ is 0.205 for the normal model and 0.015 for the DPP model whereas the corresponding quantities for Case 1 (0.4 true failure rate for new outliers) are 0.167 and 0.028. *Second*, we added to the original dataset two hospitals with population rates of 0.05 and $n_i = 80$ for each one. Comparing the original and new 25 hospital samples, there are smaller changes (i.e., smaller $|\triangle_i|$ and $|\triangle_i(p)|$) when using the DPP model than the normal model, although the magnitudes of all of these changes are small (see Case 2 in Table 3). *Third*, we deleted hospital ID=16, i.e., the one with the smallest failure rate, 0.116. Here, there are minimal differences in the results between the original and new samples (23 and 22 hospitals, respectively) for *each* of the normal and DPP analyses. *Fourth*, we deleted hospital ID=5, i.e., the one with the largest failure rate, 0.347. Here, the two models yield similar results.

We have also created new 23 hospital and 27 hospital datasets by reducing the sample sizes (to values ranging from 5 to 20), but retaining the failure rates presented in Tables 1 and 2. By doing so, the sampling variances ϕ_i are larger in the newly created datasets. Here, the dominance of δ^2 over the ϕ_i seen in the original sample of 23 hospitals (Table 1) is now reduced, so the posterior mean of θ_i, (3.2), is a weighted average of y_i and $\bar{y}$ with $E(\lambda_i|y)$ not near 1. Then, for the new datasets (with larger ϕ_i) and the normal model, the posterior means for the original 23 hospitals are, as expected, much larger in the 27 hospital dataset than in the 23 hospital dataset. By contrast, the DPP model is robust; i.e., the values of the posterior means are similar in the two datasets (see Table 3, Case 5).

4. Discussion and Summary

In applications such as “provider profiling” finding “extrema” is of greatest interest. A common way to identify those medical providers (e.g., hospitals, doctors) who merit further investigation as possible outliers is to evaluate for each provider the posterior probability that the outcome variable exceeds a threshold. As seen in Section 3, these assessments may be problematic when using the standard normal model, (2.1), because the model is sensitive to outliers. For example, the probability that hospital 1 has a failure rate larger than 0.25 is 0.369 when using data from the initial sample (Table 1) but 0.82 when using data from the second sample (which has the four outlying hospitals) (Table 2). Similarly, using the first sample, only two hospitals (4 and 5) have posterior probabilities greater than 0.50 of exceeding the threshold while using the second sample there are five such hospitals (among the original 23 hospitals). In this example it is clear that the normal model is heavily data-dependent.

By contrast when using the DPP model the results (probabilities of exceeding the threshold) for both samples are essentially the same, as one would hope they would be. That is, in each sample only hospitals 4 and 5 (among the original 23 hospitals) have posterior probabilities greater than 0.50 of exceeding the threshold. Concerning inference for the original 23 hospitals, the DPP inferences are not only similar for the two different samples, but they are also similar to those from the normal model when there are no added outliers. That is, the DPP model in (2.2) adapts appropriately to the introduction of the outlying hospitals: It is much less likely that inferences will be influenced by the presence of spurious data.

In investigations such as clinical trials, meta analyses and many others there is often a choice of the units to be included, e.g., patients in longitudinal studies, hospitals in profiling studies, studies in meta analyses, possibly leading to a biased sample. In such circumstances one of the desirable statistical properties is that inferences be less influenced by the composition of the study sample. Our results indicate that inferences from the Dirichlet process specification in (2.2) are much less dependent on the composition of the study samples than the standard analysis using (2.1). Thus, under this criterion using the Dirichlet process specification is preferable.

A second, related, criterion for selecting a method to analyze data from investigations like those just noted is that the statistical model be sufficiently flexible to accommodate the heterogeneity inherent in such investigations. Using the Dirichlet process prior generalizes the standard method by accommodating the possibility that the data consists of several distinct groups of units. For this reason it is a more suitable technique than using the standard normal model.

An alternative procedure is to use the standard normal model, (2.1), and then employ diagnostic methods that may reveal the presence of distinct groups of units. Unfortunately, assessing the fit of a mixed effect model is challenging; there are limited diagnostic methods that have been shown to be effective (e.g., Verbeke and Molenberghs, 2000; Yan and Sedransk, 2007). However, if the diagnostics are successful in detecting outlying units (e.g., outlying hospitals) a common procedure is to exclude them from the analysis and then use the standard normal model to analyze the remaining data. This, too, may not be appropriate because it may ignore the true structure. We have shown that results from the Dirichlet process specification usually agree with those from using the standard method when the assumptions underlying the standard method are met. On the other hand, analyses using the Dirichlet process treat data appropriately when the units (e.g., hospitals) are in distinct groups. Thus, using a more flexible model, (2.2), reduces the importance of the diagnostics by modeling the true structure.

In summary, our investigation reveals that conclusions from the normal model are not robust to different study samples and are vulnerable to the presence of outlying data. Conversely, using a DPP model is effective in reducing the dependence of inferences on the composition of the study sample. With advances in computation and methodology it is straightforward to apply models such as (2.2), and analogous extensions of (1.2), routinely to medical studies.

References

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis.* John Wiley and Sons.

Dey, D. , Müller, P. , and Sinha, D. (eds). (1998). *Practical Nonparametric and Semiparametric Bayesian Statistics.* Springer-Verlag.

Diggle, P. J. , Liang, K. Y. , and Zeger, S. L. (1994). *Analysis of Longitudinal Data.* Oxford Science Publications.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* **90**, 577-588.

Escobar, M. D. and West, M. (1998). Computing nonparametric hierarchical models. In *Practical Nonparametric and Semiparametric Bayesian Statistics* (Edited by Dey, D. , Müller, P. , and Sinha, D.), Springer-Verlag.

Frühwirth-Schnatter, S. , Tüchler, R. , and Otter, T. (2004). Bayesian analysis of the heterogeneity model. *Journal of Business and Economic Statistics* **22**, 1-15.

Gelman, A. *et al.* (2004). *Bayesian Data Analysis*, 2nd Edition. London: Chapman & Hall/CRC.

Hui, S. L. and Berger, J. O. (1983). Empirical Bayes Estimation of Rates in Longitudinal Studies. *Journal of the American Statistical Association* **78**, 753-760.

Jara, A. (2007). Applied Bayesian Non- and Semi-parametric Inference using DPpackage. *Rnews* **7**, 17-26.

Kleinman, K. and Ibrahim, J. (1998). A semiparametric Bayesian approach to the random effects model. *Biometrics* **54**, 921-938.

Krnjajic, M. , Kottas, A. , and Draper, D. (2008). Parametric and nonparametric Bayesian model specification: A case study involving models for count data. *Computational Statistics & Data Analysis* **52**, 2110-2128.

Liu, J. S. (1996). Nonparametric hierarchical Bayes via sequential imputations. *Annals of Statistics* **24**, 911-930.

MacEachern, S. N. and Müller, P. (2000). Efficient MCMC schemes for robust model extensions using encompassing Dirichlet process mixture models. In *Robust Bayesian Analysis. Lecture Notes in Statist* **152**, 295-316. New York.

Morris, C. N. and Christiansen, C. L. (1996). Hierarchcial models for ranking and for identifying extremes, with applications. In *Bayesian Statistics 5* (Edited by Bernardo, J. M. , Berger, J. O. , Dawid A. P. and Smith A. F. M.) Oxford University Press.

Müller, P. and Quintana, F. A. (2004). Nonparametric Bayesian data analysis. *Statistical Science* **19**, 95-110.

Müller, P. , Quintana, F. , and Rosner, G. (2007). Semiparametric Bayesian Inference for Multilevel Repeated Measurement Data. *Biometrics* **63**, 280-289.

Normand, S-L, Glickman, M. , and Gatsonis, C. (1997). Statistical methods for profiling providers of medical care: Issues and applications. *Journal of the American Statistical Association* **92**, 803-814.

Ohlssen, D. I. , Sharples, L. D. , and Spiegelhalter, D. J. (2007). Flexible random-effects models using Bayesian semi-parametric models: applications to institutional comparisons. *Statistics in Medicine* **26**, 2088-2112.

Pinheiro, J. C. , Liu, C. H. , and Wu, Y. N. (2001). Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution. *Journal of Computational and Graphical Statistics* **10**, 249-276.

Rosa, G. J. M. , Gianola, D. , and Padovani, C. R. (2004). Bayesian longitudinal data analysis with mixed models and thick-tailed distributions using MCMC. *Journal of Applied Statistics* **31**, 855-873.

Tao, H. *et al.* (1999). An estimation method for the semiparametric mixed effects model. *Biometrics* **55**, 102-110.

van der Merwe, A. J. and Pretorius, A. L. (2003). Bayesian estimation in animal breeding using the Dirichlet process prior for correlated random effects. *Genetics Selection Evolution* **35**, 137-158.

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.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data.* Springer-Verlag.

Yan, G. and Sedransk, J. (2007). Bayesian diagnostic techniques for detecting hierarchical structure. *Bayesian Analysis* **2**, 735-760.

Received March 2, 2009; accepted June 14, 2009.

Guofen Yan
Division of Biostatistics and Epidemiology
Department of Public Health Sciences
University of Virginia
PO Box 800717
Charlottesville, Virginia 22908-0717, USA
guofen.yan@virginia.edu

J. Sedransk
Department of Statistics
Case Western Reserve University
10900 Euclid Avenue
Cleveland, OH 44106-7054, USA
jxs123@case.edu