Statistics and Its Interface Volume 7 (2014) 477–486

Inference functions in high dimensional Bayesian inference

Juhee Lee* and Steven N. MacEachern†

Nonparametric Bayesian models, such as those based on the Dirichlet process or its many variants, provide a flexible class of models that allow us to fit widely varying patterns in data. Typical uses of the models include relatively low-dimensional driving terms to capture global features of the data along with a nonparametric structure to capture local features. The models are particularly good at handling outliers, a common form of local behavior, and examination of the posterior often shows that a portion of the model is chasing the outliers. This suggests the need for robust inference to discount the impact of the outliers on the overall analysis. We advocate the use of inference functions to define relevant parameters that are robust to the deficiencies in the model and illustrate their use in two examples.

Keywords and phrases: Nonparametric Bayes, Dirichlet process, Loss function.

1. INTRODUCTION

The timeless question of how to handle outliers in a data set has been debated since the earliest days of Statistics. One approach involves screening the data and ripping out cases that appear to be outliers before a subsequent, typically non-robust, analysis is performed. Equivalently, a model is expanded through inclusion of enough parameters to "knock out" the outlying cases. A second approach focuses on the use of inferential techniques that are resistant to the presence of outliers. These two approaches are embodied in the work of Gauss (least squares) and Laplace (least absolute deviations) on regression [24]. The two approaches have traditionally been viewed as opposites, but recent work shows that they can be encompassed in a single framework through penalized likelihood techniques [13, 22]. While much work on how to handle outliers has been classical in spirit, Bayesians have pursued these two approaches and added a third.

The primary Bayesian approach to handling outliers involves creation of a generative model for both the "good cases" and the "outliers". This view is in keeping with the purest of Bayesian philosophies, expressed, for example, in [20], where a Bayesian should, in principle, be able to express uncertainty about *all* unknowns in a single, comprehensive model. Taking a simple, normal theory inference problem concerning a single mean as an exemplar, typical models for outliers include the mean-shift models or variance inflation models that we describe in Section 2. These models mimic the classical approach of including extra parameters for the outlying cases and add a prior distribution on the parameters. Not knowing which cases are outliers, the model formally becomes a mixture model with a good component and an outlier component. In practice, it is hoped that the models will assign the outliers to the outlier component and that their impact will be eliminated for inference for the mean of the good component. The success of this method rests on the analyst's ability to properly model the distribution of the outliers as well as that of the good data.

The second Bayesian approach departs from the modelling tradition of Bayesian methods and instead focuses on producing an inferential strategy that performs well. With our exemplar, this is generally accomplished by placing a thick-tailed sampling density on the data. The presumed normal sampling distribution is replaced by a distribution that is not log-concave. The resulting update with Bayes' Theorem discounts the outliers, effectively removing them from the posterior calculation if they are extreme enough. With a focus on inference, we might expect this method to work well for some inferences but not for others, breaking the cohesiveness of a collection of Bayesian inferences. Indeed, the implementation we have just described focuses on estimation of the mean. We would not expect it to work well for probability statements about individual cases and, as a consequence would not expect good performance for measures that require a description of case-specific distributions such as a predictive distribution or the Bayes factor.

The third Bayesian approach takes a different tack. Rather than attempting to model good and bad components of the data or to use a relatively inflexible model that targets a specific inference, the problem is recast as density estimation. A Bayesian version of a density estimator is created, with the goal of estimating the density from which the data come, both good and bad. The models used for density estimation are high- or infinite-dimensional and can fit a wide array of patterns in the data. Typical models fall under the heading of nonparametric Bayesian models of one variety or another. Popular models for density estimation

*Corresponding author.
†This work was supported by the NSF under grant numbers DMS-1007682, DMS-1209194 and H98230-10-1-0202. The views in this paper are not necessarily those of the NSF.

include those based on the Dirichlet process [16], the Pólya tree [11], or a log-gaussian process [14].

In this work, we focus on the third approach, considering nonparametric Bayesian approaches which we have found to provide flexible, outlier-accommodating models for the data. Formally, these methods place a prior distribution on the space of distribution functions to express a full range of uncertainty about the form of the distribution function. The methods are designed to allow consistent estimation of essentially any distribution. Consequently, when outlying data are observed in a tail area (e.g., some cases are far from the bulk of the data), the posterior distribution assigns most of its mass to distributions that have a bump (or bumps) near the outlying data. The resulting density estimates derived from the posterior distribution show both the general pattern of the good data and the outliers' departure from this pattern [2, 17]. Since the resulting posterior distribution provides a comprehensive view of both good and bad data, but does not split the two parts into separate components, one might surmise that the traditional inference functions inherited from low-dimensional parametric Bayesian models will lead to suboptimal inference. This is exactly the pattern that we observe, with clarity in simple settings and more opaquely in more complicated settings. To repair our inferential paradigm, we advocate a more tailored use of inference functions in the analysis.

Classical statistics provides a wealth of information about robust inference functions. [9] is the classic reference for a wide variety of robust inference functions, including linear combinations of order statistics, trimmed means, and M-estimators. In the classical development, especially where M-estimators are concerned, the distribution from which the data come, F, is considered fixed but unknown and inference is cast as an optimization problem. A "loss function" is specified, and the loss function determines the inferential target. The target minimizes the expected loss. Under the squared error loss, the target is the mean; under the absolute error loss, the median; and so on. Relying on the convergence of the empirical cumulative distribution function to F, a wealth of results are derived on consistency and asymptotic normality of the resulting estimators. These inference functions can be applied to Bayesian posterior distributions. For the Bayesian, the inference function maps F into a summary, say η. The posterior distribution over F thus yields a posterior distribution over the summary η. The distribution of η is then summarized.

In the sequel, we investigate the sensitivity of Bayes decision problems to the inference function when the parameter space is high-dimensional. Our particular focus is on the contrast between simple, low-dimensional Bayesian models and flexible, high-dimensional Bayesian models. We find a difference between the two situations. For the simpler models, inference changes little as the inference function is varied; for the more flexible models, inference can vary considerably as the inference function varies. In general, inference under the outlier-accommodating (high-dimensional) models changes dramatically when outliers are removed from a data set and the inference function is not robust. Inference changes little when the inference function is robust. The pattern differs for typical parametric models where removal of outliers either (i) changes the inference considerably under a wide array of inference functions or (ii) changes the inference little under a wide array of inference functions. Whether pattern (i) or pattern (ii) appears depends on the details of the parametric model. Transitional cases which lie between these two patterns can also occur.

The remainder of the paper is organized as follows. Section 2 describes standard Bayesian approaches for outliers. Section 3 describes inference functions and our general approach to their use in Bayesian problems. Sections 4 and 5 report illustrative analyses of Newcomb's data and a set of longitudinal data on exercise histories. The last section concludes with a final discussion.

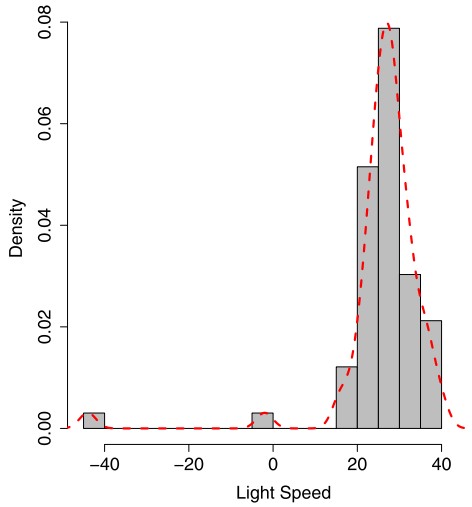
2. STANDARD APPROACHES

The standard Bayesian approaches to handling outliers is either to model them or to discount them by using a thick-tailed likelihood. We make these procedures more concrete with a simple illustrative model. Let $\boldsymbol{Y}_n$ be a vector of data consisting of n observations which are ideally conceived of as a random sample from some population. A Bayesian model for $\boldsymbol{Y}_n$ is specified by a likelihood for Y_i, $i = 1, \ldots, n$, conditional on parameters and a prior distribution for the parameters:

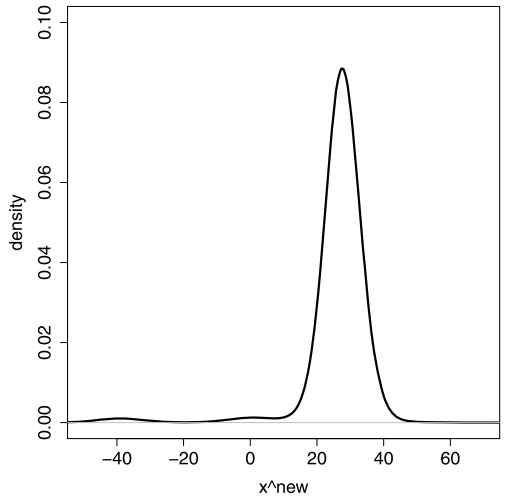
$$\begin{aligned} &\theta \mid G \sim G(\cdot), \\ &Y_i \mid \theta \sim F(\cdot \mid \theta), \text{ for } i = 1, \ldots, n. \end{aligned} \tag{1}$$

In a parametric model, F and G lie in parametric families, and they are often taken to be a conjugate pair. That is, they have a specific form such as F following a normal distribution with mean θ and known variance σ^2, leading to the conditional statement $Y_i \overset{indep}{\sim} \mathrm{N}(\theta, \sigma^2)$. The distribution G addresses uncertainty about the overall center of F, say $\mathrm{N}(\mu, \tau^2)$. This simple model leads to a posterior distribution for θ which is $\mathrm{N}((\tau^{-2}\mu + n\sigma^{-2}\bar{y})/(\tau^{-2} + n\sigma^{-2}), (\tau^{-2}+n\sigma^{-2})^{-1})$ where $\bar{y}$ represents the sample mean of $\boldsymbol{y}_n$. Whether we supplement the posterior distribution with one symmetric loss function or another makes little difference to the inference on the center. Quadratic loss, absolute loss, and others with thin enough tails all lead to θ as the center of F. This model does not admit the possibility of outliers in the data.

In practice, data often contain outliers. Aberrant points in the data arise for many reasons including "failure" of the experiment, data recording errors, and unexplained causes. Figure 1(a) presents a histogram and kernel density estimate of a famous data set from Newcomb's experiments on the passage time of light. Panel (b) shows a predictive density estimate from a mixture of Dirichlet processes (MDP)



(a) Histograms of observed times



(b) Predictive density estimate

Figure 1. (a) Histogram of observed passage times with a kernel density estimate. (b) Predictive density estimate under a mixture of Dirichlet processes model.

model. Details of the data and various analyses will be presented in Section 4. For now, we note that the data show an extreme outlier near -40 and a more modest outlier near 0.

Few applied statisticians would hesitate to trim the extreme outlier, and we believe that many would also trim the modest outlier before proceeding with their analysis. Rather than trimming the cases, we consider approaches that are more in keeping with Bayesian principles. We do so for two reasons: First, the outliers are easy to spot in this simple setting while they are often difficult to spot in more complicated settings. Second, if modelling approaches are to work anywhere, they should certainly work here. Oddly, the Bayesian principles we allude to are most clearly described in the non-Bayesian book, [9]. Their arguments are essentially those of Bayesian model averaging [7, 19] for a model supplemented with case indicators for potential outliers.

The most basic Bayesian outlier model supplements the basic distribution with a portion of the model devoted to outliers. A standard modification adds a mixture component to account for outliers: $\epsilon \sim \text{Beta}(\alpha, \beta)$ and $v_i \sim \text{Bernoulli}(\epsilon)$. If $v_i = 0$, then Y_i is drawn from the basic model; if $v_i = 1$, then Y_i is drawn from a $\text{N}(\theta, 3\sigma^2)$ distribution. An alternate form of the model specifies a location shift for each outlying case with the location-shift distribution centered at 0. With a complete model in hand, Bayesian inference proceeds via the usual prior distribution to posterior distribution update. Inference focuses only on the non-outlying component of the model.

The second Bayesian approach avoids attempting to model the outlying process directly, but appeals to properties of thick-tailed likelihoods. These likelihoods, particularly those which are not log-concave, downweight outliers. A traditional choice replaces the $\text{N}(\theta, \sigma^2)$ distribution in the basic model with a t-distribution with location θ, scale σ, and, say ν degrees of freedom. The resulting posterior distribution for θ is less influenced by outliers than is the posterior distribution from the basic model. There is generally no claim that this model captures the data-generating mechanism, but rather that it provides useful inference about the center of the distribution. It can be related to the mixture model for outlying data by noting that the t-distribution is a scale mixture of normals, setting the probability that an observation is an outlier to 1, and using a continuous mixture over variance shifts. This precludes the possibility of conditioning only on the non-outlying portion of the model for inference and implicitly mixes over "inliers" as well as outliers (corresponding to variance deflation). These first two strategies typically provide similar inference for the center of the distribution but may differ substantially for other inferences.

The third approach fits a nonparametric Bayesian model from the perspective of density estimation. Instead of modeling outliers with a distinct component of the model, we can accommodate them by using a flexible prior distribution that is able to capture nearly any pattern in the data. A standard model of this form is a mixture of Dirichlet processes (MDP) model. The model is referred to as "nonparametric" because it does not assume any specific form for G, that is, G is constructed with infinite number of parameters. Formally, instead of restricting the form of G to be in a parametric family, we consider the Dirichlet process (DP) prior for G [5].

$$G \sim \text{DP}(\alpha), \tag{2}$$

where α consists of two parts, the total mass parameter M and the base distribution G_0. To take advantage of the flexibility of G, we introduce a latent parameter θ_i into the model. Conditional on G, the θ_i are a random sample from G. The G in (2) is almost surely a discrete distribution function and it yields positive probability for ties in θ_i's. To

smooth the distribution, $Y_i \mid \theta_i \sim \mathrm{N}(\theta_i, \sigma^2)$, where σ^2 accounts for only a fraction of the variation in the basic model. The posterior distribution for G under this model is a mixture of Dirichlet processes [1].

We introduce a n-dimensional random cluster membership indicator vector $\boldsymbol{s} = (s_1, \ldots, s_n)$ to denote a partition of θ_i's into p ($\leq n$) clusters. Define the vector $\boldsymbol{s}$ by letting $s_i = j$ if and only if θ_i is in cluster j. We define clusters of θ_i's by matching their values. That is, we let $\theta_i = \theta_j^\star$ for all i's in cluster j, $j = 1, \ldots, p$. The common cluster-specific parameter $\theta_j^\star$ represents the location of cluster j. For partition $\boldsymbol{s}$, we let c_j be the size of cluster j and represent the sizes of the p clusters as a p-dimensional vector, $\boldsymbol{c} = (c_1, \ldots, c_p)$. The posterior simulation can be implemented using a Pólya urn scheme. That is, given $\theta_1, \ldots, \theta_n$, the parameter θ_{n+1} may either join cluster j with probability, $c_j/(M+n)$, or open a new singleton cluster by itself with probability, $M/(M+n)$. Given $\boldsymbol{s}$, $\theta_j^\star \overset{iid}{\sim} G_0$. The standard inference from a fit of this model proceeds through the predictive distribution, most commonly with its mean. [18] contains a recent review of DP based models, including details of computational methods.

3. ROBUST INFERENCE FUNCTIONS

We take as our goal estimation of the loosely defined center of the distribution from which the good data come. To do so, we focus on describing the posterior distribution of this center. To measure the center, we formalize our definition through the use of an inference function. The inference function is defined as a function that maps F into the target of inference, say η. That is, η is a measure of the center implied by a chosen inference function. The posterior distribution on F then determines a posterior distribution on η, and this opens the door to the usual summaries, both graphical and numerical, of the distribution of η.

The choice of inference function, $\mathrm{I}\{F\}$, implicitly defines our measure of the center of a distribution. The inference function is specified by a criterion which is to be minimized. For example, the quadratic loss criterion $\rho(y-\eta) = (y-\eta)^2$ leads to $\eta = \theta$ through the minimization;

$$\eta = \mathrm{argmin}_{\eta^*} \int \rho(y - \eta^*) dF(y). \tag{3}$$

Assuming that the mean exists (we henceforth assume that our inference functions all lead to unique minimizers), our measure of the center becomes the mean. In a similar fashion, an absolute loss inference function targets the median of F. A robust inference function based on Huber's loss targets the minimizer of equation (3) under a quadratic loss which is linearized in the tails [9]. For these inference functions and others, the inference provides a mapping, $\mathrm{I}\{F\} \to \eta$.

For the basic model or the standard Bayesian outlier model (mixture or t-distribution), the inference functions above all match η to θ. That is, when F has mean θ, $\mathrm{I}\{F\} \to \theta$. The consequence of this under the basic model is that the posterior distribution on the center is the posterior distribution of θ, that is, $\mathrm{N}((\tau^{-2}\mu + n\sigma^{-2}\bar{y})/(\tau^{-2} + n\sigma^{-2}), (\tau^{-2} + n\sigma^{-2})^{-1})$. The specific choice of inference function is immaterial, so long as it is symmetric.

The mapping becomes much more important when we work with a flexible model, as is the MDP model. We retain our focus on a future observation, Y^{new} which, under our model is presumed to come from the convolution of G with a normal kernel. This distribution, call it H, has a posterior distribution. Conceptually, we wish to describe the posterior distribution of $\mathrm{I}\{H\}$. In practice, we fit the MDP model with a Markov chain Monte Carlo method and approximate a draw of H with a finite representation of H.

The details of inference for the hierarchical model with (1) and (2) are as follows. The Markov chain produces a sequence of draws of a finite-dimensional object connected to G. For the t-th sample from the Markov chain, we have a finite number of mixture components associated with observations in the data set and an infinite number of mixture components not associated with any observation. Making use of our earlier notation, the standard representation for a draw of the chain,

$$\begin{aligned} f^{(t)}(y^{new} \mid \boldsymbol{y}, \boldsymbol{s}^{(t)}, \boldsymbol{\theta}^{\star(t)}) = \\ \sum_{j=1}^{p} \frac{c_j^{(t)}}{M+n} f(y^{new} \mid \theta_j^{\star(t)}) \\ + \frac{M}{M+n} \int f(y^{new} \mid \theta^\star) g_0(\theta^\star) \mathrm{d}\theta^\star, \end{aligned} \tag{4}$$

splits the t-th sample into two terms. The first captures the components associated with data while the second concerns the remaining mixture components. To handle the second term, we sample $\theta^\star$ using the stick-breaking process [21] to provide a close approximation to H. The inference function is then used to map H into a value for its center, η. This process is repeated for many draws of the Markov chain, resulting in many values of η.

The flexibility of the distribution H implies that the choice of inference function matters. Individual draws of H will upweight or downweight different components of the mixture. If the tail of a distribution is upweighted, the mean will be pulled strongly in the direction of that tail while a more robust measure of center will move less. This is particularly relevant in the case of Newcomb's data, as the weight given to the two outliers will have large or small impact on η. The divergence of the different measures will be shown in the next section.

We note that our method differs from computing the center of the predictive distribution (which would lead to a single summary number). Our interest focuses on the center of the distribution from which the data arise, and under a Bayesian model this center has a posterior distribution. The distribution can be summarized in the usual ways, either

graphically through kernel density estimates or histograms, or numerically by moments or quantiles. When concerned with outliers, a good measure of center will be robust to modest changes in the data set. In particular, a good measure will have a posterior distribution that changes little when outliers are included in or removed from the data.

4. EXAMPLE 1: NEWCOMB'S DATA

We analyzed Newcomb's classic data set on the speed of light [23]. 66 measurements were taken on the passage time of light which were then transformed to create the standard data set. See Figure 1(a) for a histogram of the data with a kernel density estimate. The center of this distribution is around 30 nanoseconds and there appear to be one or possibly two outliers at -44 and -2 in the left tail. These two measurements are distant from many of the other measurements recorded.

Let Y_i denote the i-th measurement of passage time. The basic model consists of a normal likelihood with unknown mean and variance and is completed with a normal prior distribution for the unknown mean and an inverse-gamma prior distribution for the unknown variance.

$$\begin{aligned} \theta &\sim \mathrm{N}(\mu, \tau^2), \qquad & \sigma^2 &\sim \mathrm{IG}(a_\sigma, b_\sigma) \\ Y_i \mid \theta, \sigma^2 &\overset{iid}{\sim} & \mathrm{N}(\theta, \sigma^2). \end{aligned} \tag{5}$$

We contrast the behavior of several approaches, considering two main models and a variety of inference functions. The first model follows from the thick-tailed strategy. Instead of the normal likelihood in (5), we define $W_i = \sqrt{\nu/(\nu-2)}(Y_i - \theta)/\sigma$ for any $\nu > 2$ and let $W_i \sim t_\nu$ where ν is the degrees of freedom of a t distribution.

The second model pursues the flexible modelling strategy by placing a Dirichlet process on G. In this model, we use a split of the variation in σ^2, allocating one portion to the base cdf G_0 and the other portion to the smoothing kernel for $Y_i \mid \theta_i$. We have found this split to work well in many circumstances. It is due to [8].

$$\begin{aligned} \theta_0 &\sim \mathrm{N}(\mu, \tau^2), \qquad & k &\sim \mathrm{Unif}(0,1),\ \sigma^2 \sim \mathrm{IG}(a_\sigma, b_\sigma) \\ G &\sim & \mathrm{D}(MG_0) & \\ \text{where } M \text{ is fixed} & & \text{and } G_0 \text{ is } \mathrm{N}(\theta_0, (1-k)\sigma^2) & \\ \theta_i &\overset{iid}{\sim} & G & \\ Y_i \mid \theta_i, k, \sigma^2 &\overset{iid}{\sim} & \mathrm{N}(\theta_i, k\sigma^2), & \end{aligned}$$

where θ_0, σ^2 and k can be interpreted as measures of location, scale and smoothness, respectively. The model with the DP implicitly assumes an infinite mixture of normals for G. We note that although θ_0 is the mean of G_0, it is not the mean of the realized G. In this and other variants of the MDP model, techniques to handle this discrepancy must be used [3, 15].

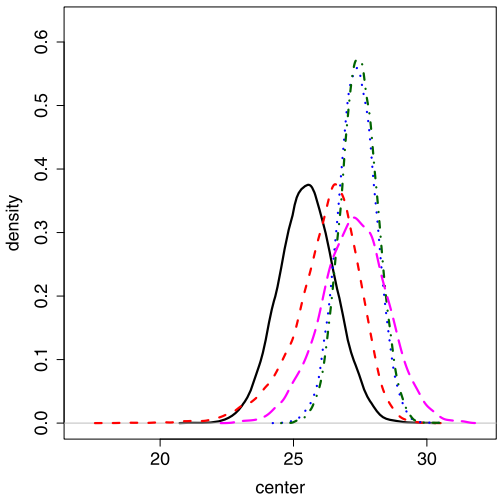
All of these models require values for the hyperparameters. For the parametric Bayesian model with the normal likelihood, let $\mu = 23.60$, $\tau^2 = 2.04^2$, $a_\sigma = 5$ and $b_\sigma = 10$. For the parametric Bayesian model with the t-distribution, we let $\nu = 4$ and used the same values of μ, τ^2, a_σ and b_σ. Similarly, we used the same hyperparameter values and fixed $M = 1$ for the MDP model.

The three models were fit to the data with standard Markov chain Monte Carlo methods. For each model, the fit resulted in T parameter vectors, harvested after each iterate of the Markov chain. After supplying an inference function, this resulted in T values for the center of the distribution, $\eta_1, \ldots, \eta_T$. For the MDP model, the computational technique mentioned in Section 2 was used to handle the infinite sum. Three inference functions were considered: those that generate the mean, the median, and Huber's estimator with a bending constant of 1.5 standard deviations. Under the normal model and the t model, all three inference functions map the distribution into θ. Under the more flexible MDP model, the inference functions result in different centers.

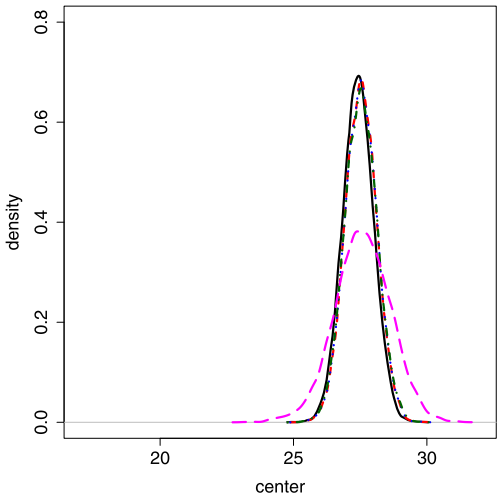
Figure 2(a) provides the posterior distributions of the center of the distribution of Y^{new} under the models and inference functions described above. The figure presents density estimates arising from the T center samples produced by the Markov chain Monte Carlo sampling algorithm. As expected, the posterior distribution of the center under the normal model is heavily impacted by the outliers on the low side. The thick-tailed t-distribution performs as advertised, effectively discounting the outliers and moving the distribution of the center to the right. For these two models, there are no differences between our three inference functions.

The MDP model shows the impact of the inference function. When the center is defined as the mean, the MDP model discounts outliers less than the thick-tailed t-distribution. This is a consequence of the small fraction of outliers (2 cases out of 66) and the mechanics of the MDP model which involve shrinking the latent θ_i for the outliers toward θ_0. The impact of the inference function is seen when we turn to the Huber estimator and the median as the measure of center. The distributions of these measures of center are stacked on top of one another and are difficult to distinguish in the figure. They are more concentrated than the other distributions and are moved toward the right due to a heavier discounting of the outliers.

The difference between the three inference functions under the MDP model is easily explained. The posterior distribution on H, the convolution of G and the smoothing kernel, assigns its probability to H that have two bumps in the far left tail. These bumps correspond to the two outlying observations. The H in keeping with the data vary in the weight given to these bumps, and the precise location of the bumps also varies. This variation results in a relatively large spread under the quadratic inference function (i.e., for the mean). In contrast, under an absolute inference function (i.e., for the median), the exact amount of mass assigned to the bumps and their precise location is immaterial. It is enough to know that a small amount of mass is located in



(a) All the observations



(b) Without two outliers

Figure 2. Distribution of the center of the distribution of Y_i with all the 66 observations in (a) and with 64 observations after deleting two outliers in (b). The black solid line is for the parametric Bayesian model with normal likelihood. The pink long-dashed line is for the parametric Bayesian model with the $t-$ likelihood. The red dashed, blue dotted, and dark green dash-dotted lines are for the MDP model with inference functions that yield the mean, the Huber estimate and the median, respectively.

the left tail. The result is a measure of center that is quite stable across plausible (in the posterior) distributions H.

Figure 2(b) illustrates the impact of removing the two outliers on the posterior distributions of the center. The distributions of the center do not move from panel (a) to panel (b) under the MDP model with the median and with Huber's estimator. In contrast, the distributions of the center under the normal model and under the MDP model with the mean are shifted greatly toward the right and become almost identical to those under the MDP model with the median and with Huber's estimator. Note that the spread of the distributions of the center decreases a lot under the normal model and the MDP model with the mean.

5. EXAMPLE 2: THE EXERCISE DATA

Blackmoor and Davis studied the connection between exercise and eating disorders [6]. Their study involved longitudinal tracking of 231 teenage girls. Among the 231 subjects, 138 subjects were hospitalized for an eating disorder and the remaining 93 subjects served as a control group. The subjects were followed over time and multiple observations from a subject were obtained at intervals of two years starting at age eight. Along with age, the estimated amount of exercise in hours per week was recorded. Since the subjects were hospitalized at different ages, the number of observations varies. Let Y'_{ij}, x'_{ij} and z_i represent the amount of exercise of subject i at the j-th time point, the age of the subject at the j-th time point and the binary group indicator for subject i (0 for control and 1 for the hospitalized group), respectively. Following the suggestion in [6] and consistent with diagnostic plots, we took a logarithmic transformation of Y'_{ij}, $Y_{ij} = \log_2(Y'_{ij} + 5/60)$. The addition of 5/60 follows [6] and is included to handle zero values of Y'_{ij}. Also, let $x_{ij} = x'_{ij} - 8$, subtracting the smallest age in the data set from all ages. We consider a hierarchical linear mixed model which allows subject-specific intercepts and slopes. These subject-specific parameters are random effects, allowing for subject-specific departures from the global pattern. The model consists of two levels: within a subject, we regress the amount of exercise on the subject-specific covariate, age (x_{ij}). At the subject level, the intercept and slope of subject i depend on its group indicator, z_i.

$$\begin{aligned} \text{Subject level} \quad &\Rightarrow \quad \begin{bmatrix} \beta_{i0} &= \gamma_{00} + \gamma_{01} z_i + u_{i0} \\ \beta_{i1} &= \gamma_{10} + \gamma_{11} z_i + u_{i1} \end{bmatrix} \\ \text{Within-subject level} \quad &\Rightarrow \quad Y_{ij} = \beta_{i0} + \beta_{i1} x_{ij} + r_{ij}, \end{aligned}$$

where $\boldsymbol{u}_i = (u_{i0}, u_{i1})$ represents the random effects for subject i and r_{ij} represents a random error for the j-th time point of subject i. The r_{ij} is drawn from $\mathrm{N}(0, \sigma^2)$ and $\boldsymbol{u}_i$ is drawn from $\mathrm{MVN}_2(\mathbf{0}_2, \Psi)$ where $\mathbf{0}_2$ represents $(0,0)^T$. Following the convention when working with random effects, the r_{ij} and $\boldsymbol{u}_i$ are (conditional on variance parameters) mutually independent. To complete the model, we assign prior distributions to the unknown regression parameters, $\boldsymbol{\gamma}_0 = (\gamma_{00}, \gamma_{01})$ and $\boldsymbol{\gamma}_1 = (\gamma_{10}, \gamma_{11})$, and to the variance parameters σ^2 and Ψ.

$$\begin{aligned} \boldsymbol{\gamma}_0 &\sim \mathrm{MVN}_2(\bar{\boldsymbol{\gamma}}_0, \Sigma_0), \qquad \boldsymbol{\gamma}_1 \sim \mathrm{MVN}_2(\bar{\boldsymbol{\gamma}}_1, \Sigma_1), \\ \sigma^2 &\sim \mathrm{IG}(a, b), \qquad \Psi \sim \mathrm{Inv-Wishart}(\Psi_0, \nu). \end{aligned} \tag{6}$$

We first apply the low dimensional model described above and examine whether or not there is evidence that a more complex model allowing full support for the distribution of

the random effects $\boldsymbol{u}_i$, G, is needed. To fit the parametric Bayesian model, we set the priors of $\boldsymbol{\gamma}_0$, $\boldsymbol{\gamma}_1$ and σ^2 to be centered at the estimates from a frequentist linear mixed model using the R function, *lme*. Thus, $\bar{\boldsymbol{\gamma}}_0 = (-0.276, -0.354)$, $\bar{\boldsymbol{\gamma}}_1 = (0.064, 0.240)$, $a = 3$ and $b = 0.77$. We use dispersed priors for the variances, taking $\Sigma_0 = \Sigma_1 = \Psi_0 = diag(1)$ and setting $\nu = 4$.

To examine the adequacy of the latent random effects distribution in the low dimensional parametric model, we use a diagnostic driven by preposterior expectation. The heart of the diagnostic is to note that, if one has written the correct model and has a large number of cases, two distributions should be approximately the same—the prior distribution of the random effects and the average of the posterior distributions of the effects, with the average taken across subjects. To pin this distribution down, we must condition on values for the fixed effects and the hyperparameters. Let $\boldsymbol{\xi} = (\boldsymbol{\gamma}_0, \boldsymbol{\gamma}_1, \Psi, \sigma^2)$ and fix $\boldsymbol{\xi}$. Let $Y_{ij}^u = Y_{ij} - (\gamma_{00} + \gamma_{01} z_i) - (\gamma_{10} + \gamma_{11} z_i) x_{ij} = u_{0i} + u_{1i} x_{ij} + r_{ij}$. Our model states that $\boldsymbol{u}_i \overset{indep}{\sim} \mathrm{MVN}_2(\mathbf{0}_2, \Psi)$ and $Y_i^u \mid \boldsymbol{u}_i \overset{indep}{\sim} \mathrm{N}(u_{0i} + u_{1i} x_{ij}, \sigma^2)$. The consequence of conditioning is that the data and latent variates for different subjects are now independent of one another.

The following calculation applies to each subject, and so we drop the subscript denoting the subject from the formulas. We use $g(\cdot)$ to denote the density of $\boldsymbol{u}$ and $f(\cdot \mid \boldsymbol{u})$ to denote the conditional density of $Y^u \mid \boldsymbol{u}$. Fix an arbitrary (measureable) set $\mathcal{A} \in \mathbf{R}^2$.

$$
\begin{aligned}
\mathrm{E}^{Y^u}\{\mathrm{P}(\boldsymbol{u} \in \mathcal{A} \mid Y^u)\} &= \int \int_{\mathcal{A}} g(\boldsymbol{u} \mid Y^u) \mathrm{d}\boldsymbol{u} h(Y^u) \mathrm{d}Y^u \\
&= \int_{\mathcal{A}} \int g(\boldsymbol{u}) f(Y^u \mid \boldsymbol{u}) \mathrm{d}Y^u \mathrm{d}\boldsymbol{u} \\
&= \int_{\mathcal{A}} g(\boldsymbol{u}) \mathrm{d}\boldsymbol{u} \\
&= \mathrm{P}(\boldsymbol{u} \in \mathcal{A}), \qquad (7)
\end{aligned}
$$

where $h(Y^u) = \int f(Y^u \mid \boldsymbol{u}) g(\boldsymbol{u}) \mathrm{d}\boldsymbol{u}$ is the marginal distribution of Y^u. This simple calculation establishes that the expected posterior probability content of $\mathcal{A}$ equals its prior probability content.

In this particular data set, we have 231 observations. The preposterior expectation result in (7) holds conditional on $\boldsymbol{\xi} = \boldsymbol{\xi}_0$ for each $\boldsymbol{u}_i$. The assumption which we intend to assess is that the form of the model is correct. Under this assumption, the form of the sampling distribution is correct and there also exists a value of $\boldsymbol{\xi}$, $\boldsymbol{\xi}_0$, for which the $\boldsymbol{u}_i$ form a random sample from $G(\cdot \mid \boldsymbol{\xi}_0)$. Under mild conditions and this assumption that the form of the model is true, the posterior distribution on $\boldsymbol{\xi}$ tends to a degenerate distribution at the true value [2], and so we fixed $\boldsymbol{\xi}$ at their posterior means based on all the observations. Alternatively, the estimates from the equivalent mixed linear effect model can be used for $\boldsymbol{\xi}_0$. In addition to providing a value of $\boldsymbol{\xi}$ to condition on, our large number of subjects enables us to appeal to the law of large numbers to establish the following approximate equality.

$$
\mathrm{E}^{Y^u}\{\mathrm{P}(\boldsymbol{u} \in \mathcal{A} \mid Y^u)\} \approx \sum_{i=1}^{n} \mathrm{P}(\boldsymbol{u}_i \in \mathcal{A} \mid Y_i^u)/n,
$$

where the calculation is understood to condition on $\boldsymbol{\xi}$. Applying this result to rectangular sets $\mathcal{A}$ implies that a histogrammed version of the prior distribution should match a histogrammed version of the average posterior distribution (here, using the average of the prior density over $\mathcal{A}$ and the average of the average of subject specific posterior densities over $\mathcal{A}$ to create the histograms). Extending this, we find that the prior density for $\boldsymbol{u}$ should be a near match for the average posterior density for $\boldsymbol{u}$.

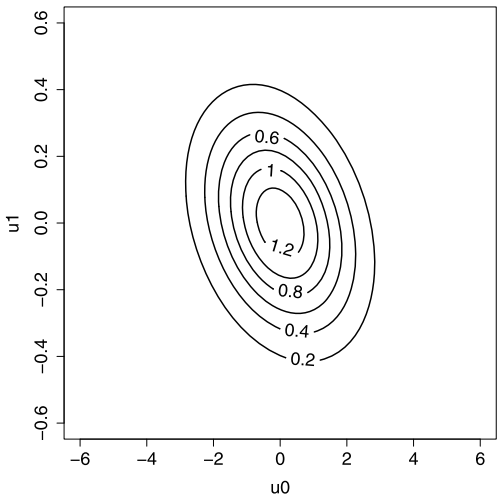
Figure 3 presents a graphical diagnostic motivated by the previous argument. The figure shows a discrepancy between the two densities, $g(\boldsymbol{u}_i \mid \boldsymbol{\xi}_0)$ and $\sum_{i=1}^{n} g(\boldsymbol{u} \mid Y_i^u, \boldsymbol{\xi}_0)/n$, which suggests that the presumed form is inappropriate for the data. The average posterior density shows left skewness for u_0, a slow "rotation" of contours as the density declines, and irregularity of contours in the low-density region. In other words, the posterior distribution of $\boldsymbol{u}_i$ departs from a bivariate normal distribution, and so we must consider a more elaborate model for its distribution, G. We have found this diagnostic to be useful in a range of problems involving latent random effects.

The model we consider stretches that given in (6) to a more complex (infinite dimensional) model through use of the nonparametric Bayesian technology. We assume the following MDP model

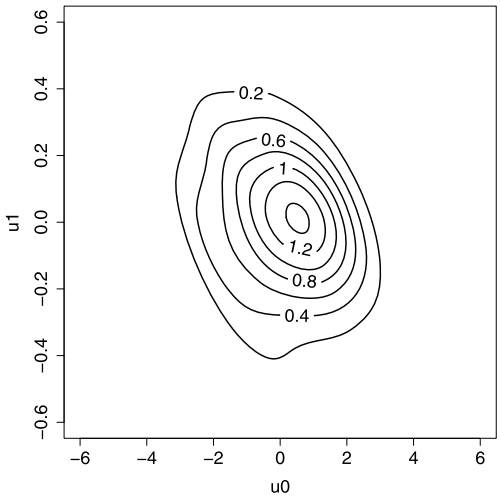
$$
\begin{aligned}
M \sim \mathrm{Gamma}(a_M, b_M), \qquad & \Psi \sim \mathrm{Inv-Wishart}(\Psi_0, \nu), \\
G &\sim \mathrm{DP}(M \mathrm{G}_0) \\
\text{where } G_0 &\text{ is } \mathrm{MVN}_2(\mathbf{0}_2, \Psi), \\
\boldsymbol{u}_i &\overset{iid}{\sim} G.
\end{aligned}
$$

We use the same hyperparameter values as in the parametric Bayeisan model. We calibrate the prior for M using function *DPelicit* in the R package, *DPpackage* [10] by setting the mean number of clusters and the standard deviation of the number of clusters among our 231 subjects to be 20 and 10. This yields $M \sim \mathrm{Gamma}(2.46, 2.06)$. The model was then fit via Markov chain Monte Carlo.

The estimated marginal predictive distributions of u_0 and u_1 have longer left than right tails as shown in Figure 4. Multimodality is evident in the predictive distribution of u_0. These features match the departures from normality suggested by the contour plots in Figure 3. They also suggest the need to consider a variety of inference functions and to examine the impact of the choice of inference function on the center of the distribution of Y^{new} at different covariate values. We consider three different ages, $x' = 8, 10, 12$. As in Section 4, we consider quadratic, Huber, and absolute inference functions, leading to the mean, Huber's summary and

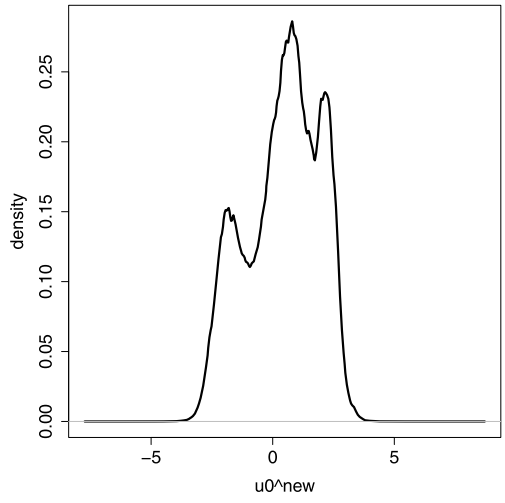


(a) $g(\boldsymbol{u} \mid \boldsymbol{\xi}_0)$

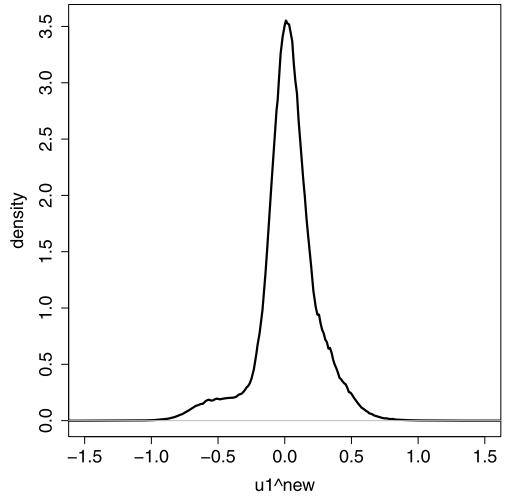


(b) Average of $g(\boldsymbol{u}_i \mid \boldsymbol{\xi}_0, Y_i^u)$

Figure 3. Contour plots of $g(\boldsymbol{u} \mid \boldsymbol{\xi}_0)$ in (a) and the average of $g(\boldsymbol{u}_i \mid \boldsymbol{\xi}_0, Y_i^u)$ over $n = 231$ subjects under the model assuming that $\boldsymbol{u}_i \overset{iid}{\sim} \mathrm{MVN}_2(\mathbf{0}_2, \Psi)$.



(a) predictive distribution of u_0



(b) predictive distribution of u_1

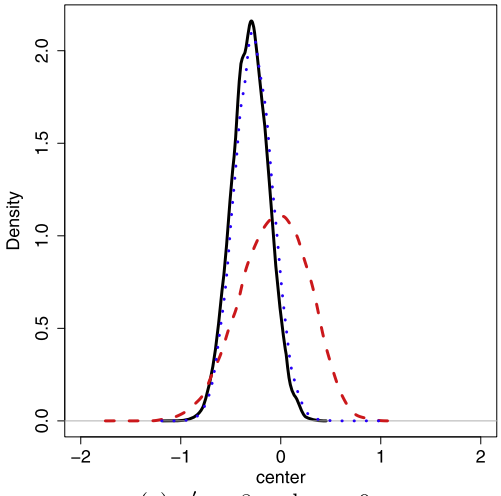
Figure 4. Estimates of the marginal predictive distributions of u_0 in (a) and u_1 in (b).

the median as definitions of the center. Figure 5 presents the posterior densties of the three measures of center at $x' = 8, 10, 12$ for the control group. Interestingly, we find that the distribution of the median is considerably less concentrated than that of the mean or Huber's summary. These data have no clear outliers to trim, and we believe the moderate non-normality of the random effects distribution is in keeping with use of the mean as a summary of the center of the distribution of Y^{new}.

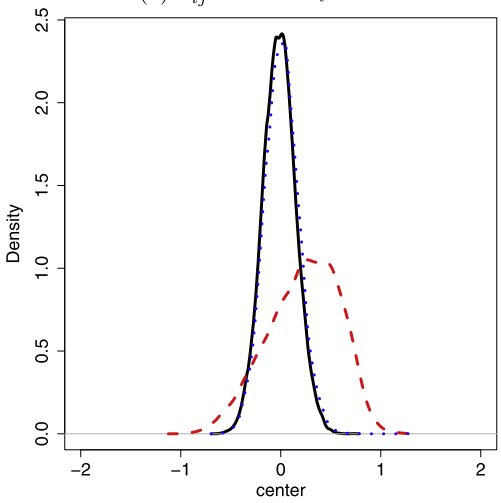
6. CONCLUSIONS

Today's data sets differ in fundamental ways from data sets in the past. They often fall in the realm of "big data", both in terms of sample size and in terms of the complexity of structures within the data. In these settings, it is often difficult to visually identify outliers. The models used to analyze those data have reacted to the novel features of the data, and the most successful of them are high- or infinite-dimensional (e.g., nonparametric Bayesian). These models do a superb job of picking up the global patterns in the data. They often operate much like a structured version of a density estimator, reacting to the local wiggles and other unusual features found in the data, whether these features represent the phenomenon under study (i.e., are good data) or do not (i.e., are outliers). The sheer size and complexity of the data also prevent us from realistically believing that we can write a formal model that both accounts for and distinguishes between the good data and the outliers.

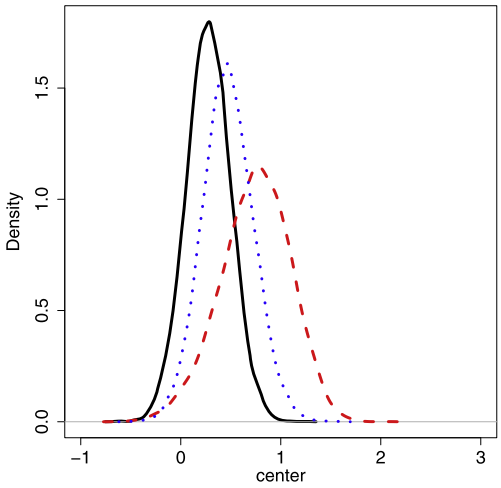
There are many arguments that support the use of flexible Bayesian models. One of the most compelling arguments



(a) $x'_{ij} = 8$ and $z_i = 0$



(b) $x'_{ij} = 10$ and $z_i = 0$



(c) $x'_{ij} = 12$ and $z_i = 0$

Figure 5. The posterior distribution of the center of the distribution of Y^{new} at $(x'_{ij}, z_i = (8, 0), (10, 0)$ and $(12, 0)$. The black solid lines, blue dotted lines, and red dashed lines show the distributions for the MDP model under quadratic, Huber, and absolute inference functions, respectively.

taps in to the value in enlarging the support of the prior distribution. This applies directly to models that include random effects, where the standard "independent draws from some parametric distribution" can easily be replaced by "independent draws from some distribution" [3]. Diagnostics such as the one we present in Section 5 are useful for confirming the need to move to a model with larger support. However, the value that comes from large support also comes at a cost—the inability to discount outliers in the fitting process as is done with a thick-tailed likelihood, for example. This leaves us in a quandary, with a preference for use of the flexible models but with a susceptibility of standard inferential methods to deficiencies in the data. The problem is aggravated by our inability to easily identify the outliers in complex-data settings.

Identification of the problems that come with the use of outlier-tracking models demands that we rethink how we make inference. Our recommendation is to think long and hard about the choice of inference function, and consequently, the target of inference. We note that our use of inference functions differs greatly from applying an inference function to the predictive distribution of a future observable or applying it as a loss function in a decision-theoretic context. In the former case, the inference function generates a single estimate; in the latter, the inference function is applied to a parameter which is already defined. In our context, the inference function defines the parameter of interest, and our methods allow us to examine the distribution of this parameter. Explicit consideration of the inference function broadens the collection of parameters that one might consider (e.g. expectiles [4]). Modern computational tools allow us to quickly explore a variety of inference functions, as shown in the examples.

The flexible models used in this paper have been based on the Dirichlet process. The details of these models do not drive our results, and similar results would be obtained for the many competing models, including Pólya trees, normalized random measures, and log-Gaussian processes. A promising direction for these models is the creation of a hybrid between use of a purely flexible model and writing a model that includes a component for outliers. We have seen the success of a head-and-tail strategy whereby the "head" of the model captures the good data and the "tail" of the model is used to sweep up the departures from the model [12]. These models suggest the use of hybrid inference strategies that make use of conditioning (on the head of the model) and inference functions (applied to the head).

Our work conveys the important message through the two examples that a good choice of the inference function can properly handle local features of the posterior distribution under a high-dimensional model. A robust inference function prevents inferences from being driven by local features but still reacts to the large-scale patterns in the data. One of remaining challenges is to develop classes of inference

functions appropriate for use in even more complicated situations.

Received 5 November 2013

REFERENCES

[1] Antoniak, C. E. (1974). Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The Annals of Statistics* 1152–1174. MR0365969

[2] Berger, J. O. (1993). *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag Inc. MR1234489

[3] Bush, C. A. and MacEachern, S. N. (1996). A semiparametric bayesian model for randomised block designs. *Biometrika* **83**, 2, 275–285.

[4] Efron, B. (1991). Regression percentiles using asymmetric squared error loss. *Statistica Sinica* **1**, 93, 125. MR1101317

[5] Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. *The Annals of Statistics* 209–230. MR0350949

[6] Fox, J. (2002). Linear mixed models. *Appendix to An R and S-PLUS Companion to Applied Regression*, URL http://cran.r-project.org/doc/contrib/Fox-Companion/appendix-mixed-models.pdf.

[7] George, E. I. and McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association* **88**, 423, 881–889.

[8] Griffin, J. (2010). Default priors for density estimation with mixture models. *Bayesian Analysis* **5**, 1, 45–64. MR2596435

[9] Huber, P. J. and Ronchetti, E. M. (2011). *Robust Statistics*. Wiley, 2nd edn. MR2488795

[10] Jara, A., Hanson, T. E., Quintana, F. A., Müller, P., and Rosner, G. L. (2011). Dppackage: Bayesian non-and semi-parametric modelling in r. *Journal of Statistical Software* **40**, 5, 1.

[11] Lavine, M. (1992). Some aspects of polya tree distributions for statistical modelling. *The Annals of Statistics* 1222–1235. MR1186248

[12] Lee, J. (2010). *Robust Statistical Modeling through Nonparametric Bayesian Methods*. Ph.D. thesis, The Ohio State University. MR2813988

[13] Lee, Y., MacEachern, S. N., and Jung, Y. (2012). Regularization of case-specific parameters for robustness and efficiency. *Statistical Science* **27**, 3, 350–372. MR3012431

[14] Lenk, P. J. (1988). The logistic normal distribution for bayesian, nonparametric, predictive densities. *Journal of the American Statistical Association* **83**, 402, 509–516. MR0971380

[15] Li, Y., Müller, P., and Lin, X. (2011). Center-adjusted inference for a nonparametric bayesian random effect distribution. *Statistica Sinica* **21**, 3, 1201. MR2827521

[16] Lo, A. Y. (1984). On a class of bayesian nonparametric estimates: I. density estimates. *The Annals of Statistics* **12**, 1, 351–357. MR0733519

[17] Müller, P. and Quintana, F. A. (2004). Nonparametric Bayesian data analysis. *Statistical Science* **19**, 1, 95–110. MR2082149

[18] Müller, P. and Rodriguez, A. (2013). Nonparametric Bayesian inference, vol. 9 of *NSF-CBMS Regional Conference Series in Probability and Statistics*, Chapter 3: Dirichlet Process, 23–41. Beachwood, Ohio, USA; and Alexandria, Virginia, USA: Institute of Mathematical Statistics and American Statistical Assocation. MR3113683

[19] Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association* **92**, 437, 179–191. MR1436107

[20] Savage, L. J. (1972). *The Foundations Statistics*. Dover Publications Inc., New York. MR0348870

[21] Sethuraman, J. (1994). A constructive definition of dirichlet priors. *Statistica Sinica* **4**, 639–650. MR1309433

[22] She, Y. and Owen, A. B. (2011). Outlier detection using nonconvex penalized regression. *Journal of the American Statistical Association* **106**, 494. MR2847975

[23] Stigler, S. M. (1977). Do robust estimators work with real data? *The Annals of Statistics* 1055–1098. MR0455205

[24] Stigler, S. M. (1986). *The History of Statistics: The Measurement of Uncertainty before 1900*. The Belknap Press of the Harvard University Press, Cambridge, Massachusetts. MR0852410

Juhee Lee
Department of Applied Mathematics and Statistics
University of California Santa Cruz
1156 High Street
Santa Cruz, CA 95064
USA
E-mail address: juheelee@soe.ucsc.edu

Steven N. MacEachern
Department of Statistics
The Ohio State University
1958 Neil Avenue
Columbus, OH 43210-1247
USA
E-mail address: snm@stat.osu.edu