Bayesian Analysis (2007) **2**, Number 2, pp. 365–384

Assessment of Locally Influential Observations in Bayesian Models

Russell B. Millar[1] and Wayne S. Stewart[2]

Abstract. In models with conditionally independent observations, it is shown that the posterior variance of the log-likelihood from observation i is a measure of that observation's local influence. This result is obtained by considering the Kullback-Leibler divergence between baseline and case-weight perturbed posteriors, with local influence being the curvature of this divergence evaluated at the baseline posterior. Case-weighting is formulated using quasi-likelihood and hence for binomial or Poisson observations, the posterior variance of an observation's log-likelihood provides a measure of sensitivity to mild mis-specification of its dispersion. In general, the case-weighted posteriors are quasi-posteriors because they do not arise from a formal sampling model. Their propriety is established under a simple sufficient condition. A second local measure of posterior change, the curvature of the Kullback-Leibler divergence between predictive densities, is seen to be the posterior variance (over future observations) of the expected log-likelihood, and can easily be estimated using importance sampling. Suggestions for identifying locally influential observations are given. The methodology is applied to a well known simple linear model dataset, to a nonlinear state-space model, and to a random-effects binary response model.

Keywords: Case sensitivity, Kullback-Leibler divergence, influence, local sensitivity, predictive density, posterior density, quasi-posterior

1 Introduction

Bayesian case-influence analysis requires some measure of the change in the posterior distribution when an observation is removed or down-weighted. For the normal linear model, Johnson and Geisser (1983) chose as their measure the Kullback-Leibler divergence between the predictive densities of the full-data and reduced-data posteriors. Johnson and Geisser (1985) and Guttman and Pẽna (1988, 1993) used the Kullback-Leibler divergence between the respective posterior densities. These works employed analytical approximations to the Kullback-Leibler divergence because of its intractability outside of the known variance case.

In the more general setting, Carlin and Polson (1991) demonstrated use of the Gibbs sampler to estimate the Kullback-Leibler divergence between full-data and reduced-data posteriors using an approach that required samples from both posteriors. Weiss (1996) and Weiss and Cho (1998) demonstrated estimation of Kullback-Leibler divergences (and other f-divergence measures, Csiszár [1967]) using a sample from the full-data posterior only. However, even in the linear model case, this procedure can be numerically

[1]Department of Statistics, Uni. Auckland, New Zealand, mailto:millar@stat.auckland.ac.nz
[2]Department of Statistics, Uni. Auckland, New Zealand, mailto:stewart@stat.auckland.ac.nz

unstable due to infinite variance of the sample average estimator of the conditional predictive ordinate (Weiss 1996; Peruggia 1997).

McCulloch (1989) took a local influence approach to model perturbation, in either the prior or likelihood. This approach assumed a family of models indexed by hyperparameters and investigated sensitivity to a small change in these. McCulloch quantified sensitivity by the curvature of the Kullback-Leibler divergence (between perturbed and unperturbed posteriors) with respect to the hyperparameters and evaluated at the unperturbed model. This curvature is well known to be the posterior Fisher information with respect to the hyperparameters (Kullback and Leibler 1951). McCulloch applied this approach to local case-influence in the normal linear model by considering sensitivity to hyperparameter w_i, where $Y_i|\boldsymbol{\beta}, \sigma^2 \sim N(\mathbf{x}_i^T\boldsymbol{\beta}, w_i^{-1}\sigma^2)$.

Here, we utilize a local influence approach to data-sensitivity that combines the approach of McCulloch (1989) with the weighted log-likelihood approach to local influence employed by Cook (1986). The focus will be on case-sensitivity, but we remark that our approach is applicable to any conditionally independent subset of the data. In the context of case-sensitivity, we evaluate the local influence of observation y_i by using a geometric weighting of the likelihood contribution from y_i. That is, y_i contributes $f_i(y_i|\boldsymbol{\theta})^{w_i}$ to the model for w_i in an open interval containing unity. Unlike the local perturbations considered in McCulloch (1989), these geometrically weighted likelihood terms do not, in general, correspond to density functions, or may not even be integrable. Consequently, the posterior densities obtained from using these modified likelihoods do not derive from a proper joint density and hence we refer to them as quasi-posteriors. In Section 2 it is shown that the quasi-posteriors are well defined under very mild conditions. The geometrically weighted likelihood is shown to be a natural way to alter case weight and the inverse weight is seen to correspond to an overdispersion term when the observations are from a binomial or Poisson distribution.

In Section 3 it is shown that, in the context of local sensitivity to observation i, the curvature of the Kullback-Leibler divergence between quasi-posteriors and the baseline posterior, (evaluated at $w_i = 1$) is simply the posterior variance of $\log f_i(y_i|\boldsymbol{\theta})$. Suggestions for identification of locally influential observations are given in Section 4. In Section 5 the curvature of the Kullback-Leibler divergence between quasi-predictive densities is seen to be the posterior variance (over future observations) of the expected log-likelihood conditional on both the observed and future observations. Section 6 includes analytical formulae in the context of multiple linear regression, with application to the Gesell adaptive score data of Mickey et al. (1967). This is followed by applications to a nonlinear state-space model of tuna biomass, and a repeated measures model of binary behavioral responses.

2 Perturbations of the likelihood

It will be assumed that the observations $y_i, i = 1, ..., n$ (possibly vector valued) are conditionally independent given $\boldsymbol{\theta} \in \Theta \subseteq R^p$. The density function (with respect to measure ν) for observation i will be denoted $f_i(y_i|\boldsymbol{\theta})$. Local case-sensitivity will be

assessed by evaluating the consequences of a small change in $f_i(y_i|\boldsymbol{\theta})$.

Change in $f_i(y_i|\boldsymbol{\theta})$ can be expressed in a variety of ways. For example, in the normal data models, specifying $\text{var}(Y_i) = w_i^{-1}\sigma^2, w_i > 0$, (e.g., McCulloch 1989) is an obvious way to alter the influence of observation i. More general perturbations of the sampling model for Y_i can be obtained, for example, by mixing f_i with some other suitable density (giving rise to a contamination class). Alternatively, the likelihood can be perturbed without the notion of an alternative sampling model. For example, in location-family models it may simply be of interest to evaluate the consequences of changing the datum value from y_i to $y_i + \epsilon$.

The case-weight approach taken below uses a quasi-likelihood type perturbation to directly alter the weight of information provided by the observation of datum y_i. These quasi-likelihoods need not correspond to a sampling model, but nonetheless, the resulting quasi-posterior is readily interpretable and is seen to be well defined under the weak sufficiency condition provided in Proposition 1 (Section 2.2).

2.1 Geometrically weighted likelihood

For a given i, we formulate dependence on w_i via weighted log-likelihood. That is, the contribution to the likelihood function from observation i is

$$L_i^{(q)}(\boldsymbol{\theta}; w_i) = f_i(y_i|\boldsymbol{\theta})^{w_i} \ . \tag{1}$$

The likelihood arising from the data with geometric weight w_i on observation i is denoted

$$L(\boldsymbol{\theta}; w_i) \quad = \quad f_i(y_i|\boldsymbol{\theta})^{w_i} \prod_{k \neq i} f_k(y_k|\boldsymbol{\theta}) \ .$$

The q superscript in (1) is used to denote quasi-likelihood because, as a function of y_i, $f_i(y_i|\boldsymbol{\theta})^{w_i}$ will not in general be a density function and may not be finitely integrable. For $0 \leq w_i \leq 1$, this quasi-likelihood has a natural interpretation as a geometric mixture of the likelihood $L_i(\boldsymbol{\theta}) = f_i(y_i|\boldsymbol{\theta})$ and the non-informative likelihood (uniform over the entire parameter space).

When Y_i is univariate with density that is of one-parameter exponential family form

$$f(y_i|\theta) = \exp\left\{[y_i\theta - b(\theta)]/a + c(y_i)\right\} \ , \tag{2}$$

the geometric weighted likelihood corresponds to the familiar quasi-likelihood of generalized linear modelling. To within a multiplicative constant $f_i(y_i|\boldsymbol{\theta})^{w_i}$ is simply given by replacing a by aw_i^{-1} in (2). Thus, if $L(p; y_i, n_i)$ is the likelihood function for observing proportion $\hat{p}_i = y_i/n_i$ successes from a Binomial experiment with n_i trials and success probability p_i, then the geometric weighted likelihood is equivalent to the same Binomial likelihood function but with the number of "trials" set to $w_i n_i$. For normal observations (with known variance σ^2), the geometric weighting is equivalent to the weighting given by changing the observation variance to $w_i^{-1}\sigma^2$. For Poisson(λ_i) observations, it is equivalent to evaluating the likelihood with "observed" value $w_i y_i$ and mean $w_i\lambda_i$.

In normal data models with unknown variance the geometrically weighted likelihood, $f_i(y_i|\boldsymbol{\theta})^{w_i}$ differs subtly from that arising from the $N(\mu_i, w_i^{-1}\sigma^2)$ model. The geometrically weighted likelihood attenuates the information content (of observation i) about all unknown parameters, including σ^2, and w_i close to zero approximates removal of y_i from the data. This is not the case for the $N(\mu_i, w_i^{-1}\sigma^2)$ model, because of the information content in this model about σ.

Some readers of an earlier draft suggested that, to avoid working with quasi-likelihood, the geometrically weighted likelihood could be normalized

$$f_i^*(y_i|\boldsymbol{\theta}; w_i) = \frac{f_i(y_i|\boldsymbol{\theta})^{w_i}}{\int f_i(y_i|\boldsymbol{\theta})^{w_i} d\nu(y_i)}$$

assuming existence of the denominator. However, this normalization violates the likelihood principle and we find it difficult to ascribe any meaningful interpretation to $f_i^*(y_i|\boldsymbol{\theta}; w_i)$ as a likelihood. For example, if Y_i is exponentially distributed with mean λ, then the normalized likelihood $f_i^*(y_i|\boldsymbol{\theta}; w_i)$ corresponds to observing y_i from an exponential density with mean $w_i^{-1}\lambda$. This makes no sense in the context of case sensitivity.

2.2 Propriety of quasi posteriors

For convenience, it will be assumed that the prior $\pi(\boldsymbol{\theta})$ is defined with respect to Lebesgue measure. Then, defining

$$\pi^{(q)}(\mathbf{y}, \boldsymbol{\theta}; w_i) = \pi(\boldsymbol{\theta}) f_i(y_i|\boldsymbol{\theta})^{w_i} \prod_{k \neq i} f_k(y_k|\boldsymbol{\theta})$$

and assuming that the integral

$$f^{(q)}(\mathbf{y}; w_i) = \int_\Theta \pi^{(q)}(\mathbf{y}, \boldsymbol{\theta}; w_i) d\boldsymbol{\theta}$$

is finite, the weighted-likelihood quasi-posterior

$$\pi^{(q)}(\boldsymbol{\theta}|\mathbf{y}; w_i) = \frac{\pi^{(q)}(\mathbf{y}, \boldsymbol{\theta}; w_i)}{f^{(q)}(\mathbf{y}; w_i)} \tag{3}$$

is a proper density function. The following proposition provides a very simple and natural sufficient condition for the propriety of (3).

Proposition 1. For any w_i, $0 \leq w_i \leq 2$, $f^{(q)}(\mathbf{y}; w_i)$ will be finite if both $\pi(\boldsymbol{\theta}|\mathbf{y}_{(-i)})$ and $\pi(\boldsymbol{\theta}|\mathbf{y}_{(+i)})$ are proper, where $\mathbf{y}_{(-i)}$ denotes the data with observation i removed and $\mathbf{y}_{(+i)}$ denotes the data consisting of $\mathbf{y}$ plus an identical copy of the datum value y_i with sampling model $f_i(y_i|\boldsymbol{\theta})$.

Proof. With y_i removed the data contribute the likelihood $\prod_{k \neq i} f_k(y_k|\boldsymbol{\theta})$, and with the addition of an identical copy of y_i, the data contribute the likelihood

$$f_i(y_i|\boldsymbol{\theta})^2 \prod_{k \neq i} f_k(y_k|\boldsymbol{\theta}) \ .$$

Now, for any $0 \leq w_i \leq 2$,

$$f_i(y_i|\boldsymbol{\theta})^{w_i} \leq 1 + f_i(y_i|\boldsymbol{\theta})^2, \quad \forall\ \boldsymbol{\theta} \in \Theta$$

and hence,

$$\begin{aligned} f^{(q)}(\mathbf{y}; w_i) &= \int_\Theta \pi(\boldsymbol{\theta}) f_i(y_i|\boldsymbol{\theta})^{w_i} \prod_{k \neq i} f_k(y_k|\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &\leq \int_\Theta \pi(\boldsymbol{\theta}) \prod_{k \neq i} f_k(y_k|\boldsymbol{\theta}) d\boldsymbol{\theta} + \int_\Theta \pi(\boldsymbol{\theta}) f_i(y_i|\boldsymbol{\theta})^2 \prod_{k \neq i} f_k(y_k|\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &< \infty\ . \end{aligned}$$

Proposition 1 extends in an obvious way to $w_i,\ 0 \leq w_i \leq n$ for any integer $n > 2$.

Note that, by application of Bayes rule, propriety of the data-reduced posterior $\pi(\boldsymbol{\theta}|\mathbf{y}_{(-i)})$ implies propriety of $\pi(\boldsymbol{\theta}|\mathbf{y})$ (with probability one). However, in the case of continuous data, it does not necessarily imply propriety of $\pi(\boldsymbol{\theta}|\mathbf{y}_{(+i)})$ because adding a copy of y_i corresponds to conditioning on a zero-probability subspace of the sample space (see Appendix A). In practice, notwithstanding pathological exceptions, we feel that the conditions of Proposition 1 are very weak and will be satisfied.

3 Local sensitivity to likelihood perturbation

3.1 Kullback-Leibler divergence and Fisher information

The Kullback-Leibler divergence (Kullback and Leibler 1951) is used here as a measure of the difference between two density functions. The directed Kullback-Leibler divergence between densities $\pi_a(\theta)$ and $\pi_b(\theta)$ (with respect to Lebesgue measure) is

$$\begin{aligned} I(\pi_a, \pi_b) &= E_{\pi_a}\left[\log\left(\frac{\pi_a(\boldsymbol{\theta})}{\pi_b(\boldsymbol{\theta})}\right)\right] \\ &= \int_\Theta \pi_a(\boldsymbol{\theta}) \log\left(\frac{\pi_a(\boldsymbol{\theta})}{\pi_b(\boldsymbol{\theta})}\right) d\boldsymbol{\theta} \end{aligned} \tag{4}$$

and can be interpreted as the information lost when π_b is used to approximate π_a (Burnham and Anderson 2002). The directed Kullback-Leibler divergence is not symmetric in its arguments and some authors (e.g., Pettit and Smith 1985; Guttman and Pena 1988, 1993) prefer to work with the symmetric Kullback-Leibler divergence given by

$$J(\pi_a, \pi_b) = I(\pi_a, \pi_b) + I(\pi_b, \pi_a)\ .$$

Assume a family of densities defined on $\Theta \subseteq R^p$ and indexed by $w \in R$ of the form

$$\{\pi_w(\boldsymbol{\theta});\quad 1 - \epsilon < w < 1 + \epsilon\} \tag{5}$$

for some ϵ greater than 0. Density π_1 will be referred to as the baseline density. The difference between π_w and the baseline density can be measured by the directed Kullback-Leibler divergence $I(\pi_w, \pi_1)$. In the context of local case-sensitivity, and in the spirit of McCulloch (1989), we consider the shape of the function $i(w) = I(\pi_w, \pi_1)$ at $w = 1$.

Kullback-Leibler divergences are non-negative and $i(w)$ is zero when $w = 1$. Consequently, assuming the appropriate third-order regularity conditions on derivatives of $\log \pi_w$ with respect to w (Kullback 1959, p. 26–27), the first derivative of $i(w)$ is zero at $w = 1$ and the second derivative of $i(w) = I(\pi_w, \pi_1)$, evaluated at $w = 1$, is

$$\ddot{I}^{\boldsymbol{\theta}} = \mathrm{var}_{\boldsymbol{\theta}} \left(\left. \frac{\partial \log \pi_w(\boldsymbol{\theta})}{\partial w} \right|_{w=1} \right) . \tag{6}$$

This second derivative is the Fisher information, with respect to w, evaluated at the baseline density. The Fisher information (6) is also the the second derivative of the directed Kullback-Leibler divergence $i^*(w) = I(\pi_1, \pi_w)$ (Kullback 1959) and hence the second derivative of the symmetric Kullback Leibler divergence $J(\pi_w, \pi_1)$ is twice the Fisher information.

3.2 Local case sensitivity

In the context of posterior sensitivity to observation i, the family of densities under consideration is of the form given in (3) for w_i in some open interval containing unity, for which Proposition 1 (Section 2.2) provided a sufficient condition. The directed Kullback-Leibler divergence between $\pi^{(q)}(\boldsymbol{\theta}|\mathbf{y}; w_i)$ and the baseline posterior is $i(w_i) = I(\pi^{(q)}(\boldsymbol{\theta}|\mathbf{y}; w_i), \pi(\boldsymbol{\theta}|\mathbf{y}))$. From (6), the second derivative of this divergence is

$$\ddot{I}_i^{\boldsymbol{\theta}} = \mathrm{var}_{\boldsymbol{\theta}|\mathbf{y}} \left(\left. \frac{\partial \log \pi^{(q)}(\boldsymbol{\theta}|\mathbf{y}; w_i)}{\partial w_i} \right|_{w_i=1} \right) \tag{7}$$

which is the Fisher information,with respect to w_i, evaluated at the baseline posterior. Regularity conditions for this result are considered in Appendix B.

Kullback-Leibler divergences between posterior distributions, and the corresponding curvatures, are not generally tractable. However, note that the normalizing constant $f^{(q)}(\mathbf{y}; w_i)$ in (3) does not depend on $\boldsymbol{\theta}$. Thus, (7) can be simplified to

$$\begin{aligned} \ddot{I}_i^{\boldsymbol{\theta}} &= \mathrm{var}_{\boldsymbol{\theta}|\mathbf{y}} \left(\left. \frac{\partial \log \pi^{(q)}(\mathbf{y}, \boldsymbol{\theta}; w_i)}{\partial w_i} \right|_{w_i=1} \right) \\ &= \mathrm{var}_{\boldsymbol{\theta}|\mathbf{y}} \left(\left. \frac{\partial \log (f_i(y_i|\boldsymbol{\theta})^{w_i})}{\partial w_i} \right|_{w_i=1} \right) \\ &= \mathrm{var}_{\boldsymbol{\theta}|\mathbf{y}}(l_i(\boldsymbol{\theta})) \end{aligned} \tag{8}$$

where $l_i(\boldsymbol{\theta}) = \log f_i(y_i|\boldsymbol{\theta})$.

We remark that (8) holds more generally for smooth perturbations of the baseline joint density $\pi(\mathbf{y}, \boldsymbol{\theta})$, in either prior or likelihood.

4 Identification of locally influential observations

Observation i can be identified as being of high relative local influence if $\ddot{I}_i^{\boldsymbol{\theta}}$ is substantially greater than the average local influence over all observations. However, it is also necessary to consider the magnitude of $\ddot{I}_i^{\boldsymbol{\theta}}$ because it is not necessarily the case that observations with high relative influence will be of concern. Conversely, it is possible that all observations could have high local influence without any one being substantively higher than the others. (In the known-variance linear regression example, equation (16) shows that under-specification of the variance will inflate $\ddot{I}_i^{\boldsymbol{\theta}}$ for all observations.)

The Kullback-Leibler calibration of McCulloch (1989) provides one approach to assessing the magnitude of $\ddot{I}_i^{\boldsymbol{\theta}}$. McCulloch (1989) showed that, for any $k \geq 0$, $q(k) = 0.5(1 + (1 - e^{-2k})^{1/2})$ is the unique value in the interval $[0.5, 1)$ such that the Kullback-Leibler divergence between Bern(0.5) and Bern(q) densities is k. $\ddot{I}_i^{\boldsymbol{\theta}}$ can be used to obtain the second order approximation of $i(w_i) = I(\pi_{w_i}, \pi_1)$ in a neighbourhood of unity and McCulloch's calibration can then be applied to this approximation evaluated at some w_i value close to unity. For example, if $\ddot{I}_i^{\boldsymbol{\theta}} = 0.5$ and we take $w_i = 0.8$, then the second-order approximation gives $I(\pi_{0.8}, \pi_1) \approx 0.5(1-0.8)^2/2 = 0.01$. A divergence of 0.01 corresponds to the information lost when a Bern(0.57) distribution is used to approximate the Bern(0.5) distribution.

However, we recommend that model complexity should be taken into consideration when deciding whether an observation has unduly high local influence. Gelman et al. (2004, p. 182) discuss two measures of model complexity. The first is that presented by Spiegelhalter et al. (2002), which can be written

$$p_D^{(1)} = 2\left(l(E_{\boldsymbol{\theta}|\mathbf{y}}[\boldsymbol{\theta}]) - E_{\boldsymbol{\theta}|\mathbf{y}}[l(\boldsymbol{\theta})]\right)$$

where $l(\boldsymbol{\theta}) = \log f(\mathbf{y}|\boldsymbol{\theta})$. The second measure is

$$p_D^{(2)} = 2\mathrm{var}_{\boldsymbol{\theta}|\mathbf{y}}(l(\boldsymbol{\theta}))\ . \tag{9}$$

Spiegelhalter et al. (2002, p. 591) show that $p_D^{(1)} \approx p$ in a model with p parameters when the prior information is negligible and the posterior is approximately normal. In the known-variance iid normal linear model of Section 6.1, $p_D^{(1)}$ and $p_D^{(2)}$ are both exactly equal to p.

Noting, in (9), that the variance of the log-likelihood is influenced by model complexity, we suggest assessing the local influence of the observations using

$$M_i = \frac{\mathrm{var}_{\boldsymbol{\theta}|\mathbf{y}}(l_i(\boldsymbol{\theta}))}{\mathrm{var}_{\boldsymbol{\theta}|\mathbf{y}}(l(\boldsymbol{\theta}))}\ . \tag{10}$$

(Note that $M_i, i = 1, ..., n$ do not in general sum to unity because the $l_i(\boldsymbol{\theta}), i = 1, ..., n$ are correlated.) Another possibility, based on using $p_D^{(1)}$ as the measure of model complexity, would be to use $l(E_{\boldsymbol{\theta}|\mathbf{y}}[\boldsymbol{\theta}]) - E_{\boldsymbol{\theta}|\mathbf{y}}[l(\boldsymbol{\theta})]$ in place of $\mathrm{var}_{\boldsymbol{\theta}|\mathbf{y}}(l(\boldsymbol{\theta}))$ in (10). However, we feel that the standardization in (10) is particularly natural. Note that $\mathrm{var}_{\boldsymbol{\theta}|\mathbf{y}}(l(\boldsymbol{\theta}))$ is the curvature of the Kullback-Leibler divergence with respect to a local weighting of all

observations, of the form $L(\boldsymbol{\theta}; w) = \prod_{i=1}^{n} f_i(y_i|\boldsymbol{\theta})^w$. Thus M_i is the local case-sensitivity due to y_i, standardized by the local sensitivity to the full data. In the examples in Section 6 we deemed an observation to be of high local influence if M_i exceeded $4/n$. That is, if $\ddot{I}_i^{\boldsymbol{\theta}}$ exceeded $2p_D^{(2)}/n$.

5 Local sensitivity of marginal and predictive distributions

More generally, interest may lie in posterior sensitivity of a measurable transformation $T(\boldsymbol{\theta})$ on Θ. Kullback and Leibler (1951) showed that the divergence between two probability spaces is at least as great as the divergence between any measurable-onto transformation of those probability spaces. That is, the Kullback-Leibler divergence $i(w_i) = I(\pi^{(q)}(\boldsymbol{\theta}|\mathbf{y}; w_i), \pi(\boldsymbol{\theta}|\mathbf{y}))$ can not increase under measurable-onto transformations of these posteriors, and consequently the curvature can not increase under such transformations. It follows that the curvature is invariant to invertible transformations of the parameters.

In the case that $\boldsymbol{\theta} = (\boldsymbol{\psi}, \boldsymbol{\lambda})$ where $\boldsymbol{\lambda}$ are nuisance parameters, then it will be the Kullback-Leibler divergence between $\pi(\boldsymbol{\psi}|\mathbf{y}; w_i)$ and $\pi(\boldsymbol{\psi}|\mathbf{y})$ that is of interest. In this case,

$$\ddot{I}_i^{\boldsymbol{\psi}} = \mathrm{var}_{\boldsymbol{\psi}|\mathbf{y}} \left(\left. \frac{\partial \log \pi^{(q)}(\boldsymbol{\psi}|\mathbf{y}; w_i)}{\partial w_i} \right|_{w_i=1} \right). \tag{11}$$

Applying eq. (6) of Millar (2004) in the context of likelihood perturbation, and assuming that $\pi^{(q)}(\boldsymbol{\theta}|\mathbf{y}; w_i)$ can be differentiated under the integral sign,

$$\left. \frac{\partial \pi^{(q)}(\boldsymbol{\theta}|\mathbf{y}; w_i)}{\partial w_i} \right|_{w_i=1} = \pi^{(q)}(\boldsymbol{\theta}|\mathbf{y})[l_i(\boldsymbol{\theta}) - E_{\boldsymbol{\theta}|\mathbf{y}}(l_i(\boldsymbol{\theta}))]$$

and hence

$$\begin{aligned} \left. \frac{\partial \pi^{(q)}(\boldsymbol{\psi}|\mathbf{y}; w_i)}{\partial w_i} \right|_{w_i=1} &= \left. \frac{\partial \int \pi^{(q)}(\boldsymbol{\theta}|\mathbf{y}; w_i) d\boldsymbol{\lambda}}{\partial w_i} \right|_{w_i=1} \\ &= \int [l_i(\boldsymbol{\theta}) - E_{\boldsymbol{\theta}|\mathbf{y}}(l_i(\boldsymbol{\theta}))] \pi^{(q)}(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\lambda} \, . \end{aligned}$$

Therefore,

$$\begin{aligned} \left. \frac{\partial \log \pi^{(q)}(\boldsymbol{\psi}|\mathbf{y}; w_i)}{\partial w_i} \right|_{w_i=1} &= \int [l_i(\boldsymbol{\theta}) - E_{\boldsymbol{\theta}|\mathbf{y}}(l_i(\boldsymbol{\theta}))] \frac{\pi^{(q)}(\boldsymbol{\theta}|\mathbf{y})}{\pi^{(q)}(\boldsymbol{\psi}|\mathbf{y})} d\boldsymbol{\lambda} \\ &= E_{\boldsymbol{\theta}|\boldsymbol{\psi},\mathbf{y}}[l_i(\boldsymbol{\theta}) - E_{\boldsymbol{\theta}|\mathbf{y}}(l_i(\boldsymbol{\theta}))] \end{aligned}$$

and so

$$\begin{aligned} \ddot{I}_i^{\boldsymbol{\psi}} &= \mathrm{var}_{\boldsymbol{\psi}|\mathbf{y}} \left(E_{\boldsymbol{\theta}|\boldsymbol{\psi},\mathbf{y}}[l_i(\boldsymbol{\theta}) - E_{\boldsymbol{\theta}|\mathbf{y}}(l_i(\boldsymbol{\theta}))] \right) \\ &= \mathrm{var}_{\boldsymbol{\psi}|\mathbf{y}} \left(E_{\boldsymbol{\theta}|\boldsymbol{\psi},\mathbf{y}}[l_i(\boldsymbol{\theta})] \right) \\ &= \ddot{I}_i^{\boldsymbol{\theta}} - E_{\boldsymbol{\psi}|\mathbf{y}} \left(\mathrm{var}_{\boldsymbol{\theta}|\boldsymbol{\psi},\mathbf{y}}[l_i(\boldsymbol{\theta})] \right) \, . \end{aligned} \tag{12}$$

Other times, sensitivity of predictive distributions may be of interest. Letting $\mathbf{y}_{rep}$ denote future observation(s) from $f_{rep}(\mathbf{y}|\boldsymbol{\theta})$, the predictive density of $\mathbf{y}_{rep}$ is given by

$$f(\mathbf{y}_{rep}|\mathbf{y}) = \int f_{rep}(\mathbf{y}_{rep};\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} \ .$$

The curvature of the Kullback-Leibler divergence between case-weight perturbed quasi-predictive densities and unperturbed predictive densities is given by the Fisher information

$$\ddot{I}_i^{\mathbf{y}_{rep}} = \mathrm{var}_{\mathbf{y}_{rep}|\mathbf{y}}\left(\left.\frac{\partial \log f_{rep}(\mathbf{y}_{rep}|\mathbf{y};w_i)}{\partial w_i}\right|_{w_i=1}\right) \ . \tag{13}$$

Using similar calculations to those used to obtain (12), it can be shown that

$$\begin{aligned}\ddot{I}_i^{\mathbf{y}_{rep}} &= \mathrm{var}_{\mathbf{y}_{rep}|\mathbf{y}}\left(E_{\boldsymbol{\theta}|\mathbf{y}_{rep},\mathbf{y}}[l_i(\boldsymbol{\theta})]\right) \\ &= \ddot{I}_i^{\boldsymbol{\theta}} - E_{\mathbf{y}_{rep}|\mathbf{y}}\left[\mathrm{var}_{\boldsymbol{\theta}|\mathbf{y}_{rep},\mathbf{y}}(l_i(\boldsymbol{\theta}))\right] \ .\end{aligned} \tag{14}$$

Moreover, from Weiss (1996, Theorem 3) it can be concluded that, under weak conditions, $I(\pi^{(q)}(\mathbf{y}_{rep}|\mathbf{y};w_i),\pi(\mathbf{y}_{rep}|\mathbf{y}))$ converges monotonically to $I(\pi^{(q)}(\boldsymbol{\theta}|\mathbf{y};w_i),\pi(\boldsymbol{\theta}|\mathbf{y}))$ as the dimension of $\mathbf{y}_{rep}$ increases. Note that the conditional variance in the second term of (14) would, under appropriate regularity conditions, become arbitrarily small as the dimension of $\mathbf{y}_{rep}$ was increased.

6 Examples

In Section 6.1, explicit formulae for $\ddot{I}_i^{\boldsymbol{\theta}}, \ddot{I}_i^{\mathbf{y}_{rep}}$ and $\ddot{I}_i^{y_{rep,i}}$ are obtained for multiple linear regression with known variance. In the unknown variance case, and in the nonlinear state-space (Section 6.2) and correlated Bernoulli data (Section 6.3) examples, calculation of the curvatures is obtained using posterior sampling.

To estimate $\ddot{I}_i^{\mathbf{y}_{rep}}$, note that

$$\pi(\boldsymbol{\theta}|\mathbf{y}_{rep},\mathbf{y}) \propto f(\mathbf{y}_{rep}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y}) \ .$$

For any value $\mathbf{y}_{rep}$, importance sampling can therefore be used to estimate $E_{\boldsymbol{\theta}|\mathbf{y}_{rep},\mathbf{y}}[l_i(\boldsymbol{\theta})]$ by

$$\frac{\sum_{k=1}^m f(\mathbf{y}_{rep}|\boldsymbol{\theta}^{(k)})\log f_i(y_i|\boldsymbol{\theta}^{(k)})}{\sum_{k=1}^m f(\mathbf{y}_{rep}|\boldsymbol{\theta}^{(k)})}$$

where $\boldsymbol{\theta}^{(k)}$ is a sample from $\pi(\boldsymbol{\theta}|\mathbf{y})$. We do not consider estimation of marginal local influence $\ddot{I}_i^{\boldsymbol{\psi}}$, but remark that this could be implemented using nested sampling (Weiss and Cho, 1998).

6.1 Linear regression

Consider the linear model with $Y \sim N_n(X\boldsymbol{\beta},\sigma^2 I_n)$, where X is $n \times p$ of rank $p < n$ and $\mathbf{x}_i^T$ denotes the ith row of X. With a uniform prior distribution on $\boldsymbol{\beta}$, and σ^2

assumed known, the Kullback-Leibler divergences and their curvatures may be obtained explicitly. Specifically, the distribution of $\boldsymbol{\beta}|\mathbf{y}$ is $N_p(Bb, B)$ where $B = \sigma^2(X^TX)^{-1}$ and $b = X^T\mathbf{y}/\sigma^2$, and $E[Y_i|\boldsymbol{\beta}] = \mathbf{x}_i^T\boldsymbol{\beta}$ has posterior mean and variance $\mu_i = \mathbf{x}_i^TBb$ and $\sigma_i^2 = \mathbf{x}_i^TB\mathbf{x}_i$. Therefore

$$\begin{aligned} \ddot{I}_i^{\boldsymbol{\beta}} &= \mathrm{var}_{\boldsymbol{\beta}|\mathbf{y}}\left(\log f_i(y_i|\boldsymbol{\beta})\right) \\ &= \frac{1}{4\sigma^4}\mathrm{var}_{\boldsymbol{\beta}|\mathbf{y}}\left((y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2\right) \qquad (15) \\ &= \sigma_i^2[2(y_i-\mu_i)^2 + \sigma_i^2]/(2\sigma^4) \\ &= h_i[(y_i-\mu_i)^2/\sigma^2 + h_i/2] \qquad (16) \end{aligned}$$

where $h_i = \mathbf{x}_i^T(X^TX)^{-1}\mathbf{x}_i$ is the leverage of covariate vector $\mathbf{x}_i$.

Denoting

$$D^{-1} = B^{-1} - \frac{1}{\sigma^2}\mathbf{x}_i\mathbf{x}_i^T \ , \qquad\qquad d = b - \frac{1}{\sigma^2}\mathbf{x}_i y_i \ ,$$

the Kullback-Leibler divergence between $\pi(\boldsymbol{\beta}|\mathbf{y}_{(-i)})$ and $\pi(\boldsymbol{\beta}|\mathbf{y})$ is

$$\begin{aligned} i_{0,i} &= I(\pi(\boldsymbol{\beta}|\mathbf{y}_{(-i)}), \pi(\boldsymbol{\beta}|\mathbf{y})) \\ &= -\frac{k}{2} - \frac{1}{2}\log\left(\frac{|D|}{|B|}\right) + \frac{1}{2}(\mathrm{tr}(B^{-1}D) + (Bb-Dd)^TB^{-1}(Bb-Dd)) \ . \end{aligned}$$

After some algebra this simplifies to

$$i_{0,i} = \frac{1}{2}\left[-1 + \log(1-h_i) + \frac{1}{1-h_i}\left(\frac{h_i(y_i-\mu)^2}{(1-h_i)\sigma^2} + 1\right)\right] \ . \qquad (17)$$

Both (16) and (17) combine measures of leverage and lack of fit. However, note that extremely high leverage observations (i.e., h_i close to unity) have a greater impact on case-deletion sensitivity than on local sensitivity.

To determine the local sensitivity of the predictive distribution, note that by definition, $\mathbf{y}|\boldsymbol{\beta}$ and $\mathbf{y}_{rep}|\boldsymbol{\beta}$ are independent and identically distributed $N_n(X\boldsymbol{\beta}, \sigma^2I_n)$, and therefore, conditional on $\mathbf{y}$ and $\mathbf{y}_{rep}$, $E[Y_i|\boldsymbol{\beta}] = \mathbf{x}_i^T\boldsymbol{\beta}$ has mean

$$\mu_{i,rep} = \mathbf{x}_i^T(X^TX)^{-1}X^T(\mathbf{y} + \mathbf{y}_{rep})$$

and variance $\sigma_{i,rep}^2 = \sigma_i^2/2$. From (15)

$$\mathrm{var}_{\boldsymbol{\beta}|\mathbf{y},\mathbf{y}_{rep}}\left(\log f_i(y_i|\boldsymbol{\beta})\right) = \frac{\sigma_i^2}{2}[2(y_i-\mu_{i,rep})^2 + \sigma_i^2/2]/(2\sigma^4) \ .$$

It is straightforward to show that $E_{\mathbf{y}_{rep}|\mathbf{y}}[\mu_{i,rep}] = \mu_i$ and $\mathrm{var}_{\mathbf{y}_{rep}|\mathbf{y}}(\mu_{i,rep}) = \sigma_i^2/2$, giving

$$E_{\mathbf{y}_{rep}|\mathbf{y}}\left[\mathrm{var}_{\boldsymbol{\beta}|\mathbf{y},\mathbf{y}_{rep}}\left(\log f_i(y_i|\boldsymbol{\beta})\right)\right] = \frac{\sigma_i^2}{4\sigma^2}[2(y_i-\mu_i)^2 + 3\sigma_i^2/2] \ .$$

After a bit of simplification we obtain from (14)

$$\ddot{I}_i^{\mathbf{y}_{rep}} = \frac{\ddot{I}_i^{\boldsymbol{\beta}}}{2} - \frac{h_i^2}{8} \,. \tag{18}$$

With similar calculations it can be shown that

$$\ddot{I}_i^{y_{rep,i}} = \frac{h_i}{1+h_i}\left(\ddot{I}_i^{\boldsymbol{\beta}} - \frac{h_i^2}{2(1+h_i)}\right) \tag{19}$$

where $y_{rep,i}$ denotes the ith element of $\mathbf{y}_{rep}$.

Gesell Data. Figure 1 shows Gesell adaptive score plotted against child's age (months) of first spoken word (Mickey et al. 1967). For the known variance scenario, we set σ^2 equal to the classical residual mean square, $s^2 = \sum_i^{21}(y_i - \mu_i)^2/19 = 121.5045$, where $\mu_i = \mathbf{x}_i^T Bb$, and the values of $i_{0,i}, \ddot{I}_i^{\boldsymbol{\beta}}, \ddot{I}_i^{\mathbf{y}_{rep}}$ and $\ddot{I}_i^{y_{rep,i}}$ were calculated analytically using equations (16)–(19). In the unknown variance case $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$ and the reference prior $\pi(\boldsymbol{\theta}) = \sigma^{-2}$, $\sigma > 0$ was used. The Kullback-Leibler divergence, $i_{0,i} = I(\pi(\boldsymbol{\theta}|\mathbf{y}_{(-i)}), \pi(\boldsymbol{\theta}|\mathbf{y}))$ was calculated using the approximation of Guttman and Pẽna (1988, eq. 5.3).

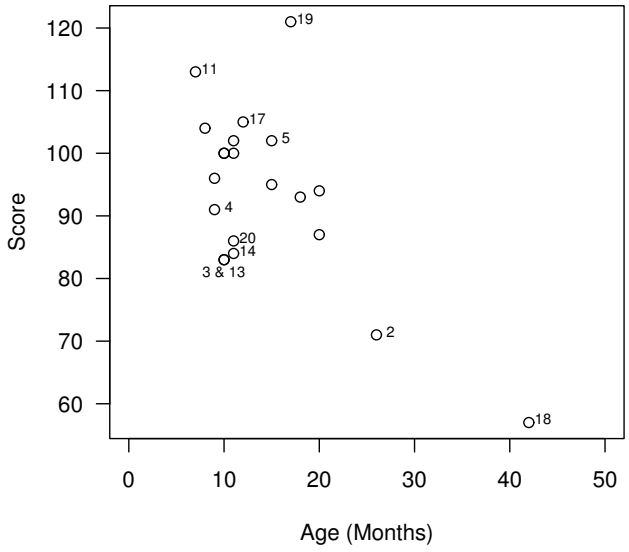

Figure 1: Gesell data from Mickey et al. (1967)

For the known σ^2 model, both $p_D^{(1)}$ and $p_D^{(2)}$ are exactly 2. Using the criterion of Section 4, observations with $\ddot{I}_i^{\boldsymbol{\beta}}$ exceeding $4/21 \approx 0.19$ are identified as having high local

influence. This identifies the high-influence observation 18 and the outlying observation 19 (Fig. 1, Table 1). The Kullback-Leibler divergence for case removal, $i_{0,i}$, is highest for observation 18. However, with regard to small changes in case weight, observation 19 ($\ddot{I}^{\boldsymbol{\beta}}_{19} = 0.40$) is slightly more influential than observation 18 ($\ddot{I}^{\boldsymbol{\beta}}_{18} = 0.38$). The high leverage of observation 18 is reflected in its relatively large value of $\ddot{I}^{y_{rep,i}}_{i}$.

				$\sigma^2 = s^2$				$\pi(\sigma^2) = \sigma^{-2}$			
i	x_i	y_i	h_i	$i_{0,i}$	$\ddot{I}^{\boldsymbol{\beta}}_{i}$	$\ddot{I}^{\mathbf{y}_{rep}}_{i}$	$\ddot{I}^{y_{rep,i}}_{i}$	$i_{0,i}$	$\ddot{I}^{\boldsymbol{\theta}}_{i}$	$\ddot{I}^{\mathbf{y}_{rep}}_{i}$	$\ddot{I}^{y_{rep,i}}_{i}$
1	15	95	0.05	0.00	0.00	0.00	0.00	0.00	0.03	0.02	0.00
2	26	71	0.15	0.09	0.13	0.06	0.02	0.06	0.13	0.06	0.01
3	10	83	0.06	0.07	0.13	0.06	0.01	0.07	0.16	0.08	0.01
4	9	91	0.07	0.03	0.05	0.02	0.00	0.00	0.05	0.03	0.00
5	15	102	0.05	0.02	0.03	0.02	0.00	0.00	0.04	0.02	0.00
6	20	87	0.07	0.00	0.00	0.00	0.00	0.00	0.03	0.02	0.00
7	18	93	0.06	0.00	0.01	0.00	0.00	0.00	0.03	0.02	0.00
8	11	100	0.06	0.00	0.00	0.00	0.00	0.00	0.03	0.02	0.00
9	8	104	0.08	0.01	0.01	0.00	0.00	0.00	0.03	0.02	0.00
10	20	94	0.07	0.02	0.03	0.01	0.00	0.00	0.04	0.02	0.00
11	7	113	0.09	0.06	0.09	0.05	0.01	0.03	0.09	0.05	0.01
12	9	96	0.07	0.01	0.01	0.00	0.00	0.00	0.03	0.02	0.00
13	10	83	0.06	0.07	0.13	0.06	0.01	0.07	0.16	0.08	0.01
14	11	84	0.06	0.05	0.09	0.04	0.00	0.03	0.09	0.05	0.01
15	11	102	0.06	0.01	0.01	0.01	0.00	0.00	0.03	0.02	0.00
16	10	100	0.06	0.00	0.00	0.00	0.00	0.00	0.03	0.02	0.00
17	12	105	0.05	0.02	0.03	0.02	0.00	0.00	0.04	0.02	0.00
18	42	57	0.65	1.09	0.38	0.14	0.10	1.05	0.39	0.12	0.08
19	17	121	0.05	0.22	0.40	0.20	0.02	1.69	1.53	0.71	0.06
20	11	86	0.06	0.04	0.06	0.03	0.00	0.01	0.06	0.03	0.00
21	10	100	0.06	0.00	0.00	0.00	0.00	0.00	0.03	0.02	0.00

Table 1: Local influence measures for the Gesell data, for known and unknown σ^2.

In the unknown σ^2 case, $p_D^{(1)} \approx 3.12$ and $p_D^{(1)} \approx 3.47$, resulting in a threshold of approximately 0.330 for identification of high local influence. Again, this identifies observations 18 and 19. Observation 19 now has largest value of both $i_{0,i}$ and $\ddot{I}^{\boldsymbol{\theta}}_{i}$, which may be attributed to the influence that this observation has on the posterior distribution of σ^2.

6.2 Nonlinear state-space model

Millar and Meyer (2000) applied a Bayesian state-space implementation of a Schaefer surplus production model to the South Atlantic albacore tuna catch-rate data of Polacheck et al. (1993). These data span the years 1967 to 1989, with 1967 being the first year of the fishery (Table 2). Under the Schaefer model, the biomass in year t is given by

$$\begin{aligned} B_{1967} &= K e^{u_{1967}} \quad , \quad t = 1967 \\ B_t &= (B_{t-1} + rB_{t-1}(1 - B_{t-1}/K) - C_{t-1})\, e^{u_t} \quad , \quad 1968 \leq t \ , \end{aligned} \tag{20}$$

where r is the intrinsic growth rate of the population, K is virgin biomass, C_t is the catch (assumed known) in year t, and u_t are iid Normal$(0, \sigma^2)$.

The catch rates, y_t, are assumed to have expected value that is proportional to biomass. Specifically, the catch rates are modelled as

$$y_t = qB_t e^{v_t} \ ,$$

where parameter q is the "catchability coefficient", and v_t are iid Normal$(0, \tau^2)$.

The model parameters are $\boldsymbol{\theta} = (K, r, q, \sigma^2, \tau^2 B_{1967}, ...B_{1989})$ and hence $y_t, t = 1967, ..., 1989$, are conditionally independent. Parameters K, σ^2, and τ^2 were given vaguely informative priors derived from expert knowledge. The prior for catchability was $\pi(q) = q^{-1}$. Intrinsic growth rate, r, was given an informative log-normal prior that was derived from analysis of other tuna stocks. These priors were assumed independent. The conditional prior $\pi(B_{1967}, ...B_{1989}|K, r, q, \sigma^2, \tau^2)$ is induced from the priors on K, r and σ^2, using (20).

The posterior distribution of the model parameters has greatest local sensitivity to the catch rates in years 1968, 1971 and 1984 (Table 1). These are all years in which the log catch rate, $\log y_t$, is considerably higher than expected under the model (Fig. 2, Millar and Meyer [2000]).

The model complexity estimates, $p_D^{(1)}$ and $p_D^{(2)}$ were vastly different, taking values 2.3 and 14.8 respectively. Using $p_D^{(2)}$, the high-influence threshold is approximately 1.29, and this is exceeded by only the 1968 catch-rate data. While it is the case that other years give high absolute values of local sensitivity (e.g., $\ddot{I}^{\boldsymbol{\theta}}_{1971} = 0.87$), they are not of sufficient magnitude to exceed the threshold.

In fisheries management, the unknown of greatest interest is the current biomass. In the context of these historical tuna data, spanning the years 1967 to 1989, the biomass of interest is B_{1990}. The posterior predictive density of B_{1990} has greatest local sensitivity to the catch rates in the earliest two years, 1967 and 1968, and the last year, 1989. The relatively high local sensitivity to the first two years can be attributed to these years being especially informative about virgin biomass, K.

6.3 Repeated Bernoulli Data

Lindsey (1993) developed a model for the repeated Bernoulli data from the behavioral experiment of Solomon and Wynne (1954). This experiment recorded whether or not a dog received an electric shock through the floor of its cage. The shock was avoidable if the dog jumped over a partition in its compartment within 10 s of a barrier being removed. The data are a binary sequence of results from 25 trials applied to 30 dogs. Lindsey (1993) models the probability that dog i receives a shock on trial k by

$$p_{ik} = a^{x_{ik}} b^{k-1-x_{ik}}$$

where x_{ik} is the number of avoidances that the dog has made in trials 1 to $k-1$. The dogs are assumed (conditionally) independent.

Year	Catch (1000s t)	Catch rate (kg/100 hooks)	$\ddot{I}_i^{\boldsymbol{\theta}}$	$\ddot{I}_i^{y_{1990}}$
1967	15.9	61.89	0.15	0.02
1968	25.7	78.98	1.51	0.09
1969	28.5	55.59	0.07	0.00
1970	23.7	44.61	0.34	0.01
1971	25.0	56.89	0.87	0.01
1972	33.3	38.27	0.35	0.01
1973	28.2	33.84	0.46	0.00
1974	19.7	36.13	0.07	0.00
1975	17.5	41.95	0.35	0.00
1976	19.3	36.63	0.07	0.00
1977	21.6	36.33	0.07	0.00
1978	23.1	38.82	0.10	0.00
1979	22.5	34.32	0.09	0.00
1980	22.5	37.64	0.14	0.00
1981	23.6	34.01	0.06	0.00
1982	29.1	32.16	0.06	0.00
1983	14.4	26.88	0.30	0.00
1984	13.2	36.61	0.55	0.00
1985	28.4	30.07	0.21	0.00
1986	34.6	30.75	0.07	0.00
1987	37.5	23.36	0.37	0.00
1988	25.9	22.36	0.08	0.01
1989	25.3	21.91	0.38	0.18

Table 2: Local influence measures for the tuna data

We used this example as implemented in the distribution of the WinBUGS package (Spiegelhalter et al. 1995), where it uses highly dispersed normal priors, truncated to $(-\infty, 0)$, on $\log a$ and $\log b$. That is, the prior is proper, but can be considered an approximation to the improper prior $\pi(a, b) = a^{-1}b^{-1}, 0 < a, b < 1$.

Local sensitivity was evaluated with respect to the vector $\mathbf{y}_i$ of 25 responses recorded for dog i. The model complexities, $p_D^{(1)}$ and $p_D^{(2)}$, were approximately 1.98 and 2.00, respectively, corresponding to a high-influence threshold of 0.133. Dogs 2 and 9 have the highest local sensitivities, $\ddot{I}_i^{\boldsymbol{\theta}}$, of approximately 0.51 and 0.56 respectively. Inspection of the data shows that dog 2 received 14 shocks in the first 15 trials (the highest of any dog), but no further shocks in the remaining 10 trials. Dog 9 appears to be the most "training-resistant" of the 30 dogs. It was the only dog to receive a shock in the last five trials, receiving shocks on trials 26 and 30. Dogs 22, 24 and 29 have local sensitivities

less than 0.24, but in excess of the threshold of 0.133. These three dogs, and dog 9, were the only four dogs to receive a shock during the last eight trials.

7 Conclusions

Local case-sensitivity provides an assessment of the sensitivity of posterior inference to a modest change in the weight given to an observation. The geometric weighted likelihood provides a natural method to alter observation weight. It corresponds to a linear weighting of the individual log-likelihood term and has the interpretation that the information content, measured as the Fisher information about $\boldsymbol{\theta}$ provided by observation i, is proportional to w_i.

Our measure of the local influence of observation i on $\pi(\boldsymbol{\theta}|\mathbf{y})$ is $\ddot{I}_i^{\boldsymbol{\theta}}$, the curvature (evaluated at the baseline posterior) of the Kullback-Leibler divergence between baseline and perturbed posteriors. This is easily estimated from a posterior sample, as the posterior variance of $\log f_i(y_i|\boldsymbol{\theta})$. Local influence on predictive densities is the curvature, $\ddot{I}_i^{\mathbf{y}_{rep}}$, of the Kullback-Leibler divergence between baseline and perturbed predictive densities, and is the posterior variance of $E_{\boldsymbol{\theta}|\mathbf{y}_{rep},\mathbf{y}}[\log f_i(y_i|\boldsymbol{\theta})]$. This can be estimated using importance sampling. Weiss (1996) showed that (under weak conditions) $\ddot{I}_i^{\mathbf{y}_{rep}}$ converges to $\ddot{I}_i^{\boldsymbol{\theta}}$ as the number of future predictions increases. Thus, even if interest is not specifically in sensitivity of $\pi(\boldsymbol{\theta}|\mathbf{y})$, $\ddot{I}_i^{\boldsymbol{\theta}}$ will nonetheless be a relevant measure of local sensitivity if the model is to be used for a large number of predictions.

Appendix A: Propriety counterexample

The following counterexample shows that propriety of $\pi(\boldsymbol{\theta}|\mathbf{y})$ does not imply propriety of $\pi(\boldsymbol{\theta}|\mathbf{y}_{(+i)})$.

Let Y be distributed $N(\mu, \sigma^2)$, and let the independent priors be

$$\begin{aligned} \pi(\mu) &\propto 1, \quad \mu \in R \\ \pi(\sigma) &= 1, \quad 0 < \sigma < 1 \ . \end{aligned}$$

For any outcome, $y_1 \in R$, $\pi(\mu, \sigma|y_1)$ is proper because the integral

$$\int_0^1 \frac{1}{\sigma} \int_{-\infty}^{\infty} e^{-\frac{(y_1-\mu)^2}{2\sigma^2}} d\mu \ d\sigma.$$

is finite (and equal to $\sqrt{2\pi}$).

Now, set $y_2 = y_1$. Under the model that y_1 and y_2 are independent observations from a $N(\mu, \sigma^2)$, propriety of $\pi(\mu, \sigma|y_1 = y_2)$ requires finiteness of the integral

$$\begin{aligned} \int_0^1 \frac{1}{\sigma^2} \int_{-\infty}^{\infty} e^{-\frac{(y_1-\mu)^2}{\sigma^2}} d\mu \ d\sigma &= \int_0^1 \frac{1}{\sigma^2} \sqrt{\pi}\sigma d\sigma \\ &= \sqrt{\pi} \int_0^1 \frac{1}{\sigma} \ d\sigma \ . \end{aligned}$$

This integral does not exist finitely.

Appendix B: Regularity conditions

Kullback (1959) gives third-order regularity conditions for the curvature result that was used to obtain equation (7). In the present context, these are conditions on the derivatives of $\pi^{(q)}(\boldsymbol{\theta}|\mathbf{y}; w_i)$ and $\log \pi^{(q)}(\boldsymbol{\theta}|\mathbf{y}; w_i)$. These can be established by appropriate conditions upon $\pi^{(q)}(\mathbf{y}, \boldsymbol{\theta}; w_i)$.

Note that $\pi^{(q)}(\mathbf{y}, \boldsymbol{\theta}; w_i)$ is continuous as a function of $w_i \in (0, 2)$, and assuming the conditions of Proposition 1, is absolutely bounded by an integrable $(d\boldsymbol{\theta})$ function for all $w_i \in (0, 2)$. Thus, $f^{(q)}(\mathbf{y}; w_i)$ is continuous (Billingsley 1979, p. 181) on $(0, 2)$. Let $A = (1 - \epsilon, 1 + \epsilon) \subseteq (0, 2)$ be such that $f^{(q)}(\mathbf{y}; w_i)$ is uniformly bounded away from zero for all $w_i \in A$.

We shall require that for all $w_i \in A$,

$$\left| \frac{\partial^j \pi^{(q)}(\mathbf{y}, \boldsymbol{\theta}; w_i)}{\partial w_i^j} \right| \leq M_j(\boldsymbol{\theta}) \quad , j = 1, 2, 3, \tag{21}$$

where each $M_j(\boldsymbol{\theta})$ is finitely integrable $(d\boldsymbol{\theta})$. From these conditions we have that $f^{(q)}(\mathbf{y}; w_i)$ can be differentiated to third-order under the integral sign (Billingsley 1979),

$$\frac{\partial^j f^{(q)}(\mathbf{y}; w_i)}{\partial w_i^j} = \int_\Theta \frac{\partial^j \pi^{(q)}(\mathbf{y}, \boldsymbol{\theta}; w_i)}{\partial w_i^j} d\boldsymbol{\theta} \ . \tag{22}$$

Assuming that $\pi^{(q)}(\mathbf{y}, \boldsymbol{\theta}; w_i)$ is strictly positive for all $\boldsymbol{\theta} \in \Theta$ and all $w_i \in A$, it follows that the derivatives (∂w_i) of $\log \pi^{(q)}(\boldsymbol{\theta}|\mathbf{y}; w_i) = \log \pi^{(q)}(\mathbf{y}, \boldsymbol{\theta}; w_i) - \log f^{(q)}(\mathbf{y}; w_i)$ exist up to third-order for $w_i \in A$. This is the first regularity condition.

The second set of regularity conditions requires that for all $w_i \in A$,

$$\left| \frac{\partial^j \pi^{(q)}(\boldsymbol{\theta}|\mathbf{y}; w_i)}{\partial w_i^j} \right| \leq G_i(\boldsymbol{\theta}) \ , i = 1, 2, 3, \tag{23}$$

where $G_i(\boldsymbol{\theta}), i = 1, 2$ are integrable over Θ. This is immediately satisfied because $f^{(q)}(\mathbf{y}; w_i)$ is uniformly bounded away from zero on A, and its first and second derivative are uniformly bounded, by (21) and (22). The final part of this second set of regularity conditions is that $E_{\boldsymbol{\theta}|\mathbf{y}}[G_3(\boldsymbol{\theta})]$ is absolutely bounded.

The third set of regularity conditions require that

$$\int_\Theta \frac{\partial^j \pi^{(q)}(\boldsymbol{\theta}|\mathbf{y}; w_i)}{\partial w_i^j} = 0, \quad j = 1, 2.$$

This follows from (23) (Billingsley 1979) and the fact that, under Proposition 1, $\pi^{(q)}(\boldsymbol{\theta}|\mathbf{y}; w_i)$ is a density function for all $w_i \in A$.

Note that

$$\begin{aligned}\left|\frac{\partial^j \pi^{(q)}(\mathbf{y},\boldsymbol{\theta};w_i)}{\partial w_i^j}\right| &= \left|\frac{\partial^j \pi(\boldsymbol{\theta}) f(\mathbf{y}_{(-i)}|\boldsymbol{\theta}) f_i(y_i|\boldsymbol{\theta})^{w_i}}{\partial w_i^j}\right| \\ &= |\pi^{(q)}(\mathbf{y},\boldsymbol{\theta};w_i) l_i(\boldsymbol{\theta})^j| \\ &\leq |\pi(\boldsymbol{\theta}) f(\mathbf{y}_{(-i)}|\boldsymbol{\theta}) l_i(\boldsymbol{\theta})^j (1 + f_i(y_i|\boldsymbol{\theta})^2)|, \quad j = 1,2,3,\end{aligned}$$

and so (21) will be satisfied if the integrals

$$\int_\Theta \pi(\boldsymbol{\theta}) f(\mathbf{y}_{(-i)}|\boldsymbol{\theta}) l_i(\boldsymbol{\theta})^3 d\boldsymbol{\theta} \tag{24}$$

and

$$\int_\Theta \pi(\boldsymbol{\theta}) f(\mathbf{y}_{(+i)}|\boldsymbol{\theta}) l_i(\boldsymbol{\theta})^3 d\boldsymbol{\theta} \tag{25}$$

exist finitely, where $\mathbf{y}_{(+i)}$ is defined in Proposition 1. These conditions are specifying finiteness of the posterior third moment of the log-likelihood, when y_i is removed, and when an additional copy of y_i (from sampling model $f_i(y_i|\boldsymbol{\theta})$) is added.

References

Billingsley, P. (1979). *Probability and Measure*. New York: Wiley.

Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference*. 2nd ed. New York: Springer-Verlag.

Carlin, B. P. and Polson, N. G. (1991). "An expected utility approach to influence diagnostics." *Journal of the American Statistical Association*, 86: 1013–1021.

Cook, R. D. (1986). "Assessment of local influence (with discussion)." *Journal of the Royal Statistical Society, Series B*, 48: 133–169.

Csiszár, I. (1967). "Information-type measures of difference of probability distributions and indirect observations." *Studia Scientiarum Mathematicarum Hungarica*, 2: 299–318.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004). *Bayesian Data Analysis*. 2nd ed. New York: Chapman and Hall.

Guttman, I. and Peña, D. (1988). "Outliers and influence: evaluation by posteriors of parameters in the linear model." In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, (eds.), *Bayesian Statistics 3*, 631–640. Oxford: Oxford University Press.

Guttman, I. and Peña, D. (1993). "A Bayesian look at diagnostics in the univariate linear model." *Statistica Sinica*, 3: 367–390.

Johnson, W. and Geisser, S. (1983). "A predictive view of the detection and characterization of influential observations in regression analysis." *Journal of the American Statistical Association,* 78: 137–144.

Johnson, W. and Geisser, S. (1985). "Estimative influence measures for the multivariate general linear model." *Journal of Statistical Planning and Inference*, 11: 33–56.

Kullback, S. (1959). *Information Theory and Statistics.* New York: John Wiley.

Kullback, S. and Leibler, R. A. (1951). "On information and sufficiency." *Annals of Mathematical Statistics*, 22: 79–86.

Lindsey, J. K. (1993). *Models for Repeated Measurements.* Oxford: Clarendon Press.

McCulloch, R. E. (1989). "Local model influence." *Journal of the American Statistical Association,* 84: 473–478.

Mickey, M. R, Dunn, O. J. and Clark, V. (1967). "Note on the use of stepwise regression in detecting outliers." *Computers and Biomedical Research,* 1: 105–111.

Millar, R. B. (2004). "Sensitivity of Bayes estimators to hyper-parameters with an application to maximum yield from fisheries." *Biometrics,* 60: 536–542.

Millar, R. B. and Meyer, R. (2000). "Non-linear state space modelling of fisheries biomass dynamics by using Metropolis-Hastings within-Gibbs sampling." *Journal of the Royal Statistical Society, Series C,* 49: 327–342.

Pettit, L. I. and Smith, A. F. M. (1985). "Outliers and influential observations in linear models (with discussion)." In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, (eds.), *Bayesian Statistics 2*, 473–494. Amsterdam: North Holland.

Peruggia, M. (1997). "On the variability of case-deletion importance sampling weights in the Bayesian linear model." *Journal of the American Statistical Association,* 92: 199–207.

Polackeck, T., Hilborn, R. and Punt, A. E. (1993). "Fitting surplus production models: comparing methods and measuring uncertainty." *Canadian Journal of Fisheries and Aquatic Science*, 50: 2597–2607.

Solomon, R. L. and Wynne, L. C. (1954). "Traumatic avoidance learning: the principles of anxiety conservation and partial irreversibility." *Psychological Review*, 61: 353–385.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002). "Bayesian measures of model complexity and fit (with discussion)." *Journal of the Royal Statistical Society, Series B,* 64: 583–639.

Spiegelhalter, D. J., Thomas, A., Best, N. G. and Gilks, W. R. (1995). *BUGS: Examples, Version 0.50,* vol. 1. Cambridge: Cambridge Medical Research Council Biostatistics Unit.

Weiss, R. (1996). "An approach to Bayesian sensitivity analysis." *Journal of the Royal Statistical Society, Series B,* 58: 739–750.

Weiss, R. E. and Cho, M. (1998). "Bayesian marginal influence assessment." *Journal of Statistical Planning and Inference,* 71: 163–177.