1st Symposium on Advances in Approximate Bayesian Inference, 2018 1–6

Sensitivity of Bayesian Inference to Data Perturbations

Lorenzo Masoero* LOM@MIT.EDU
William T. Stephenson* WTSTEPHE@MIT.EDU
Tamara Broderick TBRODERICK@CSAIL.MIT.EDU

1. Introduction

Any model for data analysis is necessarily an approximation of reality; in particular, then, in any realistic data analysis, data are not generally drawn from the posited model. Therefore, it behooves the data analyst to understand how robust their inferences and resulting decisions are to perturbations in the data. In early work, Tukey (1959) and Huber (1964) set out to quantify how much a statistic could change if an adversary arbitrarily changes an ε fraction of the data. In this case, the data is said to lie in an *ε-contamination neighborhood* of the observed data. Such extreme changes lead to pessimistic conclusions; for example, the mean of N datapoints is infinitely sensitive, in that it can vary by an infinite amount within an ε-contamination neighborhood for any $\varepsilon > 0$. Often a less dramatic perturbation may seem more natural. For instance, continuous data may be discretized by rounding or binning (see, e.g., the data used by Gelfand et al. (1990); Scollnik (2007)); this change corresponds to a bounded perturbation, rather than the arbitrary one of ε-contamination. Koh and Liang (2017) recently considered shifting each data point within a ball of small radius centered at its original location, but that work focused on shifts in an optimization framework. We here propose similar computable estimates of local sensitivity but for Bayesian analysis.

Bayesian inference is particularly appealing for modern data analysis in part due its desirable statistical properties and coherent aggregation and quantification of known, modeled sources of uncertainty. But it is often computationally intensive to use in practice, and these computational challenges are only compounded when attempting to assess data sensitivity. There is a large and rich literature on Bayesian robustness, but it has traditionally focused on assessing, and adapting to, changes in the model itself – including, but not limited to, the expression of prior beliefs. See Insua and Ruggeri (2000) for an overview. A useful concept from this literature that we repeat here is the distinction between *global* sensitivity, in which the output varies over inputs in some subspace, and *local* sensitivity, in which the inputs are perturbed infinitesimally and the change in output is assessed; see Sivaganesan (2000) for a discussion. Here we think of the inputs as the data rather than the model. As local approaches tend to lend themselves to more generically and efficiently computable measures, we focus here on proposing and evaluating a local data sensitivity measure for Bayesian analysis.

The only other local Bayesian data sensitivity measure that we are aware of is given by Clarke and Gustafson (1998), who assess the effect of data shifts by forming a Taylor

. * Denotes equal contribution

expansion around the observed data; however, they consider perturbations to the data that scale with the dataset size, for which a local approximation is unlikely to be accurate given modern dataset sizes. In Section 2, we instead consider perturbing each datapoint by a small fixed amount and give a tractable approximation to the resulting optimization problem. We go on to show that such an approximation is useful for understanding the effects of rounded data in Section 3. In Section 4 we give a conjecture, backed up by an experiment, that our metric is also useful for understanding the effects of likelihood misspecification.

2. Methods

Suppose we have a dataset $\boldsymbol{X} = \{X_1, \ldots, X_N\}$, and let $\theta \in \Theta$ be an unknown parameter of interest. In Bayesian inference, the model is described by a prior and likelihood; the output of Bayesian inference is the posterior distribution $p(\theta \mid \boldsymbol{X})$ over θ given by Bayes' Theorem – or more typically, an approximation of the posterior. The posterior expresses the knowledge of the practitioner about θ after having seen the data. Typically, it is summarized with functionals such as the posterior mean for point estimation and (co)variance for uncertainty.

Consider a scalar $\delta > 0$ and perturbed data $\boldsymbol{X'} = \{X'_1, \ldots, X'_N\} \in B_\delta(\boldsymbol{X}) \triangleq \{\boldsymbol{X'} : \forall n = \{1, \ldots, N\}, \|X'_n - X_n\|_2 \leq \delta\}$. Let $h : \Theta \to \mathbb{R}$ be a function describing a functional of interest $E(\boldsymbol{X'}; h) \triangleq \mathbb{E}_{p(\theta|\boldsymbol{X'})}[h(\theta)]$. E.g., $h(\theta) = \theta$ yields the mean. We want to compute how much this functional varies over $\boldsymbol{X'}$ within $B_\delta(\boldsymbol{X})$:

$$\max_{\boldsymbol{X'} \in B_\delta(\boldsymbol{X})} \left\{ |E(\boldsymbol{X}; h) - E(\boldsymbol{X'}; h)| \right\}. \tag{1}$$

There are three challenges with the optimization in Eq. (1): (A) The objective is typically non-convex and evaluating the objective at any point is computationally expensive, so finding the argmax $\boldsymbol{X'}$ is very difficult. (B) Even given access to the argmax, evaluating the objective would still require running posterior inference at least twice, and even twice is typically prohibitive for practitioners. (C) As stated, the optimization needs to be performed anew for each new h. To address all of these issues, we use a linear approximation:

$$\max_{\boldsymbol{X'} \in B_\delta(\boldsymbol{X})} \left\{ |E(\boldsymbol{X}; h) - E(\boldsymbol{X'}; h)| \right\} \approx \max_{\boldsymbol{X'} \in B_\delta(\boldsymbol{X})} \left\{ \sum_{n=1}^{N} \frac{\mathrm{d}E(\boldsymbol{X}; h)}{\mathrm{d}X_n}\Big|_{X_n} (X'_n - X_n) \right\}. \tag{2}$$

This immediately answers two of the above complaints about Eq. (1): (A) This linear function is maximized by moving each datapoint X_n by an amount δ in the direction $\mathrm{d}E(\boldsymbol{X}; h)/\mathrm{d}X_n$, giving us the approximate worst-case dataset $\boldsymbol{X'}$ in Eq. (1). (B) Eq. (2) then provides a straightforward way to approximate the value of the maximum in Eq. (1). This approximation defines our sensitivity measure:

$$S_\delta(\boldsymbol{X}; h) \triangleq \delta \sum_{n=1}^{N} \left\| \frac{\mathrm{d}E(\boldsymbol{X}; h)}{\mathrm{d}X_n} \right\|_2 = \delta \sum_{n=1}^{N} \left\| \mathrm{Cov}_{p(\theta|\boldsymbol{X})} \left[h(\theta), \frac{\partial \log p(X_n \mid \theta)}{\partial X_n} \right] \right\|_2. \tag{3}$$

The final equality follows from a calculation identical to the proof of Theorem 2.1 in (Giordano et al., 2018). Finally, point (C) is addressed by the fact that the covariance in Eq. (3) is taken under the *original* posterior. So if we compute $E(\boldsymbol{X}; h)$ using posterior samples $\theta^{(1)}, \ldots, \theta^{(T)} \sim p(\theta \mid \boldsymbol{X})$, we can use those same samples to compute $S_\delta(\boldsymbol{X}; h)$ for any h and δ of interest.

Implementation details. The main issues in computing Eq. (3) are sampling from the posterior $p(\theta \mid \boldsymbol{X})$ and computation of $\partial \log p(X_n \mid \theta)/\partial X_n$. These are easily automated using the probabilistic programming language Stan (Carpenter et al., 2017) and the automatic differentiation package autograd (Maclaurin et al., 2015). Although easily automated, we have observed that accurately estimating the covariance in Eq. (3) can require over an order of magnitude more posterior samples than needed to estimate $E(\boldsymbol{X}; h)$. Gustafson (1996) suggests one technique for reducing the variance of this Monte Carlo estimator; investigating this and other control variate techniques is a major direction for future work.

Approximation accuracy. One concern is that the approximations made above may be inaccurate. Specifically, we may fail to correctly find the worst-case data perturbations, and even if we do, the linear approximation of Eq. (2) may be of poor quality; we explore this issue in Appendix A, where we show the above approximations to be empirically accurate on a simple conjugate model.

3. An Application to Rat Growth Rate

As previously discussed, our approximation in Section 2 has an immediate application to understanding the effect of rounded data. The birats dataset from Gelfand et al. (1990)[1] contains $N = 30$ rats observed at $J = 5$ different times. The weight of rat n at time t_j is recorded as Y_{nt}, and we assume $Y_{nt} \sim N(\beta_{n1} + \beta_{n2} t_j, \sigma^2)$. We further assume that the coefficients β_n are distributed according to a multivariate normal $\beta_n \sim N(\mu_\beta, \boldsymbol{\Omega})$. A notable feature of this dataset is that the rats' weights have been rounded to the nearest gram.

To understand the effect of this rounding, we consider $\delta \in [-0.5, 0.5]$ and examine the sensitivity of the mean growth rate over time $\mathbb{E}[\beta_{n2}]$ for each rat $n = 1, \ldots, N$. Fig. 1 shows how S_δ from Eq. (3) predicts $\mathbb{E}[\beta_{n2}]$ to vary with this rounding. We discover an interesting fact about the first rat (rat number 0) in the dataset: the posterior under the original dataset predicts that the rat's weight is nearly constant over time ($\beta_{02} \approx 0$); however, under worst-case roundings, the expected slope of this rat's growth can actually vary from positive to negative, giving differing conclusions about this rat's growth. Note that the methods in Section 2 have two benefits in this case: first, they help us to discover the worst-case rounding of the data, and second, they allow us to compute the effect of this rounding without re-running inference for every rat.

4. Likelihood Misspecification in Gaussian Mixture Models

While rounded data is a common element in many applications, we conjecture our metric can be used more broadly to assess robustness to likelihood misspecification. Our conjecture is motivated by the following non-rigorous argument: simulated data $\boldsymbol{X}$ from a model of interest forms an empirical distribution that resembles the model's likelihood, while a perturbed dataset $\boldsymbol{X}' \in B_\delta(\boldsymbol{X})$ resembles data from a likelihood slightly outside of the specified model. Intuitively then, our sensitivity measure, which is exploring the worst-case shift of the data, is really exploring the worst-case misspecification of the likelihood. We illustrate this idea

1. Data is available on the Stan repository:
https://github.com/stan-dev/example-models/tree/master/bugs_examples/vol2/birats

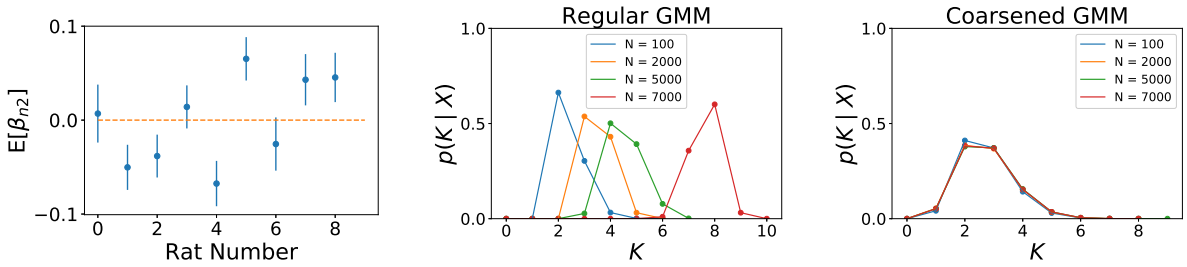

Figure 1: (*Left:*) Change in rats' growth rates from Section 3. We plot $\mathbb{E}[\beta_{n2}]$ for a subset of rats with small means, along with error bars showing the range of values predicted by our approximation with $\delta \in [-.5, .5]$. (*Right:*) GMM experiments from Section 4. Shown is the posterior over the number of active components $p(K \mid X')$ under the worst-case perturbed dataset $\boldsymbol{X}'$ for the regular (uncoarsened) and coarsened (with $\alpha = 10.0$) posteriors. Note the results in Fig. 1 come from re-running MCMC on the perturbed datasets $\boldsymbol{X}'$.

empirically on a recently proposed robustification of Gaussian mixture models (GMMs). Then, under the intuition that we are really measuring likelihood robustness, our metric empirically assess how robust the proposed model is. Consider the following generative model for a GMM:

$$\mu_{1:K} \overset{i.i.d.}{\sim} \mathcal{N}(\mu_0, \sigma_0^2),\ \sigma_{1:K} \overset{i.i.d.}{\sim} \text{Gamma}(a, b),\ \pi_{1:K} \overset{i.i.d.}{\sim} \text{Dir}(1/K),\ x_{1:N} \overset{i.i.d.}{\sim} \sum_{k=1}^{K} \pi_k \mathcal{N}(\mu_k, \sigma_k^2).$$

A potentially important question is whether, given an upper bound on the number of components, K_{max}, we correctly identify the number of significant components in the data, as measured by the function $h(\theta) = \sum_{k=1}^{K_{max}} \mathbb{1}[\pi_k > 0.05]$. Miller and Dunson (2018) attempt to robustify Bayesian models against misspecification of the likelihood by defining a *coarsened* posterior, which uses the same prior with a down-weighted likelihood $p(x_n \mid \theta)^{\alpha/(\alpha+N)}$ for a fixed $\alpha > 0$.

We generate data from a GMM with $K = 2$ components, and run the Metropolis-Hastings sampler from Miller and Dunson (2018) with $K_{max} = 20$. We compute the worst-case data shift predicted by Eq. (3) for the regular and coarsened posterior. The results in Fig. 1 reveal a trend close to that shown in Figure 3 of Miller and Dunson (2018): under our worst-case perturbation, the regular posterior $p(K \mid \boldsymbol{X}')$ concentrates on a larger K as N increases, whereas the coarsened posterior remains unchanged. We view this as a verification of the robustness of coarsening; while Miller and Dunson (2018) show coarsening to be empirically successful against ad-hoc perturbations, we empirically show that it is also robust against worst-case perturbations, which, for the simple model in Appendix A, appear to be significantly worse than random perturbations.

Acknowledgments

This research is supported in part by an NSF CAREER Award, an ARO YIP Award, and DARPA.

References

B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1), 2017.

B. Clarke and P. Gustafson. On the overall sensitivity of the posterior distribution to its inputs. *Journal of Statistical Planning and Inference*, 71, 1998.

A. E. Gelfand, S. E. Hills, A. Racine-Poon, and A. F. M. Smith. Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association*, 85(412):972–985, 1990.

R. Giordano, T. Broderick, and M. I. Jordan. Covariances, robustness, and variational Bayes. *Journal of Machine Learning Research*, 19(51):1–49, 2018.

P. Gustafson. Local sensitivity of inferences to prior marginals. *Journal of the American Statistical Association*, 91:774–781, 1996.

P. J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

David Rios Insua and Fabrizio Ruggeri. *Robust Bayesian Analysis*. Springer-Verlag New York, 2000.

P. W. Koh and P. Liang. Understanding black-box predictions via influence functions. In *International Conference in Machine Learning (ICML)*, 2017.

D. Maclaurin, D. Duvenaud, and R. P. Adams. Autograd: Effortless gradients in numpy. In *International Conference on Machine Learning 2015 AutoML Workshop*, 2015.

J. W. Miller and D. B. Dunson. Robust Bayesian inference via coarsening. *Journal of the American Statistical Association*, 2018.

D. P. M. Scollnik. On composite lognormal-Pareto models. *Scandinavian Actuarial Journal*, 2007.

S.. Sivaganesan. Global and local robustness approaches: uses and limitations. In D. R. Insua and F. Ruggeri, editors, *Robust Bayesian Analysis*. Springer-Verlag New York, 2000.

J. W. Tukey. A survey of sampling from contaminated distributions. STRG Technical report 33, Princeton, 1959.

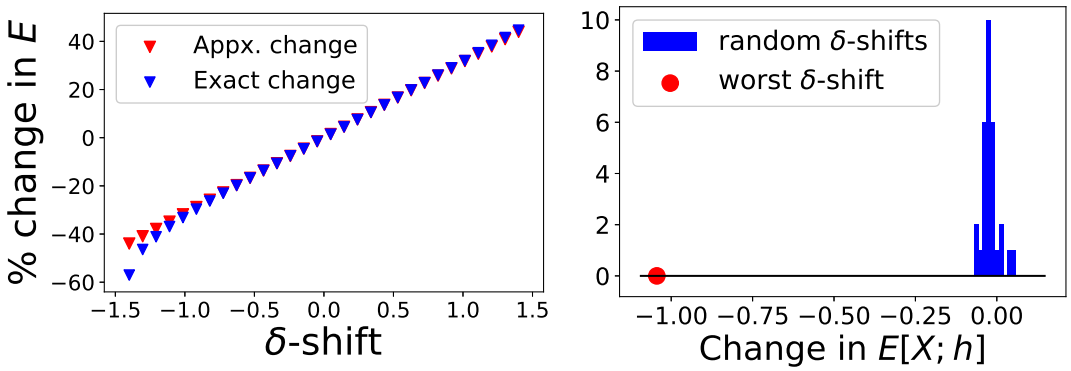

Figure 2: Results for the gamma-lognormal model in Appendix A. (*Left*): changes in $E(X;h)$ for different values of δ. Under the worst-case data shift predicted by our approximation, we compare the change in $E(X;h)$ predicted by our linear approximation (red) and the closed form in Eq. (4). (*Right*): We perturb the data by $\delta = 0.25$ in 100 directions randomly drawn from the unit sphere and re-run MCMC to compute $E(X';h)$ for each perturbation. The blue histogram shows the distribution of the change $E(X;h) - E(X';h)$; none of the random changes comes close to the result of re-running MCMC on our predicted worst-case dataset, which we show in red.

Appendix A. Approximation Accuracy on Conjugate Models

Before applying our technique in practice, it is important to check the accuracy of the various approximations made in Section 2. Specifically, we may fail to correctly find the worst-case data perturbations, and even if we do, the linear approximation of Eq. (2) may be of poor quality. To examine these potential issues, we test on two simple conjugate models for which we have access to a closed form expression for $E(\boldsymbol{X};h)$. Specifically, we consider data $X_1, \ldots, X_N \sim \text{LogNormal}(\mu, \tau)$ for known μ and place a $\text{Gamma}(\alpha, \beta)$ prior on τ. We choose the second moment of τ under the posterior as our quantity of interest, i.e. we set $h(\tau) = \tau^2$. We can compute $E(X;h)$ as:

$$\mathbb{E}[\tau^2|X] = \left(\left(\alpha + \frac{N}{2}\right) + \left(\alpha + \frac{N}{2}\right)^2\right)\left(\beta + \frac{\sum_n (\log X_n) - \mu)^2}{2}\right)^{-2}. \tag{4}$$

Fig. 2 shows that our concerns listed at the beginning of this section are not realized for this simple model. Overall, the results are encouraging in that we seem to discover a "worst-case" that is significantly different from a random perturbation and that the linear approximation in Eq. (2) is accurate even for $\delta \approx 1.0$.