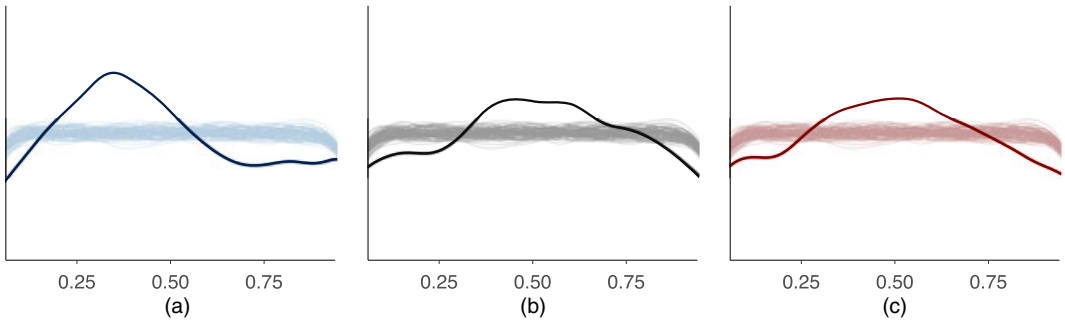

Fig. 9. Graphical check of the LOO cross-validated probability integral transform (——, simulations from the standard uniform distribution; ——, density of the computed LOO probability integral transforms) (similar plots can be made using `ppc_dens_overlay` and `ppc_loo_pit` in the `bayesplot` package; the downward slope near 0 and 1 on the 'uniform' histograms is an edge effect due to the density estimator used and can be safely discounted): (a) model 1; (b) model 2; (c) model 3

We can also perform similar checks within levels of a grouping variable. For example, in Fig. 8 we split both the outcome and the posterior predictive distribution according to region and check the median values. The two hierarchical models give a better fit to the data at the group level, which in this case is unsurprising.

In cross-validation, double use of data is partially avoided and test statistics can be better calibrated. When performing leave-one-out (LOO) cross-validation we usually work with univariate posterior predictive distributions, and thus we cannot examine properties of the joint predictive distribution. To check specifically that predictions are calibrated, the usual test is to look at the LOO cross-validation predictive cumulative density function values, which are asymptotically uniform (for continuous data) if the model is calibrated (Gelfand *et al.*, 1992; Gelman *et al.*, 2013).

The plots that are shown in Fig. 9 compare the density of the computed LOO probability integral transforms (the thick dark curve) *versus* 100 simulated data sets from a standard uniform distribution (the thin light curves). We can see that, although there is some clear miscalibration in all cases, the hierarchical models are an improvement over the single-level model.

The shape of the miscalibration in Fig. 9 is also meaningful. The frown shapes that are exhibited by models 2 and 3 indicate that the univariate predictive distributions are too broad compared with the data, which suggests that further modelling will be necessary to reflect the uncertainty accurately. One possibility would be to subdivide the super-regions further to capture within-region variability better (Shaddick *et al.*, 2018).

6. Pointwise plots for predictive model comparison

Visual posterior predictive checks are also useful for identifying unusual points in the data. Unusual data points come in two flavours: outliers and points with high leverage. In this section, we show that visualization can be useful for identifying both types of data point. Examining these unusual observations is a critical part of any statistical workflow, as these observations give hints about how the model may need to be modified. For example, they may indicate that the model should use non-linear instead of linear regression, or that the observation error should be modelled with a heavier-tailed distribution.

The main tool in this section is the one-dimensional cross-validated LOO predictive distribution $p(y_i \mid y_{-i})$. Gelfand *et al.* (1992) suggested examining the LOO log-predictive density values

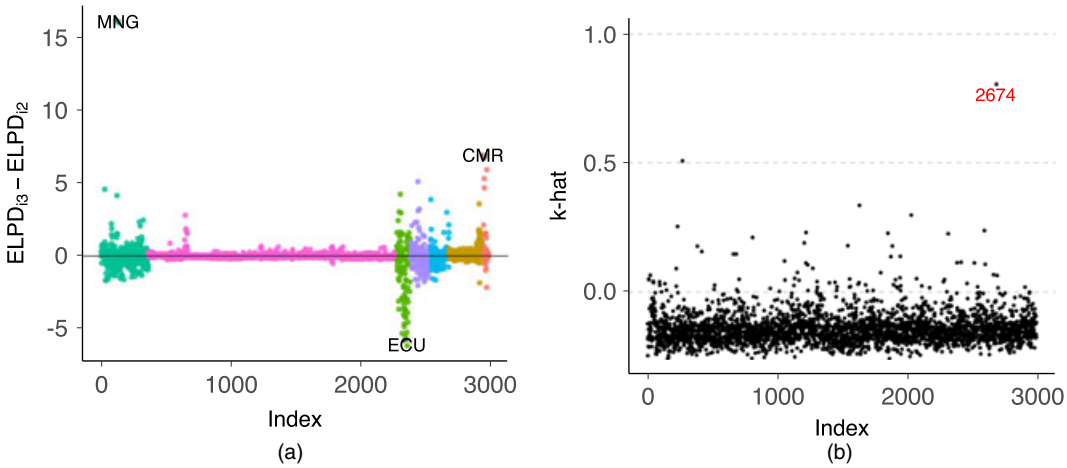

Fig. 10. Model comparisons by using LOO cross-validation (a) the difference in pointwise values obtained from LOO PSIS for model 3 compared with model 2 coloured by World Health Organization cluster (see Fig. 1(b) for the key; positive values indicate that model 3 outperformed model 2); (b) $\hat{k}$-diagnostics from LOO PSIS for model 2 (the 2674th data point (the only data point from Mongolia) is highlighted by the $\hat{k}$-diagnostic as being influential on the posterior)

(they called them conditional predictive ordinates) to find observations that are difficult to predict. This idea can be extended to model comparison by looking at which model best captures each left-out data point. Fig. 10(a) shows the difference between the expected log-predictive densities ELPD for the individual data points estimated by using Pareto-smoothed importance sampling (PSIS), (Vehtari *et al.*, (2017a,b)). Model 3 appears to be slightly better than model 2, especially for difficult observations like the station in Mongolia.

In addition to looking at the individual LOO log-predictive densities, it is useful to look at how influential each observation is. Some of the data points may be difficult to predict but not necessarily influential, i.e. the predictive distribution does not change much when they are left out. One way to look at the influence is to look at the difference between the full data log-posterior predictive density and the LOO log-predictive density.

We recommend computing the LOO log-predictive densities by using the PSIS LOO method as implemented in the `loo` package (Vehtari *et al.*, 2017c). A key advantage of using the PSIS LOO method to compute the LOO densities is that it automatically computes an empirical estimate of how similar the full data predictive distribution is to the LOO predictive distribution for each left-out point. Specifically, it computes an empirical estimate $\hat{k}$ of $k = \inf\{k' > 0 : D_{1/k'}(p||q) < \infty\}$, where

$$D_\alpha(p||q) = \frac{1}{\alpha - 1} \log \left\{ \int_\Theta p(\theta)^\alpha q(\theta)^{1-\alpha} \mathrm{d}\theta \right\}$$

is the α-Rényi divergence (Yao *et al.*, 2018). If the jth LOO predictive distribution has a large $\hat{k}$-value when used as a proposal distribution for the full data predictive distribution, it suggests that y_j is a highly inferential observation.

Fig. 10(b) shows the $\hat{k}$-diagnostics from the PSIS LOO method for our model 2. The 2674th data point is highlighted by the $\hat{k}$-diagnostic as being influential on the posterior. If we examine the data we find that this point is the only observation from Mongolia and corresponds to a measurement $(x, y) = (\log(\text{satellite}), \log(\text{PM}_{2.5})) = (1.95, 4.32)$, which would look like an outlier if highlighted in the scatter plot in Fig. 1(b). By contrast, under model 3 the $\hat{k}$-value for the

Mongolian observation is significantly lower ($\hat{k} \approx 0.5$) indicating that that point is better resolved in model 3.

7. Discussion

Visualization is probably the most important tool in an applied statistician's toolbox and is an important complement to quantitative statistical procedures (Buja *et al.*, 2009). In this paper, we have demonstrated that it can be used as part of a strategy to compare models, to identify ways in which a model fails to fit, to check how well our computational methods have resolved the model, to understand the model sufficiently well to be able to set priors and to improve the model iteratively.

The last of these tasks is a little controversial as using the measured data to guide model building raises the concern that the resulting model will generalize poorly to new data sets. A different objection to using the data twice (or even more) comes from ideas around hypothesis testing and unbiased estimation, but we are of the opinion that the danger of overfitting the data is much more concerning (Gelman and Loken, 2014).

In the visual workflow that we have outlined in this paper, we have used the data to improve the model in two places. In Section 3 we proposed prior predictive checks with the recommendation that the data-generating mechanism should be broader than the distribution of the observed data in line with the principle of weakly informative priors. In Section 5 we recommended undertaking careful calibration checks as well as checks based on summary statistics, and then updating the model accordingly to cover the deficiencies that are exposed by this procedure. In both of these cases, we have made recommendations that aim to reduce the danger. For the prior predictive checks, we recommend not cleaving too closely to the observed data and instead aiming for a prior data-generating process that can produce plausible data sets, not necessarily data sets that are indistinguishable from observed data. For the posterior predictive checks, we ameliorate the concerns by checking carefully for influential measurements and proposing that model extensions be weakly informative extensions that are still centred on the previous model (Simpson *et al.*, 2017).

Regardless of concerns that we have about using the data twice, the workflow that we have described in this paper (perhaps without the stringent prior and posterior predictive checks) is common in applied statistics. As academic statisticians, we have a duty to understand the consequences of this workflow and offer concrete suggestions to make the practice of applied statistics more robust.

Acknowledegements

The authors thank Gavin Shaddick and Matthew Thomas for their help with the $PM_{2.5}$ example, Ari Hartikainen for suggesting the parallel co-ordinates plot, Ghazal Fazelnia for finding an error in our map of ground monitor locations, Eren Metin Elçi for alerting us to a discrepancy between our text and code, and the Sloan Foundation, Columbia University, US. National Science Foundation, Institute for Education Sciences, Office of Naval Research and Defense Advanced Research Projects Agency for financial support.

References

Betancourt, M. (2017) A conceptual introduction to Hamiltonian Monte Carlo. *Preprint arxiv:1701.02434*.

Buja, A., Cook, D., Hofmann, H., Lawrence, M., Lee, E.-K., Swayne, D. F. and Wickham, H. (2009) Statistical inference for exploratory data analysis and model diagnostics. *Philos. Trans. R. Soc. Lond.*, **367**, 4361–4383.

Forouzanfar, M. H., Alexander, L., Anderson, H. R., Bachman, V. F., Biryukov, S., Brauer, M., Burnett, R., Casey, D., Coates, M. M., Cohen, A. *et al.* (2015) Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks in 188 countries, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet*, **386**, 2287–2323.

Gabry, J. (2017) bayesplot: plotting for Bayesian models. *R Package Version 1.3.0.* (Available from `http://mc-stan.org/bayesplot`.)

Gelfand, A. E., Dey, D. K. and Chang, H. (1992) Model determination using predictive distributions with implementation via sampling-based methods (with discussion). In *Bayesian Statistics 4*. (eds. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 147–167. Oxford: Clarendon.

Gelman, A. (2004) Exploratory data analysis for complex models. *J. Comptnl Graph. Statist.*, **13**, 755–779.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2013) Marginal predictive checks. In *Bayesian Data Analysis* 3rd edn, ch. 6. Boca Raton: Chapman and Hall-CRC.

Gelman, A., Jakulin A., Pittau M. G., Su Y.-S., *et al.* (2008) A weakly informative default prior distribution for logistic and other regression models. *Ann. Appl. Statist.*, **2**, 1360–1383.

Gelman, A. and Loken, E. (2014) The statistical crisis in science: data-dependent analysis—a "garden of forking paths"—explains why many statistically significant comparisons don't hold up. *Am. Scient.*, **102**, no. 6, 460.

Gelman, A., Simpson, D. and Betancourt, M. (2017) The prior can generally only be understood in the context of the likelihood. *Preprint arXiv:1708.07487*.

R Core Team (2017) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Shaddick, G., Thomas, M. L., Green, A., Brauer, M., van Donkelaar, A., Burnett, R., Chang, H. H., Cohen, A., van Dingenen, R. V., Dora, C., Gumy, S., Liu, Y., Martin, R., Waller, L. A., West, J., Zidek, J. V. and Prüss-Ustün, A. (2018) Data integration model for air quality: a hierarchical approach to the global estimation of exposures to ambient air pollution. *Appl. Statist.*, **68**, 231–253.

Simpson, D., Rue, H., Riebler, A., Martins, T. G., Sørbye, S. H. *et al.* (2017) Penalising model component complexity: a principled, practical approach to constructing priors. *Statist. Sci.*, **32**, 1–28.

Stan Development Team (2017a) RStan: the R interface to Stan, version 2.16.1. Stan Development Team. (Available from `http://mc-stan.org`.)

Stan Development Team (2017b) *Stan Modeling Language User's Guide and Reference Manual*, *Version 2.16.0*. Stan Development Team. (Available from `http://mc-stan.org`.)

Vehtari, A., Gelman, A. and Gabry, J. (2017a) Pareto smoothed importance sampling. *Preprint arXiv:1507.02646*.

Vehtari, A., Gelman, A. and Gabry, J. (2017b) Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statist. Comput.*, **27**, 1413–1432.

Vehtari, A., Gelman, A. and Gabry, J. (2017c) loo: efficient leave-one-out cross-validation and WAIC for Bayesian models. *R Package Version 1.0.0.* (Available from `http://mc-stan.org/loo`.)

Wickham, H. (2009) *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer.

Yao, Y., Vehtari, A., Simpson, D. and Gelman, A. (2018) Yes, but did it work?: Evaluating variational inference. *Preprint arXiv:1802.02538*.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Supplementary material: Visualization in Bayesian workflow'.