Visualization in Bayesian Workflow, a Summary

Quentin Le Roux

Abstract. Submitted in 2017, J. Gabry et al.'s paper[1] develops a new perspective on data visualization techniques as part of a Bayesian statistics workflow. It proposes that such techniques should be integrated throughout an iterative modeling process rather than be relegated to data exploration and post-modeling phases.

The paper's thesis

Catching rising issues and improving one's modeling are among statisticians' key objectives as part of their workflow. This workflow is described in the paper in five steps: data exploration, computational and prior checking, inference reliability checking, model fitness checking, and model selection.

Data visualization techniques (DVT) represent a valuable toolbox from which statisticians can draw insights to inform their process. However, DVT are often limited to data exploration, and post-modeling checks. The paper defends that DVT should be further integrated throughout modeling worflows as a type of qualitative input.

The problem setup

The paper explores the Bayesian estimation of $PM_{2.5}$ concentration across the world – $PM_{2.5}$ are air pollutants smaller than 2.5 microns. However, this estimation faces the hurdle that relying on sparsely distributed ground monitors yields incomplete and imbalanced data.

The paper offers that DVT can aid a Bayesian workflow in building generative models of $PM_{2.5}$ concentration, based on approximate satellite measures calibrated with observed ground data. Valuable spatial resolution of $PM_{2.5}$ concentration could thus be generated, which would inform a better modeling of human $PM_{2.5}$ exposure.

Data visualization techniques to improve a Bayesian workflow

With the $PM_{2.5}$ concentration estimation problem stated, the paper proposes valuable DVT applications for each of the five previously cited Bayesian workflow phases:

Exploratory data analysis helps identify trends that may highlight modeling prospects and issues down the line. The paper proposes the use of scatterplots to highlight data sparsity and imbalance, and regional clustering. In the given setup, the over-representation of developed countries is easily identifiable in the data, while the African continent is under-represented in terms of $PM_{2.5}$ concentration.

As such, DVT can inform statisticians' next steps and, in the $PM_{2.5}$ case, three simple models to explore: a linear regression and two multilevel models dependent on a type of regional clustering (WHO partition or ground $PM_{2.5}$ concentration-based).

[1] Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., Gelman, A.: Visualization in Bayesian workflow. http://dx.doi.org/10.1111/rssa.12378. In: Journal of the Royal Statistical Society: Series A (Statistics in Society). (2019)

Prior predictive distribution checking can be performed visually when building a generative data model. In the paper's case, generating artificial $PM_{2.5}$ ground observations from satellite data in a multivariate model could greatly help achieve better results as part of a Bayesian workflow.

DVT can help select priors, discarding so-called "vague" priors for "weakly informative" ones by scatterplotting generated data against observed data for instance. In the paper's context, weakly informative generative processes are shown to generate faithful data and several plausible datasets that can be used as references, which the paper abstracts as a "flipbook", to assess prior distributions' predictive capability.

Inference reliability can be assessed and its issues diagnosed with DVT. In the paper's case, Markov Chain Monte Carlo (and specifically the Hamiltonian kind) are investigated with DVT. Using scatter and coordinate plots, the paper shows that it is possible to identify divergent transition clusters: regions of posterior distribution that do not match the observed data, and whose concentration in a parameter space indicates an estimation problem.

Though no such cluster is identified in the case of $PM_{2.5}$ concentration, the paper's annex provides an example with randomized experiments conducted in eight schools that tested for the effectiveness of SAT-coaching programs[2].

Fitness and model evaluation can also be assessed with visual tools. As the predictive power of a generative model is paramount in model evaluation, the paper insists on the qualitative help provided by DVT: With regards to $PM_{2.5}$ concentration distributions, comparing kernel density estimates (via KDE plots) between observed and generated data can prove the efficacy of multilevel models over linear regression.

This step relies on using the same data twice, however, a controversial idea the authors themselves note, mentioning a previous paper by one of them[3]. Double use can nudge model selection towards overfitting and less-generalizing results.

To alleviate this concern, the paper proposes that DVT focuses on statistics not included earlier in a Bayesian workflow. The paper uses skewness and the median as examples. Cross-validation can further be used to lift concerns over the double use.

Model selection can finally be informed by DVT, helping highlight models' blind spots such as outliers and highly-influential data points. Investigating the latter is yet another way to navigate models as part of a Bayesian workflow.

The paper uses Mongolia's single data point as an example, which only the second of the two multilevel models can properly incorporate in its predictive distribution.

Parting words

The authors assert that DVT complement quantitative approaches to Bayesian modeling workflows. They are intuitive approaches that help identify points of failure and which should be exploited instead of relegated to only few modeling steps.

[2] Rubin, D.B.: Estimation in Parallel Randomized Experiments. In: Journal of Educational Statistics. https://doi.org/doi:10.3102/10769986006004377. (1981)

[3] Gelman, A., Loken, E.: The statistical crisis in science. American Scientist 102 (6). (2014)