Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2018-61-RC1, 2018

Geoscientific Model Development Discussions

Interactive comment on "GemPy 1.0: open-source stochastic geological modeling and inversion" *by* Miguel de la Varga et al.

Anonymous Referee #1

Received and published: 8 May 2018

Gempy.bib

[a4paper]article [english]babel [utf8x]inputenc [T1]fontenc

[a4paper,top=3cm,bottom=2cm,left=3cm,right=3cm,marginparwidth=1.75cm]geometry

url amsmath graphicx [colorinlistoftodos]todonotes [colorlinks=true, allcolors=blue]hyperref amssymb siunitx tabularx algorithm,algorithmicx amsmath,algpseudocodealgcompatible amsthm graphicx lscape verbatim color,soul xcolor subfigure tabularx,ragged2e,booktabs,caption,array,multirow,multicol csquotes mathtools lineno amsthm listings amssymb latexsym epsfig float xspace float

Review of Manuscript
GemPy 1.0: open source stochastic geological modeling and inversion

May 8, 2018

The aim of this paper is to develop a framework for fusing many sources of information into a coherent probablistic model to allow for estimation and inference in geological inversion problems. It is an ambitious task and the authors are to be commended on tackling it, and making code publically available. I am not an expert in geological inversion problems, although I have some familiarity with them, so I will confine my comments to a discussion on the probabilistic model, Bayesian inference and probabilistic programming sections.

1. The paper would benefit from a clear and concise description of an example of at least one probabilistic model.
   (a) For example in section **3.4.2 Geological Inversion: Gravity and Topology** the authors say that they construct a specific likelihood function for a topol-

ogy, but no likelihood function is given. The authors correctly state that the Jaccard index varies between 0 and 1, but then go on to state that it is *a single number we can evaluate using a a probability density function. The type of probability density function used will determine the strength or likelihood that the mean graph represent.* What does this sentence mean? They go on to say *Here we use a half Cauchy, due to tolerance for outliers*. Why? A half Cauchy has support on the interval $[0, \infty)$, whereas this statistic has support on the interval $[0, 1]$. What is meant by *its tolerance for outliers*?

(b) What is needed is a joint likelihood function on both the topology and the gravity to be specifically stated. See for example () and (). The authors should at least reference ().

(c) The authors statement *The use of likelihood functions in a Bayesian inference in opposition to simply rejection sampling has been explored by the authors during the recent years (de la Varga and Wellmann, 2016; Wellmann et al., 2017; Schaaf, 2017).* is confusing. Are the authors referring to likelihood free methods such as Approximate Bayesian Computation, ABC, where rejection sampling can be used to obtain draws from the approximate posterior? The use of likelihood function in Bayesian inference is typically not related to rejection sampling. Rejection sampling is a method to obtain draws from a non-standard distribution, in this case the posterior distribution, usually for the purpose of numerical integration. A likelihood is an assumption about the data generation process which, together with the prior, result in inference via the posterior. If the likelihood is unavailable in closed form, or if we do not wish to make assumptions about the data generating process, then the issue of how to approximate the posterior may involve rejection sampling. The authors need to articulate clearly the point they are making and provide a justification.

(d) A gravity likelihood is referred to on page 31. What is this likelihood? Are the authors assuming that the *observed* data is related to the *simulated*

data as a signal plus noise model of the form, $y_i = g(x_i) + e_i$, where e_i is independently and identically distributed (i.i.d)? If so why do they model $(y_i - g(x_i))^2$ as a folded Cauchy (i.e a folded t_1)? What is this saying about the data generating process? Surely there is geophysical knowledge about the distribution of gravity measurements? From a statistical point of view gravity is an integral, a sum of things, in which case the central limit theorem (CLT) would make the assumption of Gaussian errors, i.e. $e_i \sim N(0, \sigma^2)$, reasonable. If this were so then and the observations independent (which I'm not convinced they would be), then $\sum_{i=1}^{n} (y_i - g(x_i))^2 \sim \chi^2_n$. Perhaps this is what they do, but it is not clear from the paper.

2. **MCMC convergence** The authors need to show that the MCMC scheme converges. Convergence in geophysical inversion problems is non trivial. Posterior distributions of geophysical inversion problems are notoriously difficult to explore, for a discussion see (). and for a demonstration of how difficult they are to explore see (). The NUTS algorithm used in python works well when the derivative exists and is well behaved, but as the posterior distribution in () shows, these distributions can have many modes and derivatives which are difficult, if not impossible to compute. Parallel tempering is probably the best way to explore these multi-model distributions, as shown in ().

3. Minor points

   - The Jaccard index given by equation 13 is not a likelihood function, nor, as it is written, is it even a measure. The authors correctly state that the Jaccard index is a statistic used to compare sets, in this case topologies. It is the ratio of the *size* of the intersection over *size* of the union. It should be written as

   $$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

   where the notation $|.|$ denotes a measure of size to be defined.

- change the phrase *due to tolerance for outliers* to *because parameter estimates based on Cauchy likelihoods are more robust to outliers than parameter estimates based on, say, Gaussian likelihoods*.

References

Beardsmore, G and Durrant-Whyte, H and McCalman, L and O'Callaghan, S and Reid, A. (2016) *A Bayesian Inference Tool for Geophysical Joint Inversions*, ASEG Extended Abstracts 2016: 25th International Geophysical Conference and Exhibition, 509–518. https://library.seg.org/doi/abs/10.1071/ASEG2016ab089.

Beskos, A. and Girolami, M. and Lan, S. and Farrell, E. and Stuart, M.(2017) *Geometric MCMC for infinite-dimensional inverse problems*, Journal of Computational Physics, 335, pp 327-351

Chandra, R and Azam, D and Muller, D and Lasalles T and Cripps, S. (2018) *Bayesian Dynamic Earth models, landscape dynamics and basin evolution*, In Review (Computers and Geoscience), https://github.com/rohitash-chandra/research/blob/master/2018/BayesLands.pdf