Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2018-61-AC1, 2018

Geoscientific Model Development Discussions
Open Access
EGU

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

Miguel de la Varga et al.

miguel.de.la.varga@rwth-aachen.de

Received and published: 6 July 2018

Answer Review 1:

Thank you very much for your suggestions, detailed comments and questions, as well as the positive feedback. We carefully revised the manuscript and provide below a detailed reply to all comments. We also attached a pdf with highlighted changes between the original submission and the revised version.

We agree with the reviewer that the explanation of the Bayesian network was not clear and probably insufficient. However, the scope of this particular paper was on the gen-

Printer-friendly version

Discussion paper

erative model of the Bayesian inference. Since the whole library has been written to be part of a Bayesian inference, we consider it appropriate to just to show a vertical slice of what is possible. Nevertheless, we have reworked quite that part of the paper quite a bit and now include the whole probabilistic graphical model into the main text in order to be more consistent and clear. Also, we have added the convergence analysis in the appendix.

Regarding the reviewers general concerns about the construction of the likelihood functions and samplers, we shared many of them and we are working hard to find the optimal Bayesian model for different data sets. The comments were very insightful and hopefully we will be able to explore them in a more focused paper in the future.

RESPONSES

[1] The paper would benefit from a clear and concise description of an example of at least one probabilistic model.

Authors response: We have extended and moved to the main text the specific probabilistic graphical model

Change in manuscript:

[1a] For example in section 3.4.2 Geological Inversion: Gravity and Topology the authors say that they construct a specific likelihood function for a topolC2 GMDD Interactive comment Printer-friendly version Discussion paper ogy, but no likelihood function is given. The authors correctly state that the Jaccard index varies between 0 and 1, but

Printer-friendly version

Discussion paper

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, ∞), whereas this statistic has support on the interval [0, 1]. What is meant by its tolerance for outliers?

Authors response: We use a half-Cauchy distribution to evaluate the likelihood of the Jaccard index of the simulated model topology due to its wider tail. It is not used directly as a likelihood, but rather as a factor potential which allows us to incorporate the "soft data" of our topology information into the Bayesian inference. As the Jaccard index results in values within the interval $ [0,1]$, the half-Cauchy function is only evaluated for those given values. The shape parameter $ \ beta$ was chosen empirically, as it showed promising results for effective parameter space exploration in the used MCMC scheme.

Change in manuscript: *To evaluate the likelihood of the simulated model topology we use a factor potential with a half-Cauchy parametrization (shape parameter $ \ alpha=0$ and rate parameter $ \ beta==10^ { -3} $) to constrain our model using the "soft data" of our topological knowledge \ citep{ lauritzen1990independence, jordan1998learning, christakos2002assimilation} . This specific parametrization was chosen due to empirical evidence from different model runs to allow for effective parameter space exploration in the used MCMC scheme.*

[1b] 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 ().

Authors response: In pymc when you specify more than one likelihood/potential, it automatically starts sampling on the joint likelihood.

Change in manuscript: [1010] *Defining the topology potential and gravity likelihood on the same Bayesian network creates a join likelihood value that we need to sample from*

[1c] 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.

Authors response: In the sentence we were confusing rejection sampling as a parameter space exploration method to approximate a posterior with just forward simulating the priors. We agree with the reviewers comments and adjusted the sentence in line 855.

Change in manuscript [855]: "*The use of likelihood functions in a Bayesian inference in comparison to simple forward simulation has been explored by the authors during recent years (de la Varga and Wellmann, 2016; Wellmann et al., 2017; Schaaf, 2017).*"

[1d] 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, yi = g(xi) + ei , where ei is independently and identically distributed (i.i.d)? If so why do they model (yi − g(xi))2 as a folded Cauchy (i.e a folded t1)? 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. ei ∼ N(0, σ 2), reasonable. If this were so then and the observations independent (which I'm not convinced they would be), then Pn i=1 (yi − g(xi))2 ∼ χ 2 n . Perhaps this is what they do, but it is not clear from the paper

Authors response: We agree with the reviewer that the explanation was not sufficiently clear, in part because it was not the main purpose of this paper and in part because we did not find yet a convincing way to construct the likelihood function. The model suggested by the reviewer is quite close to what we presented here and we agree with his concerns about the correlation but again we prefer to focus around the generative model since the paper is already too large. Nevertheless, e have extended the explanation of this part and adding the complete PGM figure into the main body text

[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 ().

Authors response: We added a traces plot and a Geweke test into the appendix. Again, we share the concerns about the samplers and we are actively studying the best combinations of them for this type of problems. We agree that a combination of gradient based and parallel methods could be the best solution in the middle term. However, we also consider that that is beyond the scope of this publication.

[3a] 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

$J(A, B) = |A \cap B| / |A \cup B|$

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

Printer-friendly version

Discussion paper

Authors response: Adjusted Jaccard index to correctly contain the absolutes. Thanks for catching this!

Change in manuscript: Changed Eq. 13 to J(A, B) = | A ∩ B| / | A ∪ B|

[3b] 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.

Authors response: We agree to the change in wording, but adjusted the whole paragraph to more precisely state the use of the topology information as a factor potential in the Bayesian inference and the second reviewers comments on line [811].

Change in manuscript: *To evaluate the likelihood of the simulated model topology we use a factor potential with a half-Cauchy parametrization (shape parameter $ \ alpha=0$ and rate parameter $ \ beta==10ˆ { -3} $) to constrain our model using the "soft data" of our topological knowledge \ citep{ lauritzen1990independence, jordan1998learning, christakos2002assimilation} . This specific parametrization was chosen due to empirical evidence from different model runs to allow for effective parameter space exploration in the used MCMC scheme and due to the Cauchy distribution being more robust to outliers than parameter estimates based on, say, Gaussian likelihoods.*

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