Comments

Peter K. Kitanidis
Stanford University
Stanford, CA 94305, USA
peterk@stanford.edu

May 7, 2024

1 General Comments

The authors are to be commended for their efforts to provide a platform to evaluate and compare inverse methods in Hydrogeology. A challenging job, to say the least. I liked the paper, which I found interesting and stimulating, but my comments will emphasize my concerns or points I would like to see better clarified.

It seems that the authors' approach is to provide some "benchmarking scenarios" then evaluate methods by comparing them with reference solutions. The term reference solution is first mentioned without explanation in the abstract. My first impression was that a reference solution is the ground truth, meaning the true or correct answer. It was later made clear that the term referred to a presumably best possible solution. (That's my interpretation. It would help immensely if the authors could explain the meaning right at the beginning.) However, one may wonder whether the reference solutions are best. Even 20000 samples from the posterior, which could be correlated, may not completely explore the probability space of the highly multivariate distribution. As a consequence, I find that some of the comparison metrics may be an overkill.

I am concerned that the methodology will benchmark only part of the solution of an inverse problem, which is much more than the algorithm to apply Bayes' theorem. I will explain what I mean in the next three paragraphs.

For me, an "inverse problem" is a problem in which the forward map and the data do not suffice to give a unique answer. Thus, using prior information (or regularization, structural information, or whatever else one may call it) is essential. I applaud the authors' emphasis on the Bayesian approach. Inverse modeling is a data science problem with all the consequences.

The first consequence is that inverse modeling is an iterative process in which data, other information, and the modeling objectives are considered; an approach is proposed, hyperparameters estimated, and the overall method is tested; then parts or all are modified; and so on, until convergence. This aspect

is seldom discussed in published papers, which promote the illusion of a one-way (noniterative) workflow: Put the data, then the forward map, run, and get results. The idea of having a one-way (noniterative) workflow is appealing but dangerous. How can one know what formulation to use unless one tries several? In the approach proposed in this paper, this aspect of inverse modeling is not considered or benchmarked.

Bayesian methods are very powerful but also tricky. There are reasons why a great scientist like R. A. Fisher was a lifelong critic. Bayesian methods have come a long way since Fisher criticized them, though not every scientist or engineer who applies them is aware of the advances. Many still think of Bayesian inference as one where Bayes theorem is used and nothing else. Furthermore, that prior information is "subjective" and "preordained" or somehow given, while the accuracy with which to reproduce the data is known beforehand. These are fallacies that I mention because they make the task of evaluating and benchmarking inverse-problem methods so much more challenging. My concern is that this paper is not helpful in dispelling these common misconceptions.

Turning my attention to benchmarking metrics, one general comment is that they are numerous, but all are what I call "point-centric" or "pixel-centic". In other words, they focus on the accuracy or errors at the smallest discretization scale. Consider the following:

1. The quantities we deal with in Hydrogeology, like the log-conductivity, have support volumes. It is understood that different models and applications may require parameters with different support volumes.

2. Computing the estimation variance at the finest scale is fraught with difficulties as it depends solely on the behavior near the origin of the chosen covariance function. For example, consider the difference between an exponential versus a Gaussian covariance. They result in very different computed variances of point estimates. Because small-scale variability is in the nullspace of the forward map, the assumed covariance dominates, while the assumed covariance has much less effect on computing large-scale estimates and variances of estimation.

3. In my experience, practitioners are not interested in the complete a posteriori pdf of, say, the log-conductivity because it is understood that not only is it hard to compute, but it also relies on many assumptions whose usefulness can go only so far. Instead, the question they ask is "What scales of variability are resolved?"

In my view, when it comes to inverse problems, one must consider "scale-centric" accuracy benchmarks, such as the accuracy of discrete cosine transform (DCT) coefficients. The DCT is an excellent tool for evaluating what scales are resolved. For example, we can evaluate whether the correct average is computed or variability at scales larger than 100 meter is resolved.

It is good to find a role for conditional simulations, meaning samples from the *a posteriori* distribution. I increasingly find conditional realizations, which

reveal the scales at which there is uncertainty, much more useful than MSE or confidence intervals.

My final comments are related to using the Gaussian distribution to generate log-conductivity fields and data. Some may consider that this invalidates the results because the model used in the inversion is the same as the "true model". I am coming to this issue from a different angle. The Gaussian model cleverly deployed has been successful in Hydrogeology, River Bathymetry, and Face Recognition, but this does not mean that the true unknown is somehow Gaussian. The strength of Gaussian models (with variable transformations) is their versatility and robustness, regardless of the unknowns. I remember that George E. P. Box's, a pioneer of modern Bayesianism, gave us the memorable aphorism that "All models are wrong, but some are useful." I teach my students that modeling assumptions may be appropriate and useful *up to a point*.

Stochastic methods in inverse modeling can easily get out of hand and ignore the basic premise that inverse modeling means *estimation with limited information* and consequently should have limited objectives. For example, we can determine a mean value or a variance of certain unknown components, but we cannot estimate everything. Just because one can compute something does not suggest that one should! And just because a modeling assumption is useful for some purposes does not mean it is useful for all purposes! I question estimating something that relies critically on assumptions that cannot be verified, such as the complete distribution at the smallest scale.

The preoccupation with complete distributions is a legacy of classical (frequentist) methods and has no place in inverse modeling with Bayesian methods. In Bayesian methods, the unknowns have distributions that represent a state of knowledge, not the true distribution of the unknown.

With these thoughts in mind, I find criteria such as the K-S distance an overkill. Also, I would suggest including cases not generated by Gaussian distributions.

2 Specific Comments

I take issue with the statement that the objective is to learn about the parameters "by matching" (line 142). I would rephrase with the perhaps loftier "by assimilating information in the hard and other data (or the prior)". Too much emphasis on data matching is wrong. Many parameter sets can match data, yet they can be poor estimates. One of the key questions in solving an inverse problem is how closely to match the data.

Add "and uniform variance" (line 161).

I am unsure if the readers understand "best estimate and its variance" (line 174) since they have not been defined.

I am not excited about the normalization (line 184). The no-error case is defined as full agreement with the presented reference case. However, both the candidate and the reference solution must have some error, particularly at the point scale. The criterion will be too dependent on the small-scale errors that

are of limited interest and too dependent on assumptions.

Specifically for estimating log-conductivity, mean square and mean absolute errors are dimensionless numbers that, in my view, hardly need normalization.

I would suggest "hydraulic head or pressure, as appropriate" (Line 226).

The normalization through a sigmoid (Eq. 3) is interesting. I have no experience with it, but I suspect it will result in bunching together the bad solutions near the highest value, 1.

Is there a mathematical guarantee that the metrics of Eq. 5 or 8 will not exceed 1?

It remains to be seen whether the K-S distance at each node is useful. My first reaction is that it is an overkill. I am unsure how this will be applied and, most importantly, is it really important?

Regarding Eq. 12, is the symbol E supposed to mean numerical average? Also, I do not know what D_E signifies.

Regarding the number of forward-map calls and using Eq. 3, expensive solutions will bunch together. For example, for 10000 calls that take a day, the normalized value is 0.8. For 1000000 calls that take 100 days, the normalized value is 0.8571. The normalized-value difference seems small, while the actual difference between 1 and 100 days of computations is consequential.

In my experience, using wall clock time (or CPU time, for that matter) is pointless. It depends so much on the computer system and how busy it is with other jobs.

One important issue regarding evaluating computational effort is related to the iterative nature of most methods. One can game the effort metric by using a "starting" solution that is cleverly selected close to the final, thus reducing the required iterations.

The value -2.5 for the mean (Line 400) implies some units for conductivity (like meters per second). Please clarify. Same for σ_e (Line 409).

By observation error standard deviation, I assume it is to be used to introduce errors in generated data. However, I assume it is not to be used in solving the inverse problem. It is a pity that this information is given; in more realistic cases, deciding what to use is one of the most important steps in a method.

The sentence in Lines 433-436 is unclear.

Regarding the "thinning out" of samples in the chain, I would have expected that the main reason would be to have independent samples. Are the 20000 samples independent? I have enough experience with MCMC methods to know that they are tricky. What assurances are there that the solution is a "high-quality" solution?

On Line 502, I suggest "the better the identification that can be achieved."

3 Technical Corrections

No comments here.

I would like to take the opportunity to thank the authors for their work.