Author’s reply to reviews for manuscript “Fully probabilistic seismic source inversion – Part 2: Modelling errors and station covariances” by Simon C. Stähler and Karin Sigloch

Italic indented blocks are referee statements, red text is new text.

Reply to review by Thomas Bodin

> *This paper proposes a new likelihood function for Bayesian seismic source inversion. First, a decorrelation misfit function is presented to quantify the distance between estimated and observed waveforms. The decorrelation is defined as D=1-CC, where CC is the waveform cross-correlation coefficient. It is demonstrated that this misfit function performs better than more classical L1 or L2 “point to point” norms, when trying to infer the depth of an earthquake.*
> *In a second time, it is observed that the ensemble of waveform decorrelation coefficients for a large set of high-quality deterministic source solutions, follow a log normal distribution. The parameters of this log-normal distribution (mean and covariance) are then used to construct an empirical likelihood function for Bayesian inference.*
>
> *This is a very well written paper. It addresses an important problem that is often overlooked in global seismology : data noise (both observational and theoretical errors) in seismic waveforms may be strongly correlated, and it is important to use a proper model for data noise to avoid biases in waveform inversion.*
> *Bayesian inversion is based on having observed data d and model parameters m such that p(d|m), the conditional distribution of the observed data given parameters m, follows the statistics of data errors. This function can be interpreted as a function of m for fixed d to produce the likelihood. In this, the observed data d are fixed, measured quantities, independent of the parameters m. Then Baye’s theorem can be used to combine p(d|m) with prior information to produce the posterior p(m|d). But here p(\phi) can’t be interpreted this way as it does not strictly represent the probability of observing the data.*
>
> *Major comment*
> *I have a major comment about how the likelihood function is presented. The authors define a convenient misfit measure \phi, and then use the distribution p(\phi) as a likelihood function. However, this distribution does not reflect the distribution of data errors, but instead the distribution of misfit values.*
> *A likelihood function must be derived from an assumption about the distribution of data errors and residuals must be defined as a difference between observed and predicted data vectors. In this way, the distribution of residuals follows the statistics of data noise. This is not the case for decorrelation residuals.*

The referee makes an important point about the definition of a likelihood, which we didn’t spell out in our original submission. We have clarified throughout the revised paper that our Likelihood functional L is not a strict Likelihood, but that it provides several advantages over a strict one, which is now also discussed in the Discussion section.

We introduced the symbol L* for the new “empirical” Likelihood, in order to distinguish it from a Likelihood in a strict sense.

> *Page 4 Line 1: a probability distribution on the misfit => a probability distribution on the data.*

done

> *Page 1 line 15: the phrase “the likelihood function of misfit D” does not make sense. A likelihood function depends on a data noise model, not on a choice of misfit.*

Replaced

“its likelihood” with “this likelihood”

> *Page 7 line 12: This is not correct. Taking the distribution of any functional of observed and predicted waveforms does not give a likelihood function. Only the distribution of the difference between observed and predicted data gives a likelihood function.*

As stated above, this issue was not treated with sufficient precision in the original manuscript and was revised.

Replaced “Likelihood” with “empirical Likelihood” where appropriate and added (p.11. l.11):

“We preface the term ``Likelihood" by ``empirical" because strictly speaking the Likelihood would be associated with the noise model on the raw samples i, rather than with the noise on the composite measure D. A similar approach has been chosen independently by Bodin et al (2016) in the context of receiver-function inversion. Please note that the term “empirical Likelihood” has been used differently in statistics (Owen, 1988).”

and a section in the discussion:

“As noted, the proposed ``empirical Likelihood" function L* is no Likelihood function in a strict sense because it is not derived from the noise on the raw data samples, but rather from the noise (i.e., residual) of misfit functional D. For other inverse problems, it has to be evaluated separately, whether or not a noise model exists that can describe the difference between modelled and measured seismograms completely as an additive term. If that is the case, a classical Likelihood can be used, but many inverse problems in seismology are similar to the one presented here and the proposed empirical Likelihood offers a path to Bayesian treatment.”

> *Minor comments*
> *1. Maybe the authors should refer to the recent work of Zacharie Duputel about characterizing uncertainties in source inversions.*
> *Z. Duputel, P. S. Agram, M. Simons, S. E. Minson and J.L. Beck, 2014. Accounting for prediction uncertainty when inferring subsurface fault slip. Geophys. J. Int., v. 197, p. 464-482*
> *Z. Duputel, L. Rivera, Y. Fukahata, H. Kanamori, 2012. Uncertainty estimations for seismic source inversions, Geophysical Journal International, v. 190, iss. 2, p. 1243-1256.*
> *2. See also this recent paper:*
> *Point source moment tensor inversion through a Bayesian hierarchical model M Mustać, H Tkalčić Geophysical Journal International 204 (1), 311-323*

Done, citations added to the Introduction and discussion.

New sentence:

"For finite-fault inversion of large earthquakes, Bayesian methods have beendeveloped in the recent years (Duputel et al., 2012, 2014; Dettmer et al., 2014), and also for non-kinematic inversions of regional events (Mustać and Tkalčić, 2016) but we focus on the inversion of source time functions of intermediate-sized events (m_b 5.5 to 7.5) from broadband, teleseismic waveforms"

> *3. The log-normal likelihood goes to zero when the similarity is maximized (when D=0), right? Can this be seen as a way to penalize overfitting solutions?*

That observation is true. The distribution of misfit values converges against the distribution of the deterministic solutions in the reference catalogue. Therefore, perfectly matching waveforms are penalized. However, they are sufficiently unlikely anyway.

> *4. To verify the validity of the synthetic noise added to waveforms in (16) and (17), it could be possible to check whether the distribution of decorrelation values between different realizations of u_pert and u_iˆc is log-normal, right?*

That is a very good observation, which creates a nice bridge between the two parts of the paper. We added a figure showing that this is indeed the case to the electronic supplement and refer to it in the section where the empirical likelihood is derived.

Added text on p.12:

"The log-normal distribution also fits best our synthetic data from sect. 2.4, as calculated with the perturbations in eqs 16, 17. See fig. S4 in the supplement for a corresponding QQ-plot."

> *5. The cumulative histogram for observed misfit values is difficult to see on Figure 5b.*

Actually, there is no cumulative histogram in the strict sense in plot 5b. It is a Quantile-Quantile plot, so the cumulative histogram for the observed values is on the Y-axis. We added the explanation to the caption of figure 5:

The values on the x-axis are percentiles of the cumulative histogram of *D* in our dataset. The y-axis shows the percentiles of the best fitting distribution of each class. The closer the percentiles are to a x=y line, the better the fit of the distribution to the underlying data over the whole range of values. Both subfigures indicate that a log-normal distribution best fits the values of *D = 1-CC*.

> *6. The empirical likelihood function is constructed from a set of pre-computed deterministic source solutions. Does the Bayesian solutions differ a lot from the deterministic solutions? It would be interesting to compare the distribution of residuals obtained from both methods. Are the Bayesian residuals also log-normally distributed ?*

As stated in the response to point 3, the residuals are indeed log-normally distributed, since this is what the sampling algorithm (whatever it is, in our case the *Neighborhood Algorithm*) tries to enforce.

Review 2 (Carl Tape)

> *This is a nice paper that (to me) offers two important parts: (1) the traditional cross-correlation (CC) provides a more stable measurement of waveform misfit than either the L1 or L2 norms (2) a complete data covariance matrix can be estimated (prior to the inversion) that takes into account the station spacing and statistical properties of noise in the data.*
> *The source is described as in Stahler and Sigloch (2014): hypocenter, magnitude, moment tensor, and coefficient for a source time function. The target is on P and SH teleseismic waveforms, and it is unclear how extensible this is to, say, surface waves. The paper is very well written and well presented, with the exception of a few points (below). Data covariances tend to be messy (and ignored), but I think this paper brings a certain amount of elegance to the problem. Using a full data covariance seems tractable (if CC is the misfit function), and I suspect there are cases where it will really impact the estimated parameters of the source.*
> *I recommend the paper be published after minor revisions.*
>
> *General points*
> *+ I think there needs to be more emphasis on the point that this study applies to teleseismic P wave source inversion. It is true that the main concepts are general, but some of these main concepts (or choices) could be overwhelmed by other factors for an inversion based on surface waves. Moment tensor inversion codes "standards" for global (GCMT, USGS) and regional (Dreger) scales are all based on surface waves, where cross-correlation measurements may be inappropriate in the presence of frequency-dependent dispersion. It might be helpful to have something in the discussion about what the authors expect might be needed if one were to consider surface waves in addition to body waves. Cycle skipping can be a nightmare with surface waves – even an excellent CC value can doom an inversion, for example. What will the noise characteristics be for surface wave measurements – do we expect them to follow a log-normal distribution?*

The referee raises an important point, which is the applicability of the derived empirical Likelihood function to other inverse problems with different data sets. As he correctly mentions, the proposed misfit function based on the cross-correlation coefficient will not be very useful in some contexts. Therefore, this paper limits itself to the source inversion problem with body-wave seismograms. I suspect that the CC values of surface waves will follow a similar distribution, but as the referee correctly points out, its discriminative power may belimited in this context.

> *+ I had some challenges determining exactly what was done. Search for the label 900, for example, to see the different descriptions. Are these 900 different earthquakes or could they be 900 different solutions for the same earthquake? (I'm assuming it's 900 different earthquakes, but please try to make this more clear, if so.)*

Indeed we meant 900 different earthquakes. The new version contains more precise wording.

p.3, l.17ff:

To reduce its dimensionality for Bayesian sampling, Stähler and Sigloch (2014) selected the 900 best-constrained deterministic STF solutions from the total of > 2, 000 obtained by Sigloch and Nolet (2006), and decomposed them into empirical orthogonal functions (EOFS), denoted s_l (t).

To reduce its dimensionality for Bayesian sampling, (Staehler2014) made use of a dataset of >2,000 deterministic earthquakes source solutions (depth, moment tensor and STF) obtained by (Sigloch2006). We selected the 900 best-constrained STFs and composed this set into empirical orthogonal functions (EOFS), denoted $s_l(t)$.

p.4, l.15:

replaced “900 deterministic source solutions” with “deterministic source solutions for 900 earthquakes”

p.24, caption to Fig. 5:

From 900 source inversions based on 200,000 broadbanded, teleseismic P-waveforms

replaced with

From 200,000 broadbanded, teleseismic P-waveforms for 900 earthquakes,

> *+ On a related point, is ns really the number of seismograms or is it the number of time windows used? If you have a P and S measured on two different components (Z and R), would that be ns = 4 or ns = 2?*

That would be ns = 4. Since we limited ourselves to measuring P-waves on the vertical component and S on the transverse component, we have only 2 time windows per seismogram, but this number can and should be increased.

> *+ In essence, the authors are proposing a misfit function based on the cross-correlation. The implication is that such a misfit function could be used within adjoint-based source or structure inversions. If so, perhaps make a statement that one could start with the misfit function and derive an adjoint source (e.g., Tromp et al. 2005).*

In my understanding of the Adjoint method, the problem with the maximum correlation coefficient is that the max{}-function contained in it is not Fréchet-differentiable. Therefore we would rather not add this discussion.

> *+ P15, L23. Somewhere in Step 5 you should at least mention that you generate synthetic waveforms and subject them to (some of) the same processing steps as the data. For example, D never appears in the description (though it is there).*
> *Add this clause: "... equation 20, which contains N decorrelations between observed and synthetic waveforms, and combine…"*

We added a step 6 for the actual sampling stage:

For each source model m proposed by the sampling algorithm, calculate synthetic seismograms, and pass them through the filters of step 2. Calculate the empirical Likelihood L*(m|d) (eq.31), which is multiplied with a suitable prior to obtain a posteriour probability for m. Parametrization of m, Bayesian sampling strategy, and construction of the posterior distribution of m are described in the companion paper (Stähler and Sigloch, 2014).

> *+ P17, L32. "where n is the number of stations used to estimate the source parameters of one earthquake" Would it be easier to generalize this to the number of time windows used for one earthquake? You have up to 3 components and up to X windows per time series.*

Agreed, replaced “stations” with “time windows on different seismogram components”

> *+ P3, L8. "Depth is one parameter and a normalized description of the moment tensor requires five more (a more rigorous and uniform parameterisation of the moment tensor has been derived by Tape and Tape, 2015, 2016)."*
> *Thanks for the nod. Stahler and Sigloch (2014) is distinguished by their use of a uniform*
> *parameterizaion for moment tensors, something that is seldom done in grid searches. (They also used an ad hoc constraint on the prior toward double couples.) You might want to replace the word "focal mechanism" (which people may interpret as double couple moment tensors, by which I mean eigenvalues -lam, 0, lam) with "moment tensor" (which includes "full" moment tensors). Your approach can probably handle full moment tensors even more easily than double couple moment tensors, since no constraint is needed in the case of full moment tensors.*

I absolutely agree, but I could not find the term “focal mechanism” anywhere close to this text position.

> *+ A fraction of your audience will invariably wonder, "So does it matter whether we use a sophisticated data covariance?" You did this for the Virginia earthquake. Do you want to comment on how the use of SD might impact the determination of the mean model or the associated uncertainties (the "Bayesian beachball" in the authors’ parlance).*

The referee is correct in pointing this out. However, the application of the method to a wider set of earthquakes is current and ongoing work. A systematic comparison of the Bayesian beachballs with classical ones will require much more effort and is beyond the scope of this technical paper.

> ==
> *Miscellaneous points*
> *+ Figure 3 is an excellent demonstration. A few comments:*
> *The text (P9, L24) says alpha=0.9, beta=0.5 for Figure 3. The text about the figure says alpha=0.4, beta=0.8.*

The text above the figure is correct. p.9, l.24 has been corrected.

> *"in 40deg distance" to "at an epicentral distance of 40deg"*
> *How many stations were used for this synthetic test? I don’t have a sense for the distribution of stations by distance or azimuth (or even how many).*

Only one station was used at an epicentral distance of 40 degree. Since an explosion source was used, the problem is azimuth-independent.

Replaced “measured in 40 deg distance” with
“measured at a station at 40° epicentral distance.”

It says "The normalization coefficients are plotted in the top right corner." Are they? (Or do they need to be?)

Since they are not very pertinent, we removed this referring phrase.

separate –> separately

done.

+ Eq. 28. Why log10? Using ln, you would have a number that is equivalent to percent difference in the limit of small values. Traditional amplitude measurements have been delta ln A (e.g., Baig and Dahlen, 2002).

The referee is absolutely correct in that it should be *ln*. The wrong latex symbol was used. Thanks for the careful checking!

+ There are several occurrences of words separated by /: uncorrelated/correlated 95%/2 sigma Gaussian/l^2 "l^1/l^2" Please find a way to eliminate the usage of /, since it is ambiguous (and, or, division?).

Has been replaced, mostly by “or” throughout the text.

+ P1, L10. "of the broadband fits" add": between observed and modeled waveforms.
+ Table 1. Might want alpha and beta here. They seem important.

done

+ Eq. 3. Describe k’ (presumably normalization factor).

done

+ P5, L28. "of how to construct" what?

The covariance matrix S_D. Has been added

+ I’ve come across some studies that weight stations by voronoi cells, so a big voronoi cell would have a big weight. Not sure if this is worth mentioning.

While it sounds like an interesting approach, it would require a whole new discussion of different station weighting methods, which are beyond the scope of this article.

+ Figure 4. I think it’d be better to just have six labels in the legend. Or delete the legend and extend the labels at right to "D - weak pert," D - strong pert.," etc. Add label to ""the decorrelation D has a higher...".

Good suggestion! We modified it according to the second idea.

+ Figure 8. I think the dotted lon-lat lines are more harmful than helpful in the small globe plots.
Please state if this is a subset of stations used in the inversion or whether this includes all stations in the inversion (how many for P and SH?). Please explain the two epicentral small circles plotted (32 and 85?).

Removed the lon-lat-lines. And added an explanation to the caption of Fig. 8:

For the analysis, only stations between 32° and 85°epicentral distance have been used, as marked by the dashed lines.