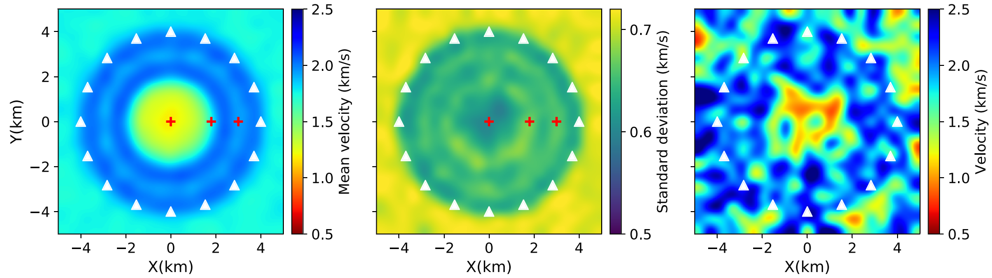

Figure 6. The mean (left), standard deviation (middle), and an individual realization from the approximate posterior distribution (right) obtained using SVGD. The red pluses show locations which are referred to in the main text.

3.1. Results

Figure 5 shows the mean, standard deviation, and an individual realization from the approximate posterior distribution calculated using ADVI. The mean model successfully recovers the low velocity anomaly within the receiver array except that the velocity value is slightly higher (~1.2 km/s) than the true value (1.0 km/s). Between the location of the central anomaly and that of the receiver array there is a slightly lower velocity loop. The standard deviation map shows standard deviations similar to that of the prior (0.72 km/s) outside of the array and clearly higher uncertainties at the location of the central anomaly. The standard deviations around the central anomaly are slightly higher than those at the center. Figure 6 shows the results from SVGD. Similarly, the velocity of the low velocity anomaly (~1.2 km/s) is slightly higher than the true value and a slightly lower velocity loop is also observed between the central anomaly and the receiver array. There is a clear higher uncertainty loop around the central anomaly; this has been observed previously and represent uncertainty due to the trade-off between the velocity of the anomaly and its shape (Galetti et al., 2015; Zhang et al., 2018). There is also another higher uncertainty loop associated with the lower velocity loop between the central anomaly and the receiver array. In contrast to this result, the loop cannot be observed in the results of ADVI.

To validate and better understand these results, Figure 7 shows the results from MH-McMC. The mean velocity model is very similar to the results from ADVI and SVGD. For example, the velocity value of the low velocity anomaly is higher than the true value, which suggests that the mean value of the posterior under the specified parameterization is genuinely biased toward higher values than the true value. A lower velocity loop is also observed between the circular anomaly and the receiver array. The standard deviation map shows similar results to those from SVGD: There is a higher uncertainty loop around the central anomaly and another one associated with the lower velocity loop between the circular anomaly and the receiver array. The latter loop suggests that this area is not well constrained by the data, and therefore, the mean velocity tends toward the mean value of the prior, which is lower than the true value. We do not observe the clear higher uncertainty loops in the result of ADVI, which may be due to the Gaussian approximation which is used to fit a non-Gaussian posterior in that method. In Figure 8 we show the results from rj-McMC. Compared to the results from the fixed-parameterization inversions, the mean velocity is a more accurate estimate of the true model and uncertainty across the model is also lower. For example, the middle low velocity anomaly has almost the same value as the true model and has standard deviation of only ~0.3 km/s compared to values significantly greater than 0.3 km/s for all other methods. Between the middle anomaly and the receivers, the

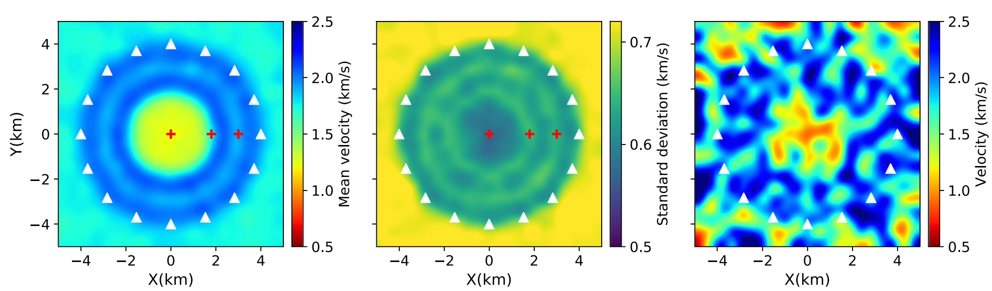

Figure 7. The mean (left), standard deviation (middle), and an individual realization from the approximate posterior distribution (right) obtained using MH-McMC. The red pluses show the point location which are referred to in the text.

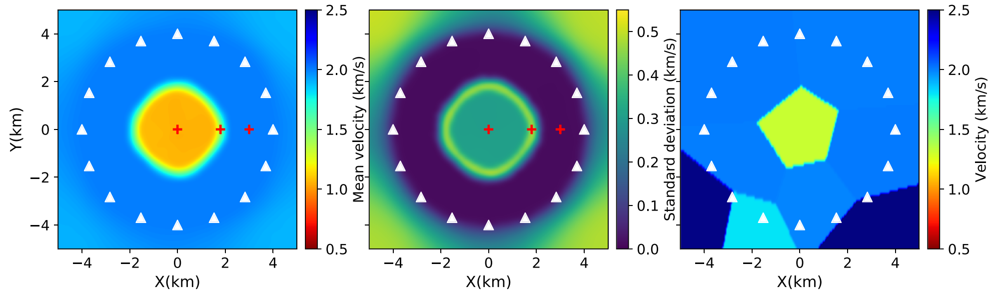

Figure 8. The mean (left), standard deviation (middle), and an individual realization from the approximate posterior distribution (right) obtained using transdimensional rj-McMC. The red pluses show the point location which are referred to in the text.

model is determined better than in the fixed-parameterization inversions (with a standard deviation smaller than 0.1 km/s). This is because in rj-McMC the model parameterization adapts to the data, which usually results in a lower-dimensional parameter space due to the natural parsimony of the method. For example, the average dimensionality of the parameter space in the rj-McMC inversion is around 10; for comparison the fixed-parameterization inversions all have dimensionality fixed to be 441. The standard deviation map from the rj-McMC also shows a clear higher uncertainty loop within the array around the low velocity anomaly and high uncertainties outside of the array where there is no data coverage.

Note that individual models from fixed-parameterization inversions (ADVI, SVGD, and MH-McMC) show complex structures because of their higher dimensionality and the simple Uniform prior distribution that we adopted (right panels in Figure 5–7). This might not be appropriate since the real Earth may have a smoother structure (de Pasquale & Linde, 2016; Ray & Myer, 2019). In that case, more informative prior information including some form of regularization might be used to produce smoother individual models (MacKay, 2003).

The results in Figure 8 do not show the double-loop uncertainty structure that is observed in the SVGD and MH-McMC results. The rj-McMC method contains an implicit natural parsimony—the method tends to use fewer rather than more cells whenever possible. While this may be useful in order to reduce the dimensionality of parameter space, it is also possible that it causes some detailed features of the velocity or uncertainty structure to be omitted, much like a smoothing regularization condition in other tomographic methods. Since the double-loop structure appears to be a robust feature of the image uncertainty, we assume that the parsimony has indeed regularized some of the image structure out of the rj-McMC results.

Note that the result from rj-McMC is fundamentally different from results obtained using the fixed-parameterization inversions (ADVI, SVGD, and MH-McMC) because of its entirely different parameterization. While the other inversion results are parameterized over a regular grid and can themselves be regarded as pixelated images, rj-McMC produces a set of models that are vectors containing positions and velocities of Voronoi cells, which can be transformed to an image on a regular grid (right panel in Figure 8). However, the Voronoi parametrization imposes prior restrictions on the pixelated form of models, for example, all pixels within each Voronoi cell have identical velocities. As a result rj-McMC produces very different results to those obtained using the other methods. In fact, the choice of parameterizaiton in rj-McMC can impose a variety of restrictions on models, and different parameterizations impose different prior information and so can produce very different standard deviation structures (Hawkins et al., 2019). Thus, the results of rj-McMC must always be interpreted in the light of the specific prior information imposed by the parameterization deployed, and whether this is expected to match the target structure.

To further analyze the results, in Figure 9 we show marginal probability distributions from the different inversion methods at three points (plus signs in Figures 5–8): Point (0, 0) at the middle of the model, point (1.8, 0) at the boundary of the low velocity anomaly which has higher uncertainties, and point (3, 0) which also has higher uncertainties in the results from SVGD and MH-McMC. Due to symmetries of the model, marginal distributions at these three points are sufficient to reflect much of the entire set of single-parameter marginal probability distributions. At point (0, 0), the three fixed-parameterization methods produce similar marginal probability distributions. However, the marginal distribution from rj-McMC is narrower and concentrates around the true solution (1.0 km/s). This is likely due to the fact that in rj-McMC we have a

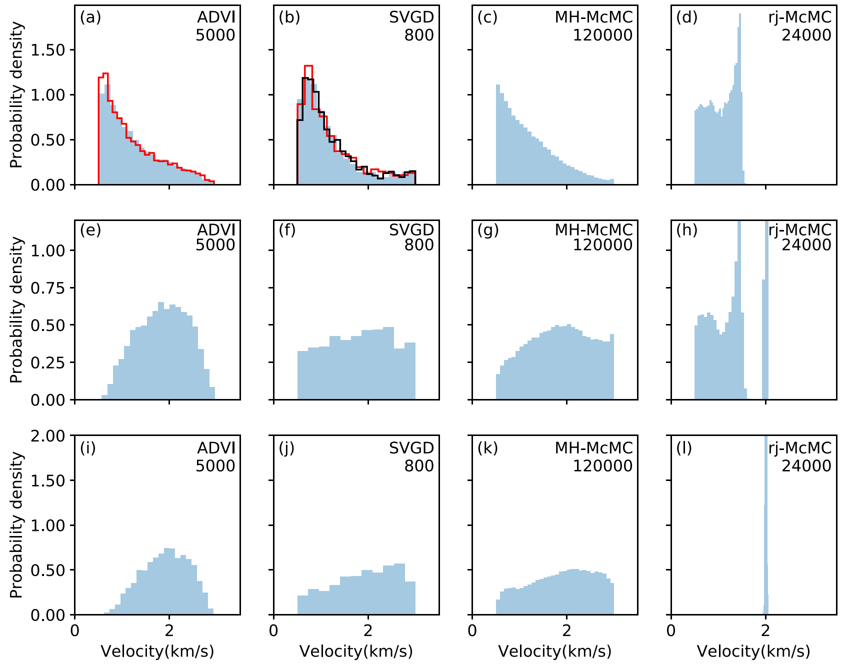

Figure 9. The marginal posterior pdfs of velocity at three points (pluses in Figures 3–6) derived using different methods. (a–d) show the marginal posterior distributions of velocity at the Point (0,0) from ADVI, SVGD, MH-McMC, and rj-McMC respectively. (e–h) show the marginal distributions at the Point (1.8,0) from the four methods respectively, and (i–l) show the marginal distributions at the Point (3,0) from the four methods, respectively. The red lines in (a) and (b) are marginal distributions obtained by doubling the number of iterations, and the black line in (b) shows the marginal distribution obtained using 1,600 particles. The number at the top right of each figure shows the number of Monte Carlo samples used for ADVI results and for the two McMC methods, and the number of particles used for SVGD.

much smaller parameter space than in the fixed-parameterization inversions. To assess the convergence, we show the marginal distributions obtained by doubling the number of iterations in ADVI and SVGD with a red line in Figures 9a and 9b. The results show that increasing iterations only slightly improves the marginal distributions, suggesting that they have nearly converged. The black line in Figure 9b shows the marginal distribution obtained using more particles (1,600) with the same number of iterations (500). The result is almost the same as the result obtained using the original set of particles, which suggests that 800 particles are sufficient in this case. At point (1.8, 0), the marginal distributions from the three fixed-parameterization inversions become broader, which explains the higher uncertainty loops observed in the standard deviation maps. The distribution from ADVI is more centrally focused than the other two, which is again suggestive of the limitations of that method caused by the Gaussian approximation. The distributions from SVGD and MH-McMC are more similar to each other and are close to the prior—a Uniform distribution—which suggests that the area is not well constrained by the data. By contrast, the result from rj-McMC shows a clearly multimodal distribution with one mode centered around the velocity of the anomaly (1 km/s) and the other around the background velocity (2 km/s) as discussed in Galetti et al. (2015). This multimodal distribution reflects the fact that it is not clear whether this point is inside or outside of the anomaly, which produces the higher uncertainty loop in the standard deviation map. This suggests that there are different causes of the higher uncertainty loops in the different models. In the fixed-parameterization inversions (ADVI, SVGD, and MH-McMC) the higher uncertainty loops are mainly caused by the low resolution of the data at the boundary of the low velocity anomaly, which produces broader marginal distributions. In the rj-McMC inversion, the higher uncertainty loops are mainly caused by multimodality in the posterior pdf. At point (3.0, 0) similarly to the point (0, 0), the marginal distributions from the three fixed-parameterization inversions have similar shape and are much broader than the result from rj-McMC. Compared to the results from SVGD and MH-McMC, the result from ADVI again shows a more centrally focused distribution reminiscent of the Gaussian limitation implicit in ADVI. In the result of rj-McMC the marginal distribution concentrates to a very narrow distribution around the true value. Overall, the marginal distributions from the

10.1029/2019JB018589

Table 1
The Comparison of Computational Cost for All Four Methods

Method	Number of simulations	CPU hours
ADVI	10,000	0.45
SVGD	400,000	8.53
MH-McMC	12,000,000	480.3
rj-McMC	3,000,000	102.6

fixed-parameterization inversions are broader than the result from rj-McMC due to their far larger parameter space. Note that although the marginal distributions from SVGD and MH-McMC have slightly different shape, which causes differences in the magnitudes of their standard deviation maps, the maps are essentially similar from these quite different methods which suggests that the results are (approximately) correct.

3.2. Computational Cost

Table 1 summarizes the computational cost of the different methods. ADVI involves 10,000 forward simulations which takes 0.45 CPU hours. However, note that in ADVI we used the full-rank covariance matrix, which becomes huge in high-dimensional parameter spaces, which could make the method inefficient. SVGD involves 400,000 forward simulations, which takes 8.53 CPU hours. This appears to make it less efficient than ADVI; however, SVGD can produce a more accurate approximation to the posterior pdf than ADVI which is limited by the Gaussian approximation. Note that SVGD can easily be parallelized by computing the gradients in equation (19) in parallel, making the method more time efficient. For example, the above example takes 0.97 hr when parallelized using 10 cores. In comparison, MH-McMC requires 2,000,000 simulations for one chain to stabilize, which takes about 80.05 CPU hours, so for all six chains it requires 480.3 CPU hours in total. The rj-McMC run involved 500,000 simulations for one chain, which takes about 17.1 CPU hours, so 102.6 CPU hours in total for six chains. The MC methods use evaluations of the likelihood and prior distribution at each sample, whereas both variational methods also deploy the information in the various gradients in equations (9), (10), and (19). The number of simulations is therefore not a good metric to compare the four methods, since the gradients in this case are calculated by ray tracing, which require more calculations per simulation in Table 1 compared to those required for MC. CPU hours is a fairer metric for comparison, but of course, this depends on the mechanism by which gradients are obtained: In other forward or inverse problems it is even possible that the variational methods take longer than MC if estimating gradients requires extensive computation.

In the comparison in Table 1, rj-McMC is more efficient than MH-McMC due to the fact that rj-McMC explores a much smaller parameter space than the fixed parameterization in MH-McMC. However, note that this might not always be true since transdimensional steps in rj-McMC usually have a very low probability of being accepted (Bodin & Sambridge, 2009; Zhang et al., 2018) and the method is generally significantly more difficult to tune (Green & Hastie, 2009). Overall, obtaining solutions from variational methods (ADVI and SVGD) is more efficient than MC methods since they turn the Bayesian inference problem into an optimization problem. This also makes variational inference methods applicable to larger data sets and offers the advantage that very large data sets can be divided into random minibatches and inverted using stochastic optimization (Kubrusly & Gravier, 1973; Robbins & Monro, 1951) together with distributed computation. MC methods can be very computationally expensive for large data sets. Of course, the above comparison depends on the methods used to assess convergence for each method, which introduces some subjectivity in the comparison so that the absolute time required by each method may not be entirely accurate. Nevertheless, from all tests that we have conducted it is clear that variational methods produce stable solutions to the above problem far more efficiently than Metropolis-Hastings and rj-McMC methods. Note that some other MC sampling methods, for example, Hamiltonian MC, also use gradient information and may be more efficient than Metropolis-Hastings methods (Fichtner et al., 2018; Neal et al., 2011; Sen & Biswas, 2017).

4. Application to Grane Field

The Grane field is situated in the North Sea and contains a permanent monitoring system composed of 3,458 four-component sensors measuring three orthogonal components of particle velocity and water pressure variations due to passing seismic waves. Zhang et al. (2020) used beamforming to show that the noise

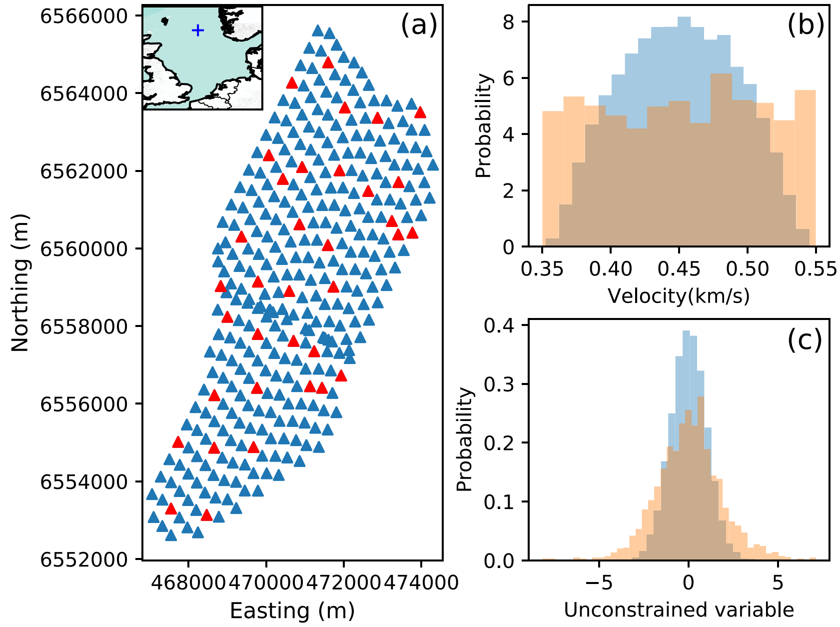

Figure 10. (a) The distribution of receivers (blue and red triangles) across the Grane field used in this study. Red triangles show the receivers that were used as virtual sources. The blue plus in the inset map shows the location of Grane field in the North Sea. The histograms show the initial distributions of each parameter in the (b) original space (velocity) and (c) transformed unconstrained space for ADVI (blue) and for SVGD (orange). Similar to Figure 4, we used 5,000 Monte Carlo samples to show probability distributions in both the original and the unconstrained space for ADVI. The initial distribution for SVGD is approximated using 1,000 particles generated from the prior (a Uniform distribution) in the original space and transformed to the unconstrained space.

sources measured in the Grane field are nearly omnidirectional, which allows us to use ambient seismic noise tomography to study the subsurface of the field. To reduce the computational cost, in this study we downsampled the number of receivers by a factor of 10, which results in 346 receivers, and we only used 35 receivers as virtual sources (Figure 10a). Cross correlations are computed between vertical component recordings at pairs consisting of a virtual source and a receiver using half-hour time segments, and the set of correlations for each pair were stacked over 6.5 hr. This process produces approximate virtual-source seismograms of Rayleigh-type Scholte waves (Campillo & Paul, 2003; Curtis et al., 2006; Shapiro et al., 2005; Wapenaar & Fokkema, 2006). Phase velocity dispersion curves for each (virtual) source-receiver pair are then automatically picked using an image transformation technique: For all processing details see Zhang et al. (2020), which presents a more complete ambient noise analysis of the field and presents tomographic phase velocity maps at various frequencies as well as estimated shear velocity structure of the near seabed subsurface. Here we use the recording phase velocity data at 0.9-s period.

We apply the variational inference methods ADVI and SVGD, and rj-McMC to the data to obtain phase velocity maps at 0.9 s and compare the results. For variational methods, the field is parametrized using a regular 26×71 grid with a spacing of 0.2 km in both x and y directions giving a velocity model dimensionality of 1,846. Due to its computational cost in such high-dimensional space, we do not apply MH-McMC. The data noise level is set to be 0.05 s, which is an average value estimated by the hierarchical Bayesian MC inversion of Zhang et al. (2020). The prior pdf of phase velocity in each model cell is set to be a Uniform distribution between 0.35 and 0.55 km/s, which is selected to be wider than the minimum (0.4 km/s) and maximum (0.5 km/s) phase velocity picked from cross correlations. The initial probability distribution for ADVI is chosen similarly to that in the synthetic tests: A standard Gaussian distribution in the unconstrained space (blue histogram in Figure 10c) and its shape in the original space is shown in Figure 10b (blue histogram). For SVGD, the initial distribution is approximated using 1,000 particles generated from the prior in the original space (orange histogram in Figure 10b) and transformed to the unconstrained space (orange histogram in Figure 10c). We then applied 10,000 iterations for ADVI and 500 iterations for SVGD. Similarly to the synthetic test above, for rj-McMC we use Voronoi cells to parameterize the model. The prior pdf of the number of cells is set to be a discrete Uniform distribution between 30 and 200, and the data noise level

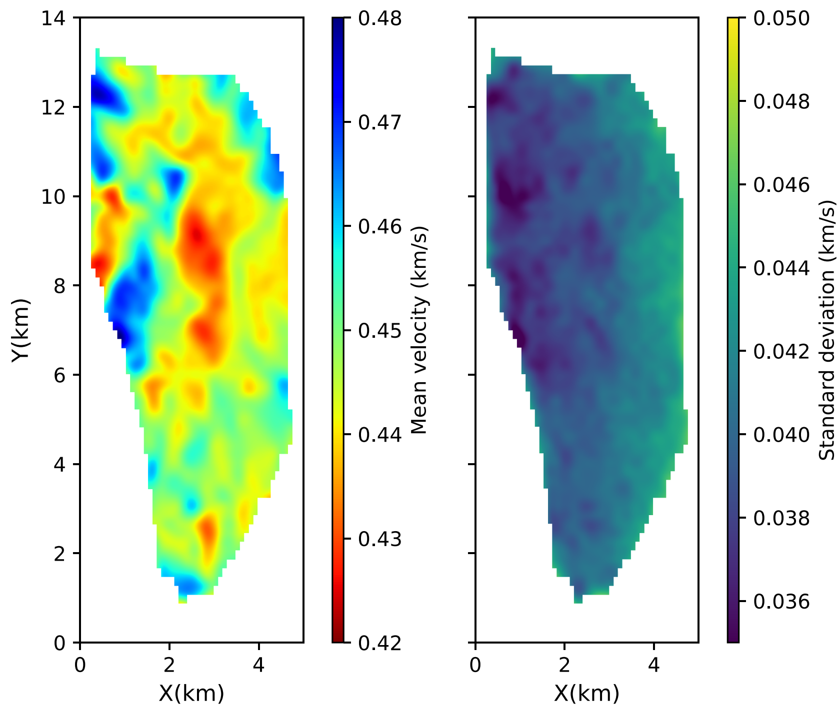

Figure 11. The mean (left) and standard deviation map (right) obtained for Grane using ADVI.

is estimated hierarchically during the inversion (Zhang et al., 2018). Proposal distributions are the same as in the synthetic test above. We used a total of 16 chains, each of which contains 800,000 iterations including a burn-in period of 400,000. To reduce the correlation between samples we only retain every fiftieth sample post burn-in for our final ensemble.

Figure 11 shows the mean and standard deviation maps from ADVI. The mean phase velocity map shows a clear low velocity anomaly around the center of the field from $Y = 6$ km to $Y = 10$ km and another at the western edge between $Y = 8$ km and $Y = 10$ km. These were also observed by (Zhang et al., 2020) using Eikonal tomography, who showed that they are correlated with areas of higher density of pockmarks on the seabed, suggesting that they are caused by near-surface fluid flow effects. At the western edge between $Y = 6$ km and $Y = 8$ km and at the northwestern edge there are high velocity anomalies which were also observed in the results of Zhang et al. (2020). In the north between $Y = 11$ km and $Y = 12$ km and along the eastern edge between $Y = 7$ km and $Y = 10$ km the model shows some low velocity anomalies. Moreover, there are some small anomalies distributed across the field. For example, to the south of the central low velocity anomaly around $Y = 6$ km there are several other low velocity anomalies. Similarly, there is a small low velocity anomaly and a small high velocity anomaly in the south of the field around $Y = 2.5$ km and a small high velocity anomaly in the north around $Y = 10.5$ km.

Overall, the standard deviation map shows that uncertainty in the west is lower than in the east. At the western edge there are some low uncertainty areas, which are associated with velocity anomalies. For example, the low uncertainty area between $Y = 6$ km and $Y = 8$ km is associated with the high velocity anomaly at the same location. Similarly, the high velocity anomaly at the northwestern edge around $Y = 12$ km shows a lower uncertainty, and the middle low velocity anomaly also shows slightly lower uncertainties. This might suggest that these velocity structures are well constrained by the data. However, in the synthetic tests we noticed that the ADVI can produce biased standard deviation maps due to the Gaussian approximation, so these uncertainty properties may not be robust.

We show the mean and standard deviation maps obtained using SVGD in Figure 12. The mean velocity map shows very similar structures to the result from ADVI, except that the velocity magnitudes are slightly different. For example, we observe the central low velocity anomaly and one at the western edge, which appeared in the mean velocity map from ADVI and are related to the density distribution of pockmarks. Similarly, there are high velocity anomalies at the western edge and a low velocity anomaly at the eastern

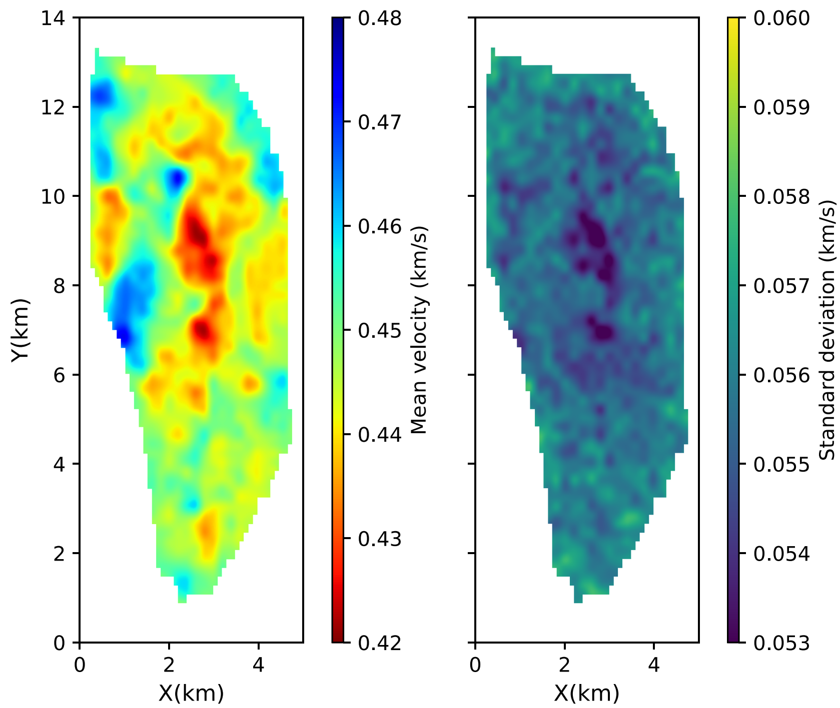

Figure 12. The mean (left) and standard deviation map (right) obtained for Grane using SVGD.

edge. Even for more detailed structure, for example, the low velocity anomalies at the north ($Y > 10$ km), the low velocity anomalies around $Y = 6$ km and the small velocity anomalies around $Y = 2.5$ km, the two results show highly consistent properties between the two methods. This suggests that we have obtained accurate mean phase velocity maps given the fixed, gridded model parameterization and the observed data.

Despite the similarity in the mean results, the standard deviation map from SVGD is quite different from the results from ADVI, which is consistent with variations that we observed in the synthetic tests. For example, there is no clear magnitude difference between the west and the east as appeared in the result from ADVI. There is a clear low uncertainty area associated with the central low velocity anomaly, which is slightly lower in magnitude than the result from ADVI. Similarly, there is a slightly lower uncertainty area at the western edge associated with the low velocity anomaly at the same location. The south-central low velocity anomaly around $Y = 6$ km also exhibits relatively low uncertainties, which suggests that those small low velocity anomalies in this area may reflect true properties of the subsurface. Similarly, there are some low uncertainty structures at the north around $Y = 11$ km, which are associated with low velocity anomalies. Note that due to the Gaussian approximation in ADVI, the standard deviation results from SVGD show different magnitudes as we saw in the synthetic tests.

Figure 13 shows the mean and standard deviation maps obtained from rj-McMC. The mean velocity map shows broadly similar structures to the results from ADVI and SVGD. For example, we also observed the middle low velocity anomaly, the low velocity anomalies at the western and eastern edges, and the high velocity anomalies at the western edge. However, compared to the previous results these structures are smoother, which is probably caused by the constant-velocity Voronoi cell parameterization and the natural parsimony that is implicit within the rj-McMC inversion method (Bodin & Sambridge, 2009; Green, 1995) similarly to the synthetic tests above. The small velocity anomalies in the previous results disappear in the result from rj-McMC; this may also be caused by the natural parsimony of rj-McMC or by overfitting of data in the variational methods due to the fixed parameterization. However, the small high and low velocity anomalies around $Y = 2.5$ km and around $Y = 10.5$ km still exist, which suggests that these detailed velocity structures may represent real properties of the subsurface (or are caused by a consistent bias in the data).

Similarly to the synthetic tests, the standard deviation map from rj-McMC shows significantly smaller uncertainties (<0.01 km/s) than the results from ADVI (~0.04 km/s) and SVGD (~0.055 km/s), which is probably caused by a lower dimensionality of parameter space used in rj-McMC (around 60 Voronoi cells were used)

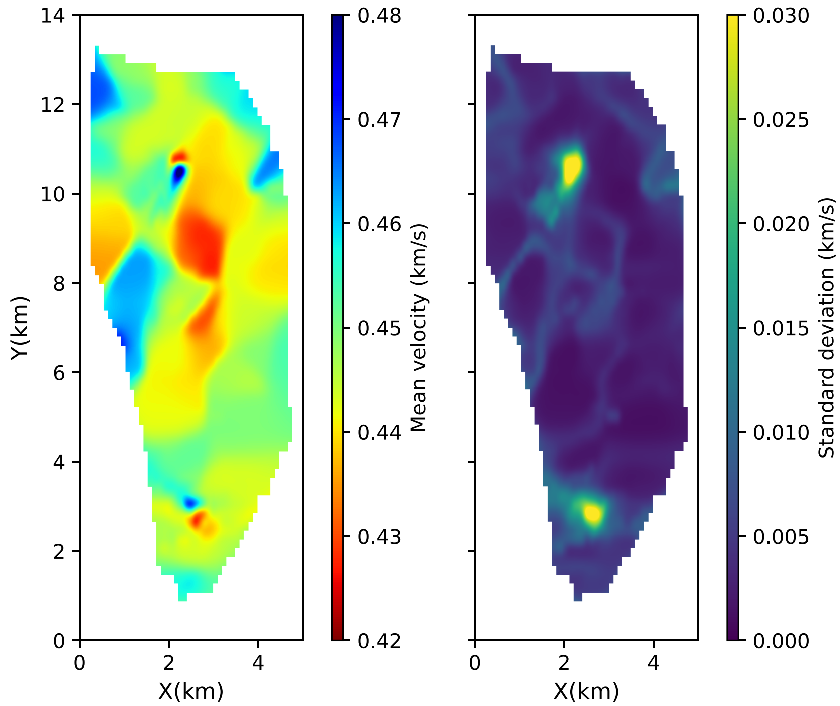

Figure 13. The mean (left) and standard deviation map (right) obtained for Grane using rj-McMC.

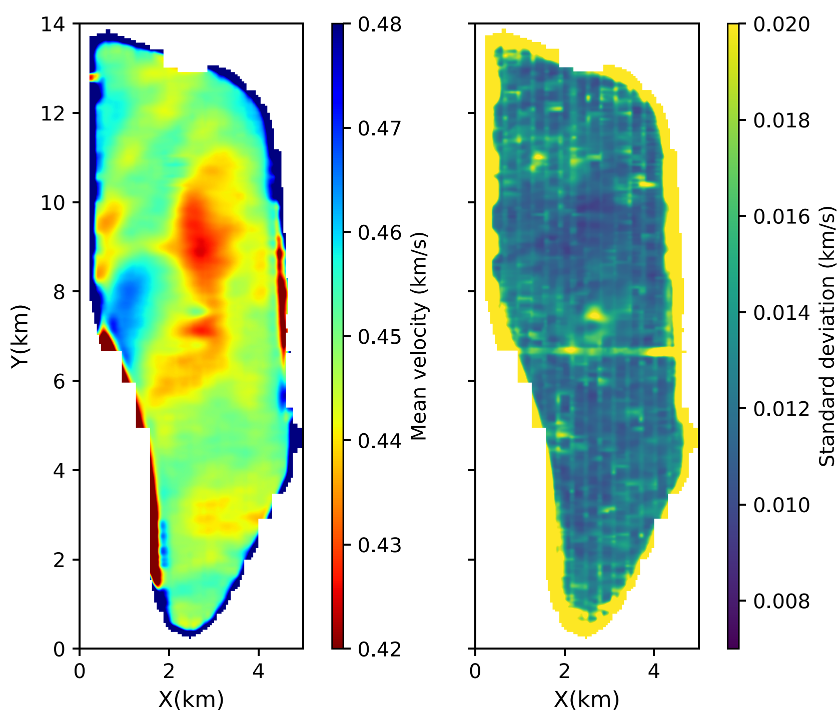

Figure 14. The mean (left) and standard deviation map (right) obtained for Grane using Eikonal tomography by Zhang et al. (2020).

than in variational methods (1,846 parameters), resulting in fewer trade-offs between parameters. However, there are higher uncertainties at the location of the small velocity anomalies at $Y = 2.5$ km and at $Y = 10.5$ km, which is probably due to the fact that not all chains found these small structures. Loop-like structures are also observed to trace out the most abrupt velocity transitions, similarly to Figure 8.

To compare our results with traditional methods, Figure 14 shows the mean and standard deviation maps obtained using Eikonal tomography by Zhang et al. (2020) using all of the available data (3,458 virtual sources and 3,458 receivers). The mean velocity model shows similar but slightly smoother structures compared to those obtained using ADVI and SVGD. This may be because the larger quantity of data used in Eikonal tomography reduces the noise and stabilizes the results, or because the interpolation used in Eikonal tomography regularizes (smooths) small-scale structure. The standard deviation map shows lower uncertainties at the location of the middle low velocity anomaly, which is similar to that obtained using SVGD. This again suggests that SVGD can produce a more accurate standard deviation estimate than ADVI. The mean velocity model from rj-McMC shows smoother structures than that from Eikonal tomography, which may suggest that rj-McMC omits small-scale structure due to its implicit parsimony. The standard deviation map from rj-McMC also does not show similar structures to those obtained using Eikonal tomography or SVGD due to the completely different parameterizations employed.

In the inversion, ADVI involved 10,000 forward simulations which took 5.1 CPU hours and SVGD involved 500,000 forward simulations, which required 141.8 CPU hours. By contrast the rj-McMC involved 12,800,000 forward simulations to obtain an acceptable result, which required 1,866.1 CPU hours. In real time, SVGD was in fact parallelized using 12 cores, which took 12.1 hr to run, while rj-McMC was parallelized using 16 cores, which therefore took about 5 days. We conclude that, although the variational methods produce higher uncertainty estimates, they can produce similar parameter estimates (mean velocity) at hugely reduced computational cost, and indeed, our synthetic tests suggest that the variational SVGD image uncertainty results may in fact be the more correct.

5. Discussion

We have shown that variational methods (ADVI and SVGD) can be applied to seismic tomography problems and provide efficient alternatives to McMC. ADVI produces biased posterior pdfs because of its implicit Gaussian approximation and cannot be applied to problems with multimodal posteriors. However, it still generates an accurate estimate of the mean model. Given that it is very efficient (only requiring 10,000 forward simulations), the method could be useful in scenarios where efficiency is important and a Gaussian approximation is sufficient for uncertainty analysis. Alternatively, a mixture of Gaussians approximation might be used to improve the accuracy of the algorithm (Arenz et al., 2018; Zobay, 2014). In a very high dimensional case, ADVI could become less efficient because of the increased size of the Gaussian covariance matrix. In that case one could use a mean-field approximation (setting model covariances to 0) or use a sparse covariance matrix to reduce computational cost since seismic velocity in any cell is often most strongly correlated with that in neighboring cells.

SVGD can produce a good approximation to posterior pdfs. However, since it is based on a number of particles, the method is more computationally costly than ADVI. In this study we parallelized the computation of gradients to improve the efficiency, and for large data sets further improvements can be obtained by using random minibatches to perform the inversion (Liu & Wang, 2016). Such a strategy can be applied to any variational inference method (e.g., also ADVI) since variational methods solve an optimization rather than a stochastic sampling problem. In comparison, this strategy cannot easily be used in McMC-based methods since it may break the detailed balance requirement of McMC (Blei et al., 2017). Though it has been shown that SVGD requires fewer particles than particle-based sampling methods (e.g., sequential MC) in the sense that it reduces to finding the MAP model if only one particle is used, the optimal choice of the number of particles remains unclear, especially for very high dimensional spaces. In the case of very high dimensionality another possibility is to use normalizing flows—a variational method based on a series of specific invertible transforms (Rezende & Mohamed, 2015).

MC and variational inference are different types of methods that solve the same problem. MC simulates a set of Markov chains and uses samples of those chains to approximate the posterior pdf, while variational inference solves an optimization problem to find the closest pdf to the posterior within a given family of probability distributions. MC methods provide guarantees that samples are asymptotically distributed according to the posterior pdf as the number of samples tends to infinity (Robert & Casella, 2013), while the statistical properties of variational inference algorithms are still unknown (Blei et al., 2017). It is possible to combine the two methods to capitalize on the merits of both. For example, the approximate posterior pdf from an efficient variational method (e.g., ADVI) can be used as a proposal distribution for Metropolis-Hastings (De Freitas et al., 2001) to improve the efficiency of McMC, or McMC steps can be integrated to the variational approximation to improve the accuracy of variational methods (Salimans et al., 2015).

We used a fixed regular grid of cells to parameterize the tomographic model in the variational methods, which might introduce overfitting of the data. For example, the mean velocity models in the synthetic tests show a slightly lower velocity loop between the low velocity anomaly and the receivers, and the uncertainties obtained from fixed-parameterization inversions are significantly higher than the results from rj-McMC. However, although rj-McMC produces lower uncertainty estimates, this is because small-scale structures are omitted in the results of rj-McMC due to their implicitly imposed a priori information and natural parsimony. For example, in our real data example, small-scale structures in the results of variational inference methods and Eikonal tomography are smoothed out in the results of rj-McMC. Indeed, the parameterization used in rj-McMC imposes restrictions on models, and different parameterizations can produce different uncertainties (Hawkins et al., 2019). This makes the interpretation and use of uncertainties from rj-McMC difficult.

It is not easy to determine an optimal grid in variational inference methods since this introduces a trade-off between resolution of the model and overfitting of the data. Therefore, it might be necessary to use a more flexible parameterization, for example, Voronoi cells (Bodin & Sambridge, 2009; Zhang et al., 2018) or wavelet parameterization (Fang & Zhang, 2014; Hawkins & Sambridge, 2015; Zhang & Zhang, 2015). It may also be possible to apply a series of different parameterizations and select the best one using model selection theory (Arnold & Curtis, 2018; Curtis & Snieder, 1997; Walter & Pronzato, 1997). Note that this might make the methods less computationally efficient to find an optimal parameterization because we may need to run a series of optimization problems with different parameterizations. However, in cases with very large data sets which may more suitably be solved by variational inference methods, it might instead be sufficient to use a parameterization with the highest resolution that the frequency of the data could resolve. Instead, some more informative prior or regularization may be used to reduce the magnitude of uncertainty estimates and to better constrain the model (MacKay, 2003; Ray & Myer, 2019).

In our experiments the results from rj-McMC are significantly different from the results obtained using variational methods or MH-McMC. This is essentially caused by different parameterizations. In ADVI, SVGD, and MH-McMC we invert for a pixelated image, while in rj-McMC we invert for a distribution of parameters that represent locations and shapes of cells and their spatially constant velocities, the pointwise spatial mean of which is visualized as an image. Therefore, even though we visualized them in the same way, the results are essentially not directly comparable. Nevertheless, the comparison with rj-McMC is interesting because until now a quite different alternative probabilistic method was never used to estimate the posterior of images from the same realistic tomography problem. The results here demonstrate that the rj-McMC method as applied in most tomography papers gives significantly different solutions than we might previously have thought; specifically, it does not produce the posterior distribution of the pixelated image that is usually shown in scientific papers (e.g., Bodin & Sambridge, 2009; Crowder et al., 2019; Galetti et al., 2015; Zulfakriza et al., 2014). Rather, it samples a probability distribution in a particular irregular and variably parametrized model space, and results should be interpreted as such. Note that some other methods, for example, rj-McMC with Gaussian processes, may provide results that can be compared between all sampling methods, and provide a means of injecting prior information with adaptable complexity into the sampling scheme (Ray & Myer, 2019).

In this study we used a fixed data noise level in the variational methods. It has been shown that an improper noise level can introduce biases in tomographic results (Bodin & Sambridge, 2009; Zhang et al., 2020), so in our example we used the noise level estimated by hierarchical McMC. It can also be estimated by a variety of other methods (Bensen et al., 2009; Nicolson et al., 2012, 2014; Weaver et al., 2011; Yao & Van Der Hilst, 2009) and maximum likelihood methods (Ray et al., 2016; Ray & Myer, 2019; Sambridge, 2013). In future it might also be possible to include the noise parameters in variational methods in a hierarchical way.

In this study we applied variational inference methods to simple 2-D tomography problems, but it is straightforward to apply the methods to any geophysical inverse problems whose gradients with respect to the model can be computed efficiently. For example, variational methods could be applied to 3-D seismic tomography problems to provide an efficient approximation, which generally demands enormous computational resources using McMC methods (Hawkins & Sambridge, 2015; Zhang et al., 2018, 2020). The methods also provide possibilities to perform Bayesian inference for full waveform inversion, which is generally very expensive for McMC (Ray et al., 2017) and suffers from notorious multimodality in the likelihoods. SVGD provides a possible way to approximate these complex distributions given that theoretically it can approximate arbitrary distributions.

6. Conclusion

We introduced two variational inference methods to geophysical tomography—ADVI and SVGD, and applied them to 2-D seismic tomography problems using both synthetic and real data. Compared to the McMC method, ADVI provides an efficient but biased approximation to Bayesian posterior probability density functions and cannot be applied to find multimodal posteriors because of its implicit Gaussian assumption. In contrast, SVGD is slightly slower than ADVI but produces a more accurate approximation. The real data example shows that ADVI and SVGD produce very similar mean velocity models, even though their uncertainty estimates are different. The mean velocity models are very similar to those produced by reversible jump McMC (rj-McMC), except that the mean model from rj-McMC is smoother because of the much lower dimensionality of its parameter space. Variational methods thus can provide efficient approximate alternatives to McMC methods and can be applied to many geophysical inverse problems.

Appendix A: The Entropy of a Gaussian Distribution

The entropy $\mathrm{H}\left[q(\boldsymbol{\theta};\phi)\right]$ of a Gaussian distribution $\mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu},\mathbf{LL}^{\mathrm{T}})$ is as follows:

$$\begin{aligned}\mathrm{H}\left[q(\boldsymbol{\theta};\boldsymbol{\phi})\right] &= -\mathrm{E}_q[\log q(\boldsymbol{\theta})]\\ &= -\int \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu},\mathbf{LL}^{\mathrm{T}})\log\mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu},\mathbf{LL}^{\mathrm{T}})d\boldsymbol{\theta}\\ &= \frac{k}{2}+\frac{k}{2}\log(2\pi)+\frac{1}{2}\log|\det(\mathbf{LL}^{\mathrm{T}})|\end{aligned}$$

where k is the dimension of vector $\boldsymbol{\theta}$. The gradients with respect to $\boldsymbol{\mu}$ and $\mathbf{L}$ can be easily calculated (see Appendix B).

Appendix B: Gradients of the ELBO in ADVI

We first describe the dominated convergence theorem (DCT) (Çınlar, 2011):

Theorem Assume $X \in \mathcal{X}$ is a random variable and $f : \mathbb{R}\times\mathcal{X}\to\mathbb{R}$ is a function such that $f(t,X)$ is integrable for all t and $\frac{\partial f(t,X)}{\partial t}$ exists for each t. Assume that there is a random variable Z such that $|\frac{\partial f(t,X)}{\partial t}| \leq Z$ for all t and $\mathrm{E}(Z) < \infty$. Then

$$\frac{\partial}{\partial t}\mathrm{E}(f(t,X)) = \mathrm{E}(\frac{\partial}{\partial t}f(t,X))$$

The proof of this theorem is given in Çınlar (2011).

We then calculate the gradients in equations (9) and (10) based on Kucukelbir et al. (2017). The ELBO $\mathcal{L}$ is:

$$\mathcal{L} = \mathrm{E}_{\mathcal{N}(\boldsymbol{\eta}|\mathbf{0},\mathbf{I})}\left[\log p\left(T^{-1}\left(R_{\phi}^{-1}(\boldsymbol{\eta})\right),\mathbf{d}_{obs}\right)+\log|\det\mathbf{J}_{T^{-1}}\left(R_{\phi}^{-1}(\boldsymbol{\eta})\right)|\right]+\mathrm{H}\left[q(\boldsymbol{\theta};\phi)\right]$$

where $\mathrm{H}[q(\boldsymbol{\theta};\phi)] = -\mathrm{E}_q[\log q(\boldsymbol{\theta})]$ is the entropy of distribution q. Assume $\frac{\partial}{\partial\boldsymbol{\phi}}\log p$ is bounded where ϕ represents variational parameters $\boldsymbol{\mu}$ and $\mathbf{L}$, then the gradients can be computed by exchanging the derivative and the expectation using the DCT and applying the chain rule:

$$\nabla_{\boldsymbol{\mu}}\mathcal{L} = \nabla_{\boldsymbol{\mu}}\left\{\mathrm{E}_{\mathcal{N}(\boldsymbol{\eta}|\mathbf{0},\mathbf{I})}\left[\log p\left(T^{-1}\left(R_{\phi}^{-1}(\boldsymbol{\eta})\right),\mathbf{d}_{obs}\right)+\log|\det\mathbf{J}_{T^{-1}}\left(R_{\phi}^{-1}(\boldsymbol{\eta})\right)|\right]+\mathrm{H}\left[q(\boldsymbol{\theta};\phi)\right]\right\}$$

Applying the DCT and since H does not depend on $\boldsymbol{\mu}$,

$$\nabla_{\boldsymbol{\mu}}\mathcal{L} = \mathrm{E}_{\mathcal{N}(\boldsymbol{\eta}|\mathbf{0},\mathbf{I})}\left[\nabla_{\boldsymbol{\mu}}\left\{\log p\left(T^{-1}\left(R_{\phi}^{-1}(\boldsymbol{\eta})\right),\mathbf{d}_{obs}\right)\right\}+\nabla_{\boldsymbol{\mu}}\left(\log|\det\mathbf{J}_{T^{-1}}\left(R_{\phi}^{-1}(\boldsymbol{\eta})\right)|\right)\right]$$

Applying the chain rule,

$$\begin{aligned}\nabla_{\boldsymbol{\mu}}\mathcal{L} &= \mathrm{E}_{\mathcal{N}(\boldsymbol{\eta}|\mathbf{0},\mathbf{I})}\left[\nabla_{\mathbf{m}}\log p(\mathbf{m},\mathbf{d}_{obs})\nabla_{\boldsymbol{\theta}}T^{-1}(\boldsymbol{\theta})\nabla_{\boldsymbol{\mu}}R_{\phi}^{-1}(\boldsymbol{\eta}) + \nabla_{\boldsymbol{\theta}}\log|\det\mathbf{J}_{T^{-1}}(\boldsymbol{\theta})|\nabla_{\boldsymbol{\mu}}R_{\phi}^{-1}(\boldsymbol{\eta})\right] \\ &= \mathrm{E}_{\mathcal{N}(\boldsymbol{\eta}|\mathbf{0},\mathbf{I})}\left[\nabla_{\mathbf{m}}\log p(\mathbf{m},\mathbf{d}_{obs})\nabla_{\boldsymbol{\theta}}T^{-1}(\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}}\log|\det\mathbf{J}_{T^{-1}}(\boldsymbol{\theta})|\right]\end{aligned}$$

The gradient with respect to $\mathbf{L}$ can be obtained similarly:

$$\begin{aligned}\nabla_{\mathbf{L}}\mathcal{L} = \nabla_{\mathbf{L}}\Big\{&\mathrm{E}_{\mathcal{N}(\boldsymbol{\eta}|\mathbf{0},\mathbf{I})}\left[\log p\left(T^{-1}\left(R_{\phi}^{-1}(\boldsymbol{\eta})\right),\mathbf{d}_{obs}\right) + \log|\det\mathbf{J}_{T^{-1}}\left(R_{\phi}^{-1}(\boldsymbol{\eta})\right)|\right] + \frac{k}{2} + \frac{k}{2}\log(2\pi) \\ &+\frac{1}{2}\log|\det(\mathbf{L}\mathbf{L}^{\mathrm{T}})|\Big\}\end{aligned}$$

Applying the DCT,

$$\begin{aligned}\nabla_{\mathbf{L}}\mathcal{L} =& \mathrm{E}_{\mathcal{N}(\boldsymbol{\eta}|\mathbf{0},\mathbf{I})}\left[\nabla_{\mathbf{L}}\left\{\log p\left(T^{-1}\left(R_{\phi}^{-1}(\boldsymbol{\eta})\right),\mathbf{d}_{obs}\right)\right\} + \nabla_{\mathbf{L}}\left(\log|\det\mathbf{J}_{T^{-1}}\left(R_{\phi}^{-1}(\boldsymbol{\eta})\right)|\right)\right] \\ &+ \nabla_{\mathbf{L}}\frac{1}{2}\log|\det(\mathbf{L}\mathbf{L}^{T})|\end{aligned}$$

and applying the chain rule, we obtain

$$\begin{aligned}\nabla_{\mathbf{L}}\mathcal{L} &= \mathrm{E}_{\mathcal{N}(\boldsymbol{\eta}|\mathbf{0},\mathbf{I})}[\nabla_{\mathbf{m}}\log p(\mathbf{m},\mathbf{d}_{obs})\nabla_{\boldsymbol{\theta}}T^{-1}(\boldsymbol{\theta})\nabla_{\mathbf{L}}R_{\phi}^{-1}(\boldsymbol{\eta}) + \nabla_{\boldsymbol{\theta}}\log|\det\mathbf{J}_{T^{-1}}(\boldsymbol{\theta})|\nabla_{\mathbf{L}}R_{\phi}^{-1}(\boldsymbol{\eta})] + (\mathbf{L}^{-1})^{T} \\ &= \mathrm{E}_{\mathcal{N}(\boldsymbol{\eta}|\mathbf{0},\mathbf{I})}[(\nabla_{\mathbf{m}}\log p(\mathbf{m},\mathbf{d}_{obs})\nabla_{\boldsymbol{\theta}}T^{-1}(\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}}\log|\det\mathbf{J}_{T^{-1}}(\boldsymbol{\theta})|)\boldsymbol{\eta}^{T}] + (\mathbf{L}^{-1})^{T}\end{aligned}$$

Appendix C: Gradients of KL Divergence in SVGD

We calculate the gradient in equation (12) following Liu and Wang (2016). Denote T^{-1} as the inverse transform of T. Then by changing the variable,

$$\mathrm{KL}[q_T||p] = \mathrm{KL}[q||p_{T^{-1}}]$$

and hence

$$\begin{aligned}\nabla_{\epsilon}\mathrm{KL}[q_T||p]\,|_{\epsilon=0} &= \nabla_{\epsilon}\mathrm{KL}[q||p_{T^{-1}}]\,|_{\epsilon=0} \\ &= \nabla_{\epsilon}\left[\mathrm{E}_q\log q(\mathbf{m}) - \mathrm{E}_q\log p_{T^{-1}}(\mathbf{m})\right]\end{aligned}$$

and since $q(\mathbf{m})$ does not depend on ϵ

$$\nabla_{\epsilon}\mathrm{KL}[q_T||p]\,|_{\epsilon=0} = -\mathrm{E}_q\left[\nabla_{\epsilon}\log p_{T^{-1}}(\mathbf{m})\right]$$

where $p_{T^{-1}}(\mathbf{m}) = p(T(\mathbf{m}))\cdot|\det\left(\nabla_{\mathbf{m}}T(\mathbf{m})\right)|$. Therefore

$$\nabla_{\epsilon}\log p_{T^{-1}}(\mathbf{m}) = \left(\nabla_{\mathbf{m}}\log\left(p(\mathbf{m})\right)\right)^{\mathrm{T}}\nabla_{\epsilon}T(\mathbf{m}) + trace\left(\left(\nabla_{\mathbf{m}}T(\mathbf{m})\right)^{-1}\cdot\nabla_{\epsilon}\nabla_{\mathbf{m}}T(\mathbf{m})\right)$$

where $T(\mathbf{m}) = \mathbf{m} + \epsilon\boldsymbol{\phi}(\mathbf{m})$, $\nabla_{\epsilon}T(\mathbf{m}) = \boldsymbol{\phi}(\mathbf{m})$ and $\nabla_{\mathbf{m}}T(\mathbf{m})|_{\epsilon=0} = \mathbf{I}$, and so

$$\begin{aligned}\nabla_{\epsilon}\mathrm{KL}[q_T||p]\,|_{\epsilon=0} &= -\mathrm{E}_q\left[\left(\nabla_{\mathbf{m}}\log\left(p(\mathbf{m})\right)\right)^{T}\boldsymbol{\phi}(\mathbf{m}) + trace\left(\nabla_{\mathbf{m}}\boldsymbol{\phi}(\mathbf{m})\right)\right] \\ &= -\mathrm{E}_q\left[trace\left(\nabla_{\mathbf{m}}\log\left(p(\mathbf{m})\right)\boldsymbol{\phi}(\mathbf{m})^{T}\right) + trace\left(\nabla_{\mathbf{m}}\boldsymbol{\phi}(\mathbf{m})\right)\right] \\ &= -\mathrm{E}_q\left[trace\left(\mathcal{A}_p\boldsymbol{\phi}(\mathbf{m})\right)\right]\end{aligned}$$

where $\mathcal{A}_p\boldsymbol{\phi}(\mathbf{m}) = \nabla_{\mathbf{m}}\log p(\mathbf{m})\boldsymbol{\phi}(\mathbf{m})^{T} + \nabla_{\mathbf{m}}\boldsymbol{\phi}(\mathbf{m})$ is the Stein operator.

Acknowledgments
The authors would like to thank the Grane license partners Equinor ASA, Petoro AS, ExxonMobil E&P Norway AS, and ConocoPhillips Skandinavia AS for allowing us to publish this work. The views and opinions expressed in this paper are those of the authors and are not necessarily shared by the license partners. The authors thank the Edinburgh Interferometry Project sponsors (Schlumberger, Equinor, and Total) for supporting this research. This work used the Cirrus UK National Tier-2 HPC Service at EPCC (http://www.cirrus.ac.uk). The data used in this study are available at Edinburgh DataShare (https://doi.org/10.7488/ds/2607).

References

Aki, K., & Lee, W. (1976). Determination of three-dimensional velocity anomalies under a seismic array using first P arrival times from local earthquakes: 1. A homogeneous initial model. *Journal of Geophysical research*, *81*(23), 4381–4399.
Arenz, O., Zhong, M., & Neumann, G. (2018). Efficient gradient-free variational inference using policy search. In *International conference on machine learning* (pp. 234–243). Stockholmsmssan, Stockholm, Sweden.
Arnold, R., & Curtis, A. (2018). Interrogation theory. *Geophysical Journal International*, *214*(3), 1830–1846.
Bensen, G., Ritzwoller, M., & Yang, Y. (2009). A 3-D shear velocity model of the crust and uppermost mantle beneath the United States from ambient seismic noise. *Geophysical Journal International*, *177*(3), 1177–1196.
Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York, USA: Springer.

Blatter, D., Key, K., Ray, A., Gustafson, C., & Evans, R. (2019). Bayesian joint inversion of controlled source electromagnetic and magnetotelluric data to image freshwater aquifer offshore new jersey. *Geophysical Journal International*, *218*(3), 1822–1837.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, *112*(518), 859–877.

Bodin, T., & Sambridge, M. (2009). Seismic tomography with the reversible jump algorithm. *Geophysical Journal International*, *178*(3), 1411–1436.

Bodin, T., Sambridge, M., Tkalčić, H, Arroucau, P., Gallagher, K., & Rawlinson, N. (2012). Transdimensional inversion of receiver functions and surface wave dispersion. *Journal of Geophysical Research*, *117*, B02301. https://doi.org/10.1029/2011JB008560

Burdick, S., & Lekić, V. (2017). Velocity variations and uncertainty from transdimensional P-wave tomography of North America. *Geophysical Journal International*, *209*(2), 1337–1351.

Çınlar, E. (2011). *Probability and stochastics* (Vol. 261): Springer Science & Business Media.

Campillo, M., & Paul, A. (2003). Long-range correlations in the diffuse seismic coda. *Science*, *299*(5606), 547–549.

Crowder, E., Rawlinson, N., Pilia, S., Cornwell, D., & Reading, A. (2019). Transdimensional ambient noise tomography of Bass Strait, southeast Australia, reveals the sedimentary basin and deep crustal structure beneath a failed continental rift. *Geophysical Journal International*, *217*(2), 970–987.

Curtis, A., Gerstoft, P., Sato, H., Snieder, R., & Wapenaar, K. (2006). Seismic interferometry—Turning noise into signal. *The Leading Edge*, *25*(9), 1082–1092.

Curtis, A., & Lomax, A. (2001). Prior information, sampling distributions, and the curse of dimensionality. *Geophysics*, *66*(2), 372–378.

Curtis, A., & Snieder, R. (1997). Reconditioning inverse problems using the genetic algorithm and revised parameterization. *Geophysics*, *62*(5), 1524–1532.

Curtis, A., & Snieder, R. (2002). Probing the Earth's interior with seismic tomography. *International Geophysics Series*, *81*(A), 861–874.

De Freitas, N., Højen-Sørensen, P., Jordan, M. I., & Russell, S. (2001). Variational MCMC. In *Proceedings of the seventeenth conference on uncertainty in artificial intelligence*, Morgan Kaufmann Publishers Inc., pp. 120–127.

de Pasquale, G., & Linde, N. (2016). On structure-based priors in Bayesian geophysical inversion. *Geophysical Journal International*, *208*(3), 1342–1358.

Detommaso, G., Cui, T., Marzouk, Y., Spantini, A., & Scheichl, R. (2018). A stein variational Newton method. In *Advances in neural information processing systems*, Montréal, Canada, pp. 9169–9179.

Devilee, R., Curtis, A., & Roy-Chowdhury, K. (1999). An efficient, probabilistic neural network approach to solving inverse problems: Inverting surface wave velocities for Eurasian crustal thickness. *Journal of Geophysical Research*, *104*(B12), 28,841–28,857.

Dziewonski, A. M., & Woodhouse, J. H. (1987). Global images of the Earth's interior. *Science*, *236*(4797), 37–48.

Earp, S., & Curtis, A. (2019). Probabilistic neural-network based 2D travel time tomography. arXiv preprint arXiv:1907.00541.

Fang, H., & Zhang, H. (2014). Wavelet-based double-difference seismic tomography with sparsity regularization. *Geophysical Journal International*, *199*(2), 944–955.

Fichtner, A., Zunino, A., & Gebraad, L. (2018). Hamiltonian monte carlo solution of tomographic inverse problems. *Geophysical Journal International*, *216*(2), 1344–1363.

Galetti, E., & Curtis, A. (2018). Transdimensional electrical resistivity tomography. *Journal of Geophysical Research: Solid Earth*, *123*, 6347–6377. https://doi.org/10.1029/2017JB015418

Galetti, E., Curtis, A., Baptie, B., Jenkins, D., & Nicolson, H. (2017). Transdimensional love-wave tomography of the British Isles and shear-velocity structure of the east Irish Sea Basin from ambient-noise interferometry. *Geophysical Journal International*, *208*(1), 36–58.

Galetti, E., Curtis, A., Meles, G. A., & Baptie, B. (2015). Uncertainty loops in travel-time tomography from nonlinear wave physics. *Physical review letters*, *114*(14), 148501.

Gorham, J., & Mackey, L. (2015). Measuring sample quality with Stein's method. In *Advances in neural information processing systems*, (pp. 226–234).

Gorham, J., & Mackey, L. (2017). Measuring sample quality with kernels. In *Proceedings of the 34th international conference on machine learning-volume 70*, JMLR. org, Sydney, NSW, Australia, pp. 1292–1301.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Byesian model determination. *Biometrika*, *82*, 711–732.

Green, P. J., & Hastie, D. I. (2009). Reversible jump MCMC. *Genetics*, *155*(3), 1391–1403.

Gretton, A. (2013). Introduction to RKHS, and some simple kernel algorithms. University College London.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97–109.

Hawkins, R., Bodin, T., Sambridge, M., Choblet, G., & Husson, L. (2019). Trans-dimensional surface reconstruction with different classes of parameterization. *Geochemistry, Geophysics, Geosystems*, *20*, 505–529. https://doi.org/10.1029/2018GC008022

Hawkins, R., & Sambridge, M. (2015). Geophysical imaging using trans-dimensional trees. *Geophysical Journal International*, *203*(2), 972–1000.

Hoffman, M. D., & Blei, D. M. (2015). Structured stochastic variational inference. Artificial intelligence and statistics.

Iyer, H., & Hirahara, K. (1993). *Seismic tomography: Theory and practice*. London, UK: Springer Science & Business Media.

Käufl, P., Valentine, A. P., O'Toole, T. B., & Trampert, J. (2013). A framework for fast probabilistic centroid-moment-tensor determination—Inversion of regional static displacement measurements. *Geophysical Journal International*, *196*(3), 1676–1693.

Karlin, S. (2014). *A first course in stochastic processes*. New York: Academic press.

Käufl, P., Valentine, A., de Wit, R., & Trampert, J. (2015). Robust and fast probabilistic source parameter estimation from near-field displacement waveforms using pattern recognition. *Bulletin of the Seismological Society of America*, *105*(4), 2299–2312.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational Byes. arXiv preprint arXiv:1312.6114.

Kubrusly, C., & Gravier, J. (1973). Stochastic approximation algorithms and applications. In *1973 ieee conference on decision and control including the 12th symposium on adaptive processes*, IEEE, San Diego, pp. 763–766.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., & Blei, D. M. (2017). Automatic differentiation variational inference. *The Journal of Machine Learning Research*, *18*(1), 430–474.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, *22*(1), 79–86.

Liu, Q. (2017). Stein variational gradient descent as gradient flow. Advances in neural information processing systems, pp. 3115-3123.

Liu, Q., Lee, J., & Jordan, M. (2016). A kernelized Stein discrepancy for goodness-of-fit tests. In *International conference on machine learning* (pp. 276–284). New York, USA.

Liu, Q., & Wang, D. (2016). Stein variational gradient descent: A general purpose Byesian inference algorithm. Advances in neural information processing systems, pp 2378-2386.

Liu, C., & Zhu, J. (2018). Riemannian stein variational gradient descent for bayesian inference. In *Thirty-second aaai conference on artificial intelligence*. New Orleans, Louisiana, USA.

MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge: Cambridge university press.
Malinverno, A. (2002). Parsimonious Byesian Markov chain Monte Carlo inversion in a nonlinear geophysical problem. *Geophysical Journal International*, *151*(3), 675–688.
Malinverno, A., & Briggs, V. A. (2004). Expanded uncertainty quantification in inverse problems: Hierarchical Byes and empirical Byes. *Geophysics*, *69*(4), 1005–1016.
Malinverno, A., & Leaney, S. (2000). A Monte Carlo method to quantify uncertainty in the inversion of zero-offset VSP data. Society of Exploration Geophysicists, 2000 seg annual meeting.
Marzouk, Y., Moselhy, T., Parno, M., & Spantini, A. (2016). An introduction to sampling via measure transport. arXiv preprint arXiv:1602.05023.
Meier, U., Curtis, A., & Trampert, J. (2007a). A global crustal model constrained by nonlinearised inversion of fundamental mode surface waves. *Geophysical Research Letters*, *34*, L16304. https://doi.org/10.1029/2007GL030989
Meier, U., Curtis, A., & Trampert, J. (2007b). Global crustal thickness from neural network inversion of surface wave data. *Geophysical Journal International*, *169*(2), 706–722.
Metropolis, N., & Ulam, S. (1949). The Monte Carlo method. *Journal of the American statistical association*, *44*(247), 335–341.
Mosegaard, K., & Tarantola, A. (1995). Monte Carlo sampling of solutions to inverse problems. *Journal of Geophysical Research*, *100*(B7), 12,431–12,447.
Nawaz, M. A., & Curtis, A. (2018). Variational Bayesian inversion (VBI) of quasi-localized seismic attributes for the spatial distribution of geological facies. *Geophysical Journal International*, *214*(2), 845–875.
Nawaz, M., & Curtis, A. (2019). Rapid discriminative variational Byesian inversion of geophysical data for the spatial distribution of geological properties. *Journal of Geophysical Research: Solid Earth*, *124*, 5867–5887. https://doi.org/10.1029/2018JB016652
Neal, R. M. (2011). Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, *2*(11), 2.
Nicolson, H., Curtis, A., & Baptie, B. (2014). Rayleigh wave tomography of the British Isles from ambient seismic noise. *Geophysical Journal International*, *198*(2), 637–655.
Nicolson, H., Curtis, A., Baptie, B., & Galetti, E. (2012). Seismic interferometry and ambient noise tomography in the British Isles. *Proceedings of the Geologists' Association*, *123*(1), 74–86.
Piana Agostinetti, N., Giacomuzzi, G., & Malinverno, A. (2015). Local three-dimensional earthquake tomography by trans-dimensional Monte Carlo sampling. *Geophysical Journal International*, *201*(3), 1598–1617.
Ranganath, R., Gerrish, S., & Blei, D. (2014). Black box variational inference. Artificial intelligence and statistics, pp 814-822.
Ranganath, R., Tran, D., & Blei, D. (2016). Hierarchical variational models. In *International conference on machine learning*, New York City, United States, pp. 324–333.
Rawlinson, N., & Sambridge, M. (2004). Multiple reflection and transmission phases in complex layered media using a multistage fast marching method. *Geophysics*, *69*(5), 1338–1350.
Ray, A., Alumbaugh, D. L., Hoversten, G. M., & Key, K. (2013). Robust and accelerated Byesian inversion of marine controlled-source electromagnetic data using parallel tempering. *Geophysics*, *78*(6), E271–E280.
Ray, A., Kaplan, S., Washbourne, J., & Albertin, U. (2017). Low frequency full waveform seismic inversion within a tree based Bayesian framework. *Geophysical Journal International*, *212*(1), 522–542.
Ray, A., & Myer, D. (2019). Bayesian geophysical inversion with trans-dimensional gaussian process machine learning. *Geophysical Journal International*, *217*(3), 1706–1726.
Ray, A., Sekar, A., Hoversten, G. M., & Albertin, U. (2016). Frequency domain full waveform elastic inversion of marine seismic data from the alba field using a bayesian trans-dimensional algorithm. *Geophysical Journal International*, *205*(2), 915–937.
Rezende, D. J., & Mohamed, S. (2015). Variational inference with normalizing flows. arXiv preprint arXiv:1505.05770.
Robbins, H., & Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, *22*, 400–407.
Robert, C., & Casella, G. (2013). *Monte Carlo statistical methods*. Berlin: Springer Science & Business Media.
Röth, G., & Tarantola, A. (1994). Neural networks and inversion of seismic data. *Journal of Geophysical Research*, *99*(B4), 6753–6768.
Salimans, T., Kingma, D., & Welling, M. (2015). Markov chain Monte Carlo and variational inference: Bridging the gap. In *International conference on machine learning* (pp. 1218–1226). Lille, France.
Sambridge, M. (1999). Geophysical inversion with a neighbourhood algorithm—I. Searching a parameter space. *Geophysical journal international*, *138*(2), 479–494.
Sambridge, M. (2013). A parallel tempering algorithm for probabilistic sampling and multimodal optimization. *Geophysical Journal International*, *196*, 357–374.
Saul, L. K., & Jordan, M. I. (1996). Exploiting tractable substructures in intractable networks. Advances in neural information processing systems, pp. 486-492.
Sen, M. K., & Biswas, R. (2017). Transdimensional seismic inversion using the reversible jump hamiltonian monte carlo algorithm. *Geophysics*, *82*(3), R119–R134.
Shahraeeni, M. S., & Curtis, A. (2011). Fast probabilistic nonlinear petrophysical inversion. *Geophysics*, *76*(2), E45–E58.
Shahraeeni, M. S., Curtis, A., & Chao, G. (2012). Fast probabilistic petrophysical mapping of reservoirs from 3D seismic data. *Geophysics*, *77*(3), O1–O19.
Shapiro, N. M., Campillo, M., Stehly, L., & Ritzwoller, M. H. (2005). High-resolution surface-wave tomography from ambient seismic noise. *Science*, *307*(5715), 1615–1618.
Shen, W., Ritzwoller, M. H., & Schulte-Pelkum, V. (2013). A 3-D model of the crust and uppermost mantle beneath the central and western US by joint inversion of receiver functions and surface wave dispersion. *Journal of Geophysical Research: Solid Earth*, *118*, 262–276. https://doi.org/10.1029/2012JB009602
Shen, W., Ritzwoller, M. H., Schulte-Pelkum, V., & Lin, F.-C. (2012). Joint inversion of surface wave dispersion and receiver functions: A Byesian Monte-Carlo approach. *Geophysical Journal International*, *192*(2), 807–836.
Sivia, D. (1996). Data analysis: A Byesian tutorial (oxford science publications).
Smith, A. (2013). *Sequential Monte Carlo methods in practice*. New York: Springer Science & Business Media.
Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the sixth berkeley symposium on mathematical statistics and probability, volume 2: Probability theory*, The Regents of the University of California, Berkeley, Calif.
Tarantola, A. (2005). *Inverse problem theory and methods for model parameter estimation* (Vol. 89). Paris, France: SIAM.
Team, S. D. (2016). Stan modeling language users guide and reference manual. Technical report.
Tran, D., Ranganath, R., & Blei, D. M. (2015). The variational Gaussian process. arXiv preprint arXiv:1511.06499.
Walter, E., & Pronzato, L. (1997). *Identification of parametric models from experimental data*. Paris: Springer Verlag.

Wapenaar, K., & Fokkema, J. (2006). Green's function representations for seismic interferometry. *Geophysics*, *71*(4), SI33–SI46.

Weaver, R. L., Hadziioannou, C., Larose, E., & Campillo, M. (2011). On the precision of noise correlation interferometry. *Geophysical Journal International*, *185*(3), 1384–1392.

Yao, H., & Van Der Hilst, R. D. (2009). Analysis of ambient noise energy distribution and phase velocity bias in ambient noise tomography, with application to SE tibet. *Geophysical Journal International*, *179*(2), 1113–1132.

Young, M. K., Rawlinson, N., & Bodin, T. (2013). Transdimensional inversion of ambient seismic noise for 3D shear velocity structure of the Tasmanian crust. *Geophysics*, *78*(3), WB49–WB62.

Zhang, X., Curtis, A., Galetti, E., & de Ridder, S. (2018). 3-D Monte Carlo surface wave tomography. *Geophysical Journal International*, *215*(3), 1644–1658.

Zhang, X., Hansteen, F., Curtis, A., & deRidder, S., (2020). 1D, 2D and 3D Monte Carlo ambient noise tomography using a dense passive seismic array installed on the North Sea seabed. *Journal of Geophysical Research: Solid Earth*, *125*, e2019JB018552. https://doi.org/10.1029/2019JB018552

Zhang, X., & Zhang, H. (2015). Wavelet-based time-dependent travel time tomography method and its application in imaging the Etna volcano in Italy. *Journal of Geophysical Research: Solid Earth*, *120*, 7068–7084. https://doi.org/10.1002/2015JB012182

Zhdanov, M. S. (2002). *Geophysical inverse theory and regularization problems*, vol. 36. New York: Elsevier.

Zheng, D., Saygin, E., Cummins, P., Ge, Z., Min, Z., Cipta, A., & Yang, R. (2017). Transdimensional Byesian seismic ambient noise tomography across SE tibet. *Journal of Asian Earth Sciences*, *134*, 86–93.

Zobay, O. (2014). Variational Bayesian inference with gaussian-mixture approximations. *Electronic Journal of Statistics*, *8*(1), 355–389.

Zulfakriza, Z., Saygin, E., Cummins, P., Widiyantoro, S., Nugraha, A. D., Lühr, B.-G., & Bodin, T. (2014). Upper crustal structure of central Java, Indonesia, from transdimensional seismic ambient noise tomography. *Geophysical Journal International*, *197*(1), 630–635.