Uncertainty quantification for Wavefield Reconstruction Inversion using a PDE free semidefinite Hessian and randomize-then-optimize method

Zhilong Fang[1]*, Chia Ying Lee[2], Curt Da Silva[1], Tristan van Leeuwen[3] and Felix J. Herrmann[1]
[1]Seismic Laboratory for Imaging and Modeling (SLIM), University of British Columbia
[2]Department of Mathematics, University of British Columbia
[3]Mathematical Institute, Utrecht University

SUMMARY

The data of full-waveform inversion often contains noise, which induces uncertainties in the inversion results. Ideally, one would like to run a number of independent inversions with different realizations of the noise and assess model-side uncertainties from the resulting models, however this is not feasible because we collect the data only once. To circumvent this restriction, various sampling schemes have been devised to generate an ensemble of models that fit the data to within the noise level. Such sampling schemes typically involve running multiple inversions or evaluating the Hessian of the cost function, both of which are computationally expensive. In this work, we propose a new method to quantify uncertainties based on a novel formulation of the full-waveform inversion problem – wavefield reconstruction inversion. Based on this formulation, we formulate a semidefinite approximation of the corresponding Hessian matrix. By precomputing certain quantities, we are able to apply this Hessian to given input vectors *without* additional solutions of the underlying partial differential equations. To generate a sample, we solve an auxiliary stochastic optimization problem involving this Hessian. The result is a computationally feasible method that, with little overhead, can generate as many samples as required at small additional cost. We test our method on the synthetic BG Compass model and compare the results to a direct-sampling approach. The results show the feasibility of applying our method to computing statistical quantities such as the mean and standard deviation in the context of wavefield reconstruction inversion.

INTRODUCTION

Full-waveform inversion (FWI) aims to reconstruct the Earth's subsurface velocity structure from data measured at the surface (Tarantola and Valette, 1982a; Virieux and Operto, 2009). The measured data of FWI often contains noise that leads to uncertainties in data. As a result of fitting this noisy data, these uncertainties propagate through the inversion procedure, which result in uncertainties in the final model (Malinverno and Briggs, 2004). The conventional approach to study these random fluctuations is to use the Bayesian inference approach to formulate the posterior distribution of the unknown parameter and generate, say, mean and standard deviation plots (Kaipio and Somersalo, 2004). Unfortunately the FWI is a strongly non-linear inverse problem and the forward problem of FWI is computationally expensive, which leads to difficulties in the explicit computation of the parameter space distribution and in calculating the mean and pointwise variance of velocity (Tarantola and Valette, 1982b).

Wavefield reconstruction inversion (WRI) (van Leeuwen and Herrmann, 2013; Peters et al., 2014) is a new approach to solve wave-equation based inversion problems. Unlike conventional FWI, which at each iteration solves the wave-equation exactly, WRI considers the wave-equation as a ℓ_2-norm penalty term in the objective function, weighted by a penalty parameter. For WRI, when either the velocity or the wavefield is fixed, the problem becomes a linear fitting problem in the other variable. As a result, WRI is "less non-linear" compared to FWI. Additionally, the WRI enjoys an approximate diagonal Hessian of the objective function that does not require additional partial-differential equation (PDE) solves (van Leeuwen and Herrmann, 2013). Given these advantages of WRI, Fang et al. (2015) proposed a posterior distribution based on WRI and produced samples with an Gaussian distribution based on the diagonal approximation of the Hessian.

The diagonal approximation of the Hessian utilized by Fang et al. (2015) does not account for the statistical correlations of the wavefields with the model vectors, resulting in a poor approximation of the probability density when the penalty parameter is large. In this work, we propose to use a more accurate semidefinite Hessian approximation (van Leeuwen and Herrmann, 2015) to approximate the posterior distribution. This Hessian does involve additional PDE-solves, but the necessary wavefields can be easily precomputed and stored for later use. Thus, we can still apply the Hessian to given input vectors at little additional computational cost.

The conventional approach to draw samples from a Gaussian distribution consists of two steps: (1) minimize the negative logarithm function of the posterior distribution to obtain the *maximum a posteriori* (MAP) estimate, i.e., the point that maximizes the posterior probability; (2) compute the Cholesky factor of the covariance matrix at the MAP estimate and use it to draw samples (Martin et al., 2012). In practice, the latter step requires us to form the full Hessian matrix explicitly and is computationally expensive (Rue, 2001). In this work, we utilize the so-called randomize-then-optimize (RTO) method (Solonen et al., 2014), which generates a sample from a Gaussian distribution by solving a stochastic least-squares problem involving the covariance matrix. By using our PDE-free Hessian, we can solve these least-squares problems efficiently using standard solvers such as CGLS or LSQR (Paige and Saunders, 1982). This results in a scalable algorithm for generating samples from the posterior distribution.

POSTERIOR DISTRIBUTION FOR WRI

Tarantola and Valette (1982b) first introduced the Bayesian approach to the seismic inverse problem and proposed the general framework of statistical FWI. Based on the Bayes' law, for a model parameter $\mathbf{m}$ with n_{grid} points, the posterior distribution $\rho_{post}(\mathbf{m}|\mathbf{d})$ of $\mathbf{m}$ given data $\mathbf{d}$ consisting n_{src} shots and n_{freq} frequencies is proportional to the product of the likeli-

hood distribution $\rho_{\text{like}}(\mathbf{d}|\mathbf{m})$ and the prior distribution $\rho_{\text{prior}}(\mathbf{m})$ yielding:

$$\rho_{\text{post}}(\mathbf{m}|\mathbf{d}) \propto \rho_{\text{like}}(\mathbf{d}|\mathbf{m})\rho_{\text{prior}}(\mathbf{m}). \qquad (1)$$

In other words, the most probable model vectors $\mathbf{m}$ that correspond to our data $\mathbf{d}$ are those that likely correspond to our prior beliefs about the underlying model, irrespective of data, (i.e., $\rho_{\text{prior}}(\mathbf{m})$) as well as those that generate data $\mathbf{d}$ that are probabilistically consistent with our underlying physical and noise model (i.e., $\rho_{\text{like}}(\mathbf{d}|\mathbf{m})$). This formulation allows us to explicitly write down expressions for the last two terms, while simultaneously ignoring the normalization constant hidden in $\propto$. Computing this constant would otherwise involve high dimensional integration, which is challenging both numerically and theoretically.

With the assumption that the noise on the data obeys a Gaussian distribution with mean 0 and variance σ^2, we arrive at the following posterior distribution for FWI (Tarantola and Valette, 1982b; Martin et al., 2012):

$$\rho_{\text{post}}(\mathbf{m}|\mathbf{d}) \propto \exp(-\frac{1}{2}\sigma^{-2}\|\mathbf{P}\mathbf{A}(\mathbf{m})^{-1}\mathbf{q} - \mathbf{d}\|^2)\rho_{\text{prior}}(\mathbf{m}). \quad (2)$$

In this expression, $\mathbf{P}$, $\mathbf{A}$ and $\mathbf{q}$ represent the receiver projection operator, Helmholtz matrix, and source term, respectively. We omit the dependence on source and frequency indices to ease the notation. As a result of its non-linear dependence on $\mathbf{m}$ through $\mathbf{A}(\mathbf{m})^{-1}\mathbf{q}$, this density is not easily approximated by a Gaussian distribution. Although samples could in principle still be generated from such complicated distributions, this is computationally expensive (Tarantola and Valette, 1982b).

To avoid these difficulties, Fang et al. (2015) considered a different formulation for FWI where both the wavefields and model are unknowns. This formulation leads to the following posterior distribution:

$$\rho_{\text{post}}(\mathbf{m},\mathbf{u}) \propto \exp(-\frac{1}{2}(\sigma^{-2}\|\mathbf{P}\mathbf{u} - \mathbf{d}\|^2 \\ -\lambda^2\|\mathbf{A}(\mathbf{m})\mathbf{u} - \mathbf{q}\|^2))\rho_{\text{prior}}(\mathbf{m}). \qquad (3)$$

Here, the additional parameter λ is a penalty parameter to balance the data and PDE misfit. Contrary to eliminating the wave-equation as a constraint by solving it (Virieux and Operto, 2009) as in Equation (2), in (3) we treat the wave-equation as additional prior information. It is readily observed that the distribution is Gaussian in $\mathbf{u}$ for fixed $\mathbf{m}$ and vice versa. This suggests that this distribution is more amenable to a Gaussian approximation than the one stated in Equation (2). Unfortunately the search space of problem (3) is significantly larger than problem (2), i.e, from n_{grid} unknown variables to $n_{\text{grid}} + n_{\text{grid}} * n_{\text{src}} * n_{\text{freq}}$ unknown variables. To reduce the dimensionality, Fang et al. (2015) proposed to utilize the variable projection method (van Leeuwen and Herrmann, 2013) to define the following conditional distribution:

$$\rho_{\text{post}}(\mathbf{m},\mathbf{u}(\mathbf{m})) \propto \exp(-\frac{1}{2}(\sigma^{-2}\|\mathbf{P}\mathbf{u}(\mathbf{m}) - \mathbf{d}\|^2 \\ -\lambda^2\|\mathbf{A}(\mathbf{m})\mathbf{u}(\mathbf{m}) - \mathbf{q}\|^2))\rho_{\text{prior}}(\mathbf{m}), \qquad (4)$$

where the wavefield $\mathbf{u}$ is generated by solving the following

data-augmented system:

$$\begin{pmatrix} \lambda\mathbf{A}(\mathbf{m}) \\ \sigma^{-1}\mathbf{P} \end{pmatrix} \mathbf{u} = \begin{pmatrix} \lambda\mathbf{q} \\ \sigma^{-1}\mathbf{d} \end{pmatrix}. \qquad (5)$$

Here, we can observe that for given $\mathbf{m}$, the solution $\mathbf{u}$ of the least-squares problem (5) maximizes the posterior distribution (3). In this work, we follow Fang et al. (2015) and quantify uncertainties for WRI based on the posterior distribution (4).

METHODOLOGY OF QUANTIFYING UNCERTAINTY

Gaussian approximation — The posterior distribution (4) is not Gaussian, but can be approximated by a Gaussian distribution *locally* around its maximum. The model that maximizes the posterior density is called the maximum a posterior (MAP) estimate, $\mathbf{m}_{\text{MAP}}$. In this work, we use the following Gaussian distribution $\rho_{\text{Gauss}}(\mathbf{m})$ to approximate the posterior distribution (4):

$$\rho_{\text{Gauss}}(\mathbf{m}) \propto \exp(-(\mathbf{m} - \mathbf{m}_{\text{MAP}})^T\mathbf{H}(\mathbf{m} - \mathbf{m}_{\text{MAP}})), \qquad (6)$$

where $\mathbf{H}$ represents the Hessian matrix of the negative logarithm function $-\log(\rho_{\text{post}}(\mathbf{m}))$ of the posterior distribution (4). As discussed by Martin et al. (2012), the Hessian $\mathbf{H}$ consists two parts: the Hessian of the WRI misfit $\mathbf{H}_{\text{mis}}$ and the Hessian related to the prior term $\mathbf{H}_{\text{prior}}$. Fang et al. (2015) proposed to approximate the Hessian $\mathbf{H}_{\text{mis}}$ by a diagonal matrix that does not require additional PDE solves to form (van Leeuwen and Herrmann, 2013). However, this diagonal approximation does not include terms that statistically correlate the wavefield to the velocity model. As a result, the approximation becomes poorer as the penalty parameter λ increases. In this work, we propose to use the following semidefinite Hessian (van Leeuwen and Herrmann, 2015), which approximates these terms,

$$\mathbf{H}_{\text{mis}} = \sigma^{-2}\mathbf{G}^T\mathbf{A}^{-T}\mathbf{P}^T(\mathbf{I} + \frac{1}{\lambda^2\sigma^2}\mathbf{P}\mathbf{A}^{-1}\mathbf{A}^{-T}\mathbf{P}^T)^{-1}\mathbf{P}\mathbf{A}^{-1}\mathbf{G}. \qquad (7)$$

In this expression, $\mathbf{G} = \frac{\partial\mathbf{A}\mathbf{u}}{\partial\mathbf{m}}$ and T indicates the adjoint operation. In practice, since we only need to compute the Hessian to form the Gaussian approximation (6) once at the very beginning of the inversion, we precompute and store the wavefields $\mathbf{u}$ and the term $\mathbf{A}^{-T}\mathbf{P}^T$. We can compute and store these terms in parallel to accelerate the computation and avoid saturating a single computational node's memory. Once all these wavefields are computed and stored, the matrix-vector product with this Hessian can be performed extremely quickly, since it only involves efficient products of matrices and vectors and does not require additional PDE solves. Moreover, we do not store the entire $n_{\text{grid}} \times n_{\text{grid}}$ matrix $\mathbf{H}_{\text{mis}}$ explicitly, only the factors above that allow us to form matrix-vector products. As a result, we can utilize any iterative least-squares solver, including CG and LSQR, to invert the Hessian.

In order to test the accuracy of this approximation (6), we compared the behavior of the negative logarithm function of the posterior distribution (4), $f_t = -\log(\rho_{\text{post}}(\mathbf{m}))$, and that of the Gaussian distribution, $f_q = -\log(\rho_{\text{Gauss}}(\mathbf{m}))$, around the MAP point with both small λ, $\lambda = 1e0$ and large λ, $\lambda = 1e6$, as shown in Figure 1. We first generate a random perturbation $d\mathbf{m}$, with a maximum absolute value of $0.8\,\text{km/s}$. Next, we calculate the objective values $f_t(\mathbf{m} + \alpha d\mathbf{m})$ and $f_q(\mathbf{m} + \alpha d\mathbf{m})$ for $\alpha \in$

$[-1:0.1:1]$. In both cases, the negative logarithm function of the Gaussian approximation closely matches that of the true posterior distribution, which indicates that our approximation is accurate.
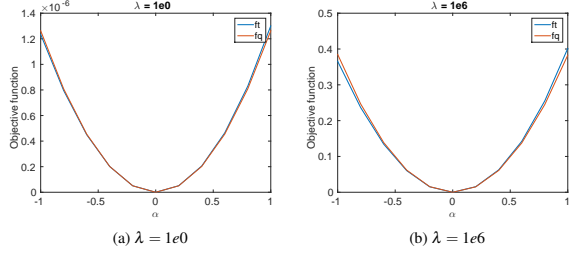

(a) $\lambda = 1e0$ (b) $\lambda = 1e6$

Figure 1: The comparison of the negative logarithm function of the Gaussian approximation $f_q(\mathbf{m} + \alpha d\mathbf{m})$ with the Hessian and the posterior distribuion $f_t(\mathbf{m} + \alpha d\mathbf{m})$ corresponding to a) $\lambda = 1e0$ and b) $\lambda = 1e6$. The parameter α ranges from -1 to 1. The maximum absolute value of the random perturbation vector is $0.8\,\mathrm{km/s}$.

Randomize then optimize method — Given our Gaussian approximation of the posterior distribution, a standard procedure to generate samples from such a distribution is to compute the Cholesky factor of its covariance matrix, followed by applying this factor to random noise (Rue, 2001). Unfortunately, since the size of the Hessian is $n_{\text{grid}} \times n_{\text{grid}}$, and n_{grid} is often in the range of 10^6 or higher, we are unable to store the Hessian explicitly, let alone compute its Cholesky factors, which is also computationally expensive.

The randomize-then-optimize (RTO) method (Solonen et al., 2014) circumvents this problem and allows us to draw samples from a given Gaussian distribution $\rho(\mathbf{m}) \propto \exp\left(-(\mathbf{m} - \mathbf{m}_{\text{MAP}})^T \mathbf{H}(\mathbf{m} - \mathbf{m}_{\text{MAP}})\right)$, by solving the following stochastic optimization problem (Solonen et al., 2014),

$$\min_{\mathbf{m}} \|\mathbf{Lm} - (\mathbf{Lm}_{\text{MAP}} + \mathbf{r})\|^2. \qquad (8)$$

Here, $\mathbf{L}$ is any full rank matrix that satisfies $\mathbf{H} = \mathbf{L}^T \mathbf{L}$ and $\mathbf{r}$ is a random realization from the Gaussian distribution $\mathcal{N}(0, \mathscr{I})$.

Fortunately, by the construction of our approximate Hessian (7), we have the following factorization of $\mathbf{H}_{\text{mis}}$:

$$\begin{aligned}
\mathbf{H}_{\text{mis}} &= \mathbf{L}_{\text{mis}}^T \mathbf{L}_{\text{mis}}, \\
\mathbf{L}_{\text{mis}} &= \sigma^{-1}(\mathbf{I} + \frac{1}{\lambda^2 \sigma^2}\mathbf{PA}^{-1}\mathbf{A}^{-T}\mathbf{P}^T)^{-\frac{1}{2}}\mathbf{PA}^{-1}\mathbf{G}.
\end{aligned} \qquad (9)$$

In this expression, computing the inverse square-root of $\mathbf{I} + \frac{1}{\lambda^2 \sigma^2}\mathbf{PA}^{-1}\mathbf{A}^{-T}\mathbf{P}^T$ is computationally tractable because it involves a $n_{\text{rcv}} \times n_{\text{rcv}}$ matrix, which is much smaller than the $n_{\text{grid}} \times n_{\text{grid}}$ matrix $\mathbf{H}_{\text{mis}}$. With the assumption that all shots share the same receivers, we only need to compute this matrix once for each frequency. Without loss of generality, we assume that the Hessian of the prior term has a simple structure (e.g., Low rank and sparse) so that we can compute its square root $\mathbf{L}_{\text{prior}}$, where $\mathbf{H}_{\text{prior}} = \mathbf{L}_{\text{prior}}^T \mathbf{L}_{\text{prior}}$, with a low complexity (Solonen et al., 2014). Under these assumptions, we

are able to utilize the RTO method (8) to draw samples from the Gaussian distribution $\mathcal{N}(\mathbf{m}_{\text{MAP}}, \mathbf{H}^{-1})$ through replacing terms $\mathbf{L}$, $\mathbf{Lm}_{\text{MAP}}$ and $\mathbf{r}$ in (8) by $\begin{pmatrix} \mathbf{L}_{\text{mis}} \\ \mathbf{L}_{\text{prior}} \end{pmatrix}$, $\begin{pmatrix} \mathbf{L}_{\text{mis}}\mathbf{m}_{\text{MAP}} \\ \mathbf{L}_{\text{prior}}\mathbf{m}_{\text{MAP}} \end{pmatrix}$, and $\begin{pmatrix} \mathbf{r}_{\text{mis}} \\ \mathbf{r}_{\text{prior}} \end{pmatrix}$. Here, $\mathbf{r}_{\text{prior}}$ is a realization from Gaussian distributions $\mathcal{N}(0, \mathscr{I}_{n_{\text{grid}} \times n_{\text{grid}}})$. $\mathbf{r}_{\text{mis}}$ is a realization from Gaussian distributions $\mathcal{N}(0, \mathscr{I}_{n_{\text{rcv}} \times n_{\text{rcv}}})$ for each shot and frequency.

NUMERICAL EXPERIMENTS

In this section, we test the feasibility of our uncertainty quantification method in a realistic seismic exploration setting. We consider the BG compass model (Li et al., 2012; Fang et al., 2015), which enjoys a large degree of variability constrained by well data. The true model and initial model are shown in the Figure 2. The model size is $2\,\mathrm{km} \times 4.5\,\mathrm{km}$. We carry out simulations with 91 shot and 451 receiver positions sampled at 50 m and 10 m, respectively, yielding a maximum offset of 4.5 km. We place all shots and receivers at the depth $z = 40\,\mathrm{m}$. According to Li et al. (2012), we carry out the inversions sequentially in 10 frequency bands ranging from $[2, 3, 4]\,\mathrm{Hz}$ to $[29, 30, 31]\,\mathrm{Hz}$ for the sake of avoiding local minima and improving convergence. The ratio of the noise and signal in the data is 1% as shown in the Figure 3 and the noise level σ is calculated correspondingly. We select $\lambda = 1000\sigma$ according to the largest singular value of $\sigma^{-1}\mathbf{A}^{-T}\mathbf{P}^T$ with the initial model in Figure 2b (van Leeuwen and Herrmann, 2015). We utilize the initial model as the prior model $\mathbf{m}_{\text{prior}}$, as is standard (Martin et al., 2012):

$$\rho_{\text{prior}}(\mathbf{m}) \propto \exp(-\gamma^{-2}\|\mathbf{m} - \mathbf{m}_{\text{prior}}\|^2). \qquad (10)$$

Here, we select the standard deviation $\gamma = 2\mathrm{km/s}$, which is larger than the maximum difference $1.5\,\mathrm{km/s}$ between the initial model and the true model.
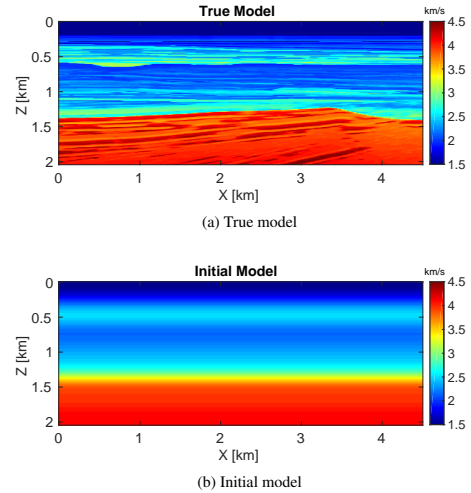

(a) True model

(b) Initial model

Figure 2: The true model (a) and the initial model (b).

To obtain the MAP estimate, we first solve an optimization problem that corresponds to minimizing the negative logarithm function of the posterior distribution (4), i.e., a standard WRI problem. The result is included in Figure 4a. Next, we compute
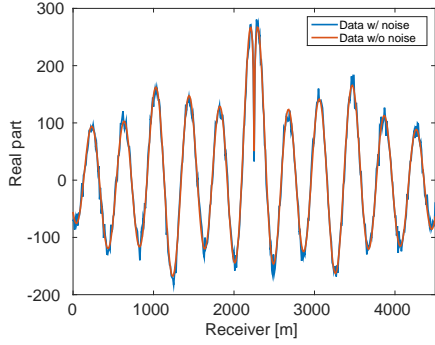
Figure 3: The real part of 5 Hz data with and without noise for the shot located at $x = 2500$ m.

(a) MAP estimate

(b) Standard deviation

Figure 4: (a) MAP estimate and (b) standard deviation.

the Hessian at this MAP estimate according to the Equation (7). Finally, we draw 1000 samples by the RTO method (8) to calculate the standard deviation (STD) of the velocity model at each grid point, shown in Figure 4b. We can observe that the small level of uncertainties in the data — the 1 % noise — leads to the small STD, which has a maximum value of 0.1 km/s. In the shallow area of the model, the STD is small, while in the deep part, it becomes large, which illustrates the uncertainties in the deep part are larger than in the shallow area. This behavior matches the fact that, compared to velocity at the deep area, the velocities in the shallow regions provide a larger contribution to the measured data. In order to validate these confidence intervals, we generate 1000 noisy realizations, add them to the data, and use them to generate 1000 independent inversion results, which should be samples from the true posterior distribution (Bardsley et al., 2015). We calculate the STD and mean value of these 1000 inversion results and compare to that we obtained by our approach (Figure 5) at the cross section $x = 1000$ m. The mean values obtained by the two approaches match each other nearly exactly. The STDs obtained by the two approaches are not as close as their corresponding mean values, but still remain reasonably stable. Considering the assumptions we made for our approach, these differences between the computed standard deviations are acceptable.

(a) Comparison of STD at $x = 1000$ m

(b) Comparison of mean value at $x = 1000$ m

Figure 5: The STD and mean value obtained by 1000 inversion results and our Gaussian approximation. (a) Standard deviation and b) mean value.

CONCLUSION

In this work, we proposed a semidefinite Hessian for WRI and applied it to perform uncertainty quantification. By precomputing and storing the necessary wavefields, we are able to compute the matrix-vector product of the Hessian without additional PDE solves. We used the randomize-then-optimize technique to draw samples from an Gaussian distribution that approximates the posterior distribution, and in doing so avoid forming the Hessian explicitly. As one would expect, our numerical experiment produces small uncertainties in the shallow part and large uncertainties in the deep part of the model, which matches the fact that the observed data is more sensitive to the velocity perturbation in the shallow part than in the deeper part. The STD and mean value obtained by our approach are close to those computed from the 1000 inversion results, which shows that our approach produces reasonable uncertainty analysis. Storing all these precomputed wavefields may still be a
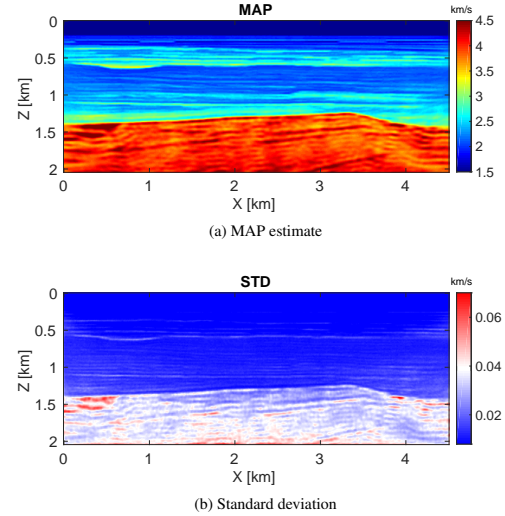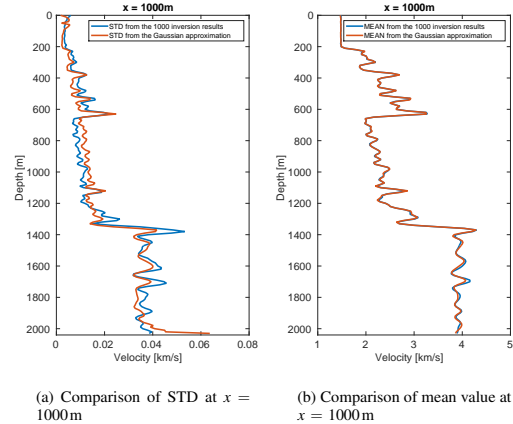
practical barrier to extending this method to the 3D case. Given the number of approximations made in the derivation of this method, further study is warranted on the approximation of the underlying density by the Gaussian density.

REFERENCES

Bardsley, J. M., A. Seppanen, A. Solonen, H. Haario, and J. Kaipio, 2015, Randomize-then-optimize for sampling and uncertainty quantification in electrical impedance tomography: SIAM/ASA Journal on Uncertainty Quantification, **3**, 1136–1158.

Fang, Z., C. Lee, C. Silva, F. Herrmann, and R. Kuske, 2015, Uncertainty quantification for wavefield reconstruction inversion: Presented at the 77th EAGE Conference and Exhibition 2015.

Kaipio, J., and E. Somersalo, 2004, Statistical and Computational Inverse Problems (Applied Mathematical Sciences) (v. 160), 1 ed.: Springer.

Li, X., A. Y. Aravkin, T. van Leeuwen, and F. J. Herrmann, 2012, Fast randomized full-waveform inversion with compressive sensing: Geophysics, **77**, A13–A17.

Malinverno, A., and V. A. Briggs, 2004, Expanded uncertainty quantification in inverse problems: Hierarchical bayes and empirical bayes: Geophysics, **69**, 1005–1016.

Martin, J., L. Wilcox, C. Burstedde, and O. Ghattas, 2012, A Stochastic Newton MCMC Method for Large-scale Statistical Inverse Problems with Application to Seismic Inversion: SIAM Journal on Scientific Computing, **34**, A1460–A1487.

Paige, C. C., and M. A. Saunders, 1982, Lsqr: An algorithm for sparse linear equations and sparse least squares: ACM Transactions on Mathematical Software (TOMS), **8**, 43–71.

Peters, B., F. J. Herrmann, and T. van Leeuwen, 2014, Wave-equation based inversion with the penalty method: adjoint-state versus wavefield-reconstruction inversion: Presented at the EAGE.

Rue, H., 2001, Fast sampling of gaussian markov random fields: Journal of the Royal Statistical Society: Series B (Statistical Methodology), **63**, 325–338.

Solonen, A., A. Bibov, J. M. Bardsley, and H. Haario, 2014, Optimization-based sampling in ensemble kalman filtering: International Journal for Uncertainty Quantification, **4**.

Tarantola, A., and B. Valette, 1982a, Generalized nonlinear inverse problems solved using the least squares criterion: Reviews of Geophysics, **20**, 219–232.

——, 1982b, Inverse problems = quest for information: Journal of Geophysics, **50**, 159–170.

van Leeuwen, T., and F. J. Herrmann, 2013, Mitigating local minima in full-waveform inversion by expanding the search space: Geophysical Journal International, **195**, 661–667.

——, 2015, A penalty method for PDE-constrained optimization in inverse problems: Inverse Problems, **32**, 015007. ((Inverse Problems)).

Virieux, J., and S. Operto, 2009, An overview of full-waveform inversion in exploration geophysics: GEOPHYSICS, **74**, WCC1–WCC26.