Linear Waveform Tomography Inversion Using Machine Learning Algorithms

Tue Holm-Jensen[1] · Thomas Mejer Hansen[2]

Abstract A linear probabilistic waveform inversion strategy is proposed for cross-hole tomographic data using machine learning algorithms that integrates a priori information described by Gaussian distributed slowness fields. A theoretical framework is outlined that combines an approximate linear forward operator describing the waveform with the ridge regression algorithm. The framework approximates arbitrary geostatistical a priori information with a Gaussian distribution, which makes it possible to obtain an analytic description of the posterior probability density, the solution of the inverse problem. The suggested inversion strategy is tested on synthetic tomographic crosshole ground-penetrating radar full-waveform data generated using multiple-point-based a priori information incorporating realistic noise. The benefits of the proposed strategy include: (I) Increased resolution of the posterior probability density compared to first arrival type inversion, due to using the wavefield. (II) Much faster execution than existing full waveform inversion algorithms based on Monte Carlo sampling or the adjoint method. (III) Greater efficiency in obtaining realizations from the posterior probability density. The main challenge in using the method is that it needs a quite large training data set from which the linear forward model can be inferred.

Keywords Machine learning · Inversion · Ground-penetrating radar (GPR) · Tomography

✉ Tue Holm-Jensen
   Tueholmjensen@gmail.com

   Thomas Mejer Hansen
   tmeha@geo.au.dk

[1] Copenhagen, Denmark

[2] Department of Geoscience, Aarhus University, Høegh-Guldbergs Gade 2, 8000 Aarhus C, Denmark

1 Introduction

Crosshole ground-penetrating-radar (GPR) tomography has the potential to produce high-spatial-resolution images of the subsurface (Hubbard et al. 1997; Eppstein and Dougherty 1998; Binley et al. 2001; Alumbaugh et al. 2002). The electromagnetic wave velocity is directly related to the relative permittivity of a material and thus allows for a spatial distinction between materials with differing permittivity. This can be useful in hydrologic applications, where it has been used extensively (Annan 2005).

The high-resolution two-dimensional tomographic images obtained from crosshole GPR data provide information about the structures of the subsurface that could otherwise be obtained only through extensive invasive sampling. Obtaining high-resolution two-dimensional images is difficult, however, as there are uncertainties due to experimental noise and assumptions in the applied inversion algorithms. These assumptions can be introduced to make the inversion computationally feasible or to allow probabilistic inversion. Common assumptions include simplifying the forward problem with a straight-ray approximation, assumptions of the measurement errors, choice of regularization method, acquisition geometry, and a priori model. All these assumptions influence wave-velocity distribution properties (Menke 1989; Peterson 2001; Kowalsky et al. 2005; Linde et al. 2006; Cordua et al. 2008; Hansen et al. 2008).

Some inversion algorithms may introduce artifacts in areas with limited ray coverage, as investigated by Day-Lewis et al. (2005), who looked into deterministic least-squares inversion algorithms that produce smooth minimum-variance estimates of the two-dimensional tomographic images in the inter-borehole region (Eppstein and Dougherty 1998; Alumbaugh et al. 2002). Day-Lewis and Lane (2004) concluded that GPR-derived tomograms may have little utility for inferring subsurface geostatistical properties based on the analysis of five synthetic models and inversion artifacts. To improve the inverse estimate, stochastic inversion approaches (Hansen et al. 2006; Gloaguen et al. 2007; Dafflon et al. 2009) or full-waveform inversion algorithms (Ernst et al. 2007; Cordua et al. 2012; Keskinen et al. 2017; Mosser et al. 2018) can be adopted.

Stochastic and/or full-waveform inversion methods can be computationally very costly. The objective of this study is to demonstrate how full-waveform crosshole GPR data can be inverted thanks to a probabilistic formulation using linear inverse Gaussian methods (Tarantola and Valette 1982b). This is achieved by making use of machine learning methods to first describe the full waveform data set by a limited number of attributes, and secondly by using machine learning to estimate a linear forward model.

First the linear Gaussian probabilistic inversion methodology is introduced. Then it is demonstrated how a seismic waveform can be represented by a limited number of parameters using principal components analysis compression. Then it is demonstrated how a linear forward model can be estimated using ridge regression. Finally the methodology is demonstrated on a synthetic case study, where the reference model is not a realization of a Gaussian model and where data have been computed using a non-linear finite-difference based full-waveform forward model.
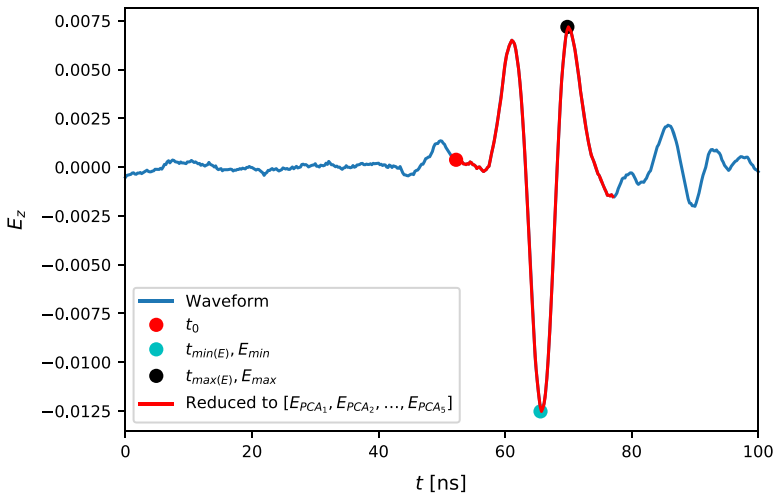
Fig. 1 A single synthetic recording of a Ricker wavelet after passing through a slowness field containing sand channels. Measured is the electric field. Also shown are the ten attributes used to quantify the waveform: $[t_0, E_{\max}, E_{\min}, t_{\max(E)}, t_{\min(E)}, E_{PCA_1}, E_{PCA_2}, E_{PCA_3}, E_{PCA_4}, E_{PCA_5}]$

2 Theory

Let **m** represent a vector containing parameters describing the subsurface. Let prior information about **m** be quantified though the a priori distribution $\rho(\mathbf{m})$. Further, information from geophysical observations $\mathbf{d}_{obs}$ is quantified by a likelihood function $\mathrm{L}(\mathbf{d}_{obs} - g(\mathbf{m}))$, which quantifies the distribution of noise or the data residuals. g refers to the forward modeling operator that maps a given model **m** into a set of data **d** as

$$\mathbf{d} = g(\mathbf{m}). \qquad (1)$$

In the probabilistic formulation of the inverse problem, the combined information about **m** from the a priori distribution and the likelihood can be combined through conjunction of information, which becomes the a posteriori probability distribution, $\sigma(\mathbf{m})$, as given by Tarantola and Valette (1982b)

$$\sigma(\mathbf{m}) \propto \rho(\mathbf{m})\, \mathrm{L}(\mathbf{d}_{obs} - g(\mathbf{m})). \qquad (2)$$

In general, the forward operator is non-linear, and it is not possible to describe the posterior distribution analytically. Instead, one can resort to sampling methods for generating a sample of the posterior distribution, and use that to characterize the solution to the inverse problem (Tarantola and Mosegaard 1995).

If, however, the forward operator is linear such that the forward problem can be evaluated using a simple linear mapping matrix **G**

$$\mathbf{d} = \mathbf{G} \cdot \mathbf{m}, \qquad (3)$$

and both the prior $\rho(\mathbf{m})$ and the likelihood $\mathrm{L}(\mathbf{d} - g(\mathbf{m}))$ can be described as a Gaussian probability density, then the posterior distribution is also a Gaussian distribution $\sigma(\mathbf{m}) \sim (\mathbf{m}^{\mathrm{est}}, \mathbf{C}_{\mathrm{m}}^{\mathrm{est}})$ which can be computed analytically as given by Tarantola and Valette (1982a)

$$\mathbf{m}^{\mathrm{est}} = \mathbf{m}_0 + \mathbf{C}_{\mathrm{m}}\mathbf{G}^T\big[\mathbf{G}\mathbf{C}_{\mathrm{m}}\mathbf{G}^T + \mathbf{C}_{\mathrm{d}}\big]^{-1}\big[\mathbf{d} - \mathbf{G}\cdot\mathbf{m}_0\big] \tag{4}$$

and

$$\mathbf{C}_m^{\mathrm{est}} = \big[\mathbf{G}^T\mathbf{C}_{\mathrm{d}}^{-1}\mathbf{G} + \mathbf{C}_{\mathrm{m}}^{-1}\big]^{-1}, \tag{5}$$

where $\mathbf{m}_0$ and $\mathbf{C}_{\mathrm{m}}$ are the prior mean and covariance of the prior probability density $\rho(\mathbf{m}) \sim \mathcal{N}(\mathbf{m}_0, \mathbf{C}_{\mathrm{m}})$, and $\mathbf{C}_{\mathrm{d}}$ is the data covariance, describing the likelihood as $\mathrm{L}(\mathbf{m}) = \mathcal{N}(0, \mathbf{C}_{\mathrm{d}})$. This approach to solving inverse problems as linear inverse problems has been used in a range of studies (Tarantola 1987; Menke 1989). Traditionally, data for crosshole GPR velocity tomography are based on the first-arrival traveltimes, $\mathbf{d}$, picked from the recorded GPR waveforms. $\mathbf{m}$ is the slowness (inverse velocity) values of a grid of cells describing the subsurface between two boreholes, and $\mathbf{G}$ contains the distances that the different rays have traveled in the individual cells. This can lead to an efficient linear inversion problem that can be solved using Eqs. 4 to 5.

Here, however, the motivation is to include information from the whole recorded waveform, as opposed to using only the first arrival traveltimes. Therefore, information about the full waveform is quantified using a sparse representation based on first arrival traveltimes, amplitudes, and compression of the wavefield using Principal Components Analysis (PCA).

2.1 Quantifying the Full Waveform Signal with a Sparse Data Set

Figure 1 shows an example of a recorded waveform between a receiver recorded at a certain time-delay with respect to the ignition of a controlled source at another location. In practice, as will be exemplified later, many such waveforms are recorded for different source and receiver locations in a typical crosshole tomographic experiment. Each of these recorded waveforms will be described in this paper using K attributes: the first five attributes are (I) the first arrival traveltime, t_0, (II) the maximum voltage, $E_{\max}$, (III) the minimum voltage, $E_{\min}$, (IV) the time of the maximum voltage, $t_{\max(E)}$, and (V) the time of the minimum voltage, $t_{\min(E)}$. The last $K - 5$ attributes are P principal components describing the shape of the waveform right after the first arrival traveltime, $[E_{\mathrm{PCA}_1}, E_{\mathrm{PCA}_2}, \ldots, E_{\mathrm{PCA}_P}]$. The method can easily be extended by using more or different attributes.

The First Arrival Traveltime was found by an automatic first arrival picking algorithm inspired by Molyneux and Schmitt (1999). The traveltime of a waveform $\mathrm{U}(t)$ was found by the following three steps: (I) convolving $\mathrm{U}(t)$ with the first part of the observed pulse obtained by sending a Ricker wavelet through a uniform velocity field $\mathrm{R}(t)$ to give $(\mathrm{U} * \mathrm{R})(t)$. (II) Regularizing the product of the convolution with $\mathrm{f}(t)$ to give $\mathrm{U}^*(t) = (\mathrm{U} * \mathrm{R})(t) \cdot \mathrm{f}(t)$ where $\mathrm{f}(t)$ is given by

$$f(t) = \exp\left(\left(\frac{t - t_{\text{mean}}^{\text{exp}}}{4(t_{\text{slow}}^{\text{exp}} - t_{\text{fast}}^{\text{exp}})}\right)^2\right). \tag{6}$$

$t_{\text{mean}}^{\text{exp}}$, $t_{\text{fast}}^{\text{exp}}$ and $t_{\text{slow}}^{\text{exp}}$ are the expected traveltimes assuming a uniform velocity field with velocity equal to the prior mean velocity, the prior channel velocity and the prior non-channel velocity, respectively. (III) Picking the first arrival traveltime to be the first time where $U^*(t)$ reaches 30% of its maximum value. The first arrival time is shown as a red dot on Fig. 1 for one waveform.

The Minimum and Maximum Amplitudes were found independently of the traveltime simply by searching the entire waveform for the minimum and maximum amplitude. The locations of the times $t_{\text{min}(E)}$ and $t_{\text{max}(E)}$, and the minimum and maximum amplitudes, E_{max} and E_{min}, are shown on Fig. 1 for one waveform.

Principal Components Analysis (PCA) is a method that can be used as a linear dimensionality-reduction scheme (Bishop 2006; Murphy 2012). It works by finding P orthonormal basis vectors that explain most of the variability of the data when given some examples to learn from.

The principals components describing the shape of the waveform were found by the following five steps: (I) creating a number N_{PCA} of model realizations from the prior. (II) Evaluating the non-linear forward for each prior realization to get $N_{\text{pairs}} \cdot N_{\text{PCA}}$ examples of waveforms, where N_{pairs} is the number of source-receiver pairs in a single experiment. (III) Computing the first arrival traveltime as described above for each waveform. (IV) Cutting out t_{PCA} ns of each waveform following the first arrival. (V) Training the principal components using the $N_{\text{pairs}} \cdot N_{\text{PCA}}$ examples of cutouts of waveforms. Thus, any waveform can now be described by the K attributes described above.

2.2 Constructing a Linear Forward Operator Using Ridge Regression

In general, electromagnetic (and elastic) waveform modeling represent a non-linear forward model. However, in order to make use of the linear Gaussian solution (Eqs. 4, 5) the forward problem needs to be linear. Therefore, a linear approximation is sought to the full non-linear forward model. To achieve this, ridge regression is proposed.

Ridge Regression (Bishop 2006; Murphy 2012) is an algorithm that can find a linear operator $\mathbf{G}^{ML}$ between a model and observed data using a data set by finding a matrix $\mathbf{G}^{ML}$ that minimizes

$$\mathscr{L} = \sum_{i=1}^{N} \left(\mathbf{d}_i - \mathbf{G}^{ML} \cdot \mathbf{m}_i\right)^2 + \alpha \cdot \sum_{i=1}^{N_d} \sum_{j=1}^{N_m} \left(\mathbf{G}_{i,j}^{ML}\right)^2, \tag{7}$$

where $\mathbf{G}_{i,j}^{ML}$ is element (i, j) in $\mathbf{G}^{ML}$ and α is a regularization parameter. The vectors $\mathbf{m}_i$ and $\mathbf{d}_i$ represent N sets of corresponding models and data, where the vectors

have lengths N_m and N_d. This represent a *training* data set from which $\mathbf{G}^{ML}$ can be inferred. The training data are obtained by generating a large, N, set of models from the assumed prior distribution, and then computing the corresponding data by evaluating an accurate and typically non-linear, forward model. This may be CPU intensive, but it needs to be done only once. In the example that follows, the accurate forward model was chosen as two-dimensional full waveform modeling, but in principle any nonlinear forward models can be approximated in this way by a linear forward model.

Minimizing the above functional is referred to as training the ridge regressor. L-fold cross-validation is a method that can be used to determine whether a specific choice of regularization α is a good value ((Bishop 2006), (Murphy 2012)). It works by (I) partitioning the data set into L subsets, (II) training the ridge regression algorithm with the chosen value of α using only $L - 1$ subsets, (III) evaluating the performance of the trained ridge regression algorithm on the Lth subset. Steps (II) to (III) are then repeated L times, where each time a new subset is left out during the training process. The average performance of the ridge regressor on the subset that was left out says how good that specific choice of α is.

A commonly used measure of the performance of a ridge regressor is the R^2-score (Tanaka and Huba 1989). It is defined as

$$R^2 = 1 - \frac{\sum_i |\mathbf{G}^{ML} \cdot \mathbf{m}_i - \mathbf{d}_i|^2}{\sum_i |\mathbf{d}_i - \bar{\mathbf{d}}|^2}, \tag{8}$$

where $\bar{\mathbf{d}}$ is the empirical mean of the examples and the sum runs over all examples in the subset. A value of α can be picked by scanning through values spaced uniformly on a log-scale, performing L-fold cross-validation for each value of α and picking the value that results in the highest average R^2-score.

2.3 Quantifying the Modeling Error

The linear operator $\mathbf{G}^{ML}$ described above is an approximation to using full wave-form modeling. This leads to modeling errors that, if not accounted for, will lead to fitting noise as data. The approximate forward $\mathbf{G}^{ML}$ can therefore not be used as it is without quantifying and accounting for the modeling error introduced by using the approximation. To achieve this, the methodology described in Hansen et al. (2014) was used.

A new set of N_{me} realizations of the prior is generated as $[\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_{N_{me}}]$. For each generated model, $\mathbf{m}_i$, the forward response is now computed using both the accurate non-linear forward model as $\mathbf{d}_{acc} = g(\mathbf{m}_i)$, and using the approximate linear forward model as $\mathbf{d}_{rr} = \mathbf{G}^{ML}\mathbf{m}_i$. The difference between the two sets of data for the same model is then a realization of the modeling error, $\Delta \mathbf{d}_i = \mathbf{d}_{acc} - \mathbf{d}_{rr}$. For a high number of N_{me} one can in this way obtain a representative sample of the (unknown) probability distribution describing the modeling error.

If the noise model is Gaussian (as it is in the linear Gaussian formulation, $\mathcal{N}(0, \mathbf{C}_d)$) and the modeling error, here through the sample of the model error, can be quantified using a Gaussian model, then Mosegaard and Tarantola (2002) demonstrated that the
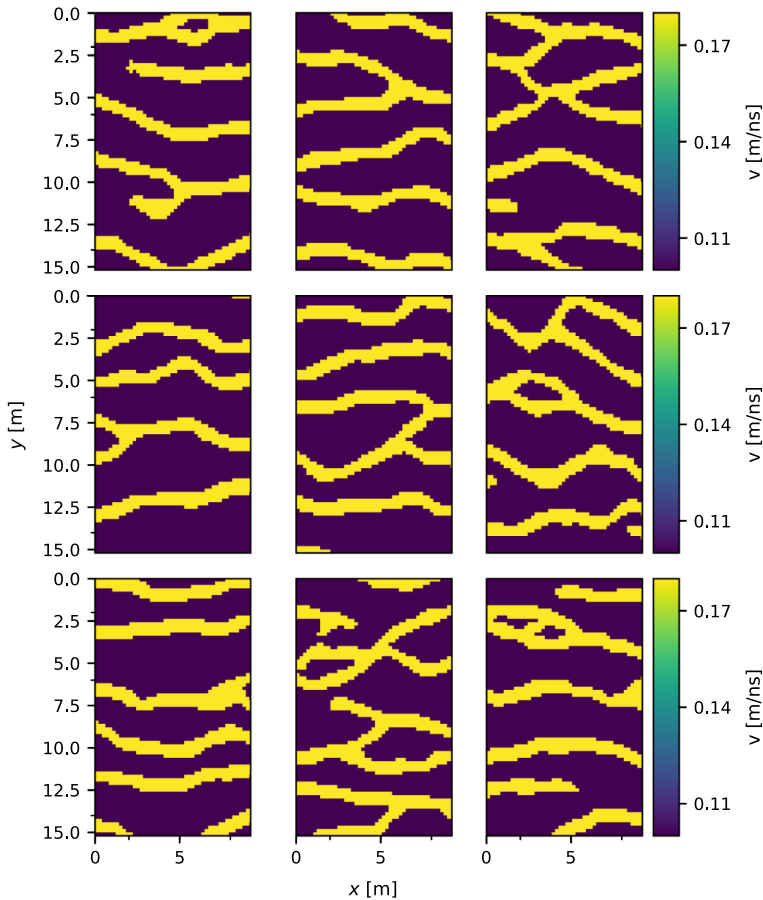
Fig. 2 Realizations of velocity field taken from a multiple-point statistics distribution. The top row shows the reference models $\mathbf{m}^1_{\text{ref}}$, $\mathbf{m}^2_{\text{ref}}$ and $\mathbf{m}^3_{\text{ref}}$

modeling error can be accounted for by addition of the covariance describing the noise and the modeling error. This is very convenient, as this allows accounting for the modeling error simply by using the Gaussian likelihood model $L(\mathbf{m}) = \mathcal{N}(\mathbf{d}_{0_T}, \mathbf{C}_d + \mathbf{C}_T)$.

When imposing a Gaussian model to describe the modeling error, it is important to consider how valid the inferred Gaussian model is to describe the sample. To this end a χ^2-test is proposed. If the chi-squared value for a specific realization of the modeling error, $\Delta\mathbf{d}_i$, is computed using

$$\chi_i^2 = -\frac{1}{2}(\Delta\mathbf{d}_i - \mathbf{d}_{0_T})\mathbf{C}_T^{-1}(\Delta\mathbf{d}_i - \mathbf{d}_{0_T})^T, \tag{9}$$
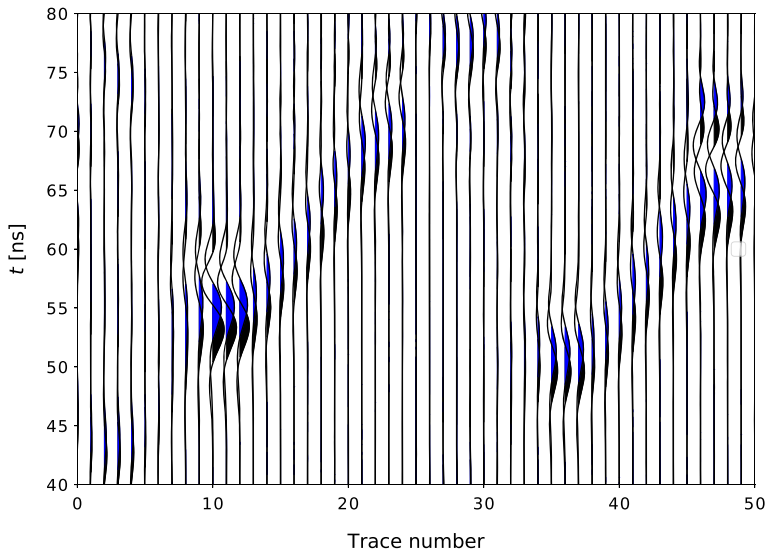
Fig. 3 Synthetic recordings of Ricker pulses with noise after passing through the subsurface slowness field. Black is a recording without noise, blue is with noise. Only 50 out of 702 recordings in a single synthetic experiment are shown here

then all χ^2 values will be distributed according to a chi-squared distribution with $N_d/2$ degrees of freedom (Tarantola 2005). This criterion will later be used to test the validity of the Gaussian assumption.

2.4 Constructing a Gaussian Prior

The least squares approach demands that a Gaussian a priori probability distribution be used (Eqs. 4, 5). When this is not the case, one must instead choose the Gaussian model whose mean and covariance best describe the available information. If one can generate a large set of examples of a preferred, non-Gaussian, model, then one can simply compute the empirical mean $\mathbf{m}_0$ and covariance $\mathbf{C}_m$ from those examples. This approximated prior, $\mathcal{N}(\mathbf{m}_0, \mathbf{C}_m)$, is different from the true prior, but it captures the two-point statistics, while maximum entropy is implicitly assumed for the higher order statistics. Hence, the information content in the inferred Gaussian prior is lower than what is represented by the non-Gaussian prior, but no additional, erroneous information is introduced by using a Gaussian prior (Journel and Zhang 2006).

3 Synthetic Case Study

The methodology described above is tested on a synthetic case. Care has been taken to make the case realistic, in the sense that a reference model is obtained from a non-Gaussian (multiple point based) statistical model, non-linear full waveform modeling is used and experimental noise is simulated to create the data set.

3.1 Creation of the Reference Models and Data

The Subsurface Model **m** describes a two-dimensional slowness field of size 9.2×15.2 m (46×76 pixels of size 0.2 m $\times$ 0.2 m). A number of reference models are constructed to mimic a reasonable geological variation, based on multiple-point statistical (MPS) simulation. In practice, the SNESIM algorithm (Strébelle 2002) is used as implemented by Hansen et al. (2016). It can generate realizations of a high order statistical model by borrowing conditional statistics from a training image, which is a type of example model believed to represent spatial variability. In practice, such a training image can, for example, be obtained from outcrops in a nearby gravel pit and/or a natural cliff. The widely used training image from Strébelle (2001) (Fig. 4.1(c)) is used as an example. This results in realizations from the resulting prior model as shown in Fig. 2. The key feature of these models is that they are inconsistent with a Gaussian model, as they produce a binary slowness field that is assumed to represent gravel channels in a sandy background model.

Electromagnetic signals in a material are sensitive to properties of the material such as the dielectric permittivity, the electrical conductivity, and the magnetic permeability. In this study we consider only the dielectric permittivity, which influences the phase velocity of the signal. The relative dielectric permittivity of the sand and the gravel channel is set to $\varepsilon_r = 8.99/\varepsilon_0$ (0.10 m/s) and $\varepsilon_r = 2.77/\epsilon_0$ (0.18 m/s), respectively (Ernst et al. 2006), where ϵ_0 is the dielectric permittivity of free space. Figure 2 shows nine realizations of this prior model, the top row shows three reference models $\mathbf{m}_{ref}^1$, $\mathbf{m}_{ref}^2$ and $\mathbf{m}_{ref}^3$, that will be considered in detail.

The Forward Model The geometrical setup is the same as in Hansen and Cordua (2017) with 2 boreholes and $N_{pairs} = 702$ source-receiver pairs, resulting in 702 waveforms for each slowness model. The forward model is an implementation of a 2-D finite-difference time-domain solution of Maxwell's equations by Ernst et al. (2006). This solves the transverse magnetic and transverse electric mode equations, which are Maxwell's equations in two-dimensional rotationally symmetric cylindrical coordinates. A Ricker pulse with a dominant frequency of 10^8 Hz is used as source. An experiment is simulated where the waveform is recorded for 100 ns for each source receiver pair. In order to avoid too-large numerical dispersion, a finer discretization is needed in the finite difference scheme than is used to describe the model space. Therefore, linear interpolation of the velocity field to a finer grid with pixel size 0.1 m $\times$ 0.1 m is performed. 5 cells at the border are used to avoid reflection from the boundaries. The sources and receivers were always at least 2 m from the boundaries to avoid artifacts. The discretization of the subsurface in the forward model introduces some error as demonstrated by Kaipio and Somersalo (2007), although this has been assumed negligible compared to the simulated experimental noise discussed below. An example of recorded waveforms is shown in Fig. 3.

Noise In an effort to add realistic noise to the simulated experiment, a combination of correlated instrument/measurement noise and uncertainty of borehole locations is considered. The uncertainty due to measurement noise is described by a correlated Gaussian noise model. The covariance of the electric noise between time t and t'
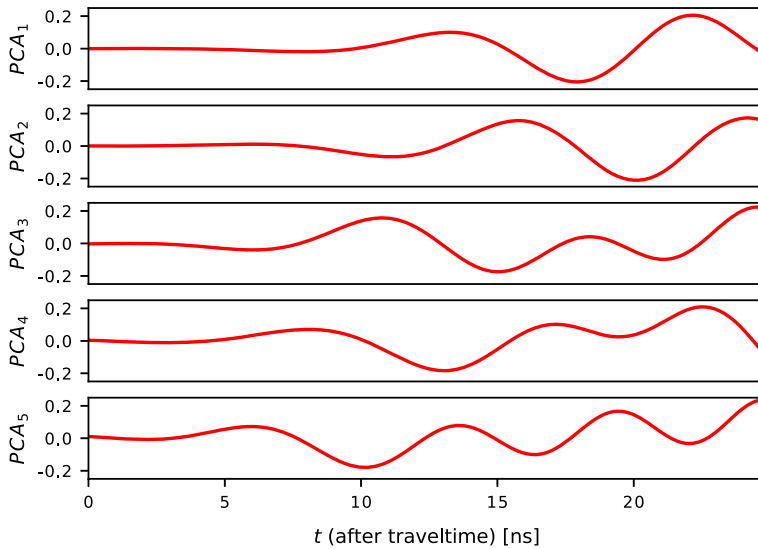
Fig. 4 Principal components describing the shape of the wavelet when it arrives at the receiver location. The 25 ns of the waveform following the first arrival time are used to obtain E_{PCA_1}, E_{PCA_2}, etc

is given by $4 \cdot 10^{-8} \cdot \exp\left(-|t - t'|/212\,\text{ns}\right)$ V^2. Lassen et al. (2015) suggest that the source-receiver location is associated with uncertainty. First the uncertainty of location of the boreholes is assumed to be distributed uniformly within ± 5 cm of the assumed location. The direction, as angle from horizontal, of the borehole is assumed to follow a normal distribution with mean 90 degrees and standard deviation of 1.5 degrees as suggested by (Lassen et al. 2015). Finally, the location of the source-receiver locations are assumed to be distributed ± 5 cm around the assumed location. Then the actual location of the source-receiver locations, for one specific model, is obtained by simulating one value from each of these three distributions prior to solving the forward problem. Then a realization of the measurement noise is added to the simulated data, to obtain an example of a simulated observed data set, contaminated with realistic noise. Figure 3 shows an example of traces with and without added noise.

Creation of the Principal Components The number of principal components was chosen to be $P = 5$, resulting in $K = 10$ attributes describing each waveform. The principal components were chosen to describe $t_{PCA} = 25$ ns following the first arrival traveltimes, describing the main part of the information in the waveform. To generate the examples to train the principal components, $N_{PCA} = 10$ realizations of the prior were generated, resulting in $N_{PCA} \cdot N_{pairs} = 7020$ examples of waveforms used for training. The resulting principal components can be seen in Fig. 4. These are the basis vectors that the cutout waveforms are projected to. A relatively short value of t_{PCA} was chosen, as the latter part of the waveform was expected to be less linear since it contains more reflections.

In this way three reference models, $\mathbf{m}_{\text{ref}}^1$, $\mathbf{m}_{\text{ref}}^2$ and $\mathbf{m}_{\text{ref}}^3$, and a corresponding set of reference data with noise, $\mathbf{d}_{\text{ref}}^1$, $\mathbf{d}_{\text{ref}}^2$ and $\mathbf{d}_{\text{ref}}^3$, have been generated. The inversion strategy is eventually tested by considering 1000 such reference data sets.

3.2 Linear Gaussian Waveform Inversion

Given a reasonably realistic data set, the goal is now to demonstrate how the proposed method can be used to solve the waveform inversion problem using linear Gaussian inversion (Eqs. 4, 5). Recall that to achieve this, it is necessary to select a Gaussian prior, a linear forward model, and a Gaussian noise model.

The prior, $\mathcal{N}(\mathbf{m}_0, \mathbf{C}_m)$. The reference models were realizations from a non-Gaussian MPS model where it is possible to generate realizations of the MPS model. Following Sect. 2.4, a large number of realizations, 7000, of the MPS prior were generated, from which an empirical prior Gaussian model was estimated. Nine realizations of the MPS model are shown in Fig. 2 and nine examples of the Gaussian prior are shown in Fig. 5 for comparison. It is clear that the full MPS information is not taken into account in the realizations from the Gaussian prior, but the second-order statistics are consistent with the MPS model.

The forward, $\mathbf{G}^{ML}$. A data set containing 7000 model realizations (from the MPS prior) and data (using the non-linear full waveform modeling approach) was generated. These were used to train the ridge regressor, as described in Sect. 2.2, with five-fold cross-validation and using α-values spaced uniformly on a log-scale $[10^3, 10^4, \ldots, 10^{13}]$. The value of α that gave the best average R^2-score was used and the ridge regressor was trained using the whole dataset using this value of α. A ridge regressor was fit separately for each attribute (t_0, $t_{\max(E)}$, $t_{\min(E)}$, ...) because the different data attributes have different scalings.

The sensitivity kernels for a few source-receiver pairs and attributes are shown in Fig. 6 with the locations of the source and receiver. A sensitivity kernel is a row in the approximate linear $\mathbf{G}^{ML}$ forward operator that is reshaped into the same shape as the subsurface slowness field. It shows how much a small change in a model parameter influences the observed data. The noise on each kernel differs but, for example, the sensitivity kernels for the first arrival time, Fig. 6 (top row), are quite similar to the theoretically obtained kernels considered by Woodward (1992) and Dahlen et al. (2000). These kernels are obtained using the linear forward model, which is numerically very efficient to evaluate.

The linear $\mathbf{G}$ describing the first arrival time, t_0, could in principle have been obtained analytically using the theory in Woodward (1992). However, that theory was developed for a very specific, theoretically defined, first arrival time. Here the forward model, and hence the linear operator, is directly related to the choice of picking method. Another choice of data could have been to pick the time defined by where the curvature is highest between the first arrival and time of maximum amplitude. The linear operator to describe such data is not easily inferred from theory, but can be trivially used in the methodology presented above. Analysis of the corresponding sensitivity kernels can afterwards be efficiently computed numerically, and analyzed as in Fig. 6.
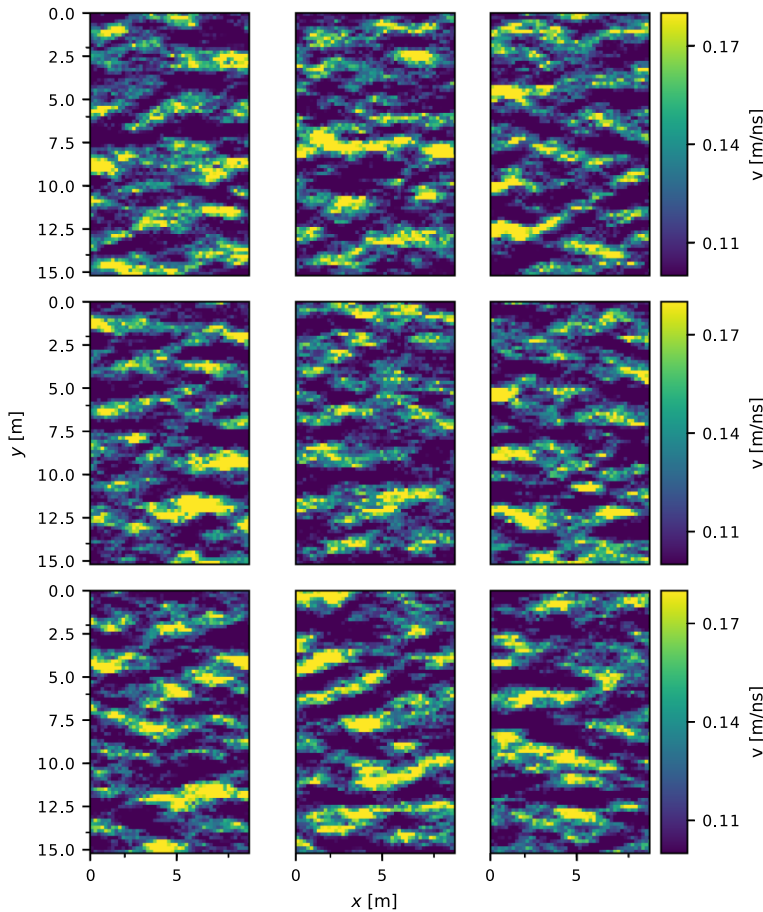
Fig. 5 Realizations of velocity field taken from a Gaussian distribution that approximates the multiple-point statistics prior (examples shown in Fig. 2)

The modeling error, $\mathcal{N}(\mathbf{d}_{0_T}, \mathbf{C}_T)$. A number of model realizations $N_{me} = 1,500$ from the MPS prior were generated and the resulting forward response computed using both the linear approximate $\mathbf{G}^{ML}$ forward operator, as well as using the non-linear g forward operator. From these, a Gaussian model of the modeling error due to using linear approximate forward models is constructed using the methodology described in Sect. 2.3.

In practice, the modeling error may not be perfectly described by a Gaussian model. If, for example, the one-dimensional marginal distribution of the modeling tends more toward a Cauchy distribution, with long tails, than a Gaussian, then enforcing the Gaussian model will lead to underestimating the probability of the tail values. If not accounted for, this will lead to fitting these extreme values as data, when in fact they represent noise. To remedy this, additional uncorrelated noise is added to the error distribution by multiplying the diagonal of $\mathbf{C}_d$ with a constant c. At worst, this will
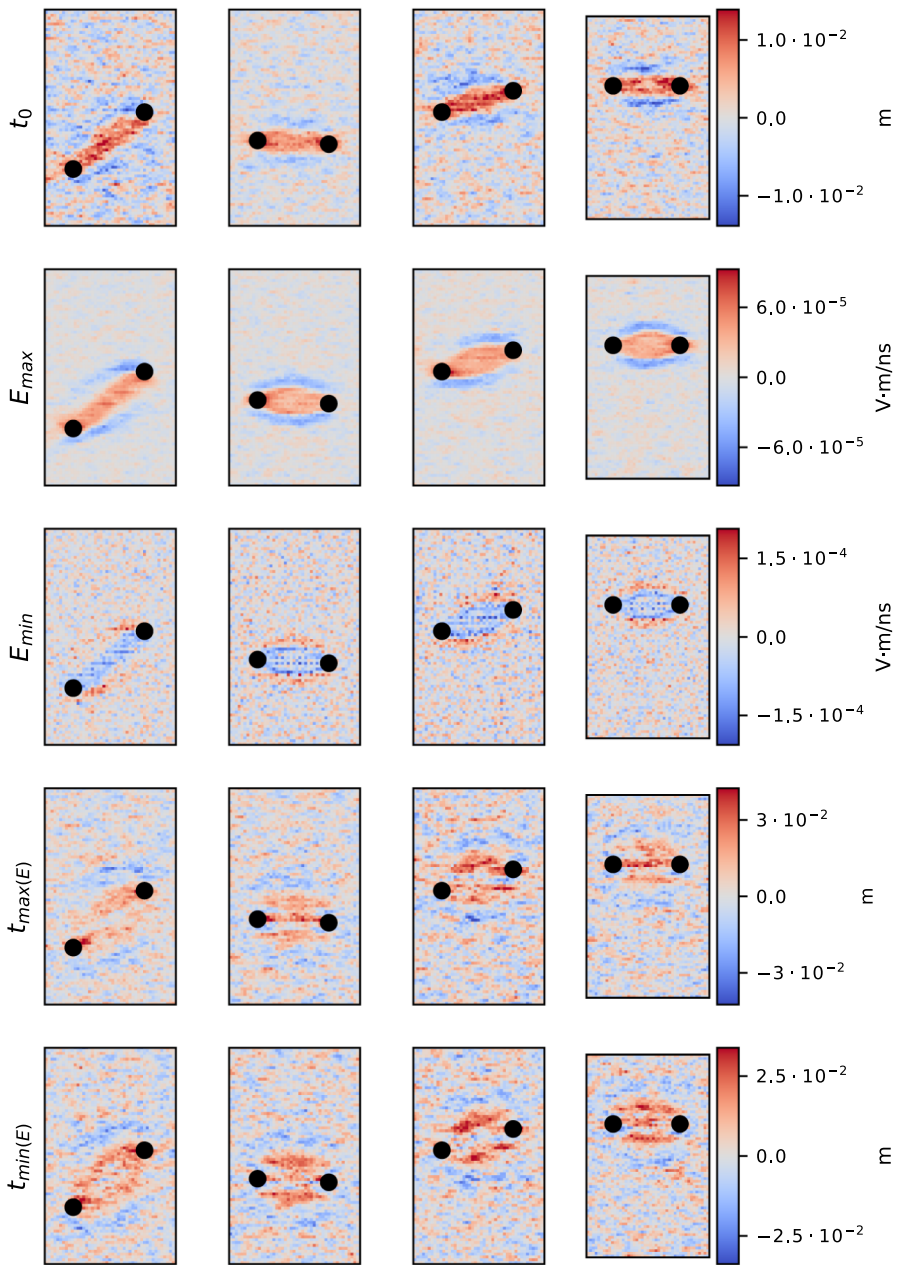
Fig. 6 A few examples of sensitivity kernels of the approximated linear operator $\mathbf{G}^{ML}$ obtained using the ridge regression algorithm. The location of the source and receiver are marked with black circles. Each row shows a different data attribute
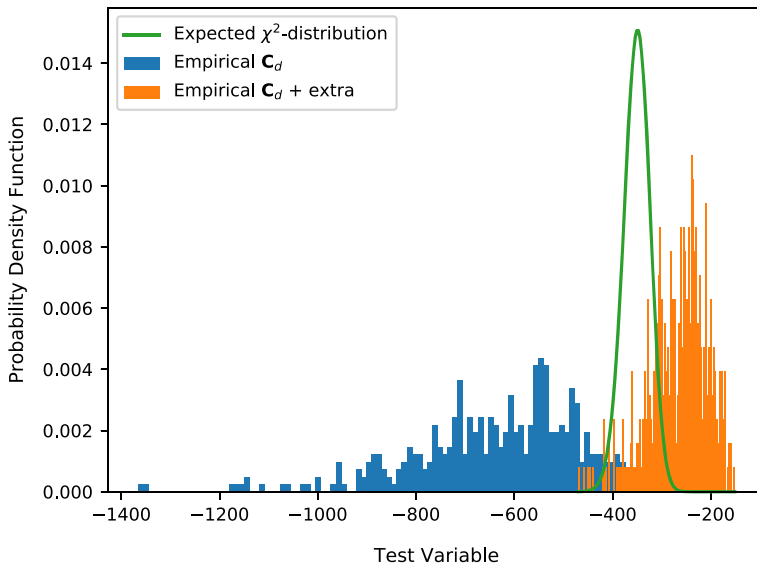
Fig. 7 The empirical distribution of the test statistic in Eq. 9 when determining the covariance of the error model. The empirical distribution is shown before and after the diagonal of $\mathbf{C}_d$ has been multiplied by 1.1. The expected χ^2-distribution is shown for comparison

only decrease the resolution of the posterior distribution, since extra noise is added. At best, this will remove unwanted inversion artifacts.

The choice of c is picked by creating 500 model realizations from the MPS prior and computing the resulting forward response by using the linear $\mathbf{G}^{ML}$ and the non-linear g as above and, finally, calculating the chi-squared test variable described in Eq. 9. If more than half of these 500 empirical chi-squared values have a lower value than the mean of the expected chi-squared distribution, a higher value of c is tried. This process is repeated until more than half of the empirical chi-squared values have a higher value than the mean of the expected chi-squared distribution. See Fig. 7 for an example of an empirical chi-squared distribution before and after $\mathbf{C}_d$ has been multiplied by $c = 1.1$.

3.3 Inversion Results

It is now computationally efficient to solve the waveform inversion problem by calculating the posterior probability density analytically using Eqs. 4, 5. This is done for the 1000 reference data described above. Three examples of analytical posterior mean models obtained using the linear Gaussian wavefield inversion methodology are shown in Fig. 8. The original reference models are shown in Fig. 8a.

The inversion strategy proposed in this paper is compared with three different inversion strategies, and the corresponding posterior mean models are computed and shown in Fig. 8b–e: (I1) The first inversion strategy is a linear forward operator based on straight rays to model the first arrival time, with assumed uncorrelated noise with a standard deviation of 1.5 ns. The modeling errors are thus assumed negligible.
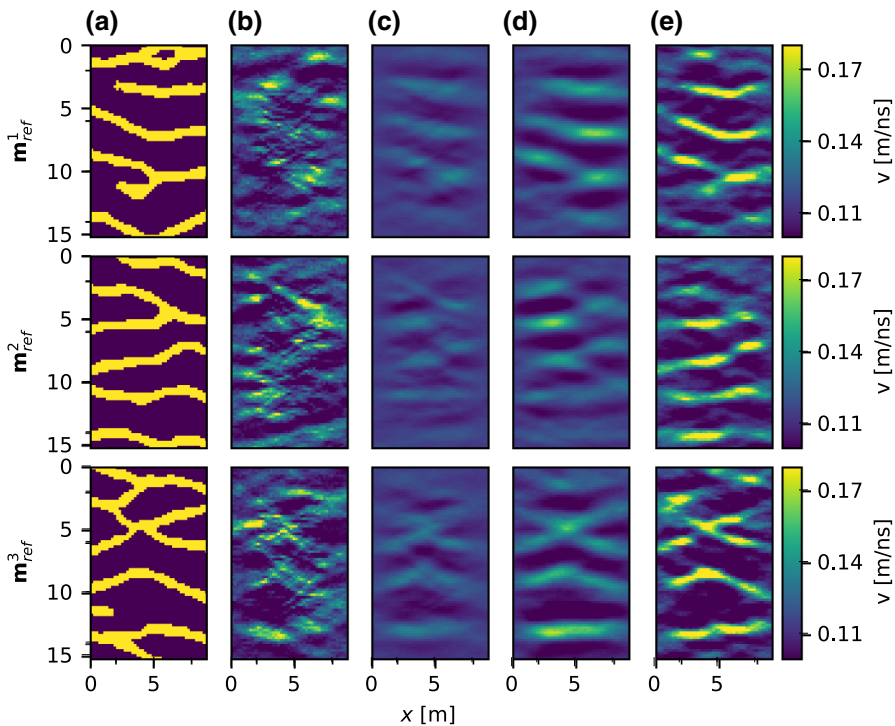
Fig. 8 Qualitative comparison of four inversion strategies on the three reference models. Each row is a new reference model with four posterior means obtained with the strategies described in Sect. 3.3. **a**: the original model that is sought to reconstruct. **b**: Strategy I1. **c** Strategy I2. **d**: Strategy I3, **e**: Strategy I4, the strategy proposed in this paper

Posterior mean is shown in Fig. 8b. (I2) The second inversion strategy is similar to I1, except that the modeling error is quantified and accounted for using the method in Sect. 2.3. Posterior mean is shown in Fig. 8c. (I3) The third inversion strategy uses only one attribute (traveltimes) and makes use of the ridge regression algorithm to estimate a linear forward modeling operator. In addition, the associated modeling error is quantified. Posterior mean is shown in Fig. 8d. (I4) The fourth inversion strategy is the one proposed in this paper, using all ten attributes to describe the wavefield, with a full description of the modeling error. Posterior mean is shown in Fig. 8e. Inversion strategies I1, I2, and I3 only make use of the first arrival time t_0, while I4 makes use of all ten attributes!

A quantitative comparison of those three inversion strategies and the strategy proposed in this paper is shown in Table 1 using 1000 generated reference models and data. The table contains (I) the correlation coefficient between the true models and the posterior means and (II) the area under a Receiver Operating Characteristics (ROC) curve (Bishop 2006) when considering the inversion strategies as linear classifiers distinguishing between the two pixel types. If the area under the ROC curve is equal to unity, it means a perfect classifier. If it is equal to 0.5, it means a totally random classifier. In general, Table 1 shows that the quality of the inversion result increases when

Table 1 Quantitative comparison of 1000 posterior means of **m** with the true values of **m** for the four different inversion strategies

Method	Correlation coefficient	Area under ROC curve
I1, Straight-ray (uncorrelated error)	0.31	0.70
I2, Straight-ray (modeling error)	0.52	0.84
I3, First arrival traveltimes (Ridge Regression)	0.56	0.86
I4, Proposed method	0.74	0.95

Only pixels between boreholes are considered

(a) the noise is being more realistically described, and (b) when more data attributes are taken into account. Nine posterior realizations of a single posterior distribution are shown in Fig. 9.

4 Discussion

The proposed methodology is general, and the the posterior distribution can be sampled as follows:

1. Estimate a Gaussian prior distribution $\mathcal{N}\left(\mathbf{m}_0, \mathbf{C}_m\right)$ from a large sample of an arbitrary probability density (possibly non-Gaussian) that represents prior information about **m**.
2. Estimate a linear forward operator $\mathbf{G}^{ML}$ from a large sample of sets of **m** and **d** using the ridge regression algorithm.
3. Estimate a Gaussian model $\mathcal{N}\left(\mathbf{d}_{0_T}, \mathbf{C}_T\right)$ of the modeling error using a large sample of **m** and **d** independent of the sample above.
4. Calculate the posterior analytically using Eqs. 4, 5.

The only thing required to use the strategy is a way to generate prior realizations and a way to simulate the experiment with noise. Only a few thousand prior realizations/simulated experiments are necessary compared to a Monte Carlo sampling strategy, where hundreds of thousands/millions of prior realizations/simulated experiments are required.

The proposed strategy can allow solving inverse problems that were, until now, very challenging due to a lack of linear forward operator with associated description of the modeling error, and a forward that is too computationally expensive for hundreds of thousands of realizations. The examples demonstrate clearly that including the information about the full waveform, using inversion strategy I4, provides by far the best results. This is both clear from Fig. 8 and quantified in Table 1.

The proposed method can also be useful even in the case of first arrival travel time inversion. The simple high frequency straight ray approximation, I2, provides significantly less resolution than using the ridge regressor, I3. As discussed above, the ridge regressor finds the linear operator $\mathbf{G}^{ML}$ that best matches how the specific data were obtained (picked/inferred). In the example considered here, the data preprocessing includes how to find the traveltimes. It does not matter if the travel time is chosen to be the first break, the arrival of the maximum amplitude, or any other attribute,
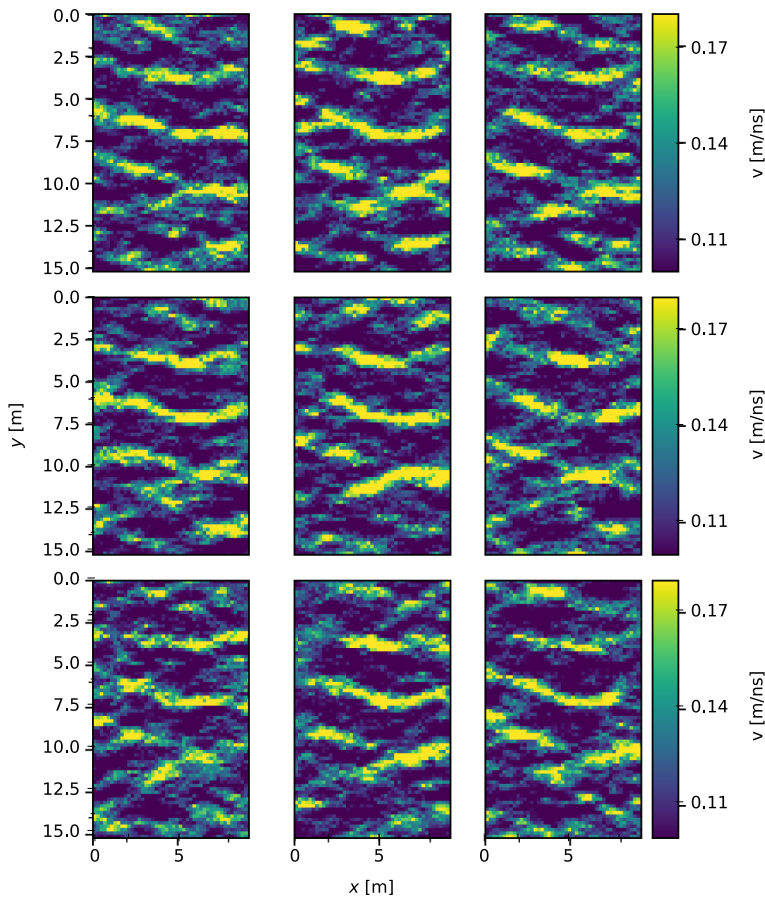
Fig. 9 Realizations drawn from the posterior distribution obtained using the proposed method in this paper for a single model $\mathbf{m}^1_{\text{ref}}$

which will directly affect the validity of any of the ray based approaches (Jacobsen et al. 2010). If only the strategy to pick the arrival time is consistent, then the ridge regression algorithm will learn an optimal linear forward model specifically for that picking strategy. This makes the method quite robust to the different ways in which data, such as the first arrival travel time, may be extracted from the full waveform data. Different users may choose slightly different approaches. The methodology presented above will naturally take this into account.

It may be surprising to observe the very poor resolution obtained using I2, which can be seen in the lack of details in Fig. 8c. It was demonstrated by Hansen et al. (2014) that the modeling error can be orders of magnitude higher than the measurement uncertainty in case of a bimodal velocity field with large contrasts between high and low velocity (as considered here). This is also the reason for the exceptionally bad results obtained when ignoring the modeling errors as in I1, Fig. 8b.

In this context, the use of attributes to describe the full wavefield is very promising. Even when there is only little apparent value on a specific attribute, for example E_{PCA_2}, it will add some information. One worry is that the found linear operator is a very bad approximation. However, as the associated modeling error is assessed, the choice of a bad approximation will lead to high modeling errors. This will then lead to little resolution of the posterior probability density. So using an attribute that has no sensitivity to any part of the waveform will just end up with a very high modeling error and will not bias the inversion result. Only the attributes used that are actually sensitive to the waveform (controlled by the modeling error) will add information to the posterior distribution. Notice from e.g. Fig. 8 that ignoring the modeling errors, as in case I1, leads to erroneous results as errors due to modeling imperfections are fitted as data. Accounting for modeling errors, as described above and used in cases I2, I3, and I4, leads to little or no artifacts due to fitting modeling errors as noise. Apparently none of the modeling error is mapped back into the posterior distribution as well resolved features. This suggests that the way the modeling error is quantified works well for the case considered here.

Further, as more information is taken into account, the resolution of the posterior distribution increases, as seen by comparing the results from I2, I3, and I4. Case I4, using the most information (including the PCA components) adds significant information to the posterior distribution, while at the at the same time this added information is consistent with the reference models, with no apparent artifacts related to modeling errors.

In the presented case study, the source wavelet was considered known. In practice this is never the case, and the wavelet should be estimated separately. The uncertainty of the wavelet can be quantified analytically as demonstrated by Buland and Omre (2003). The variability of the source wavelet can easily be included in the proposed method as an extra contribution of noise during the non-linear forward modeling, which will then propagate into the modeling error. In this way it can be taken into account in the proposed method.

The case study was constructed to be reasonably realistic using (I) reference models that are non-Gaussian, (II) noise on the recording geometry and (III) finite difference based full waveform modeling. One could argue that the effect of using a 'wrong' prior model should be considered. But for any probabilistic formulated inverse problem, the choice a 'wrong' prior model, in the sense that it is inconsistent with actual subsurface/reference model, will make very little sense, and will produce unreliable results. Instead a conservative choice of prior was made: assume as much as possible, but never more than what is known. In the present case the 'optimal' Gaussian model was chosen to describe a discrete non-Gaussian reference model.

At least two other types of waveform inversion methods exist. Using Markov chain Monte Carlo sampling methods, the posterior distribution can in principle be sampled using the full non-linear forward and and an arbitrarily complex prior model. But, while possible for a limited set of sources and receivers (Cordua et al. 2012) this leads to a computationally intractable problem for any realistically sized problem. The gradient based adjoint method is based on an implicit Gaussian prior assumption, and requires relatively few realizations of the full non-linear forward problem (see Tarantola 1984; Ernst et al. 2007 and Keskinen et al. 2017). However, the solution is

one optimal model, and not a set of realizations from the posterior distribution, and it is difficult to assess the associated uncertainty. Also, the inversion result is strongly dependent on a good starting model, as demonstrated by Keskinen et al. (2017). The method proposed here has the Gaussian limitations of the adjoint based approach, but it allows exploring the full posterior uncertainty by sampling the posterior distribution. It leads to a worse resolution of the posterior probability density than the full non-linear non-Gaussian sampling approach, but it is computationally orders of magnitude faster than the sampling approach.

The main challenges using the methodology are: (I) *Generating a large training data set*. The full forward model needs to be evaluated N times (in this example, $N = 9000$ times). This may be time-consuming, but it only needs to be done once. Monte Carlo sampling algorithms typically need many more iterations (normally around 100,000–1,000,000 iterations). (II) *Training the ridge regressor*. It takes time to train the ridge regressor, although this time is only measured in minutes. (III) *Difficulties generalizing to other geometries or priors*. The forward operators obtained using the proposed strategy are applicable only to the specific recording geometry and prior that was used in the training set. If one wishes to use a different recording geometry or prior model, then a new training data set needs to be computed.

5 Conclusion

A general linear probabilistic waveform strategy has been demonstrated that allows computing the Gaussian posterior distribution analytically by replacing a computationally complex forward model with a linear forward model using a ridge regressor. The linear forward thus becomes specific to the problem at hand, including data pre-processing. Approximating a non-linear forward with a linear operator introduces a modeling error which was quantified to allow for a consistent inversion.

The strategy has been demonstrated on a tomographic crosshole problem and compared with three other probabilistic inversion strategies that also yield an analytical description of the posterior slowness field (qualitative comparison in Fig. 8 and quantitative comparison in Table 1). Better inversion results are observed when using the proposed strategy, and no artifact related to modeling errors can be identified.

The main benefits of the proposed strategy are that it uses information about the whole waveform and can give a closed-form description of the posterior slowness field after only a few thousand simulated examples ($N = 9000$ in this example). It is thus much faster than a computationally heavy Monte Carlo full-waveform sampling algorithm and it gives better inversion results than the straight-ray traveltime inversion strategy.

The proposed strategy is general and can be applied to other problems where it is computationally expensive to do more than a few thousand simulations of the experiment, as needed by sampling methods.

References

Alumbaugh D, Chang PY, Paprocki L, Brainard JR, Glass RJ, Rautman CA (2002) Estimating moisture contents in the vadose zone using cross-borehole ground penetrating radar: a study of accuracy and repeatability. Water Resour Res 38(12):45-1

Annan AP (2005) GPR methods for hydrogeological studies. In: Rubin Y, Hubbard SS (eds) Hydrogeophysics. Springer, Berlin, pp 185–213

Binley A, Winship P, Middleton R, Pokar M, West J (2001) High-resolution characterization of vadose zone dynamics using cross-borehole radar. Water Resour Res 37(11):2639–2652

Bishop CM (2006) Pattern recognition and machine learning. Springer, Berlin

Buland A, Omre H (2003) Bayesian wavelet estimation from seismic and well databayesian wavelet estimation. Geophysics 68(6):2000–2009

Cordua KS, Hansen TM, Mosegaard K (2012) Monte carlo full-waveform inversion of crosshole GPR data using multiple-point geostatistical a priori information. Geophysics 77(2):H19–H31

Cordua KS, Looms MC, Nielsen L (2008) Accounting for correlated data errors during inversion of crossborehole ground penetrating radar data. Vadose Zone J 7(1):263–271

Dafflon B, Irving J, Holliger K (2009) Use of high-resolution geophysical data to characterize heterogeneous aquifers: Influence of data integration method on hydrological predictions. Water Resour Res 45(9):W09407. https://doi.org/10.1029/2008WR007646

Dahlen FA, Hung SH, Nolet G (2000) Fréchet kernels for finite-frequency traveltimes—I. Theory. Geophys J Int 141(1):157–174

Day-Lewis FD, Lane JW (2004) Assessing the resolution-dependent utility of tomograms for geostatistics. Geophys Res Lett 31(7):L07503

Day-Lewis FD, Singha K, Binley AM (2005) Applying petrophysical models to radar travel time and electrical resistivity tomograms: resolution-dependent limitations. J Geophys Res Solid Earth 110(B8):B08206

Eppstein MJ, Dougherty DE (1998) Efficient three-dimensional data inversion: Soil characterization and moisture monitoring from cross-well ground-penetrating radar at a Vermont test site. Water Resour Res 34(8):1889–1900

Ernst JR, Green AG, Maurer H, Holliger K (2007) Application of a new 2D time-domain full-waveform inversion scheme to crosshole radar data. Geophysics 72(5):J53–J64

Ernst JR, Holliger K, Maurer H, Green A (2006) Realistic FDTD modelling of borehole georadar antenna radiation: methodlgy and application. Near Surface Geophys 4(1):19–30

Gloaguen E, Giroux B, Marcotte D, Dimitrakopoulos R (2007) Pseudo-full-waveform inversion of borehole GPR data using stochastic tomography. Geophysics 72(5):J43–J51

Hansen TM, Cordua KS (2017) Efficient Monte Carlo sampling of inverse problems using a neural network-based forward–applied to GPR crosshole traveltime inversion. Geophys J Int 211(3):1524–1533

Hansen TM, Cordua KS, Jacobsen BH, Mosegaard K (2014) Accounting for imperfect forward modeling in geophysical inverse problems—exemplified for crosshole tomography. Geophysics 79(3):H1–H21

Hansen TM, Journel AG, Tarantola A, Mosegaard K (2006) Linear inverse Gaussian theory and geostatistics. Geophysics 71(6):R101–R111

Hansen TM, Looms M, Nielsen L (2008) Inferring the subsurface structural covariance model using crossborehole ground penetrating radar tomography. Vadose Zone J 7(7):249–262

Hansen TM, Vu LT, Bach T (2016) MPSLIB: a C++ class for sequential simulation of multiple-point statistical models. SoftwareX 5:127–133, ISSN 2352-7110

Hubbard SS, Peterson JE, Majer EL, Zawislanski PT, Williams KH, Roberts J, Wobber F (1997) Estimation of permeable pathways and water content using tomographic radar data. Lead Edge 16(11):1623–1630

Jacobsen BH, Hansen TM, Cordua KS (2010) Understanding why first-arrival travel times do not obey banana-doughnut sensitivity. In: EGU general assembly conference abstracts, volume 12 of EGU General Assembly Conference Abstracts. p 6480

Journel A, Zhang T (2006) The necessity of a multiple-point prior model. Math Geol 38(5):591–610

Kaipio J, Somersalo E (2007) Statistical inverse problems: discretization, model reduction and inverse crimes. J Comput Appl Math 198(2):493–504

Keskinen J, Klotzsche A, Looms MC, Moreau J, van der Kruk J, Holliger K, Stemmerik L, Nielsen L (2017) Full-waveform inversion of crosshole GPR data: Implications for porosity estimation in chalk. J Appl Geophys 140:102–116

Kowalsky MB, Finsterle S, Peterson J, Hubbard S, Rubin Y, Majer E, Ward A, Gee G (2005) Estimation of field-scale soil hydraulic and dielectric parameters through joint inversion of GPR and hydrological data. Water Resour Res 41(11):W11425

Lassen RN, Sonnenborg TO, Jensen KH, Looms MC (2015) Monitoring CO_2 gas-phase migration in a shallow sand aquifer using cross-borehole ground penetrating radar. Int J Greenh Gas Control 37:287–298

Linde N, Binley A, Tryggvason A, Pedersen LB, Revil A (2006) Improved hydrogeophysical characterization using joint inversion of cross-hole electrical resistance and ground-penetrating radar traveltime data. Water Resour Res 42(12):W12404

Menke W (1989) Geophysical data analysis: discrete inverse theory. Academic Press Inc, Cambridge

Molyneux J, Schmitt D (1999) First break timing: arrival onset times by direct correlation. Geophysics 64(5):1492–1501

Mosegaard K, Tarantola A (2002) Probabilistic approach to inverse problems. In: Lee W, Kanamori H, Jennings P, Kisslinger C, editors, International handbook of earthquake and engineering seismology, volume 81A, chapter 16, Academic Press, Cambridge. ISBN 978-0124406520, pp 237–265

Mosser L, Dubrule O, Blunt MJ (2018) Stochastic seismic waveform inversion using generative adversarial networks as a geological prior. arXiv preprint arXiv:1806.03720

Murphy KP (2012) Machine learning: a probabilistic perspective. The MIT Press, Cambridge ISBN 0262018020, 9780262018029

Peterson JE (2001) Pre-inversion corrections and analysis of radar tomographic data. J Environ Eng Geophys 6(1):1–18

Strébelle S (2001) Sequential simulation drawing structures from training images. Ph.D. thesis, Stanford University, an optional note

Strébelle S (2002) Conditional simulation of complex geological structures using multiple-point statistics. Math Geosci 34:1–21

Tanaka JS, Huba GJ (1989) A general coefficient of determination for covariance structure models under arbitrary GLS estimation. Br J Math Stat Psychol 42(2):233–239

Tarantola A (1984) Inversion of seismic reflection data in the acoustic approximation. Geophysics 49(8):1259–1266

Tarantola A (1987) Inverse problem theory. Methods for data fitting and model parameter estimation. Elsevier Science Publ. Co. Inc., Amsterdam

Tarantola A (2005) Inverse problem theory and methods for model parameter estimation, vol 89. SIAM, Philadelphia

Tarantola A, Mosegaard K (1995) Monte carlo sampling of solutions to inverse problems. J Geophys Res Solid Earth 100(B7):12431–12447

Tarantola A, Valette B (1982a) Generalized nonlinear inverse problems solved using the least squares criterion. Rev Geophys 20(2):219–232

Tarantola A, Valette B (1982b) Inverse problems= quest for information. J Geophys 50(3):150–170

Woodward MJ (1992) Wave-equation tomography. Geophysics 57(1):15–26