Computational Geosciences (2021) 25:1105–1121
https://doi.org/10.1007/s10596-021-10043-4

ORIGINAL PAPER

Batch seismic inversion using the iterative ensemble Kalman smoother

Michael Gineste[1] · **Jo Eidsvik**[1]

Received: 24 June 2020 / Accepted: 5 February 2021 / Published online: 8 March 2021

Abstract

An ensemble-based method for seismic inversion to estimate elastic attributes is considered, namely the iterative ensemble Kalman smoother. The main focus of this work is the challenge associated with ensemble-based inversion of seismic waveform data. The amount of seismic data is large and, depending on ensemble size, it cannot be processed in a single batch. Instead a solution strategy of partitioning the data recordings in time windows and processing these sequentially is suggested. This work demonstrates how this partitioning can be done adaptively, with a focus on reliable and efficient estimation. The adaptivity relies on an analysis of the update direction used in the iterative procedure, and an interpretation of contributions from prior and likelihood to this update. The idea is that these must balance; if the prior dominates, the estimation process is inefficient while the estimation is likely to overfit and diverge if data dominates. Two approaches to meet this balance are formulated and evaluated. One is based on an interpretation of eigenvalue distributions and how this enters and affects weighting of prior and likelihood contributions. The other is based on balancing the norm magnitude of prior and likelihood vector components in the update. Only the latter is found to sufficiently regularize the data window. Although no guarantees for avoiding ensemble divergence are provided in the paper, the results of the adaptive procedure indicate that robust estimation performance can be achieved for ensemble-based inversion of seismic waveform data.

Keywords Ensemble smoother · Iterative ensemble kalman smoother · Data assimilation · Seismic inversion

Mathematics Subject Classification (2010) 62L12 · 86-08

1 Introduction

The motivation behind this work is seismic waveform inversion, where the goal is to predict theelastic attributes of the subsurface, in the form of acoustic- and shear wave velocities and density, conditional on records of seismic reflection data. Seismic inversion thus provides an image of the subsurface and its interpretation can, combined with other geophysical analysis, be used to establish a geological model.

Phrased in a Bayesian setting, where initial knowledge is incorporated via a prior probability distribution and a likelihood model is used for the specific data, the solution to this Bayesian inversion problem is available as the posterior probability distribution. However, with the non-linearity and complexity of the forward model, there is no closed-form solution to this posterior. In theory, the posterior can be explored by Markov chain Monte Carlo (McMC) sampling, see e.g. [12] for a recent proof-of-concept contribution. However, McMC approaches are difficult to run in parallel and they would require tremendous computing time to ensure convergence and mixing of the output Markov chain.

With the growing availability of diverse data types in complex spatio-temporal systems, there is currently much focus on data assimilation methods that scale well with high-dimensional spaces. One such method is the ensemble Kalman framework [1, 8, 17] which is increasingly applied to problems in the geosciences [5] and has a succesful track record in history matching applications. In particular, the method referred to as the iterative ensemble Kalman smoother (IEnKS), introduced by Bocquet and Sakov [3], is here used for the case of static parameter estimation. The

✉ Michael Gineste
michael.gineste@ntnu.no

Jo Eidsvik
jo.eidsvik@ntnu.no

[1] Department of Mathematical Sciences, NTNU, Trondheim, Norway

IEnKS combines aspects of ensemble-based and variational approaches to data assimilation. Notably, it avoids the need for an available tangent linear model, an attractive feature when using black-box forward models. Instead the linearization is indirectly provided by the ensemble evaluation.

Recent developments indicate that ensemble-based approaches can be used for inversion of seismic tomography data [20] and seismic waveform data [13, 28]. However, it is not obvious how to assimilate the massive data in a reliable manner.

The inversion is formulated as a sequential data assimilation problem, where disjoint subsets or batches of the seismic data records, are used to update the ensemble in a series of assimilation cycles. If the batches are way too large, the ensemble collapses and the estimation procedure is stuck. If the batches are somewhat smaller but still too large, the IEnKS procedure is likely to diverge due to spurious updates to the estimate. On the other hand, if the batches are too small, the inversion run time will grow because of the computational time of the forward model. Hence a key issue is to find efficient batch sizes, and to do so automatically. An efficient batch size is maximizing the amount of data while minimizing the risk of divergence.

With the tuning of data batch windows, the work partially relates to that of [10] who used a data assimilation window in time, but their focus was on chaotic dynamics rather than high-dimensional data. The current work also relates to [19] who studied the information content of data in a history matching setting, but they used subspace pseudo inversion, data coarsening or front extraction, which is very different from the automatic batch window tuning. Notably, the approach used here is clearly different from any kind of partitioning done on the ensemble space [24]. A general issue with ensemble-based methods is rank deficiency of estimated covariance matrices, which is often addressed using localization and/or inflation. Neither of these techniques are considered here, as the focus is solely on the data dimension aspect and how to select this appropriately.

Analysis of the iterative update as a vector in the Hessian eigenbasis has been considered by others, e.g. [27] that used it to guide the choice of a Levenberg-Marquardt regularization parameter. The focus in this work is also on the update vector whereas the angle of analysis here is on how it is made up of contributions from prior and likelihood, and how these change when the amount of data increases. The prior is a regularizing component and maintaining enough of its influence is the aimed balance.

An example with synthetic seismic data is used throughout the paper to give intuition around concepts and methods. Albeit in a synthetic setting, the challenges addressed are realistic and so is the described solution.

The paper is structured as follows: In Section 2 the main building blocks for sequential seismic inversion are introduced. In Section 3 the IEnKS method is outlined. In Section 4 two alternative methods for adaptive batch size selection are presented. In Section 5 one of the methods are applied to examplify the seismic inversion problem. In Section 6 a discussion and analysis of the two methods is provided.

2 Seismic inversion by sequential data assimilation

The seismic inverse problem is that of inferring the subsurface properties from measured seismic reflection data, in the light of a physical model predicting the seismic experiment. The inverse problem is ill-posed and can have multiple solutions, as non-unique subsurface properties can result in a nearly identical seismic response in a smaller time frame. This poses problems to seismic inversion. After describing the Bayesian approach to inversion, a sequential method is presented as an important building block in the suggested ensemble-based solution.

2.1 Seismic waveform model and Bayesian inversion

The elastic properties sought inferred are acoustic wave velocity v_p, shear wave velocity v_s and density ρ. Common midpoint (CMP) seismic gathers are considered here for inferring these elastic properties. Such gathers represent partly processed waveform data, obtained by stacking shot-receiver data to a common mid-point location along the seismic acquisition line and sorted in the time-offset domain [26]. Assuming a subsurface consisting of layers, such data can be simulated by a seismic forward model that maps a depth profile of layers with associated elastic properties, to reflection seismograms at offset points from the source. A commonly used forward model is the reflectivity method [16]. Under the layered subsurface assumption, the elastic wave equation can be transformed and solved in the slowness-frequency domain, and mapped to time domain seismograms via (inverse) Fourier transformation. The full recording time of the gather is thus calculated at once. Several implementations of the reflectivity method exist, the one used here is ERZSOL3 [15]. While the reflectivity method is quite fast compared to other numerical methods for elastic wave propagation, it is still time consuming and the number of simulations as part of solving the inverse problem is a limiting factor.

The seismic gather that will provide the example in this paper is shown in Fig. 1. Here, the seismic CMP data are semi-synthetic in the sense that processed data from a well log of elastic measurements have been forward propagated

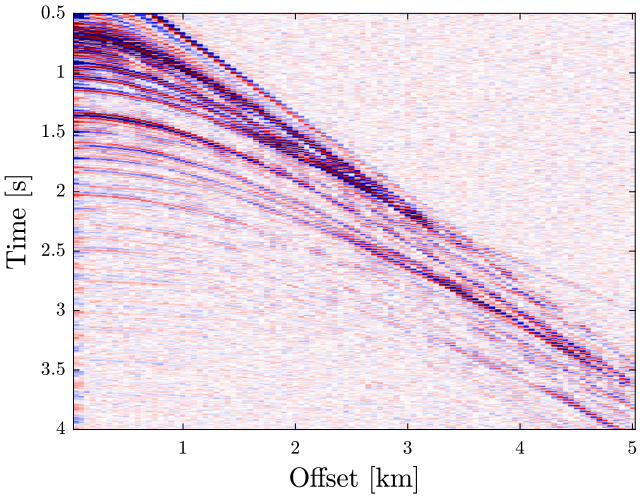

Fig. 1 Example of seismic CMP gather data

with the same forward model. The strongest reflections in Fig. 1 represent major gradients in the elastic properties at shallower layers. These reflections appear as hyperbolic lines in the time-offset plot as the seismic waves take longer time to reach far offsets. At early arrival times and far offsets, the gather contains no records of a reflected wave, and such data are typically zeroed out (muted). In this work, such data points which includes the direct source-to-receiver wave propagation, are excluded and referred to as a mute region. The seismic gather record has a set end time after which measurements are no longer used for the inversion.

Because the acoustic wave travels faster than the shear wave, the earlier parts of the data are dominated by the acoustic wave velocity differences between the subsurface layers, while the reflection from differences in shear properties arrive later in the data and with smaller amplitude. The seismic reflection data indicate the changes in products of elastic properties. For instance, the acoustic impedance is defined as the product of acoustic wave velocity and density. It is hence difficult to split a reflection event in a causal underlying change in lower wave velocity and higher density, or vice versa. The data sensitivities to density perturbations are somehow masked behind the sensitivity to wave velocities, and these must be known quite accurately before one can target density. These ambiguities are smaller with plenty of offset information available in time-offset plots as in Fig. 1, but it is still difficult to infer elastic properties from seismic CMP data.

Framed as a Bayesian inverse problem, the random variable of interest is the parameter state vector $\mathbf{x}$, with an assigned prior probability density function $p(\mathbf{x})$. This state consists of elastic properties in l homogeneous layers of the subsurface, and the prior distribution is represented by a multivariate Gaussian process with mean and covariance terms specified from initial knowledge. To ensure valid physical values of the elastic attributes, positivity of these is enforced by the state vector being the log-transform of elastic properties $\mathbf{x} = \left[\log \mathbf{v}_p, \log \mathbf{v}_s, \log \boldsymbol{\rho}\right] \in \mathbb{R}^m$. The parameter dimension m is three times the number of layers l, and the number of layers and their thickness depth profiles are held fixed.

Data are denoted $\mathbf{y}$ and are measurements of the reflected wave amplitude as function of, besides the subsurface model, arrival time and offset relative to source position, as well as the source signal and boundary conditions. The data are linked to the state via the forward model $h(\mathbf{x})$, which represents the elastic wave propagation as simulated by the reflectivity method. The observation model for data $\mathbf{y}$ is assumed to be unbiased (perfect model assumption) and with an additive noise component $\mathbf{y} = h(\mathbf{x}) + \mathbf{e}$, where $\mathbf{e}$ is an independent zero-mean Gaussian measurement noise vector with covariance matrix $\mathbf{R}$. The resulting likelihood is $p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y};\, h(\mathbf{x}),\, \mathbf{R})$. The solution to the probabilistic inverse problem is then, from Bayes' rule, the posterior probability density function $p(\mathbf{x}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{x})\, p(\mathbf{x})$. Due to the non-linear relationship between parameter state and observations, the posterior distribution is not directly available.

The non-uniqueness of the inverse problem means that the posterior distribution principally can be multimodal at certain depth regions. The ensemble Kalman method as such is incapable of updating into a multimodal posterior as the ensemble is updated using a common gradient. Thus the resulting posterior ensemble will converge to either of the local modes.

2.2 Sequential data integration

The seismic gather can be split in disjoint subsets or batches $\mathbf{y}_k, k = 1, \ldots, K$ such that $\mathbf{y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_K\}$. The partitioning of data into batches is expressed as different arrival time windows, which are referred to as partitions, windows, batches, or batch windows. Partition k of data is extracted by selecting suitable elements of the forward operator for the data, denoted by $h_k(\mathbf{x})$. Assuming conditionally independent measurement noise terms, given the state vector, the likelihood function can also be partioned as $p(\mathbf{y}|\mathbf{x}) = \prod_k p(\mathbf{y}_k|\mathbf{x})$, with kth batch likelihood $p(\mathbf{y}_k|\mathbf{x}) = \mathcal{N}(\mathbf{y}_k;\, h_k(\mathbf{x}),\, \mathbf{R}_k)$ and $\mathbf{R}_k$ is the covariance matrix of the observation error in the corresponding window. When data is partitioned in such arrival time windows, by the nature of the reflection signal, it confines the influence region to a limited depth range and thereby regularizes the problem of estimating the parameters. This influence region is here referred to as the observed depth region.

Data are assimilated sequentially over these disjoint partitions. At the first assimilation cycle, the prior forms the *forecast* model, which is updated using data $\mathbf{y}_1$ in the first batch. The *analysis* model from the first cycle is then $p(\mathbf{x}|\mathbf{y}_1)$. This procedure of going from a forecast model to an analysis model continues at the subsequent cycles. Using Bayes' rule;

$$p(\mathbf{x}|\mathbf{y}_1,\ldots,\mathbf{y}_k) \propto p(\mathbf{y}_k|\mathbf{x})\, p(\mathbf{x}|\mathbf{y}_1,\ldots,\mathbf{y}_{k-1}) \ , \tag{1}$$

repeatedly for $k = 2,\ldots,K$, and at cycle K all data has been assimilated. The main contribution of this paper is to robustly scale the size of a batch window when initiating an assimilation cycle. With this focus, the cycle index k is ignored in the following where the method is outlined for one assimilation cycle only.

3 Iterative Ensemble Kalman Smoother

This section introduces the IEnKS and its components, of which some are fundamental for the adaptive batch window selection. First, its ensemble aspect is presented followed by outlining the iterative solution to the variational problem and the stopping criteria of this iterative scheme. Finally, the method is put in the current context of elastic seismic waveform inversion.

The IEnKS shares the feature of other iterative ensemble smoothers (see e.g. [9]) of using the (negative) log-posterior $-\log p(\mathbf{x}|\mathbf{y})$ as target for minimization. The single state that minimizes such an objective function corresponds to the maximum a posteriori solution and this is used in the IEnKS as the estimate for the ensemble mean.

The method can be regarded as an iterative version of the Ensemble Transform Kalman Filter (ETKF, [2, 14]), a deterministic square-root filter whose formulation makes it efficient in high-dimensional observation spaces with respect to the required matrix inversion. The approach being deterministic means that it avoids random perturbations of observation as is used in stochastic versions of ensemble filters.

3.1 Ensemble-based Data Assimilation

The density functions in Eq. 1 are approximated by ensembles of realizations from these distributions, and their moments approximated by sample moments. In an assimilation cycle the forecast ensemble is input while the output is an analysis ensemble. For the static parameter estimation problem considered here, the analysis ensemble then forms the forecast for the next cycle, and this continues until all data batches are processed.

The members of a forecast ensemble denoted $\mathbf{x}_i^{\mathrm{f}}$, $i = 1,\ldots,n$, where n is the ensemble size, are collected as columns in the $m \times n$ forecast ensemble matrix $\mathbf{E}^{\mathrm{f}}$. The state estimate is the ensemble mean:

$$\overline{\mathbf{x}}^{\mathrm{f}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i^{\mathrm{f}} \ . \tag{2}$$

The second moment is defined via the $m \times n$ normalized state anomaly matrix

$$\mathbf{X}_{\mathrm{f}} = \left(\mathbf{E}^{\mathrm{f}} - \overline{\mathbf{x}}^{\mathrm{f}}\mathbf{1}^{\mathrm{T}}\right)/(n-1)^{1/2} \ , \tag{3}$$

which is the square root of the error covariance estimate

$$\mathbf{P}_{\mathrm{f}} = \mathbf{X}_{\mathrm{f}}\mathbf{X}_{\mathrm{f}}^{\mathrm{T}} \ . \tag{4}$$

Turning (3) around, the forecast ensemble matrix assembles the mean and anomalies as

$$\mathbf{E}^{\mathrm{f}} = \overline{\mathbf{x}}^{\mathrm{f}}\mathbf{1}^{\mathrm{T}} + (n-1)^{1/2}\,\mathbf{X}_{\mathrm{f}} \ . \tag{5}$$

Conditioned to data $\mathbf{y}$, an assimilation cycle updates the forecast ensemble to an analysis ensemble, described by the analysis mean $\mathbf{x}^{\mathrm{a}}$ and analysis anomaly matrix $\mathbf{X}_{\mathrm{a}}$.

The analysis state is found as a linear combination $\mathbf{x}^{\mathrm{a}} \in \left\{\overline{\mathbf{x}}^{\mathrm{f}} + \mathbf{X}_{\mathrm{f}}\mathbf{w} \,|\, \mathbf{w} \in \mathbb{R}^n\right\}$ within the span of the ensemble anomalies $\mathbf{X}_{\mathrm{f}}$, referred to as the ensemble subspace. With this parameterization of the analysis state, the control vector $\mathbf{w}$ replaces the state vector (the elastic parameters) as the variable of interest. The state that maximizes the posterior distribution is equivalent to the control vector that minimizes the negative log-posterior expressed in terms of this subspace reparameterization. This is the variational aspect of IEnKS. Hence, the change of variable $\mathbf{x} \to \mathbf{w}$ induces a change of the log-prior term $\|\mathbf{x} - \mathbf{x}^{\mathrm{f}}\|^2_{\mathbf{P}_{\mathrm{f}}} \to \|\mathbf{w}\|^2_{\mathbf{I}}$, with the notation $\|\mathbf{a}\|^2_{\mathbf{B}} = \mathbf{a}^{\mathrm{T}}\mathbf{B}^{-1}\mathbf{a}$. The analysis mean is then $\mathbf{x}^{\mathrm{a}} = \overline{\mathbf{x}}^{\mathrm{f}} + \mathbf{X}_{\mathrm{f}}\mathbf{w}^{\mathrm{a}}$ with the optimal weight vector being the solution $\mathbf{w}^{\mathrm{a}} = \arg\min_{\mathbf{w}} J(\mathbf{w})$, where the objective function is given as

$$J(\mathbf{w}) = \tfrac{1}{2}\|\mathbf{y} - h\big(\overline{\mathbf{x}}^{\mathrm{f}} + \mathbf{X}_{\mathrm{f}}\mathbf{w}\big)\|^2_{\mathbf{R}} + \tfrac{1}{2}\|\mathbf{w}\|^2 \ . \tag{6}$$

As in the ETKF, the analysis anomaly matrix is updated using an ensemble transform matrix $\mathbf{T}$ such that the square root update is $\mathbf{X}_{\mathrm{a}} = \mathbf{X}_{\mathrm{f}}\mathbf{T}$. The analysis covariance can be expressed as $\mathbf{X}_{\mathrm{a}}\mathbf{X}_{\mathrm{a}}^{\mathrm{T}} = \mathbf{X}_{\mathrm{f}}(\mathbb{H}|_{\mathbf{w}^{\mathrm{a}}})^{-1}\mathbf{X}_{\mathrm{f}}^{\mathrm{T}}$ where $\mathbb{H}|_{\mathbf{w}^{\mathrm{a}}}$ is the Hessian of the objective function (6) evaluated at the optimum [14]. Therefore the transform matrix $\mathbf{T} = (\mathbb{H}|_{\mathbf{w}^{\mathrm{a}}})^{-1/2}$ provides the analysis update for the covariance square root. With the analysis mean $\mathbf{x}^{\mathrm{a}}$ and anomalies $\mathbf{X}_{\mathrm{a}}$ in place, the analysis ensemble is assembled similarly to Eq. 5 and thereby closes an assimilation cycle.

3.2 Iterative procedure

Letting index j indicate iteration number, the variational problem of minimizing $J(\mathbf{w})$ is solved iteratively as $\mathbf{w}_{j+1} = \mathbf{w}_j + \Delta\mathbf{w}_j$, where the search direction is taken as the

Gauss-Newton update $\Delta \mathbf{w}_j = -\mathbb{H}_j^{-1} \nabla J_j$. This involves the $n \times 1$ gradient (Jacobian) ∇J and the $n \times n$ (approximative) Hessian $\mathbb{H}$, of the objective function:

$$\nabla J_j = \mathbf{w}_j - \mathbf{Y}_j^{\mathrm{T}} \mathbf{R}^{-1} \left(\mathbf{y} - \bar{\mathbf{y}}_j \right) , \tag{7a}$$

$$\mathbb{H}_j = \mathbf{I}_n + \mathbf{Y}_j^{\mathrm{T}} \mathbf{R}^{-1} \mathbf{Y}_j . \tag{7b}$$

The gradient and Hessian calculation notably involves ensemble evaluations only.

The value of iterating comes from a reevaluation of the forward model gradient $\nabla_{\mathbf{x}} h|_{\mathbf{x}_j}$ around the iteratively improved mean state $\mathbf{x}_j = \bar{\mathbf{x}}^{\mathrm{f}} + \mathbf{X}_{\mathrm{f}} \mathbf{w}_j$. Ensemble based assimilation does not require this model gradient explicitly, instead the prior observation anomalies $\mathbf{Y}_j$ are assumed to be the image of the prior state anomalies, mapped through an iteratively reevaluated model gradient $\mathbf{Y}_j = (\nabla_{\mathbf{x}} h|_{\mathbf{x}_j}) \mathbf{X}_{\mathrm{f}}$. This reevaluation is attained using the IEnKS transform variant [25]. Therein an intermediate ensemble is used to evaluate the observation anomalies, where this ensemble is preconditioned to data in the sense that it uses the currently available mean $\mathbf{x}_j$ and transform matrix $\mathbf{T}_j = \mathbb{H}_{j-1}^{-1/2}$. The iterative ensemble

$$\mathbf{E}_j = \mathbf{x}_j \mathbf{1}^{\mathrm{T}} + (n-1)^{1/2} \mathbf{X}_{\mathrm{f}} \mathbf{T}_j , \tag{8}$$

is consequently used to evaluate the $p \times n$ observation anomaly matrix

$$\mathbf{Y} = \left(h(\mathbf{E}_j) - \bar{\mathbf{y}}_j \mathbf{1}^{\mathrm{T}} \right) / (n-1)^{1/2} , \tag{9}$$

with $\bar{\mathbf{y}}_j = h(\mathbf{E}_j)\mathbf{1}/n$ being the mean of the observation ensemble. The observation anomalies (9) are conditional to the iterative ensemble $\mathbf{E}_j$, but should relate to the prior $\mathbf{E}^{\mathrm{f}}$. For this reason they are untransformed (or "deconditioned", [21]) as $\mathbf{Y}_j = \mathbf{Y}\mathbf{T}_j^{-1}$, before being used in the sensitivities (7).

The iterations are initialized with $\mathbf{w}_0 = \mathbf{0}$ and $\mathbf{T}_0 = \mathbf{I}$, so that $\mathbf{E}_0 = \mathbf{E}^{\mathrm{f}}$ and the prior distribution on $\mathbf{w}$ is a standard normal distribution. This will be utilized later.

The matrix power operations applied to the Hessian are facilitated when this matrix is factorized in a eigen-decomposition. In this work, the singular value decomposition (SVD) is applied to the $p \times n$ ensemble of standardized observation anomalies $\mathbf{R}^{-1/2}\mathbf{Y}$, and with $p \gg n$ generally being the case, 'economic' SVD offers significant computational savings. Ignoring the subscript j, the decomposition is

$$\mathbf{R}^{-1/2}\mathbf{Y} = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathrm{T}} = \sum_{i=1}^{n} \lambda_i \mathbf{u}_i \mathbf{v}_i^{\mathrm{T}} , \tag{10}$$

where the $p \times n$ matrix $\mathbf{U}$ has left singular vector $\mathbf{u}_i$ as ith column, and correspondingly for the $n \times n$ matrix of right singular vectors $\mathbf{V}$. The $n \times n$ diagonal matrix $\mathbf{S}$ holds the sorted singular values $(\mathbf{S})_{ii} = \lambda_i$, $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_n \geq 0$. Equation 10 involves the inverse square root of the error covariance matrix $\mathbf{R}^{-1/2}$, which is straightforward to compute when $\mathbf{R}$ is a diagonal matrix. Inserting this decomposition into Eq. 7b, and using the orthogonal properties $\mathbf{V}\mathbf{V}^{\mathrm{T}} = \mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}$, which holds when $p > n$, the Hessian becomes

$$\mathbb{H} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{\mathrm{T}} \text{ with } \mathbf{\Lambda} = \mathbf{I} + \mathbf{S}^{\mathrm{T}}\mathbf{S} .$$

Here, the diagonal matrix $\mathbf{\Lambda}$ has elements $(\mathbf{\Lambda})_{ii} = (1 + \lambda_i^2)$. The inverse and square root of the Hessian are then obtained from the corresponding operations on the diagonal matrix $\mathbf{\Lambda}$:

$$\mathbb{H}^{-1} = \mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{V}^{\mathrm{T}} \text{ and } \mathbb{H}^{-1/2} = \mathbf{V}\mathbf{\Lambda}^{-1/2}\mathbf{V}^{\mathrm{T}} ,$$

which are used in the search direction and as the transform matrix, respectively.

3.3 Stopping criteria

A termination criteria is needed to stop the iteration process when continued iteration does not improve the solution significantly. Such a criteria is most often expressed as an absolute or relative change in either objective function evaluation or some norm of the control variable, falling below a given tolerance level. In the considered application, it was found challenging to set an appropriate tolerance level for any commonly used measure (on $J(\mathbf{w}_j)$ or $\mathbf{w}_j$) that resulted in consistent termination across varying data dimension, signal-to-noise ratios and ensemble sizes.

The reflection data assimilated in an analysis cycle is related to a (local in depth) observation region. The optimal $\mathbf{w}^{\mathrm{a}}$ controls the full depth analysis mean and is supposed to form the estimate of the elastic parameters in this observed region, while keeping the prior mean more or less unchanged outside this region. The scale of the cost function is dominated by the data misfit term and is insensitive to (smaller) adjustments in $\mathbf{w}$ that acts on the state mean $\mathbf{x}$ outside the observed region. Thus basing termination on changes in $J(\mathbf{w}_j)$ does not necessarily express that a steady global mean has been reached. Norms on either $\mathbf{w}_j$ or $\Delta\mathbf{w}_j$ are very dependent on the time window span and on the "distance" between forecast and analysis mean. Suitable threshold on changes in these would vary for different depths of observed region, making them difficult to set beforehand.

One measure was seen to have a consistent behavior across the data batch window lengths and position within the gather and across different ensemble sizes. Importantly, the measure might have different scale but behaved similarly when it seemingly was a good time to stop iterating, i.e. when $\|\mathbf{w}_j\|$ or $\|\Delta\mathbf{w}_j\|$ had reached stationary levels. This measure is the mutual information (or Shannon information content), originating from information theory but also used within data assimilation [11, 23], which addresses the

reduction in entropy/improvement of knowledge. It can be evaluated from the eigenvalues of the $n \times n$ matrix $\mathbf{Y}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{Y}$, referred to as the information matrix in ensemble subspace [29], as

$$\mathrm{MI} = \frac{1}{2}\sum_{i=1}^{n} \log\left(1+\lambda_i^2\right) , \tag{11}$$

with λ_i the singular values of Eq. 10. This quantity decreases during (converging) iterations, flattens out and eventually increase slightly. The point where the measure reaches a minimum level is associated with stationarity in that the eigenvalues λ_i^2 do not change, meaning that $\mathbf{Y}_j$ and thus $\mathbf{T}_j$ do not change either. Hence the stopping rule was formulated as when $\mathrm{MI}_j > \mathrm{MI}_{j-1}$, iterations are terminated. This stopping criterion is complemented by a maximum allowed number of iterations.

3.4 Elastic inversion with IEnKS

Estimation of the elastic parameters is complicated by the different sensitivity of the reflection data to the different elastic properties. The control variable $\mathbf{w}$ determines the full depth mean of all three elastic parameters. So the focusing of its effect onto only the observed region relates to how well the cross-covariances between the parameter state and the seismic data is resolved.

The iterative ensemble (8) represents a sequence of ensembles going from forecast to analysis, where the sequence reflects the gradual change in parameter estimation and its uncertainty over iterations. The gradual change in the ensemble used to evaluate the sensitivities (7) is important. The implicit ensemble approximation of the tangent linear model again depends on the cross-covariance between each elastic parameter and the seismic waveform being well estimated. Within a batch window of seismic data, the variability in acoustic velocity affects the variability in seismic waveforms more strongly than shear velocity and density. So the preconditioning to data in the iterative ensemble gradually accounts for the stronger reflections from shallower layers, reduces their effect and enhances the sensitivity to shear velocity and density, as well as the variables in layers further down in the observed region. This is the reason for partitioning data into arrival time windows and sequentially processing these in an ordered manner. The parameters and their variability at shallower depths must be accounted for before the ensemble smoother can estimate the sensitivity to parameters at deeper layers in the seismic waveforms.

The sample estimate of cross-covariances between data and density perturbations in particular is more susceptible to rank-deficiency issues ("spurious correlations"). This occurs when either the ensemble size is too small, the data batch size too large or the observation errors are large. In these situations, the update to the density profile has a higher risk of divergence in the sense that the ensemble no longer represents the true error statistics. As the density couples with the velocities in the reflection/transmission coefficients, a diverging density estimate during iterations also has a negative influence on the estimation of velocities. Moreover, a divergent solution in the observed region leads to problems for all parameters at larger depths because the parameters above are incorrect.

4 Selection of batch window

The amount of data included in an analysis step can be a challenge for ensemble-based methods. In the application with seismic gathers, the massive waveform data available must be assimilated sensibly to avoid problems: first, the ensemble linearization becomes a limiting assumption when a large time span of data are integrated, and the iterative procedure might not converge. Second, a large dataset can lead to over-fitting as in underestimation of the state uncertainty.

When performing the kth analysis cycle, the data partition expressed through h_k must be known, so determining this is the first step when entering a cycle. The focus is on choosing a batch window that results in a stable initial iteration as this is fundamental for later convergence.

In what follows, a method for automatic selection of the batch window is presented. A window size is considered acceptable if a criteria is fulfilled, and the window expanded until this acceptance condition is broken. The acceptance criteria is formulated based on an analysis of the initial update, and two alternatives are presented, each with a different angle of interpretation.

4.1 Spectral interpretation

Inserting the singular value decomposition (10) into the gradient and Hessian expressions (7) along with the normalized innovations $\Delta\tilde{\mathbf{y}} = \mathbf{R}^{-1/2}(\mathbf{y}-\bar{\mathbf{y}})$, the Gauss-Newton update direction can be written as

$$\begin{aligned}\Delta\mathbf{w} &= -\mathbf{V}\boldsymbol{\Lambda}^{-1}\mathbf{V}^{\mathrm{T}}\left(\mathbf{w}-\mathbf{V}\boldsymbol{\Sigma}^{\mathrm{T}}\mathbf{U}^{\mathrm{T}}\Delta\tilde{\mathbf{y}}\right)\\ &= \sum_{i=1}^{n}\mathbf{v}_i\left(\frac{-\mathbf{v}_i^{\mathrm{T}}\mathbf{w}}{1+\lambda_i^2}+\frac{\lambda_i(\mathbf{u}_i^{\mathrm{T}}\Delta\tilde{\mathbf{y}})}{1+\lambda_i^2}\right).\end{aligned} \tag{12}$$

The update direction is evidently within the span of right singular vectors $\mathbf{V}$. The coefficient of each basis vector $\mathbf{v}_i$ is composed of (projected) contributions from the prior (via $\mathbf{w}$) and from the likelihood (via $\Delta\tilde{\mathbf{y}}$). The update vector

can be split into the vector components $\Delta\mathbf{w}_x = \mathbf{Va}$ and $\Delta\mathbf{w}_y = \mathbf{Vb}$, with ith coefficients

$$\mathbf{a}_i = -(\mathbf{v}_i^{\mathrm{T}}\mathbf{w})(1+\lambda_i^2)^{-1} \text{ and } \mathbf{b}_i = \lambda_i(\mathbf{u}_i^{\mathrm{T}}\Delta\tilde{\mathbf{y}})(1+\lambda_i^2)^{-1}\,.$$

The control vector is initialized with $\mathbf{w}_0 = \mathbf{0}$ hence the first update $\mathbf{w}_1 = \Delta\mathbf{w}_0$ is solely determined by the likelihood contribution with basis coefficients $\mathbf{b}$.

Each of the vector components consists of projection coefficients with a weighting factor derived from the eigenvalues of the information matrix $\mathbf{Y}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{Y}$. Figure 2 shows the eigenspectra and weighting curves as a function of an increasing window span. They are obtained by using a sequence of time windows with identical start and increasing length, and provides a useful visualization to the argumentation. The distribution in Fig. 2a is the information matrix eigenspectrum λ_i^2 from which eigenvalues of the Hessian $1+\lambda_i^2$, inverse Hessian $\left(1+\lambda_i^2\right)^{-1}$ and the transform matrix eigenvalues $\left(1+\lambda_i^2\right)^{-1/2}$ are derived (Fig. 2b). Their eigenbasis are the right singular vectors, as explained in Section 2. Figure 2c shows the weighting curve that enters the likelihood component coefficient $\mathbf{b}$.

While the specific shape of the λ_i^2 distribution and hence also of the derived spectra will depend on the prior ensemble characteristics and configuration of the forward model, the general behavior is well represented by these displays.

Referring to the information matrix, [29] divide its eigenspectrum into a region of signal, where $\lambda_i^2 \geq 1$ and the forecast errors are larger than observational noise, and a noise region with $\lambda_i^2 < 1$ and the forecast errors being smaller than the noise. According to this, the weighting curves $(1+\lambda_i^2)^{-1}$ and $\lambda_i(1+\lambda_i^2)^{-1}$ to the left and right of $\lambda_i^2 = 1$, have relation to the influence from observation and from prior. In the contribution from the prior to the basis coefficient $\mathbf{a}_i + \mathbf{b}_i$, this division into likelihood and prior influences corresponds to $(1+\lambda_i^2)^{-1} \leq 1/2$ and $> 1/2$ respectively. This interpretation of influence regions also applies to the transform matrix with eigenvalues $(1+\lambda_i^2)^{-1/2}$ (Fig. 2b), where components with eigenvalues approaching 1 contributes to preserving ensemble spread. These eigenvalues are associated with states where observations are non-informative and the prior should dominate. The opposite occurs for components approaching 0, which are related to the data conditioning, providing information for improved estimation and reduced uncertainty. For the likelihood contribution, this division into influence regions corresponds to either side of the "center" point with peak value of 1/2.

When the amount of data increases, the prior contribution to the update $\Delta\mathbf{w}$ is reduced as more and more of its projection coefficients $(\mathbf{v}_i^{\mathrm{T}}\mathbf{w})$ are weighted with values approaching zero (Fig. 2b). In the likelihood contribution, the weighting of the coefficients $(\mathbf{u}_i^{\mathrm{T}}\Delta\tilde{\mathbf{y}})$ is also shifted but not downweighted similarly, and generally increases the likelihood contribution to the update. The region of peak weighting moves to higher indices and amplifies projections onto singular components $\mathbf{u}_i$ with more high-frequency content, relative to components with lower indices. Basing the search direction $\Delta\mathbf{w}_0$ on projections of $\Delta\tilde{\mathbf{y}}$ onto higher-frequency components can overfit to noise rather than structural eigenbasis components of the observation error covariance. This could render the mean update $\mathbf{x}_1$ highly varying with values that are unacceptable for the forward solver, or so far from the true profile that the error covariances evaluated around this new mean are useless for the linearization.

The singular value index i for which $\lambda_i^2 \simeq 1$ is close in value to another information theoretic measure referred to as the degree of freedom for signal d_{s}. This measure can be viewed as the influence of observations to the analysis or entropy reduction [4, 18]. The ensemble subspace version of this quantity can be expressed as $d_{\mathrm{s}} = \sum_{i=1}^{n} \lambda_i^2 \left(1+\lambda_i^2\right)^{-1}$ [29]. Both the measure d_{s} and the particular singular value index can thus be thought of

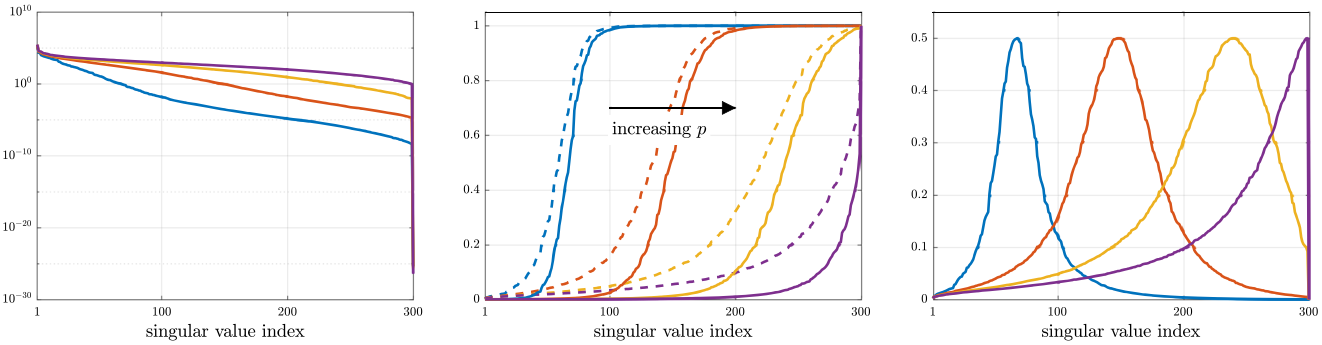

Fig. 2 Eigenvalue distributions as a function of singular values λ_i for; **a** information matrix, **b** inverse Hessian (full) and transform matrix (dashed), **c** weight factor for projected innovations, as a function of increasing data dimension p. Colors are for same p across plots. **a** λ_i^2 **b** $(1+\lambda_i^2)^{-1}$ and $(1+\lambda_i^2)^{-1/2}$ **c** $\lambda_i(1+\lambda_i^2)^{-1}$

as indicating the information content of the observations within a batch, and one can control either to restrain the information content. Approaches to obtaining a balance between ensemble degree of freedom and observation information content was also considered in [19] within the context of history matching. While seismic inversion and the history matching problem do not have the same means available of approaching this balance, the target of stable updates is the same.

Next, the algorithm to select a batch window is presented. This uses an acceptance criteria to evaluate the window with the aim of balancing likelihood influence with prior restrainment. Following this, the two alternative criteria are described.

4.2 Batch selection strategy

The batch selection searches for an end time T_{E} of a time window so that the kth data batch is $h_k : y(t,\cdot),\ t \in [T_{\mathrm{S}}, T_{\mathrm{E}}]_k$. That is, seismic data for arrival times T_{S} to T_{E} for all offsets. The exclusion of data in the mute region is implicit in h_k. The subsequent assimilation cycle then starts at $T_{\mathrm{S},k+1} = T_{\mathrm{E},k} + \Delta t$. Simplicity is sought by using a minimum of user-supplied tuning parameters. The approach involves repeated use of singular value decomposition, but this cost is negligible compared to the benefit of having selected an adequate batch of seismic data that reduces the risk of diverging iterations.

The principle of the algorithm is to keep extending the time window until a criteria is no longer respected, as illustrated by the flowchart in Fig. 3. The algorithm is independent of the particular acceptance criteria. The extension is done in step sizes of ΔT. For efficiency, a larger step size ($\Delta T_{\max}$) for the time window elongation can be used initially. Once the criteria is no longer satisfied, the latest time window increase is reverted, the step size decreased

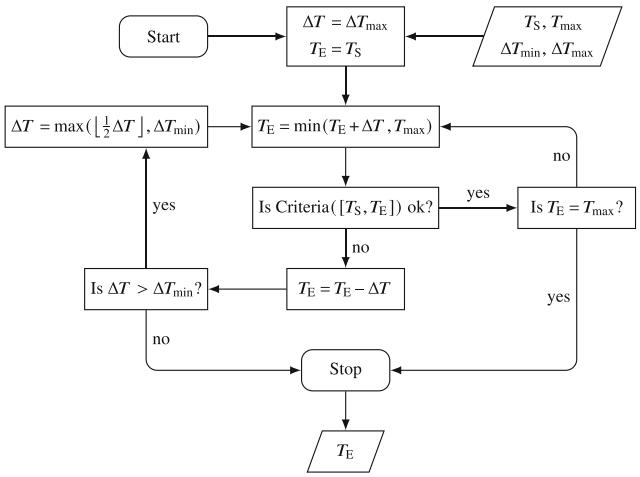

Fig. 3 Flowchart of data batch size selection

and the loop repeats until a minimum step size ($\Delta T_{\min}$) or the end time of the gather $T_{\max}$ is reached.

4.2.1 Weight criteria

This criteria uses the interpretation of the weighting curves and their division into likelihood/prior influence regions. The initial update $\Delta\mathbf{w}_0$ equals $\Delta\mathbf{w}_y$, but by controlling the initial distribution of the weight curves (Fig. 2b and c), the aim is that the initial likelihood contribution is restrained sufficiently to let the prior retain its influence at the following iteration, balancing their contributions.

As explained, the point $\lambda^2 = 1$ has a central interpretation and its location is useful for controlling the weight distributions. The relative location of this point is therefore taken as the key variable in the window acceptance criteria. Setting $i_{\mathrm{C}} = \max\left\{i \,|\, \lambda_i^2 \geq 1;\ i = 1,\ldots,n\right\}$ the weight criteria is expressed as

$$\frac{n - i_{\mathrm{C}}}{i_{\mathrm{C}}} \geq \beta\,, \tag{13}$$

where $\beta = 1$ means $i_{\mathrm{C}} \approx n/2$. The ratio parameter β must naturally be positive, and for $\beta > 1$ the window size selection will be smaller than for $\beta = 1$ and vice versa for $\beta < 1$.

4.2.2 Norm criteria

This alternative approach is based on the update direction (12), where the focus of argument is shifted from the weighting curves to the prior and likelihood vector components. With $\Delta\mathbf{w} = \Delta\mathbf{w}_x + \Delta\mathbf{w}_y$ all that is known is that the inequality $\|\Delta\mathbf{w}\| \leq \|\Delta\mathbf{w}_x\| + \|\Delta\mathbf{w}_y\|$ holds, using the standard 2-norm $\|\mathbf{z}\| = \left[\sum_i |z_i|^2\right]^{1/2}$. The second strategy to accept a batch window is based on the criteria

$$\frac{\|\Delta\mathbf{w}_x\|}{\|\Delta\mathbf{w}_y\|} \geq \beta\,, \tag{14}$$

where again β is a preset threshold parameter. This indirectly sets a bound on the ratio $\|\Delta\mathbf{w}\|/\|\Delta\mathbf{w}_y\|$ and consequently restrains the likelihood component contribution to the update vector. The initial control variable $\mathbf{w}_0$ is considered to have a standard normal prior distribution, and a Monte Carlo estimate of a fictitious prior component vector $\widehat{\mathbf{a}}$ can be generated. For a given partition selection with associated set of singular components $\{\lambda_i, \mathbf{u}_i\}$, a large batch of B samples $\mathbf{w}^b \sim \mathcal{N}(0, \mathbf{I})$ is used to form

$$\widehat{\mathbf{a}}_i = \frac{1}{B}\sum_{b=1}^{B}\left|\frac{-\mathbf{u}_i^{\mathrm{T}}\mathbf{w}^b}{1+\lambda_i^2}\right|\,, \tag{15}$$

which subsequently sets $\|\Delta\mathbf{w}_x\| = \|\widehat{\mathbf{a}}\|$ used to evaluate the criteria Eq. 14. The Monte Carlo expectation is taken

on the norm argument instead of the norm due to Jensen's inequality. The absolute value operator of the 2-norm is taken within the estimator (15) as otherwise the expectation would average out to nearly a zero vector.

Figure 4 illustrates a typical development of these norms during a batch window search. At the initial window expansion (as described in Section 4.2), the fictitious prior component norm dominates. Continuing the window expansion (with increasing data dimension p) enlarges the likelihood component norm and eventually this will dominate, consistent with the evolution of the weight curves in Fig. 2b and c. The point where these norms cross in Fig. 4 is the conditional branch where the selection algorithm (Fig. 3) reverts the window increase and decreases the stepsize.

In contrast to the weight curve approach, this approach takes into account the innovations which makes it more sensitive to their contribution to the update. The actual $\Delta \mathbf{w}_x|_{j=1}$ will still be zero but the aim is that at the following iteration, the magnitude of the prior component will be in the vicinity of that of the likelihood component and thereby reducing the risk of observations dominating the update on behalf of the prior constrainment.

5 Numerical example

Results of applying the IEnKS to seismic data inversion are presented next. First, the seismic data acquisition design is outlined, followed by the prior model description, and then results of the adaptive partitioning for sequential inversion. This section uses the norm criteria with a ratio parameter of $\beta = 1$, while results using the weight criteria is discussed in Section 6.1.

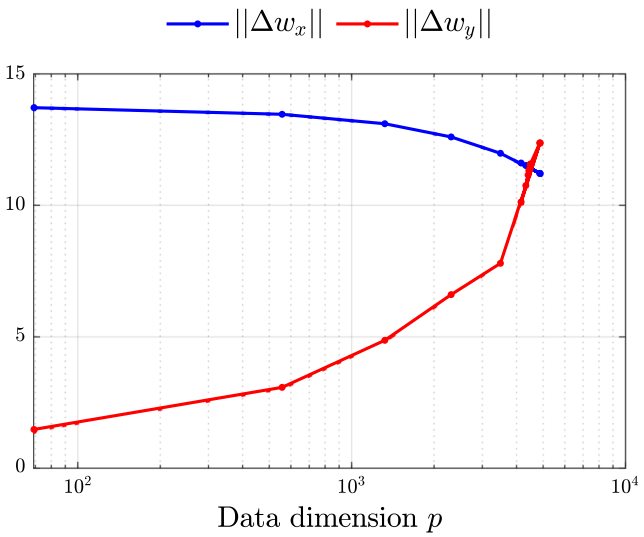

Fig. 4 Evolution of vector components norm during time window search. Prior component $\|\Delta \mathbf{w}_x\|$ in blue, likelihood component $\|\Delta \mathbf{w}_y\|$ in red. Here $\beta = 1$

5.1 Description of setup

The measurement configuration consists of 100 receiver locations, at offsets distributed in the range 50 m to 5 km with a uniform spacing of 50 m. The source is located in the top layer at 5 m below the top surface, which has the boundary condition of a free surface. This top layer of 500 m depth has fixed properties. The source time signal is a fifth order Butterworth wavelet with frequency bandpass 2–50 Hz and time sampling is 2 ms. The seismic traces has a limited frequency bandwidth compared to the source signal, where these are generated with a frequency content 5–32 Hz, with linear in- and out-tapering from 5–7 Hz and 30–32 Hz. The gather data up to 4 s is used for the inversion, excluding data in a mute region defined by normal move-out in the top layer. The total number of data points is $\sim 10^5$.

Data from a processed well log are used as the true subsurface model $\mathbf{m}^{\mathrm{t}}$ and using this as input for the forward model a simulated data set is considered the true seismic CMP gather, with a sample of measurement noise added. A constant noise level $\mathbf{R} = \sigma_{\mathrm{e}}^2 \mathbf{I}$ is used in the measurement model, where the noise variance σ_{e}^2 is set to have a signal-to-noise ratio of 13 dB with respect to a reference signal power. This reference signal power is set as an averaged power in the time interval 1–3 s and offset range 0–3 km of the true seismic gather. As the amplitude of the seismic signal decays with time, this means a very non-uniform signal-to-noise ratio will be present in the data to be assimilated.

Larger ensemble size had a tendency to make the system unstable. The reason is growth of the largest eigenvalues of the information matrix during iterations. This propagates into a corresponding largest eigenvalue of the inverse transform matrix, so that observation anomalies were upscaled unreasonably causing problems for the used SVD routine. This is a numerical issue and was handled by clipping eigenvalues of the transform matrix below a certain threshold which propagates into its inverse. This approach is the same as applied in [25].

5.2 Specification of prior

As the analysis mean is found as a linear combination within the ensemble subspace, the trends and smoothness specified in the prior structure influence the ability to form combinations of sufficient variability to fit the true underlying profile of elastic parameters. Consequently, the prior specification is a cornerstone for a succesful inversion.

The prior ensemble is here specified by samples from a multivariate Gaussian distribution of the log-elastic

parameters. This initial distribution for the ensemble is defined through a mean vector and a covariance matrix. The log-units domain is not the most intuitive domain for prior specification, and instead, linear depth trends for the mean and standard deviation are set in the physical domain of the three elastic parameters. Using the relation between arithmetic moments of normal and log-normal distributions, these trends are mapped into normal-domain mean $\boldsymbol{\mu}_x$ and standard deviation $\boldsymbol{\sigma}_x$. A cross-correlation structure between the three parameters and a spatial correlation must also be specified. Using a separable structure, the final covariance matrix is

$$\operatorname{cov}\mathbf{x} = \boldsymbol{\Sigma}_x = \operatorname{diag}(\boldsymbol{\sigma}_x)\left[\begin{bmatrix} 1 & \eta_{ps} & 0 \\ \eta_{ps} & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \otimes \boldsymbol{\Gamma}\right]\operatorname{diag}(\boldsymbol{\sigma}_x)\ ,$$

where the spatial (depth) correlation structure $\boldsymbol{\Gamma}$ is taken as a Matérn function of order 3/2 with a range parameter such that correlation is 5% at 500 m distance. The cross-correlation between velocities is $\eta_{ps} = 0.5$, and $\otimes$ is the Kronecker product.

Initially, the ensemble consists of n independent samples from this Gaussian model. For the benchmark case presented here, an ensemble size of $n = 300$ is used. In the discussion, ensemble sizes of $n = 150$ and $n = 600$ are also studied for comparison.

The support of the prior ensemble when mapped to the log-normal domain, is presented in Fig. 5. This displays the ensemble median and the span of 90% empirical coverage of the prior ensemble.

5.3 Results

The gather data were limited to the time interval 0.6−4s. By running the algorithm this interval was partitioned into 5 blocks, as displayed in Fig. 6. The time lengths of these windows were 556, 690, 798, 528, and 828 ms, with corresponding number of data points 4247, 13635, 25304, 22125, and 40809. Ensemble evaluation was performed

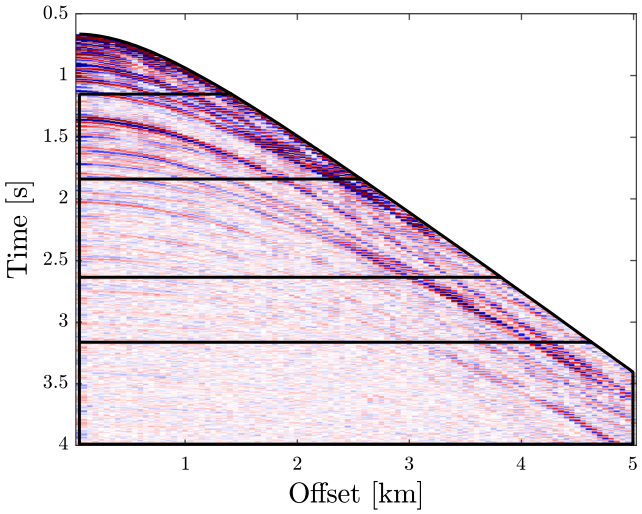

Fig. 6 Partition into batch windows of seismic gather data

in parallel using 20 cores, and computation time for the inversion was around two and a half hours.

The resulting posterior ensemble is displayed in Fig. 5, along with the truth and the prior ensemble. This shows that acoustic and shear velocities are estimated well down to around 4 km depth, whereas density is only estimated well down to 3.5 km. Generally, the density estimate is less accurate than that of velocities, consistent with the expected smaller sensitivity of the waveform data to density perturbations.

The assimilation statistics are presented further in Fig. 7. Here, the estimation bias $\left|\widehat{\mathbf{m}} - \mathbf{m}^{\mathrm{t}}\right|$ is shown. The physical state estimate $\widehat{\mathbf{m}} = \operatorname{median}(\exp(\mathbf{E}^{\mathrm{a}}))$ is seen to correlate well with the (marginal) standard deviation from the analysis ensemble. For the shear velocity in Fig. 7b, estimation results stand out in two areas at shallower depths (at 0.6 km–0.8 km and 1 km–1.2 km depth). In these areas, the bias deviates largely and this can also be seen in the standard deviation where there is a local increase. Both

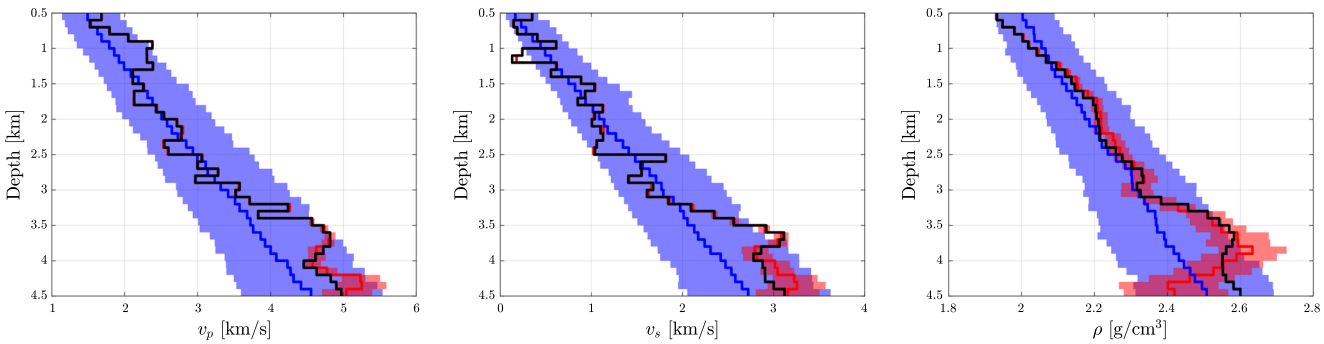

Fig. 5 Ensemble and truth; blue is prior, red is posterior, black is truth. **a** Acoustic velocity **b** Shear velocity **c** Density

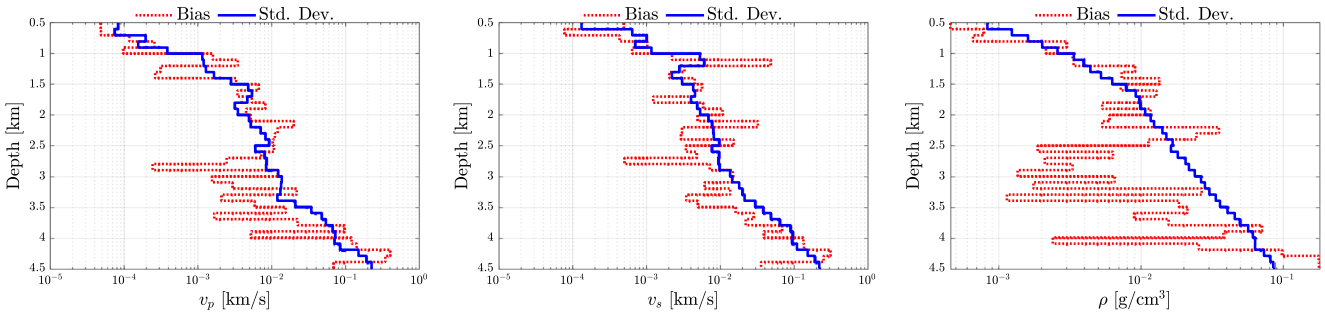

Fig. 7 Estimation statistics. **a** Acoustic velocity **b** Shear velocity **c** Density

cases are associated with very low values of true shear velocity. The estimation is there seen to be more difficult, possibly associated with challenging parts in the forward model. The consistent correlation between the (unknown) estimation bias and the ensemble spread shows that an indicative quantification of the estimation uncertainty can indeed be extracted from the ensemble solution.

Figure 8 shows how the sequential estimation proceeds, by displaying the ensemble (marginal) standard deviation for the three parameter types over the course of assimilation cycles. Each analysis cycle reduces the acoustic velocity uncertainty to slightly larger depths than shear velocity and density. This is expected due to the higher acoustic wave speed. A given window of data contains acoustic wave reflections from deeper layers, and the estimation of acoustic velocity therefore occurs slightly further in depth than for the other two parameters. The sensitivity to the three parameter types is also somehow visible from these plots, where the reduction in uncertainty of acoustic velocity is changing in more focused steps, compared to especially density. For each assimilation cycle, the ensemble spread below the observed region is maintained. This implies that the low uncertainty in Fig. 7 is not a result of ensemble collapse but indeed indicates estimation uncertainty in this synthetic example, as ensemble collapse would affect throughout the full depth and consequently the ensemble spread below an observed region. Evidently, this is not the case.

Figure 9 shows the iteration history of the objective function, the mutual information measure used as stopping criteria, and norms of control vector and its update, for each of the five batches. The objective function consistently reaches a stationary level faster than the other measures. The objective is dominated by the data misfit, and it flattens out when continued iteration does not update the ensemble mean $\mathbf{x}_j$ in the observed region. Still, the changes below the observed region could be substantial, where the analysis mean ideally should not be far from the prior mean. Hence, much of the later effort of iterating does not contribute to reduce the data misfit, but rather to focusing the analysis update to the relevant parameters in the observed region. This is reflected in the continued change in magnitude of $\|\mathbf{w}_j\|$ and its update $\|\Delta\mathbf{w}_j\|$, which continues long after a stationary level of data misfit is observed. The MI measure also reaches a stationary level later than the objective function, but the (iteration) onset of this flattening out

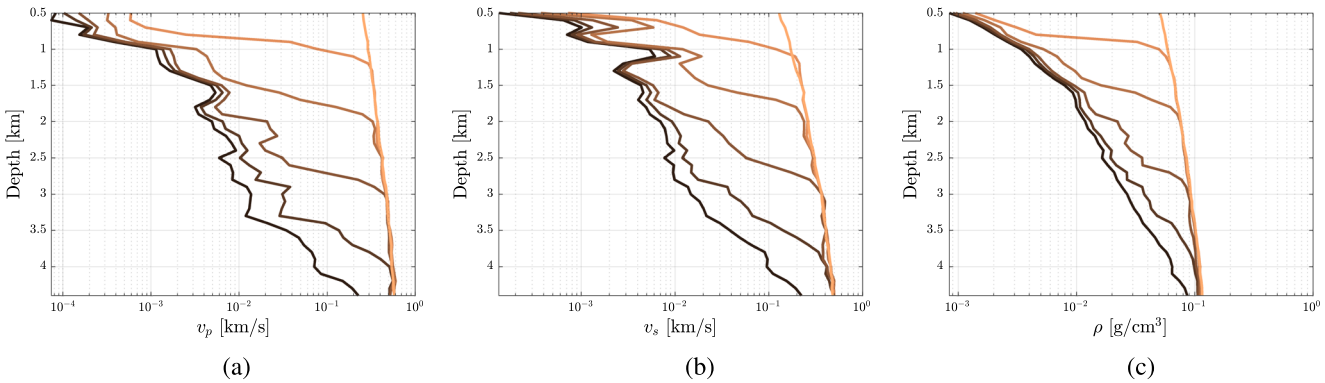

Fig. 8 Ensemble standard deviation over analysis cycles. Order is from lightest (initial ensemble) to darkest (final analysis). **a** Acoustic velocity **b** Shear velocity **c** Density

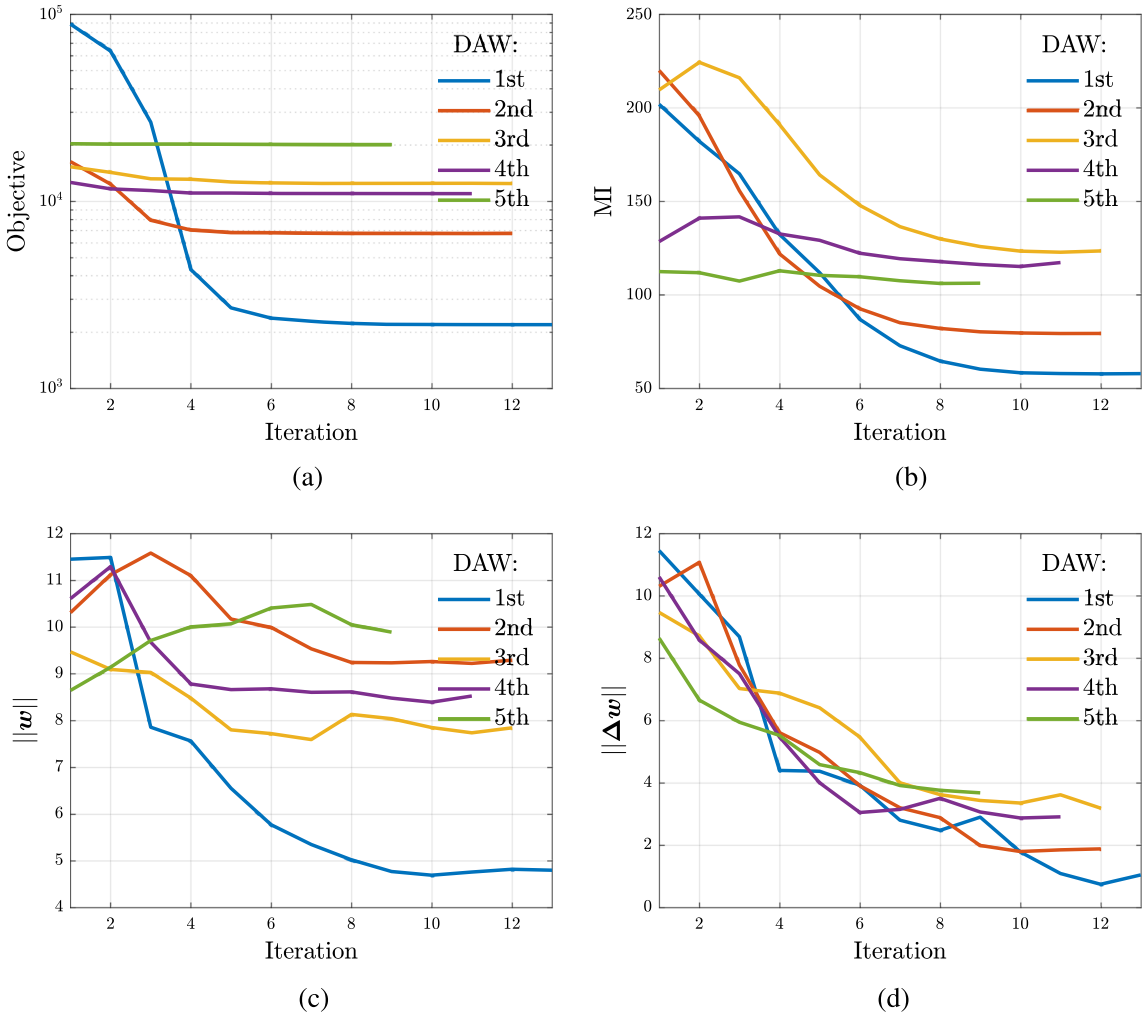

Fig. 9 Iterative history of (**a**) objective function, (**b**) mutual information and Euclidean norm of (**c**) **w** and (**d**) $\Delta\mathbf{w}$, over the 5 batch windows. **a** $J(\mathrm{w}_j)$ **b** MI_j **c** $\|\mathbf{w}_j\|$ **d** $\|\Delta\mathbf{w}_j\|$

correlates much better with $\|\mathbf{w}_j\|$ converging, than with data misfit. Which support the choice of using this measure as stopping criterion. Not shown here is the iterative evolution of the measure of degree of freedom for signal measure d_s (Section 4.1), which is very similar to that of the mutual information.

Another apparent feature is the large dependency of relative difference between initial and final level of objective value, on the overall signal-to-noise ratio within the data window. Seismograms will always have decaying amplitude with traveltime and the measurement error might not decay in a similar manner, so this issue will generally be present. In Fig. 9a the first batch window has much larger difference than the others as measurement noise is relatively low compared to signal amplitude. In contrast, the data of the fifth assimilation cycle is masked by noise to a degree that the data misfit has very little reduction, but nevertheless contributes to better estimation of parameters in the 3.5–4 km depth range.

6 Discussion

The challenge of non-unique solutions to the inverse problem are discussed first, continuing the numerical example using the weight rather than the norm criteria. Then a replicate study is presented for both the weight and norm criteria, comparing stability, batch sizes and the number of forward evaluations. The replicate study is performed to evaluate the strategies in the light of a reduced dependency on the particular initial ensemble. Stability, or rather divergence, is considered in the sense of inessential/spurious updates where the estimation fails due the ensemble no longer representing the true error or

forward model crashing due non-physical input. Divergence as in ensemble collapse was not encountered in any of the repeated trials as the batch size was never increased to a degree where collapse occurs.

6.1 Challenge of non-uniqueness

The example presented in the previous section used the norm approach to finding the window lengths. Using the same sample of prior ensemble and measurement noise, the inversion was performed using the weight curve strategy instead, with a ratio parameter $\beta = 1$. This approach generally chooses larger windows thus making it more exposed to entering another local mode of the posterior distribution.

Figure 10 shows the results after the third assimilation cycle, which indicates divergence and exemplifies a case of misestimation. The observable region starts at around 2.5 km depth, and the top of this region is well estimated. Down the observed region around 3 km, the divergence starts. The data time window covers 2.326–3.184 s with 34052 data points. This partition is larger than both the third and fourth batch in Section 5.3.

Below the depth where a wrong local mode is found, the mean is highly spurious. The mechanism driving the misestimation in this case is the ensemble linearization of the density gradient/tangent linear model. The cross-covariances between waveform data and densities are much more susceptible to being poorly estimated, i.e. "spurious correlations", than for the velocities. The density effect on the reflected waveform amplitudes is more obscure, and acts in combination with the velocities. Ambiguity in the estimation of density sensitivity can lead to an update direction $\Delta\mathbf{w}$ that points towards a local and erroneous mode. And the chance of this occuring increases with the batch size. Once a local mode is discovered through the control vector $\mathbf{w}$, the mean $\mathbf{x}$ below that local misestimation will diverge. A false mode has $\widehat{\rho} > \rho^{\mathrm{t}}$ and velocities $\widehat{v}_{p,s} < v^{\mathrm{t}}_{p,s}$ or vice versa in some localised depth region. In which "direction" the false mode is estimated is seemingly a question of the position of the prior mean $\overline{\mathbf{x}}^{\mathrm{f}}$ of the assimilation cycle.

To illustrate this, a closer look at the course of the iterative mean is displayed in Fig. 11. Focusing on the depth range 2.5–3.5 km, Fig. 11a shows the estimation bias for each of the elastic properties over the course of the first 5 iterations. The top 2-3 layers are seen to be well estimated within the first few iterations, whereas divergence takes place from layer 4 or 5 and downwards. The density bias indicates that the onset of misestimation is from the very first iteration, starting at depth around 3.2 km from where the bias evolves consistently into adjacent layers reaching a fixed value. This is in contrast to the estimation bias of the velocities that varies over iterations, and implies that it indeed is density that drives the estimation divergence. The objective function, Fig. 11b, shows that the data misfit is reduced while the mean is updated towards an erroneous local solution. So the minimization problem is converging in the sense of reducing the data misfit, just at the wrong solution.

The norm criteria seeks to assure that $\|\Delta\mathbf{w}_{x,j}\|$ and $\|\Delta\mathbf{w}_{y,j}\|$ for $j = 2$ are of comparable size (for $\beta = 1$) in the hope that this is a good start for stable iterations. In comparison, the weight criterion has no notion of this. Figure 12 shows the norm components for this divergent 3rd window, along with the corresponding components from the 4th assimilation cycle in Section 5.3. It is not a fair comparison, as the previous section's 4th cycle is shorter with fewer data points, but it highlights an observation that is fairly consistent across encountered examples of divergence and exemplifies a characteristic of convergent versus divergent solutions. The main difference is the continued dominance of $\|\Delta\mathbf{w}_{y,j}\|$ when the estimation is diverging. While the norm criterion starts out with a slightly lower $\|\Delta\mathbf{w}_{x,2}\|$ than $\|\Delta\mathbf{w}_{y,2}\|$, the following iterations has a larger prior component magnitude until they equalize,

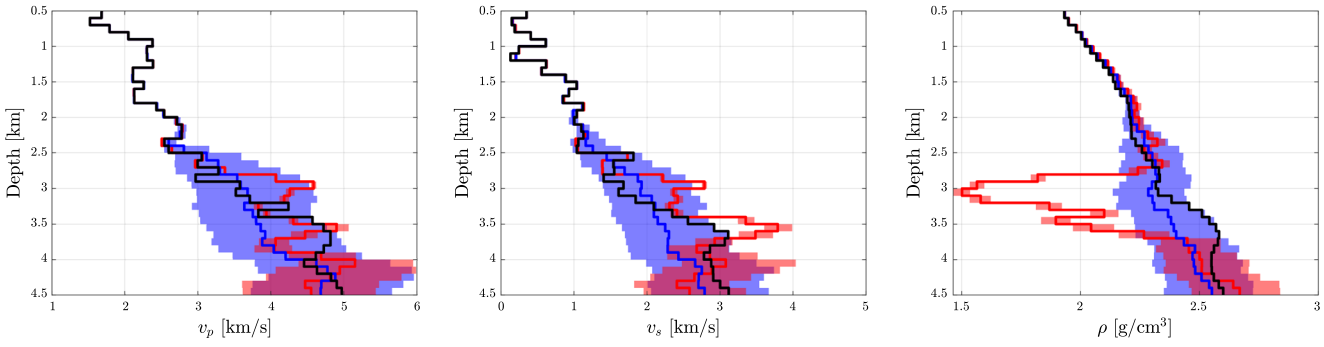

Fig. 10 Prior and analysis ensemble of the 3rd batch window; blue is prior, red is posterior, black is truth. **a** Acoustic velocity **b** Shear velocity **c** Density

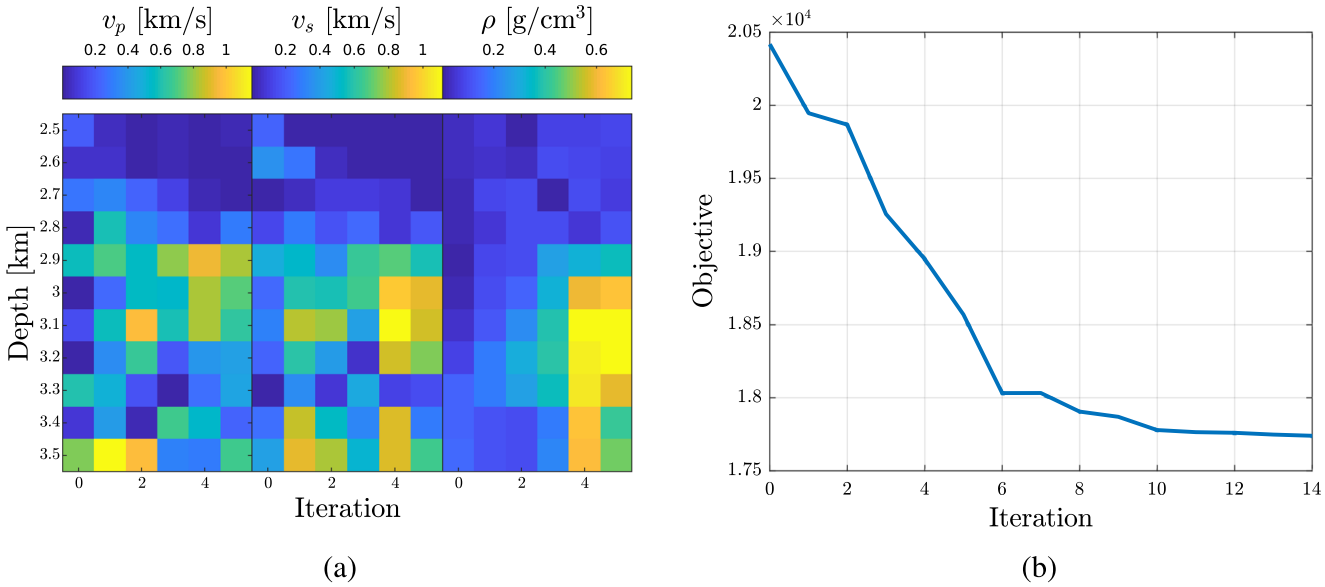

Fig. 11 Over iterations: **a** the estimation bias $|\widehat{\mathbf{m}}_j - \mathbf{m}^t|$ for the first 5 iteration for each elastic attribute and $j = 0$ is prior estimate, and **b** the objective function. **a** $|\widehat{\mathbf{m}}_j - \mathbf{m}^t|$ **b** $J(\mathbf{w}_j)$

which is the point where $\|\mathbf{w}\|$ reaches a stationary level (Fig. 9c). Contrary for the divergent weight criterion, the likelihood component keeps dominating until they equalize. From our experience, the pattern is that the relation between these vector norms in the first 2 to 4 or 5 iterations determines whether the estimation is converging or not.

We speculate whether monitoring the course of the update vector components' magnitude could effectively be used as a running diagnostic. A diagnostic that indicates divergence with the potential to stop iterating and restart the assimilation cycle with a shorter data window. The monitoring has a cost though, as the left singular vector $\mathbf{u}_i$ must be available for calculating $\Delta\mathbf{w}_y$ while not strictly necessary for elements of the IEnKS as such. Yet, this could be outweighed by the possible robustness added to an inversion routine.

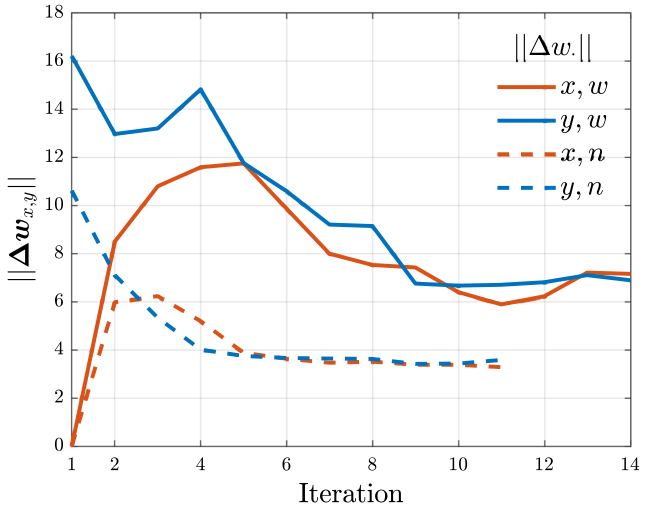

Fig. 12 Norm over iterations of vector components $\Delta\mathbf{w}_{x,j}$ and $\Delta\mathbf{w}_{y,j}$. In legend, 'w' (weight) refers to 3rd batch in Fig. 10, while 'n' (norm) refers to 4th batch of Fig. 5

As a final remark on the issue of non-uniqueness and misestimation of a gradient towards a false local mode. The general observed picture is as seen in this case, where the top of the observed region is well estimated but the mean updates towards a local mode further down, it was thought that dampening the update could help by the mean moving less (at deeper depths), giving an opportunity to reevaluate the sensitivities when the top of the observed region had been accounted for. To dampen the update step, the principle of Multiple Data Assimilation (MDA, see e.g. [6, 7]) was applied. A sequence of MDA iterations with its inflation of the observation error covariance matrix, was used for a fixed number of initial iterations. The sequence choice of inflation factors was based on a geometric serie [22]. The MDA error inflation changes both the update direction and dampen its magnitude, and it was hoped that it could downplay the contributions from observations to the control vector that locked the mean state in a wrong mode. But the results showed no effective improvement as it did not guarantee against misestimation. This might be due that MDA inflates the error for all data points equally, whereas it probably would be benefial to have inflated them differently, in order to downplay data at later time points within the window. This could be achieved through covariance (**R**-) localization, by upscaling the observation error at later time instances and reducing this inflation gradually over the initial iterations. How the upscaling should be distributed

across time and offsets is a complicated question and this approach has not been pursued any further.

The use of a comparable stochastic method (the stochastic iterative ensemble smoother [21]) was also tried out to check whether the perturbed observations would have a positive effect with respect to the false local mode estimation. While the stochastic version gives comparable results when successful, there was an observed tendency that misestimation towards a local mode occurred more frequently.

6.2 Replicate study

The example presented so far was for a particular sample of prior ensemble and measurement noise. If these were resampled and the inversion done again, another outcome is obtained. In order to examine the general robustness of the strategies, the parameter β and dependency on the ensemble size, repeated estimation trials was performed. Independent replicate trials that randomize the initial prior ensemble and the additive noise in the synthetic measurement data, are used to evaluate estimation performance. The statistical model and forward model configuration are kept fixed, so the results are in light of those. For each of the ensemble sizes used, a batch of 20 samples are used across the β parameter variation and the strategies. For each strategy, the configurations are combinations of three ensemble sizes $n = (150, 300, 600)$ and $\beta = (3/2, 1, 3/4)$. The larger $\beta = 3/2$ is a slightly more conservative choice with shorter time span of windows and smaller β increases the window lengths.

Each replicated estimation is accepted or rejected. If the solver was not able to compute with the given model input, the replicate is naturally rejected. Otherwise, to classify a posterior as an acceptable estimation, only the state subset $\mathbf{z} = \log \mathbf{v}_p|\text{depth} \leq 3.5\text{km} \in \mathbb{R}^{35}$ is considered. The reason is that the estimation of acoustic velocity will generally be better than for the other two elastic properties, especially for smaller ensemble sizes. As estimation measure the Mahalanobis distance of the true $\mathbf{z}^{\text{t}}$ subset is used, with respect to the distribution represented by the posterior ensemble:

$$\text{MD}_{\text{t}} = \left((\mathbf{z}^{\text{t}} - \widehat{\mathbf{z}})^{\text{T}} \mathbf{C}^{-1} (\mathbf{z}^{\text{t}} - \widehat{\mathbf{z}}) \right)^{1/2}, \tag{16}$$

where $\widehat{\mathbf{z}}$ is the ensemble mean. The covariance matrix $\mathbf{C}$ is the ensemble sample covariance, but with an important modification as it will use a truncated eigenbasis that retains only 75% of the total variance. The reason for doing this truncation is to make the distance measure more robust. The eigen-components with the smallest 25% of total variance are associated with the shallower layers, and smaller estimation error at these layers between different ensembles makes the measure more volatile and less useful for this purpose. This is a consequence of the low estimation uncertainty/ensemble spread at the shallower layers. Alternatively, the depth range of the state subset used for the estimation measure could be limited to e.g. 1.5–3.5 km as this would roughly give the same effect.

Table 1 Number of accepted inversion runs out of 20 replicates

n	150		300		600	
β\ Criteria	w	n	w	n	w	n
3/2	11	17	9	20	15	20
1	8	19	8	17	12	17
3/4	4	12	9	13	5	15

To set a (per replicate) threshold for the accept/reject classification, each ensemble member is distance-measured as MD_i against the same $(\widehat{\mathbf{z}}, \mathbf{C})$, which gives a level of within-ensemble distance. If $\text{MD}_{\text{t}} \leq \overline{\text{MD}_i} + 4 \times \text{std}(\text{MD}_i)$, it is accepted as a satisfactory solution. Thus each replicate has its own threshold value. No false positive, i.e. an accepted divergent solution, was confirmed by visual inspection. On the contrary, especially for the smallest ensemble size considered, some cases could have been judged acceptable but did not pass the classification rule.

According to this rule, the number of accepted runs among the 20 replicates is listed in Table 1 and the average number of resulting windows $\overline{K}$ is listed in Table 2. In the header of these tables, the strategy 'w' and 'n' refers to the weight and norm criteria respectively.

The general pattern is that the norm strategy is much more robust than the weight strategy, and that estimation performance decays with β decreasing (larger batch windows). The case $\beta = 3/2$ versus $\beta = 1$ for $n = 150$ deviates from this pattern, but the cause is more related to ensemble size than to window size.

From these results, the emphasis on controlling the norm magnitude of the likelihood vector component is definitely more influential on stable updates than the weighting curve argument. The latter, with its focus on the weighting distribution, is not addressing the mechanism that controls the potential spurious update. As the norm

Table 2 Average number $\overline{K}$, rounded to nearest integer, of batches

n	150		300		600	
β\Criteria	w	n	w	n	w	n
3/2	7	18	5	8	3	5
1	6	9	4	6	3	4
3/4	6	7	4	5	3	3

Table 3 Total number of forward model evaluation

n	150		300		600	
β\Criteria	w	n	w	n	w	n
3/2	9122 (327)	15908 (1146)	16766 (1054)	19230 (967)	29120 (903)	32640 (2548)
1	8643 (375)	10318 (762)	15450 (578)	17311 (1111)	25900 (668)	29188 (819)
3/4	8812 (495)	9200 (390)	14866 (400)	16223 (1063)	24000 (1200)	25840 (1350)

The table entries is the average over accepted replicates, while in parenthesis is the standard deviation. Numbers are rounded to nearest integer

approach utilizes the actual likelihood vector component, its better performance is expected.

The expected correlation between shorter windows and estimation stability is present for all ensemble sizes. The norm approach is generally more conservative than its alternative, resulting in a higher number of windows, as Table 2 shows. Where perturbing the β-parameter for the weight approach seemingly does not do much for the number of windows, the difference lies in their distribution, where for $\beta = 3/4$ the last window just becomes shorter (up to the gather end time). The norm criteria on the other hand, is more sensitive to the ratio β setting along with a stronger dependency on ensemble size.

The number of accepted estimation (for n-criteria) are comparable for $\beta \geq 1$ and $n = 300$ and 600, and one could get the impression that there is no benefit of using the largest ensemble size. Especially when considering the total number of forward model evaluations, as listed in Table 3. But the estimation results (not shown here) show that $n = 600$ performs much better than $n = 300$, at estimating density generally and all elastic properties at the bottom 1 km depth. So size does matter for sufficiently alleviate rank issues in resolving the gradients when data has a high level of noise.

The numbers in Table 3 are quite high and this implies and demands parallel ensemble evaluation. The number of evaluation decreases with longer windows, so it is not the case that shorter data windows results in an earlier termination of iterating, sufficient to counterbalance the larger number of windows. In terms of efficiency, the $\beta = 1$ case is preferable, while not as consistent in accepted estimations as $\beta = 3/2$ (for $n \geq 300$). If combined with a monitoring and handling of divergence as described in previous section, the more intuitive case of $\beta = 1$ is deemed a good choice.

7 Summary and conclusions

In this paper an ensemble-based sequential method for seismic inversion is presented. The iterative ensemble Kalman smoother is the core method that uses the ensemble to evaluate sensitivites, thus no tangent linear model is needed and suitable for black-box forward models.

The approach for assimilating the high-dimensional seismic data builds on a strategy of partitioning the data in windows of traveltime, and the inversion is stable if these windows are selected wisely.

A method for automatically selecting an appropriate data window when entering an assimilation cycle is introduced. The method is based on an analysis of the iterative update to the control variable of the variational problem, and on an interpretation of how this update is influenced by the prior and likelihood. Two alternative angles of interpretation are presented and their performance evaluated through a repeated trials simulation study. Only one of the alternatives showed robust with respect to estimation performance, the approach based on norms of prior and likelihood vector. This aspect was highlighted and discussed in a comparison of a converging and a divergent estimation.

A synthetic example was used in this paper. Future work includes testing on field data. In doing so, one must likely also use auxiliary data such as well logs for tuning parameters in the forward model. Well log data can also be used together with seismic recordings to estimate parameters in noise covariance matrix. An assumption of the presented work is independent measurement errors. Even though within-batch correlations could be included easily, adjustments are required to handle possible between-batch correlations.

While the motivation for this work is reliable nonlinear elastic inversion, the need to partition the data set and assimilate these sequentially is expected to be present in other types of parameter estimation problems. In the geoscience domain there is for instance potential for similar inversion methods for electromagnetic data, gravimetric data, fiber optical data and ground penetrating radar data, which all involves large-size data, complex physical forward models and static state parameters. As a consequence, the observations from this study might be applicable to and of use in other domains.

Acknowledgements We thank the Norwegian Research Council and partners of the Uncertainty in Reservoir Evaluation (URE) project and the GAMES consortium at the Norwegian University of Science and Technology (NTNU) for the financial support (grant no. 294404). We further thank BP for data. We also thank the three reviewers for constructive comments that improved the paper.

Funding Open access funding provided by NTNU Norwegian University of Science and Technology (incl St. Olavs Hospital - Trondheim University Hospital).

References

1. Asch, M., Bocquet, M., Nodet, M.: Data Assimilation: Methods, Algorithms, and Applications. Society for Industrial and Applied Mathematics, Philadelphia (2016). https://doi.org/10.1137/1.9781611974546
2. Bishop, C.H., Etherton, B.J., Majumdar, S.J.: Adaptive sampling with the ensemble transform kalman filter. part i: Theoretical aspects. Mon. Weather Rev. **129**(3), 420–436 (2001). https://doi.org/10.1175/1520-0493(2001)129<0420:ASWTET>2.0.CO;2
3. Bocquet, M., Sakov, P.: An iterative ensemble Kalman smoother. Q. J. Roy. Meteorol. Soc. **140**(682), 1521–1535 (2014). https://doi.org/10.1002/qj.2236
4. Cardinali, C., Pezzulli, S., Andersson, E.: Influence-matrix diagnostic of a data assimilation system. Q. J. Roy. Meteorol. Soc. **130**(603), 2767–2786 (2004). https://doi.org/10.1256/qj.03.205
5. Carrassi, A., Bocquet, M., Bertino, L., Evensen, G.: Data assimilation in the geosciences: an overview of methods, issues, and perspectives. WIREs Climate Change **9**(5), e535 (2018). https://doi.org/10.1002/wcc.535
6. Emerick, A.A.: Deterministic ensemble smoother with multiple data assimilation as an alternative for history-matching seismic data. Computational Geosciences, https://doi.org/10.1007/s10596-018-9745-5 (2018)
7. Emerick, A.A., Reynolds, A.C.: Ensemble smoother with multiple data assimilation. Comput. Geosci. **55**, 3–15 (2013). https://doi.org/10.1016/j.cageo.2012.03.011
8. Evensen, G.: Data Assimilation. Springer, https://doi.org/10.1007/978-3-642-03711-5 (2009)
9. Evensen, G.: Analysis of iterative ensemble smoothers for solving inverse problems. Comput. Geosci. **22**(3), 885–908 (2018). https://doi.org/10.1007/s10596-018-9731-y
10. Fillion, A., Bocquet, M., Gratton, S.: Quasi-static ensemble variational data assimilation: a theoretical and numerical study with the iterative ensemble kalman smoother. Nonlinear Process. Geophys. **25**(2), 315–334 (2018). https://doi.org/10.5194/npg-25-315-2018
11. Fowler, A., van Leeuwen, P.J.: Measures of observation impact in non-gaussian data assimilation. Tellus A: Dynamic Meteorology and Oceanography **64**(1), 17192 (2012). https://doi.org/10.3402/tellusa.v64i0.17192
12. Gebraad, L., Boehm, C., Fichtner, A.: Bayesian elastic full-waveform inversion using hamiltonian monte carlo. J. Geophys. Res. Solid Earth, 125(3). https://doi.org/10.1029/2019JB018428 (2020)
13. Gineste, M., Eidsvik, J., Zheng, Y.: Ensemble-based seismic inversion for a stratified medium. Geophysics **85**(1), R29–R39 (2020). https://doi.org/10.1190/geo2019-0017.1
14. Hunt, B.R., Kostelich, E.J., Szunyogh, I.: Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter. Physica D: Nonlinear Phenomena **230**(1-2), 112–126 (2007). https://doi.org/10.1016/j.physd.2006.11.008
15. Kennett, B.: ERZSOL3. http://www.spice-rtn.org/library/software/ERZSOL3.html, accessed on 2017-09-01 (2005)
16. Kennett, B.: Seismic Wave Propagation in Stratified Media. ANU Press (2011)
17. Leeuwen, P.J.V., Cheng, Y., Reich, S.: Nonlinear Data Assimilation, Frontiers in Applied Dynamical Systems: Reviews and Tutorials, vol 2. Springer International Publishing, https://doi.org/10.1007/978-3-319-18347-3 (2015)
18. Liu, J., Kalnay, E., Miyoshi, T., Cardinali, C.: Analysis sensitivity calculation in an ensemble kalman filter. Q. J. Roy. Meteorol. Soc. **135**(644), 1842–1851 (2009). https://doi.org/10.1002/qj.511
19. Mannseth, T., Fossum, K.: Assimilating spatially dense data for subsurface applications–balancing information and degrees of freedom. Comput. Geosci. **22**(5), 1323–1349 (2018). https://doi.org/10.1007/s10596-018-9755-3
20. Muir, J.B., Tsai, V.C.: Geometric and level set tomography using ensemble Kalman inversion. Geophys. J. Int. **220**(2), 967–980 (2020). https://doi.org/10.1093/gji/ggz472
21. Raanes, P.N., Stordal, A.S., Evensen, G.: Revising the stochastic iterative ensemble smoother. Nonlinear Process. Geophys. **26**(3), 325–338 (2019). https://doi.org/10.5194/npg-26-325-2019
22. Rafiee, J., Reynolds, A.C.: Theoretical and efficient practical procedures for the generation of inflation factors for ES-MDA. Inverse Problems **33**(11), 115003 (2017). https://doi.org/10.1088/1361-6420/aa8cb2
23. Rodgers, C.D.: Inverse Methods for Atmospheric Sounding. World Scientific Publishing, https://doi.org/10.1142/3171 (2000)
24. Sacher, W., Bartello, P.: Sampling errors in ensemble kalman filtering. part i: Theory. Mon. Weather. Rev. **136**(8), 3035–3049 (2008). https://doi.org/10.1175/2007MWR2323.1
25. Sakov, P., Oliver, D.S., Bertino, L.: An iterative enKF for strongly nonlinear systems. Mon. Weather. Rev. **140**(6), 1988–2004 (2012). https://doi.org/10.1175/mwr-d-11-00176.1
26. Sheriff, R.E., Geldart, L.P.: Exploration Seismology, 2nd edn. Cambridge University Press, https://doi.org/10.1017/CBO9781139168359 (1995)
27. Shirangi, M.G., Emerick, A.A.: An improved tsvd-based levenberg–marquardt algorithm for history matching and comparison with gauss–newton. J. Pet. Sci. Eng. **143**, 258–271 (2016). https://doi.org/10.1016/j.petrol.2016.02.026
28. Thurin, J., Brossier, R., Métivier, L.: Ensemble-based uncertainty estimation in full waveform inversion. Geophys. J. Int. **219**(3), 1613–1635 (2019). https://doi.org/10.1093/gji/ggz384
29. Zupanski, D., Hou, A.Y., Zhang, S.Q., Zupanski, M., Kummerow, C.D., Cheung, S.H.: Applications of information theory in ensemble data assimilation. Q. J. Roy. Meteorol. Soc. **133**(627), 1533–1545 (2007). https://doi.org/10.1002/qj.123