Odd Kolbjørnsen

Nonlinear Topics in the Bayesian Approach to Inverse Problems with Applications to Seismic Inversion

Dr. Ing. Thesis

Department of Mathematical Sciences
Norwegian University of Science and Technology
2002

Preface

This thesis is submitted in partial fulfillment for the requirements of the degree "Doktor Ingeniør" at the Norwegian University of Science and Technology (NTNU). The research was funded by a grant from the Research Council of Norway, and was carried out at the Department of Mathematical Sciences, NTNU.

I would like to thank my supervisor Professor Henning Omre for his patience, encouragement, guidance and support. Part of the work on this thesis was done while I was visiting Department of Statistics, Stanford University the year 1999 and I thank Professor Paul Switzer for inviting me to stay at Stanford and for several interesting discussions. I thank Statoil and the Sleipner licence for permission to use data from the Sleipner Øst Field, and Arild Buland for discussions about these data and for pleasant collaboration. I thank Professor Peter Lindqvist for enlightening discussions, Håkon Tjelmeland for valuable comments, everybody at the Department of Mathematical Sciences, in particular the Statistics group, for providing a stimulating working environment, the staff and the computer engineers for being helpful and service minded.

My warmest thanks goes to my wife Katrine for her continuous love and support through out the thesis work, and to my two sons Peter and Tobias for allowing me to think of statistics and forcing me not to.

Trondheim, May 2002

Odd Kolbjørnsen

Thesis Outline

The thesis consist of the following papers, which can be read independently of each other, though it is natural to read Paper I before Paper II and III.

Paper I	**Bayesian inversion of piecewise affine operators in a Gaussian framework.** With Henning Omre. Submitted.
Paper II	**Geostatistical approach to event migration of seismic reflection times.** In Kleingeld, W.J. and Krige D.G. eds, (2002) Proceedings of the Sixth international Geostatistics congress, Cape Town, South Africa, April 2000, Vol 1, pp 114-123, revised 2001.
Paper III	**Case specific uncertainty assessment in cross well tomography.** Report
Paper IV	**Cauchy prior for Bayesian linearized seismic AVO inversion.** Report
Paper V	**Rapid spatially coupled AVO inversion in the Fourier domain.** With Arild Buland and Henning Omre. Submitted.

In addition, the thesis contain an introduction in which inverse problems in general are discussed and comment on the content of the five papers from this general point of view.

Background

The five papers included in the thesis are motivated by challenges encountered in the Bayesian approach to inverse problems. In particular nonlinear topics are of interest.

The linear theory of Bayesian inversion can be defined by requiring a linear relation between the parameter and the observations, additive Gaussian observation errors and a Gaussian prior distribution for the parameter. In the linear theory, both the relation between the parameter and observations, and the relation between the observations and the estimator are linear. The Bayesian inference in this standard case is well known from the statistical litterature. Nonlinearities arise whenever either of the three assumptions defining the linear theory of Bayesian inversion is invalid. In the nonlinear case Bayesian inference must normally be adapted to the problem at hand, and frequently it requires

sophisticated sampling methods to explore the posterior distribution.

In Paper I, II and III the nonlinearity is introduced by assuming that the observations have a nonlinear relation to the parameter, still keeping the Gaussian assumption for the prior and the observation error. By assuming the relation between the parameter and the observations to be piecewise affine, sufficient structure of the inverse problem is imposed such that part of the analysis can be treated analytically.

Paper IV and V consider inverse problems where the likelihood model can be linearized on a logarithmic scale. In Paper IV a non Gaussian prior distribution is introduced hence the resulting estimator is a nonlinear operator on the observations. Paper V is only marginally nonlinear, since it is fully linearized on the logarithmic scale. The main concern in Paper V is however the dimensionality of the problem.

Paper I is mainly concerned with theoretical aspects of the posterior distribution for piecewise affine inverse problems. The other papers are related to various topics in seismic inversion. Paper IV and V have been developed to a stage such that real data is used in the inversion.

Summary

Paper I considers piecewise affine inverse problems. This is a large group of nonlinear inverse problems. Problems that obey certain variational structures are of this type. In inverse problems it is frequently such that some features are well determined by the observations while others are poorly resolved. In the Bayesian approach this imply that the likelihood forces the posterior distribution to be concentrated near hyper surfaces in the parameter space. In nonlinear problems this causes most generic sampling algorithms to be slow. The structure that is enforced in piecewise affine inverse problems allows the posterior distribution to be decomposed as a mixture of truncated Gaussian distributions. Under given regularity conditions this mixture distribution is non singular even if the observations are exact. A sampling algorithm that exploit this decomposition is proposed. The decomposition can however be used in a variety of sampling algorithms and is not limited to the sampling algorithm used here. Two small example problems are used to illustrate the theory as it is developed.

Paper II treats a problem in reflection seismic within the framework of piecewise affine inverse problems. Assuming a known, constant velocity in a layer, the problem is to determine the position of a reflector in the subsurface based on zero offset traveltimes. This is a standard simplification of the problem in reflection seismic. A synthetic example show that the uncertainty is well represented if there is a small number of observations, whereas the subsurface is satisfactory reconstructed when a large number of observations are considered. In the

example it is demonstrated that the current approach improve the standard approach.

In Paper III cross well tomography is discussed in a Bayesian setting. In cross well tomography the slowness field, being the inverse of the velocity, is reconstructed based on the traveltimes of a signal generated in one well and received in an other well. The travel time recorded is the shortest time that is physically possible. The inverse problem is approximated by a piecewise affine inverse problem of the form considered in Paper I. The calculations are carried through for this problem by exploiting Fermat's principle of least time. The methodology is tested for a synthetic example. In the Bayesian approach to this problem, several slowness fields are sampled from the posterior distribution. All the proposed samples honor the traveltime observations up to the specified error structure. These slowness fields are averaged to produce the Bayesian estimator. The resulting estimator does not honor the the traveltime observations as the individual samples do, but generally have larger traveltimes. This is due to the nonlinearity in the problem. This effect is carefully explained in the paper. The synthetic example further show that a linearized approach is reasonable in the sense that it capture the main features in the estimate. The nonlinear estimate does however reduce the loss with about 10 % in the synthetic example. The linearized approach does not give a realistic representation of the uncertainty. In synthetic example the linearized approach underestimate the integrated variance by 30 %.

In Paper IV the objective is inversion of seismic pressure amplitudes recorded in a marine seismic survey. After several steps of preprocessing, the seismic observations can be modeled by a linear relation to the seismic reflectivity, which again may be approximated by a linear relation to the material parameters on a logarithmic scale. The material parameters considered are pressure wave velocity, shear wave velocity and rock density. The seismic data that correspond to reflections below one location at the surface are given as angle gathers. In Paper IV each angle gather is inverted independently. The main concern in Paper IV is that the seismic reflectivity have heavier tails than what is predicted by a standard Gaussian model. A prior distribution based on superposition of a Cauchy process and Gaussian processes is proposed. As a test case material parameters observed in a well log at the Sleipner Øst Field is used to generate synthetic seismic observations. This is used as a basis for comparison between the proposed Cauchy model and a purely Gaussian model. The well log is used to estimate the parameters in the prior distribution both for the Cauchy model and for the pure Gaussian model. In a region with large variability the estimator for the pressure wave velocity resulting from the Cauchy model improves the risk by as much as 14 %. The Cauchy model also cause the uncertainty bounds to vary such that regions with low variability have shorter credibility intervals and regions with high variability have longer credibility intervals than for a pure Gaussian model. The model is also tested for real seismic observations. The results are satisfactory, although the uncertainty is large due to large observation

errors in the seismic data.

Paper V has the same objective as Paper IV, that is to estimate pressure wave velocity, shear wave velocity and rock density, based on preprocessed data from a marine seismic survey. In Paper V it is however assumed that the Gaussian assumption can be justified. The focus in this paper is to incorporate lateral dependencies in the estimates. When latteral dependencies are included, all parameters are coupled, and must be solved simultaniously. This lead to a high dimensional problem. Paper V exploits the fact that a Fourier transform of the problem yield a block diagonal form such that a small problem may be solved for each frequency component independently. Both the posterior mean and the posterior covariance can be computed and stored efficiently, due to the special structure of the problem. This opens the possibility for including additional information such that well data to obtain a refined solution around the well. The methodology is tested on a seismic cube from the Sleipner Øst Field, where 12 million parameters are estimated. The total computing time after preprocessing is 6 minutes, the posterior covariance can be computed in additional 3 minutes on a single 400 Mhz Mips R12000 CPU. Hence the algorithm is extremely rapid.

Fundamentals of inverse problems

Odd Kolbjørnsen

Department of Mathematical Sciences
Norwegian University of Science and Technology
Norway

1 Introduction

Inverse problems can be defined as problems that consist in finding the cause of an observed effect. An inverse problem is always paired with a direct problem that provide the effect of a given cause. This definition requires the formulation of any specific problem to be based on physical laws and that physics must specify what is a cause and what is an effect as well as provide the equations relating the effects to the causes (Bertero 1989). Inverse problems arise naturally if one is interested in determining the internal structure of a system based on the system's observed behavior or in determining the unknown input that give rise to an observed output (Hansen 1998).

In a mathematical language an inverse problem relate to an operator equation,

$$y = K(z) \ , \tag{1}$$

with $K : \mathcal{Z} \to \mathcal{Y}$ being a possibly nonlinear operator. The direct problem is to determine the effect, y, of a given cause, z, whereas the inverse problem is to determine the cause, z, of an observed effect y. The function space $\mathcal{Z}$ is commonly denoted the model space or parameter space, while $\mathcal{Y}$ is denoted the data space.

Expression (1) is unlikely to hold when y is a measured quantity, since measurements have finite precision. In addition Expression (1) may be inaccurate in the sense that the operator K does not model all aspects of the physical processes that produce the observations. The problem is hence more realistically stated as,

$$y = K(z) + \varepsilon \ , \tag{2}$$

with ε being an error term.

To give a comprehensive account for all aspects of inverse problems, is impossible in a short introduction since the field have so many branches spanning physical, mathematical, computational and statistical aspects. The current presentation include basic mathematical and statistical definitions that are relevant

for inverse problems, and discuss some of the philosophies that underlies the different solution methods. The presentation will concentrate on the case where $K : \mathcal{Z} \to \mathcal{Y}$ is a compact linear operator since this theory is by far the best developed. The inversion methods and the underlying philosophies are frequently generalized to solve nonlinear inverse problems. This is briefly discussed.

The presentation is organized as follows. Mathematical aspects of inverse problems are presented in in Section 2. Inversion by regularization is presented in Section 3. The statistical theory of point estimation is presented in 4. Statistical minimax inversion and Bayesian inversion are presented in Section 5 and 6 respectively. In Section 7 the three inversion methodologies are compared with respect to similarities and differences. Section 8 contains some concluding remarks and the authors personal preferences. In Section 9 the content of the thesis is discussed in light of the current introduction.

2 Mathematics of inverse problems

The presentation in this section is based on Engl, Hanke and Neubauer (1996), Kirsch (1996) and Hansen (1998). It contain basic mathematical definitions and discusses approximate solutions to inverse problems.

2.1 Problem classification

According to the informal definition above a problem is classified as direct or inverse by the physics defining the problem. From a mathematical point of view problems are more naturally labeled as being well-posed or ill-posed. A problem is well-posed if there exists a unique, stable solution. The notion of a well-posed problem is attributed to Hadamard (1902, 1923). Although there is no formal connection between the two sets of labels, it is however true, with few exceptions, that direct problems are well-posed while the corresponding inverse problems are ill-posed. A definition of a well-posed inverse problem reads,

Definition 1 (Well-posed) *Let $\mathcal{Z}$ and $\mathcal{Y}$ be normed spaces and let $K : \mathcal{Z} \to \mathcal{Y}$ be a continuous operator from $\mathcal{Z}$ into $\mathcal{Y}$. The problem $y = K(z)$ is well-posed in the sense of Hadamard if the following three conditions are satisfied:*

1. *Existence: There exist a solution $z \in \mathcal{Z}$ for any $y \in \mathcal{Y}$ with $K(z) = y$*
2. *Uniqueness: There exist at most one solution $z \in \mathcal{Z}$ for any $y \in \mathcal{Y}$ with $K(z) = y$*
3. *Stability: For every positive number ϵ, there exist a positive number $\delta(\epsilon)$ such that any pair $z_1, z_2 \in \mathcal{Z}$ for which $\|K(z_1) - K(z_2)\| < \delta(\epsilon)$, $\|z_1 - z_2\| < \epsilon$*

Problems for which at least one of the three conditions above fails to hold are termed ill-posed.

Whether a problem is well-posed or not, depend both on the operator K and the function spaces $\mathcal{Z}$ and $\mathcal{Y}$.

The simplest case of an operator equation is, a matrix equation, $y = K\,z$, for which $\mathcal{Z} = R^n$, $\mathcal{Y} = R^m$ and K is a $m \times n$ matrix. The existence criterion then imply that the rank of K is equal to m, the uniqueness criterion imply that the rank of K is equal to n. Hence to assure both existence and uniqueness the matrix must be square and have full rank. These are also sufficient conditions for a matrix equation to be well posed. Any inverse problem formulated as a square matrix equation of full rank is hence stable in a strict mathematical sense. For matrix equations the criterion of stability relates to computational aspects of the inverse, K^{-1}. If a small change in y produce a large change in $z = K^{-1}y$, the system is said to be unstable. The standard example of such a matrix is

$$K = \begin{bmatrix} 1 & 1+\epsilon \\ 1 & 1 \end{bmatrix}$$

with ϵ being a small number. Let the supscript T denote matrix transpose. The solution for $y^T = [\,2\ ,\ 2\,]$ is $z^T = [\,2\ ,\ 0\,]$, while the solution for $y^T = [\,2+\epsilon\ ,\ 2\,]$ is $z^T = [\,1\ ,\ 1\,]$, hence a change in the input of order ϵ result in a change in the answer of order one. In unstable systems, some of the equations are almost linearly dependent. These systems are therefore hard to solve numerically, see Hansen (1998) for an extensive discussion.

2.2 Singular value expansion

Consider an operator equation

$$y = K\,z\ , \tag{3}$$

with $K : \mathcal{Z} \to \mathcal{Y}$ being a compact linear operator between two Hilbert spaces. In common notation $K^* : \mathcal{Y} \to \mathcal{Z}$ denotes the adjoint of K, and is defined by the requirement that for all $z \in \mathcal{Z}$ and $y \in \mathcal{Y}$, $(K\,z, y) = (z, K^*y)$, with $(\cdot,\cdot)$ denoting inner products in $\mathcal{Y}$ and $\mathcal{Z}$ at the left and the right side of the equality respectively. For any compact linear operator $K : \mathcal{Z} \to \mathcal{Y}$, there exist a singular system $\{\sigma_i, v_i, u_i\}_{i=1}^{\infty}$, with σ_i being nonnegative numbers, $\{v_i\}_{i=1}^{\infty}$ and $\{u_i\}_{i=1}^{\infty}$ being complete orthonormal systems of basis elements for $\mathcal{Z}$, and $\mathcal{Y}$ respectively. That is, $z \in \mathcal{Z}$ and $y \in \mathcal{Y}$ can be represented by the generalized Fourier series, $z = \sum z_i\,v_i$ and $y = \sum y_i\,u_i$, with $z_i = (v_i, z)$ and $y_i = (u_i, y)$. The numbers σ_i are the singular values of K, these are usually ordered in a non increasing order, $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$. Singular systems resembles the eigensystem of compact self adjoint operators, indeed $\{\sigma_i^2, v_i\}_{i=1}^{\infty}$ and $\{\sigma_i^2, u_i\}_{i=1}^{\infty}$ are the eigensystems of the self adjoint operators K^*K and KK^* respectively.

The singular system defines the singular value expansion of K,

$$K\,z = \sum_{i=1}^{\infty} \sigma_i\,(v_i, z)\,u_i\ , \tag{4}$$

The singular value expansion diagonalize the problem such that the generalized Fourier coefficients of z can be solved independently, i.e.

$$Kz = y \ \Leftrightarrow\ \sigma_i\,z_i = y_i\ ,\ i = 1, 2, \ldots\ . \tag{5}$$

The ill-posedness of a linear inverse problem is frequently related to the decay of the singular values. As $i \to \infty$, $\sigma_i \to 0$, hence the effect of z_i in Kz diminishes as $i \to \infty$. The rate of decay of the singular values can be used to classify linear ill-posed problems. A problem is termed mildly ill-posed if $\sigma_i \sim i^{-r}$ and $0 < r \leq 1$, moderately ill-posed if $\sigma_i \sim i^{-r}$ and $r > 1$ or severely ill-posed if $\sigma_i \sim \exp\{-ri\}$ or worse.

The singular value expansion is the infinite dimensional analog of the singular value decomposition of a matrix. In the case of K being a real $m \times n$ matrix, this decomposition reads,

$$K = U\Sigma V^T = \sum_{i=1}^{\min(n,m)} \sigma_i\,u_i\,v_i^T\ ,$$

with $U = [u_1\,u_2\,\cdots\,u_m] \in R^{m\times m}$ and $V = [v_1\,v_2\,\cdots\,v_n] \in R^{n\times n}$ being matrices with orthonormal columns, and Σ being a $m \times n$ diagonal matrix with the singular values, $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{min(n,m)}$, on the diagonal. From the equations $K^TK = V\Sigma^T\,\Sigma V^T$ and $KK^T = U\Sigma\,\Sigma^T U^T$ it is seen that the singular value decomposition of K is closely linked to the eigenvalue decomposition of K^TK and KK^T.

2.3 Fundamental subspaces and the generalized inverse

The range of K, $\mathcal{R}(K)$, are those $y \in \mathcal{Y}$ that can be reached from a $z \in \mathcal{Z}$. This is in general not a proper subspace of $\mathcal{Y}$, but the closure of this set, $\overline{\mathcal{R}(K)}$, is so. $\overline{\mathcal{R}(K)}$ is spanned by the basis elements of $\mathcal{Y}$ that correspond to strictly positive singular values, i.e. $\{u_i\}_{\{i\,:\,\sigma_i>0\}}$. The orthogonal complement of $\overline{\mathcal{R}(K)}$ in $\mathcal{Y}$ is termed the null space of K^*, denoted $\mathcal{N}(K^*)$, and it is spanned by the basis elements for which the corresponding singular values are zero, i.e. $\{u_i\}_{\{i\,:\,\sigma_i=0\}}$. Similarly $\mathcal{Z}$ can be divided into two subspaces corresponding to whether the elements influence the output of K or not. From Expression (4) it is easy to see that the subspace that influence the output of K is spanned by the basis elements of $\mathcal{Z}$ for which the corresponding singular values are strictly positive, i.e. $\{v_i\}_{\{i\,:\,\sigma_i>0\}}$. This space is the closure of the range of K^*, denoted $\overline{\mathcal{R}(K^*)}$. The orthogonal complement of $\overline{\mathcal{R}(K^*)}$ in $\mathcal{Z}$ is termed the null space

of K, denoted $\mathcal{N}(K)$, and is spanned by the basis elements of $\mathcal{Z}$ for which the corresponding singular values are zero, i.e. $\{v_i\}_{\{i\,:\,\sigma_i=0\}}$. It is more natural to relate to the operator K instead of the adjoint, hence it is common to define $\overline{\mathcal{R}(K^*)}$ as the orthogonal complement of $\mathcal{N}(K)$, i.e. $\overline{\mathcal{R}(K^*)} = \mathcal{N}(K)^\perp$.

The generalized inverse, $K^\dagger$, of a compact linear operator K can be defined using the singular system of K,

$$K^\dagger y = \sum_{\{i\,:\,\sigma_i>0\}} \frac{(u_i, y)}{\sigma_i}\, v_i \ , \tag{6}$$

when $y \in \mathcal{R}(K)$ this generalized Fourier series converge. The solution is easily found by solving the sequence problem in Expression (5). For a given $y \in \mathcal{Y}$ the convergence of the series in Expression (6), is equivalent with y satisfying the Picard criterion,

$$\sum_{\{i\,:\,\sigma_i>0\}} \frac{|(u_i, y)|^2}{\sigma_i^2} < \infty \ .$$

When the Picard criterion is fulfilled, the general solution to the inverse problem is characterized by the sum of one component from the null space of K and the generalized inverse of y, i.e. for any $z_0 \in \mathcal{N}(K)$,

$$z = z_0 + K^\dagger y \ . \tag{7}$$

This decomposition of the general solution as a sum of the homogeneous solution and a particular solution, is common in differential equations and matrix algebra.

2.4 Approximate solutions

In a real case the observations are prone to contain error hence the operator equation in Expression (3), should be replaced by

$$y = K\, z + \varepsilon \ , \tag{8}$$

with ε being an error term, see Expression (2). For most inverse problems the generalized inverse, $K^\dagger$, is unstable because $\sigma_i \to 0$ as $i \to \infty$. This imply that a small error ε will contribute significantly to the series in Expression (6) because the inner product (u_i, ε) is divided by σ_i. That is, the Picard criterion is usually not fulfilled for measured data. Since the true solution of Expression (8) cannot be obtained, an approximate solution is sought. An approximate solution is denoted $\hat{z}$. Some commonly used approximations are discussed below. All the approximate solutions are parameterized by a nonnegative number α that defines the degree of approximation. The parameter, α, is defined such that $\alpha = 0$ defines the exact solution. This parameter is briefly discussed below and more throughly in Section 3.

Filtering is a common way to obtain smoother solutions. In the context of inverse problems, filter factors may be introduced in the generalized Fourier series defining the generalized inverse, see Expression (6). The approximation may then be written as

$$\hat{z} = \sum_{\{i\,:\,\sigma_i>0\}} \phi_i(\alpha)\ \frac{(u_i, y)}{\sigma_i}\, v_i\ , \tag{9}$$

with $\phi_i(\alpha)$ being filter or shrinkage factors, and $\{\sigma_i, v_i, u_i\}_{i=1}^{\infty}$ being the singular system of K. The filter factors satisfy $0 \leq \phi_i(\alpha) \leq 1$ and $\phi_i(0) = 1$, and are defined such that the series in Expression (9) converge for $\alpha > 0$. Many different approximate solutions of Expression (8) have the form of Expression (9). In fact this expression is to general and a specific choice must be made for the filter factors $\phi_i(\alpha)$. An example is the truncated singular value expansion, which can be defined by $\phi_i(\alpha) = I\{\sigma_i > \alpha\}$, where $I\{\cdot\}$ is one if the event in the brackets is true, zero otherwise. In this case only the terms where the singular value exceed α are included. The singular value expansion of a problem is generally not known this complicates the approach in practical situations.

Tikhonov regularization exploits the fact that for any $y \in \mathcal{R}(K)$ the generalized inverse is the unique solution of the least squares problem

$$z = \arg\min_{z\ \in\ \mathcal{N}(K)^{\perp}}\ \|Kz - y\|^2. \tag{10}$$

When $y \notin \mathcal{R}(K)$, the solution $K^{\dagger}y$ is unbounded, i.e. $\|K^{\dagger}y\| = \infty$. Tikhonov regularization avoid this by adding a penalty term in the minimization to keep z bounded. The approximate solution, $\hat{z}$, is defined as the unique solution of

$$\hat{z} = \arg\min_{z\in\mathcal{Z}}\ \|Kz - y\|^2 + \alpha J(z), \tag{11}$$

with $J(z)$ being a suitable penalizing functional; and α being a positive number determining the trade off between the mismatch to the data and the penalizing term. The most common choice of penalizer is the squared norm in $\mathcal{Z}$, i.e. $J(z) = \|z\|^2$. In this case the approximate solution have the form of Expression (9), with $\phi_i(\alpha) = \sigma_i^2/(\sigma_i^2 + \alpha)$. Other choices are Sobolev norms, L^1 norm and maximum entropy. The methodology can also be generalized by using other measures for deviations in the data. Tikhonov regularization may be formulated in different ways, such as minimizing the error in the data subject to an upper bound on the penalizing functional, or as minimizing the functional subject to an upper bound on the error. For any given data y there is an one to one connection between the different formulations, where the bounds can be computed in terms of α and y.

Landweber iteration is an algorithmically defined approximate solution. It can be regarded as steepest decent algorithm with a fixed step length, $\omega < 1/\sigma_1^2$. The approximate solution is defined iteratively by

$$z^n = z^{n-1} - \omega K^*(Kz^{n-1} - y)\ , \tag{12}$$

with starting point $z^0 = 0$. After $m = 1/\alpha$ iterations, the solution obtained have the form in Expression (9) with $\phi_i(\alpha) = 1 - (1 - \omega\sigma_i^2)^m$. Note that when $y \notin \mathcal{R}(K)$ the true solution is unbounded, since Expression (12) converge to the true solution it is not beneficial to iterate the expression too many times. The amount of approximation hence lies in the number of iterations.

Conjugate gradient is another iterative technique for approximating the solution in Expression (10). The conjugate gradient identify the best solution in the Krylov subspace space of order m after m iterations. The Krylov subspace of order m is defined by

$$\mathcal{K}_m(K^*K, K^*y) = \text{span}\{K^*y, (K^*K)K^*y, \ldots, (K^*K)^{m-1}K^*y\}.$$

In certain applications the Krylov subspace of order m is an approximation to the subspace spanned by the first m singular vectors, i.e. $\text{span}\{v_i\}_{i=1}^m$. In these cases the conjugate gradient method stopped after $m = 1/\alpha$ iterations, can be seen as an approximation to the truncated singular value expansion with $1/\alpha$ terms. Note again that the amount of approximation is determined by the number of iterations.

Landweber and conjugate gradient iterations as defined above are used as means to define approximate solutions, for this purpose a finite number of iterations is required. The iterations can also be used as numerical schemes to solve well posed problems such as Expression (11). For such cases the the number of iterations is a purely numerical question. Practical implementation of the methods above require discretization. Different discretization schemes can also be used to define approximate solutions of the continuous problem. Most of the methods above can be described as filtering of the singular system. The singular value expansion is however not directly accessible for a given problem. In a given situation the Landweber iteration is hence much easier to implement than the general filtering scheme.

The approximate solutions above are not fully specified but depend on the parameter α that determines the tradeoff between data adaption and the boundedness of the approximate solution, $\hat{z}$. Many techniques are developed for choosing the parameter α. The L-curve is a helpful tool in understanding the impact of a particular choice, and can be used in various ways to pick a particular value of α. Cross validation and generalized cross validation (Wahba 1990) are also used for this purpose. The parameter choice is formalized in the regularization theory to be discussed next.

3 Inversion by regularization

The basic idea of regularization theory, is that the approximate solution should be stable with respect to small deviations in the observations. The problem can be seen as a game between a scientist and a malicious opponent. For given

bounds on the parameter $z \in C$ and the error $\|\varepsilon\| < \delta$, the scientist can choose the approximate solution, $\hat{z}$, depending only on the data y. The subset $C \subset \mathcal{Z}$ should at least exclude the components of z that does not influence the data, i.e. $\mathcal{N}(K)$. The opponent can chose the parameter, z, within the restriction C. The pay off in the game is the maximum deviation between approximate solution and the parameter for errors within the error bound

$$\sup_{\|y-K\,z\|<\delta} \|\hat{z} - z\|^2$$

This measure of deviation can be interpreted as allowing the opponent to chose the error in addition to the parameter. The ultimate goal for the scientist is hence to find an approximate solution that minimize the worst case deviation.

In order to reduce the complexity of the problem an indexed family of continuous operators $R_\alpha : \mathcal{Y} \to \mathcal{Z}$ is considered. A family of operators is denoted a regularization strategy if $R_\alpha K z \to K^\dagger K z$ for all $z \in \mathcal{Z}$ when $\alpha \to 0$. That is $R_\alpha K$ converge pointwise to the projection operator onto $\mathcal{N}(K)^\perp$. All of the approximate solutions listed in Section 2.4 are valid regularization strategies. A regularization method consist of a regularization strategy, R_α, and a rule for selecting the index α. A selection rule that only depend on the error bound, δ, is denoted an apriori selection rule, if the index also depend on the data, y, it is termed an aposteriori selection rule. A pair (R_α, α) is called an admissible or convergent regularization method if

$$\sup_{\|y-K\,z\|<\delta} \alpha(\delta, y) \to 0 \quad \text{as } \delta \to 0\ ,$$

and

$$\sup_{\|y-K\,z\|<\delta} \|R_{\alpha(\delta,y)} y - K^\dagger K z\| \to 0 \quad \text{as } \delta \to 0\ .$$

That is, as the error in the data tends to zero, so should the amount of regularization and the error of the approximation. By choosing a regularization strategy, the original problem has been reduced to a one dimensional problem, selecting a value of α for a given value of δ and y.

The most widespread procedure for selection of α is the discrepancy principle of Morozov, which defines the value of the regularization parameter α to be the one that yields $\|KR_\alpha y - y\| = \delta$, with δ being the maximum bound on the error.

Regularization methods are evaluated by the rate of convergence as the error in the observation approach zero. In this respect the goal is to obtain an uniform convergence rate in $\mathcal{Z}$, this is however impossible for most inverse problems. For this reason attention is drawn to subsets of $\mathcal{Z}$ of the form

$$\mathcal{Z}_{B,\rho} = \{\, z = Bw,\ \|w\| < \rho \,\}, \tag{13}$$

with $B : \mathcal{W} \to \mathcal{Z}$ being a bounded linear operator; and ρ being a finite number. This can be interpreted as an abstract smoothness constraint on z. Typical

theorems for inversion by regularization consist of two results, the first result state the optimal convergence rate in $\mathcal{Z}_{B,\rho}$, the second result prove that one particular regularization method have the optimal rate of convergence in $\mathcal{Z}_{B,\rho}$. A typical theorem is hence that the approximate solution found by Tikhonov regularization using a quadratic regularizer and a selection rule given by the discrepancy principle, is obtained by an admissible regularization method and obtain the optimal convergence rate in $\mathcal{Z}_{K^*,\rho}$.

Inversion by regularization is a general approach and can be used to solve nonlinear inverse problems. There are however few general results for this type of problems. Most convergence theorems are of local type, or assume that the global solution of a minimization problem can be found. Further, the uniqueness problem is not as easily decomposed as for linear systems, see Expression (7). To avoid problems in this respect a new origin, z_0, that represent the best prior guess is selected. When several solutions can be chosen, the one closest to z_0 is preferred. Tikhonov regularization, with penalizer $J(z) = \|z - z_0\|^2$ is widely used in nonlinear inverse problems. Since there are few rigorous results in this area there is no general optimality results for the solution obtained in most practical cases.

4 Statistical aspects of inverse problems

In the current presentation, inverse problems will be discussed in the context of point estimation (Lehmann and Casella 1998). There are many other statistical aspects of inverse problems than those discussed here. Most important are statistical methods for estimating the regularization parameter α, for a regularization strategy R_α, without having a prior bound on the error, see O'Sullivan (1986), Wahba (1990) and Hansen (1998) for a discussion of some of these methods, see also Stark (2000) for an insightful discussion of inverse problems as statistics.

From a statistical point of view an inverse problem, as phrased in Expression (2) and (8), is no different from any other estimation problem. A parameter z in a parameter space $\mathcal{Z}$ is to be estimated based on observations, y, in the data space $\mathcal{Y}$. The statistical link between the parameter, z, and the observations, y, is described by the likelihood, $p(y|z)$. Here, and in what follows, $p(\cdot)$ is being used as a generic probability distribution. A parameter z, or a feature of z, is said to be unidentifiable if it does not influence the likelihood, otherwise it is identifiable. An estimator, $\hat{z}$, for z is a measurable function of the data, $\hat{z}(y)$, or in general an operator, $\hat{z} : \mathcal{Y} \to \mathcal{Z}$. To evaluate an estimator a loss function, $L(z, \hat{z})$, is defined. A common choice, that will be used in what follows, is the squared L^2 norm, i.e. $L(z, \hat{z}) = \|z - \hat{z}\|^2$. The statistical philosophy is that if the experiment conducted to give the observations y is repeated, a new sample y' from $p(y|z)$ is obtained. Hence the error, ε, in Expression (2) and (8) is given a random variable interpretation. The objective is now to identify the estimator

that minimizes the expected loss when observations are sampled according to the likelihood. The expected loss of an estimator, $\hat{z}(y)$, is denoted the risk, $r_{\hat{z}}(z)$, and is defined pointwise in $\mathcal{Z}$ by,

$$r_{\hat{z}}(z) = \int_{\mathcal{Y}} \|z - \hat{z}(y)\|^2 \, dp(y|z) = \mathrm{E}_{Y|z} \left\{ \|z - \hat{z}(Y)\|^2 \right\}.$$

An estimator is said to be admissible if no other estimator can improve the risk uniformly in $\mathcal{Z}$. The risk is defined pointwise in $\mathcal{Z}$, but the estimator must be chosen without knowledge of z, hence in some way or other the estimator must take the risk for all $z \in \mathcal{Z}$ into account. In the minimax risk approach, the maximum risk over $\mathcal{Z}$ is used to compare estimators. An estimator is optimal if it has the least maximum risk in comparison to any other estimator. This philosophy is used for inverse problems in Section 5 below. In the average risk approach a measure is defined on $\mathcal{Z}$ and the the optimal estimator is defined as the one that minimizes the average risk according to this measure. The average risk approach is the fundament of Bayesian statistics which is further developed in Section 6 below.

The principles of estimation are the same for inverse problems as for any other statistical problem. Most inverse problems do however have some characteristics that distinguish them from the classical statistical theory. In inverse problems the number of parameters will frequently be of the same order, most often larger, than the number of observations. This can be seen from the sequence problem in Expression (5). Inverse problems of this type have closer resemblance to problems where the number of parameters grow together with the number of observations, than to the classical large sample theory (Lehmann 1999). Stein (1956) showed that the celebrated maximum likelihood estimator is inadmissible in a sequence model when there is an equal number of observations and parameters, larger than two. It is hence not likely that the maximum likelihood methodology will succeed in solving inverse problems. Further, in inverse problems the parameter is observed through a transform, see Expression (2), and not directly as in the traditional statistical theory of function estimation.

5 Statistical minimax inversion

In the minimax approach the estimates are evaluated by the maximum risk in $\mathcal{Z}$. The problem can be seen as a game between the scientist and a malicious opponent. For given bounds on the parameter $z \in C$ and a specified likelihood model, $p(y|z)$, the scientist can choose the estimator, $\hat{z}$, depending only on the data y. The opponent can chose the parameter, z, within the restriction $C \subset \mathcal{Z}$. The pay off in the game is the risk for the opponents choice of parameter, that is the expected loss under the likelihood model,

$$r_{\hat{z}}(z) = \mathrm{E}_{Y|z} \left\{ \|z - \hat{z}(Y)\|^2 \right\}.$$

The subset C may be a smoothness constraint such as Expression (13). The ultimate goal for the scientist is hence to find an estimator that minimize the worst case expected loss.

Estimators in the minimax approach are frequently evaluated by the rate of convergence as the information content of the date increase, the zero noise limit is common. Typical theorems for statistical minimax inversion consist of two results. First the optimal rate of convergence in C is obtained next an optimal estimator is found. The case where the set C is of the quadratic type, see Expression (13) is treated in Johnstone and Silverman (1990,1991), in which a rate optimal estimator is defined. The estimator correspond to filtering of singular values, see Expression (9). The estimator truncate the singular value expansion and shrink the remaining coefficients.

In some cases a smoothness constraint on the parameters such as Expression (13) can be limiting. The resulting estimators are always linear or almost so. Resent developments in the field of computational harmonic analysis allow for using the notion of sparsity rather than smoothness. The resulting estimator being represented by the wavelet-vaguelette decomposition (Donoho 1995). The main idea is that wavelets give a sparse representation of functions. Since the functions sought have few large coefficients, the focus can be directed towards which coefficients that should be estimated, instead of trying to estimate all. The typical result for these type of estimators is that the minimax rate of convergence is obtained adaptively within a logarithmic term. The adaptivity is in contrast to the traditional approaches where the smoothness must be defined prior to the estimation. The estimators based on wavelet-vaguelette decomposition has been particular successful for mildly and moderately ill-posed inverse problems.

The concern in minimax estimation is to get the best possible estimator for z, not to assess the uncertainty. There are however statistical results that deal with the uncertainty of the estimates also in this case, Stark (2001) considers confidence intervals for linear estimators of linear functionals, and report the methods of strict bounds (Backus 1989) and minimax confidence intervals (Donoho 1994), in both cases under the assumption of K being a compact linear operator.

The minimax approach is a general principle for estimation, and would apply also to nonlinear inverse problems. It is however a complex machinery and to the knowledge of the author, which may be limited, there has been no extensive study of minimax estimation for general nonlinear inverse problems.

6 Bayesian inversion

In the Bayesian approach knowledge and uncertainty regarding the parameter, z, is summarized in probability distributions. The prior distribution, $p(z)$, represent the knowledge of z prior to observations. The average risk, commonly

denoted the Bayes risk, of an estimator, $\hat{z}(y)$, is the expected risk under the prior measure,

$$B_{\hat{z}}[p(z)] = \mathrm{E}_Z \left\{ r_{\hat{z}}(Z) \right\} = \mathrm{E}_Z \left\{ \mathrm{E}_{Y|Z} \left\{ \|Z - \hat{z}(Y)\|^2 \right\} \right\} \ . \tag{14}$$

The objective in Bayesian estimation is to find the estimator that minimizes the Bayes risk, $B_{\hat{z}}[p(z)]$, for a given prior $p(z)$. When the Bayes risk is finite, the order of integration in Expression (14) can be interchanged. The Bayes estimator, $\hat{z}_B : \mathcal{Y} \to \mathcal{Z}$, is then formally defined by

$$\hat{z}_B = \arg\min_{\hat{z}} \mathrm{E}_Y \left\{ \mathrm{E}_{Z|Y} \left\{ \|Z - \hat{z}(Y)\|^2 \right\} \right\}$$

The problem can be solved for each y separately by minimizing

$$\hat{z}_B(y) = \arg\min_{\hat{z}} \mathrm{E}_{Z|y} \left\{ \|Z - \hat{z}(y)\|^2 \right\} \tag{15}$$

The major advantage of this expression is that the estimator only need to be found for the observation, y, actually obtained. The unique minimizer of Expression (15) is known to be the posterior expectation, that is

$$\hat{z}_B(y) = \mathrm{E}_{Z|y}\{Z\} \ . \tag{16}$$

This is the classical Bayes estimator. The averaging measure in Expression (15) and (16) is denoted the posterior distribution and can formally be written as

$$p(z|y) = \frac{p(y|z)p(z)}{p(y)}. \tag{17}$$

For the Bayesian analyst the posterior distribution is the answer to the inverse problem, since this contains his updated knowledge regarding the parameter. The knowledge can be used to produce the best estimate of a parameter according to a general loss function and to assess uncertainty regarding the parameter.

Expression (16) and (17) look quite convenient, but computation of these quantities can be difficult. In order to evaluate expectations under the posterior distribution in the general case various types of Monte Carlo integration can be used. The most common approach is Markov chain based techniques like Metropolis-Hastings (Robert and Casella 1999). One important special case is however analytically tractable and will be describe in grater detail below. In the special case the observations are related to a compact linear operator with additive error, see Expression (8), and the parameter, z, and the error, ε, are modeled as Gaussian random functions.

A Gaussian random function, Z, in a separable Hilbert space can be represented by the Karhunen-Loève expansion, see Yaglom (1987),

$$Z = \sum_{i=1}^{\infty} Z_i \, v_i \ ,$$

with $\{Z_i\}_{i=1}^{\infty}$ being independent Gaussian random variables with mean μ_i and variance γ_i^2; and $\{v_i\}_{i=1}^{\infty}$ being the corresponding basis elements of unit length. The pairs $\{\gamma_i^2, v_i\}_{i=1}^{\infty}$ is the eigensystem of the covariance operator of Z. This is the infinite dimensional equivalent of the eigenvalues and eigenvectors of the covariance matrix. For simplicity let $\{Z_i\}_{i=1}^{\infty}$ be centered, i.e. $\mu_i = 0$, $\forall\, i$. The observations are $y = Kz + \varepsilon$, see Expression (8), with $K : \mathcal{Z} \to \mathcal{Y}$ being a compact operator; and ε being an error term, modeled as a Gaussian random function. Assume further that ε have the Karhunen-Loève expansion

$$\varepsilon = \sum_{i=1}^{\infty} \varepsilon_i \, u_i$$

with $\{\varepsilon_i\}_{i=1}^{\infty}$ being centered independent Gaussian random variables with variance λ_i^2, and for presentational simplicity that K have the singular system $\{\sigma_i^2, v_i, u_i\}_{i=1}^{\infty}$, with v_i and u_i being identical to the basis elements in the Karhunen-Loève expansion of Z and ε respectively. The posterior random function $(Z|Y = y)$ can then be represented by the same Karhunen-Loève expansion as the prior, only with different coefficients. Defining $y_i = (u_i, y)$ this reads

$$(Z|Y = y) = \sum_{i=1}^{\infty} (Z_i|Y_i = y_i) \, v_i \ , \tag{18}$$

with $\{(Z_i|Y_i = y_i)\}_{i=1}^{\infty}$ being independent Gaussian random variables with mean $y_i\sigma_i/(\sigma_i^2 + \lambda_i^2/\gamma_i^2)$ and variance $\gamma_i^2[1 - \sigma_i^2/(\sigma_i^2 + \lambda_i^2/\gamma_i^2)]$. The optimal estimator can in this case be found explicitly as

$$\hat{z}(y) = \sum_{i=1}^{\infty} \frac{\sigma_i^2}{\sigma_i^2 + \lambda_i^2/\gamma_i^2} \frac{(u_i, y)}{\sigma_i} \, v_i$$

Note that this result is of the form in Expression (9).

Nonlinear inverse problems fits equally well into the Bayesian methodology, as linear. The optimal estimator under quadratic loss is again the conditional expectation, and the uncertainty is again described by the posterior distribution. There is no additional problem with identifiability since the posterior is a measure on the parameter space. There is however a computational cost which may be a severe obstacle.

7 Comparison of methodologies

In the previous sections inversion by regularization, statistical minimax inversion and Bayesian inversion, are presented as methodologies to solve inverse problems. In the current section these methodologies are compared.

In the presentation above inversion by regularization is presented as a mathematical approach whereas the other two are presented as statistical approaches.

This classification focuses the observation error, ε. In the mathematical approach the error is chosen by a malicious opponent that always makes the least favorable choice, whereas in the statistical approaches the error is considered random, hence it will change if the experiment is repeated.

Historically the mathematical and statistical approaches are developed separately and different languages have emerged. The result is that different names have been given to the same effect, and similar names have been given to different effects. The first is exemplified by uniqueness in the mathematical language and identifiability in the statistical. An example of the latter is that an admissible regularization strategy relate to an effect in the zero noise limit, while an admissible estimator relates to the performance of a particular estimator regardless of the noise level.

From a mathematical point of view the important notion for the solution is stability and convergence, which is implied by an upper bound on the estimation error in terms of the observation error. These bounds are seldom tight such that tight uncertainty bounds for the approximate solution can be derived. In a discrete problem the maximum estimation error of an approximate solution, can in theory be found using constrained optimization. This is however a hard problem to solve numerically. In the statistical literature the two estimation approaches justify two different strategies to assess the uncertainty. In the minimax approach few techniques are able to assess the uncertainty in the setting of inverse problems, the few rigorous methods stated above are limited to linear inverse problems. In the Bayesian approach the uncertainty is described by the posterior distribution, any probabilistic uncertainty statement regarding the parameter can be deduced from this distribution. Stark (1992) denote the Bayesian uncertainties as formal uncertainties, because they are based on an apriori assumption about the parameter that cannot be verified.

In many respects it is more natural to classify the methodologies by their view on the parameter. Inversion by regularization and statistical minimax inversion regard the parameter as a fixed quantity, while in the Bayesian approach it is considered to be random. In Donoho (1994) a related problem is investigated, a deep connection between two fixed parameter approaches corresponding to those above is found.

Results in any of the three methodologies, require that additional information is given. In the fixed parameter approaches this is done by imposing bounds on the parameter space such as Expression (13). In the Bayesian approach the information is given in terms of a probability measure on $\mathcal{Z}$. The Bayesian approach hence requires stronger assumptions, since the relative importance of any two elements in $\mathcal{Z}$ can be measured.

Consider also the achievement of the methodologies, within their own standard. In the Bayesian approach the optimal estimator under quadratic loss is well defined, i.e. $\mathrm{E}_{Z|y}\{Z\}$, hence it is a computational question to obtain the solution for given set of data. The fixed parameter approaches are more ambitious, but

only rate optimality is established.

Tikhonov regularization is frequently given a Bayesian interpretation, by defining

$$p(z) = const \times \exp\{-\tau^2 \alpha J(z)/2\}$$

and

$$p(y|z) = const \times \exp\{-\tau^2 \|Kz - y\|^2/2\}$$

with *const* being a generic normalizing constant; τ being a scaling factor; and $J(z)$, $\|Kz-y\|^2$ and α being as as for Expression (11). The posterior distribution is then

$$p(z|y) = const \times \exp\{-\tau^2(\|Kz - y\|^2 + \alpha J(z))/2\}$$

The value of z that maximizes this distribution is denoted the maximum posterior estimate. This is identical to the solution found by Tikhonov regularization, see Expression (11). Although this estimate for computational reasons is commonly used in Bayesian analysis, it is not a proper Bayes estimate in the cases considered here, since no proper loss function correspond to this estimator.

There is also a connection between the statistical minimax inversion and Bayesian inversion. The minimax approach can be seen as a game version of the Bayesian approach. In this game the scientist pick the estimator, $\hat{z}$, whereas the opponent may pick the prior distribution, $\pi(z)$, within a restricted class of distributions. The pay off in this game is the Bayes risk with the prior from the opponent, i.e. $B_{\hat{z}}[\pi(z)]$. This connection is an essential part in the theory of statistical minimax inversion.

8 Conclusions

The three methodologies are all successful for linear inverse problems, and the solutions look surprisingly similar.

The philosophical difference between mathematical approach and the statistical approaches is the nature of the observation error. The mathematical approach considers the worst case error. The statistical approaches regard the error as random. To choose one over the other based on this criterion is a philosophical debate of the nature of the error, ε. The error is caused by many sources. If all of the sources are of equal strength, the central limit theorem, can be used to argue the case for random errors. If some of the sources are dominant, this will produce a systematic error hence the mathematical philosophy would be preferable.

A practical difference between the mathematical approach and the two statistical approaches, is that stability is focused in the mathematical approach whereas uncertainty is focused in the statistical approach. There is a fundamental difference between the two notions, i.e. a solution can be stable and have large

uncertainty. In the authors opinion assessment of uncertainty is an important issue, hence he tends to favor the statistical approaches.

The Bayesian choice of prior distribution is usually criticized in traditional statistics. The critique is not as severe when it comes to the inverse problems considered here since information about the parameter must be included apriori in any case. Non informative prior distributions is in the author opinion only of interest for hyper parameters when inverse problems are considered, since inverse problems requires additional structure to be enforced. The Bayesian methodology achieves more and is more widely applicable, but the Bayesian assumptions regarding the parameters are stronger than that of minimax estimation. Whether the Bayesian achievements are worth the price of stronger assumptions is for the practitioner to decide.

From a purely statistical point of view the minimax estimator usually have better properties and should be preferred. On the other hand minimax estimators are only found for special cases, and are hence generally not available. Phenomena that are studied in inverse problems frequently have spatio-temporal structures, hence modeling the prior by random fields seem natural and may give the Bayesian estimates an advantage. The Bayesian methodology also guarantee the estimator to be admissible. Hence the Bayesian estimators can be used also by non-Bayesian that do not fully believe in the posterior distribution.

When it comes to aspects of uncertainty, the question is whether formal uncertainties are acceptable or not, keeping in mind that the alternative might be no assessment at all. In the authors opinion formal uncertainties are acceptable in any engineering application, but can be questioned for scientific purposes.

Non of the theories are fully developed in the nonlinear case. In the Bayesian approach the optimal estimator is known in theory, but there is no general way to compute it. The choice in the nonlinear case is frequently between Tikhonov regularization and the Bayesian approach, i.e. the maximum aposteriori estimate and the conditional expectation. The authors personal preference is the conditional expectation since this account for many reasonable solutions and is the one where loss criterion carry through to the final estimate also for nonlinear problems.

9 The thesis in the current context

The thesis is fully within the framework of Bayesian inversion, but parts has been inspired by work within the other solution frameworks.

The focus in the introduction is mainly on linear estimators for linear inverse problems. Nonlinear inverse problems and nonlinear estimators are presently subject to extensive research interest, but a complete theory for these type of problems is still lacking.

Paper I is devoted to a particular type of nonlinear inverse problem an practicalities for Bayesian inversion in this case. Paper II and III are specific applications of Paper I. A feature that is of particular interest in Paper III is that the the conditional expectation does not honor the data to the same extent as individual samples. This imply that the maximum posterior estimator is better adapted to the data than the conditional expectation. The reason for this is that the conditional expectation take all of the parameter space into account when producing the final estimator. The maximum posterior estimator identify only one extreme case among many possible. The conditional expectation hence produce a more robust summary of the posterior distribution.

In Paper IV the prior is formulated by smoothing independently scattered random measures. For the Gaussian case this relates to Expression (13) with B being the smoothing kernel. When the prior measure is discretized this correspond to the version presented in Neumaier (1998). The use of both a Gaussian random process and a Cauchy process may be seen as a Bayesian version of basis selection (Donoho and Stark 1989; Donoho and Huo 2001). The choices being a spike basis defined through the Cauchy process or a harmonic basis defined through the Gaussian processes.

The essence of the Paper V is contained in Expression (18). By using the Fourier transform each set of frequency components may be solved independently. The only difference to Expression (18) is that in Paper V several parameters are solved simultaneously using information from several angles. Hence a block version of the result is required. In the context of seismic inversion Expression (18) is obtained when the impedance is estimated from the zero offset data. The fact that the fast Fourier transform correspond to the singular vectors of the discretized problem makes the approach highly efficient.

References

Backus, G.E. "Confidence set inference with a prior quadratic bound", Geophys.J. 97:pp 119-150.

Bertero, M. (1989), "Linear inverse and ill-posed problems", Advances in electronics and electron physics, Vol. 75, pp 1-120.

Donoho, D.L. and Stark, P.B. (1989) "Uncertainty principles and signal recovery", SIAM J. Appl. Math. 49, no. 3, 906–931.

Donoho, D.L. and Huo, X. (2001) "Uncertainty principles and ideal atomic decomposition", IEEE Trans. Inform. Theory 47, no. 7, 2845–2862.

Donoho, D.L. (1994) "Statistical estimation and optimal recovery", Ann. Statist. Vol 22, no. 1, pp 238-270.

Donoho, D.L. (1995) "Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition", Appl. Comput. Harmon. Anal. Vol 2, no. 2, pp

101-126.

Engl, H.W., Hanke, M. and Neubauer, A. (1996) "Regularization of inverse problems", Mathematics and its Applications, 375. Kluwer Academic Publishers Group, Dordrecht.

Hadamard, J. (1902) "Sur les problèmes aux dérivées partielles et leur signification physique", Bull. Univ. Princeton 13, 49.

Hadamard, J. (1923) "Lectures on Cauchy problem i linear partial differential equations", Yale Univ. Press, New Haven.

Hansen, P.C. (1998) "Rank-deficient and discrete ill-posed problems", SIAM Monographs on Mathematical Modeling and Computation. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.

Johnstone, I.M. and Silverman, B.W. (1990) "Speed of estimation in positron emission tomography and related inverse problems", Ann. Statist. Vol 18. no. 1, pp 251-280.

Johnstone, I.M. and Silverman, B.W. (1991) "Discretization effects in statistical inverse problems", J. Complexity, Vol 7, no. 1, pp 1-34.

Kirsch, A. (1996) "An introduction to Mathematical theory of inverse problems", Vol. 127 of Applied Mathematical Series, Springer, New York.

Lehmann, E.L. and Casella, G. (1998) "Theory of point estimation", Second edition. Springer Texts in Statistics. Springer-Verlag, New York.

Lehmann, E.L. (1999) "Elements of large-sample theory", Springer Texts in Statistics. Springer-Verlag, New York.

Neumaier, A. (1998) Solving ill-conditioned and singular linear systems: A tutorial on regularization,SIAM Rev. 40 (1998), no. 3, 636–666

Robert, C.P. and Casella, G. (1999) "Monte Carlo statistical methods", Springer Texts in Statistics. Springer-Verlag, New York.

Stark P.B. (1992) "Minimax confidence intervals in geomagnetism", Geophys. J. Int. 108:pp 329-338.

Stark, P.B. (2000) "Inverse problems as statistics", Surveys on solution methods for inverse problems, pp 253-275, Springer, Vienna.

Stein, C. (1956) "Inadmissibility of the usual estimator for the mean of a multivariate normal distribution", Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Vol. I, pp. 197–206. University of California Press, Berkeley and Los Angeles.

Wahba, G. (1990) "Spline models for observational data", CBMS-NSF Regional Conference Series in Applied Mathematics, 59. Society for Industrial and Applied Mathematics (SIAM), Philadelphia.

Yaglom, A. M. (1987) "Correlation theory of stationary and related random functions", Vol. I. Basic results. Springer Series in Statistics. Springer-Verlag, New York

I

Bayesian inversion of piecewise affine operators in a Gaussian framework

Bayesian inversion of piecewise affine operators in a Gaussian framework

Odd Kolbjørnsen and Henning Omre

Department of mathematical sciences
Norwegian university of science and technology
Norway

Abstract

Piecewise affine inverse problems, are solved in a Bayesian framework with a Gaussian random field prior on the parameter space.

The inverse problem is to reconstruct the parameter when its mapping through a piecewise affine operator is observed, possibly with errors. A piecewise affine operator is defined by partitioning the parameter space and assign a specific affine operator to each part. Both problems with a discrete finite partition and a continuous partition are considered.

Piecewise affine inverse problems is a general class of nonlinear inverse problems, in particular inverse problems obeying certain variational structures, such as Fermat's principle in traveltime tomography, is of this type.

The main result is that the posterior distribution is found to be a mixture of truncated Gaussian distributions, and the expression for the mixing distribution is partially analytical tractable. The decomposition is used to propose a sampling algorithm. The algorithm is applicable also for problems with exact observations, for which generic sampling algorithms tends to fail.

KEY WORDS: *Bayesian statistic, Nonlinear inverse problem, Piecewise linear inverse problem, Sampling algorithm.*

1 Introduction

A feature of many inverse problems, that makes them unfit for the traditional statistical setting of parameter estimation in large sample theory (Lehmann 1999), is that the number of unknown parameters is of the same order or larger than the number of observations.

In the statistical literature on inverse problems of this type, two different risk criterions are used for evaluating estimators, being the minimax risk and the Bayesian risk. For the early theory of linear inverse problems, the two approaches essentially give the same solution. A classical minimax result of Pinsker (1980) in function estimation, prove that the solution obtained by quadratic regularization, (Tikhonov 1963), have the optimal rate of convergence in the zero noise limit. This result is extended to cover the setting of inverse problems (Johnstone and Silverman 1990). In the Bayesian approach a quadratic regularizer is formally equivalent with a Gaussian random field prior as described in Tarantola (1987) and Wahba (1990). The conditional expectation in this model will again be the solution found by regularization. This solution is also denoted the maximum posterior (MAP) solution since it is the mode in the posterior distribution. The minimax theory will not be pursued any further, although there are many resent results in this field, (Donoho 1995; Abramovich and Silverman 1998; Johnstone 1999).

In this article a Bayesian approach with a Gaussian random field prior is used, to solve piecewise affine inverse problems. In the Bayesian tradition the full posterior distribution is the solution to the inverse problem. The maximum posterior (MAP) solution is commonly used also for nonlinear inverse problems (O'Sullivan 1986; Tarantola 1987; Wahba 1990). For genuine nonlinear inverse problems this choice is not obvious, since the posterior frequently have multiple modes and the mode with the highest posterior value need not be the one that is most representative. Multiple random samples from the posterior yields a common Monte Carlo representation of the posterior, and is the approach used here.

In the current article piecewise affine inverse problems both with a finite and a continuous partition of the parameter space is considered. The main result is that the posterior is calculated to be a mixture of truncated Gaussian distributions in both cases. The main contribution of the article is the solution in the case of a continuous partition. This is a non trivial extension of the results for a finite partition. An algorithm which uses the decomposition to sample the posterior is proposed. The algorithm is based on rejection sampling, but the decomposition can be used more generally and is essential for efficiently exploiting the global structure of the inverse problem in any sampling algorithm.

The class of piecewise affine inverse problems is very general by its definition. In particular an inverse problem obeying a certain type of variational structure, can be phrased as a piecewise affine inverse problem. Fermat's principle in travel

time tomography (Berryman 1997) is of this type. The authors have applied the methodology to event migration in reflection seismic and nonlinear cross well tomography. In the current article two small examples of piecewise affine inverse problems are used to illustrate the concepts as they are introduced.

Example 1: The L^1 norm
Problem: A two dimensional vector is to be recovered, based on one observation of the L^1 norm of the vector. This is a low dimensional synthetic example that gives results that is easy to illustrate graphically. □

Example 2: The meteorologist.
Problem: A meteorologist is supposed to continuously monitor the temperature during a day. After reading the thermometer in the morning, he leaves it unattended the rest of the day. The next day when he return to the office, he reads the thermometer again. In the 24 hour period between the two readings, only the day maximum and the day minimum value can be obtained from the thermometer, not at what time they occurred. The meteorologist need to recover the entire temperature field of the previous day. □

To the authors knowledge there is no previous attempt to unify the treatment of piecewise affine inverse problems. The results for the finite index case, are however generally known and special cases are encountered by several authors, see Kolessa (1986) for this type of observations in a filtering setting. Liu and Chen (1998) also consider an example of this type in the setting of a particle filter. The results for the continuous index is a nontrivial extension of the finite index case, and is the main contribution of the current paper.

2 Bayesian approach to inverse problems

The recovery of a function z based on observations of a transform of the function is termed an inverse problem. When z is assumed to be in a function space $\mathcal{Z}$, the observations are $\boldsymbol{y} = \boldsymbol{K}(z) + \boldsymbol{\varepsilon}$, with $\boldsymbol{K} : \mathcal{Z} \to \mathbf{R}^r$ being the known forward map of the inverse problem, and $\boldsymbol{\varepsilon} \in \mathbf{R}^r$ being observation error.

Normal typing is used to denote scalars and scalar functions, while bold typing denote vectors and vector valued functions. A generic probability distribution is denoted $p\{\cdot\}$ and $P\{\cdot\}$ denotes a generic probability. Both the notation $p\{z\}$ and $p\{Z = z\}$ is used to denote the distribution of Z, the latter to emphasis the random variable in question. The notation $p\{z|\boldsymbol{Y} = \boldsymbol{y}\}$ denotes the conditional distribution of Z given $\boldsymbol{Y} = \boldsymbol{y}$.

The Bayesian approach to inverse problems, is in principle no different from the Bayesian approach to any other problem (Gelman et al 1995). The analyzer must specify a prior on the parameter space and a likelihood for the observations. The Bayesian framework is very flexible and it allows for stochastic modeling of

the prior distribution to quantify lack of knowledge and to merge observations from different sources.

The prior distribution, $p\{z\}$, is modeled by a Gaussian random field (Vanmarcke 1983) in this work.

Example 1 The L^1 norm continued
The random vector is assumed to have a bi-Gaussian prior distribution, hence $p\{\boldsymbol{z}\} = N_2(\boldsymbol{0}, \boldsymbol{\Sigma_Z})$ with $\sigma_{11} = \sigma_{22} = 1$ and $\sigma_{12} = \sigma_{21} = 0.5$. The contour lines of this density is plotted as ellipses in Figure 1. □

Example 2 The meteorologist continued
The meteorologist has developed a stochastic model for the temperature during one day, $t \in [0, 24]$;

$$Z(t) = Z_0 + Z_1 \sin\frac{\pi t}{12} + Z_2 \cos\frac{\pi t}{12} + \widetilde{Z}(t),$$

with Z_0, Z_1, Z_2 and $\widetilde{Z}(t)$ being stochastically independent; $p\{z_0\} = N(15, 5^2)$, $p\{z_1\} = N(5, 2^2)$, $p\{z_2\} = N(0, 0.5^2)$, and $\widetilde{Z}(t)$ is a zero mean residual Gaussian random process with covariance function $\mathrm{C}_{\widetilde{Z}}(t, t+h) = \exp\{-3|h/4|^2\}$. Hence Z is a Gaussian random process, being twice continuous differentiable for $t \in [0, 24]$. Figure 2 shows 200 samples from this prior distribution. □

The likelihood model is assumed to contain additive Gaussian error, hence

$$p\{\boldsymbol{y}|Z = z\} = N_r(\boldsymbol{K}(z), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^z),$$

with N_r being the r dimensional multi Gaussian distribution; $\boldsymbol{K}(z)$ being the forward map of the inverse problem and $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^z$ being the covariance of the observation error, note that this in general may depend on z.

The Bayesian answer to any question is contained in the posterior distribution, $p\{z|\boldsymbol{Y} = \boldsymbol{y}\}$, being the conditional distribution of the parameters given the data. The posterior distribution is formally proportional to the product of the prior and the likelihood,

$$p\{Z = z|\boldsymbol{Y} = \boldsymbol{y}\} \propto p\{Z = z\}p\{\boldsymbol{Y} = \boldsymbol{y}|Z = z\}\ . \tag{1}$$

The object is to sample the posterior distribution to obtain a Monte Carlo representation. The samples represent the posterior uncertainty of the inversion and can be combined to a single estimate by using a loss function (Gelman et al 1995). Several generic sampling algorithms are developed. Generic approaches only require the probability distribution to be known up to a normalizing factor and are hence ideal for sampling the distribution in Expression (1). Some generic approaches are, rejection sampling (von Neumann 1951), resampling schemes (Rubin 1988) and Markov chain based methods (Hastings 1970). All of these methods have merits in solving inverse problems, and there are numerous

variants of their implementations. The Markov chain based methods appear as the most general ones and are extensively used.

Frequently in inverse problems, it is such that some features are well determined from the likelihood model while others are poorly resolved. The result being that the posterior is concentrated along hyper surfaces in the parameter space. In this case most generic approaches become slow. Normally they fail for the case of exact nonlinear observations. The current work is primarily motivated by inverse problems where the observations have high precision.

3 Piecewise affine inverse problems

In the first part of this section, piecewise affine inverse problems are defined in general. In the following two parts, attention is drawn to two special cases where the piecewise affine operator have a finite and a continuous index. For both cases the posterior is calculated as a mixing distribution, and a sampling algorithm based on rejection sampling is proposed. For the case of a finite index the algorithm produces exact independent samples from the posterior distribution. For the case of a continuous index, the mixing distribution to be sampled is only known up to a normalizing constant. Given independent samples from this mixing distribution, independent samples from the posterior may be produced. The proposed algorithm is not the only way to exploit the decomposition. It is also discussed how the decomposition can be used to produce other sampling algorithms.

3.1 Problem Definition

Piecewise affine inverse problems are defined by the forward map being a piecewise affine operator.

Definition 1 (Piecewise affine operator) *An operator $\boldsymbol{K} : \mathcal{Z} \to \mathbf{R}^r$, is said to be piecewise affine, if it can be represented in the following way:*

$$\boldsymbol{K}(z) = \boldsymbol{K}_x z + \boldsymbol{k}_x \quad \textit{for} \quad z \in \mathcal{A}_x \; ; \quad x \in \mathcal{X}$$

with $\mathcal{X}$ being an index set, $\{\mathcal{A}_x\}_{x\in\mathcal{X}}$ being a partition of $\mathcal{Z}$; $\boldsymbol{K}_x : \mathcal{Z} \to \mathbf{R}^r$ being bounded linear operators on $\mathcal{Z}$ and $\boldsymbol{k}_x$ being r dimensional vectors. The indexed set of triplets $\{\mathcal{A}_x, \boldsymbol{K}_x, \boldsymbol{k}_x\}_{x\in\mathcal{X}}$ are the parameters of the piecewise affine operator.

This definition of piecewise affine operators is very general. Its usefulness depends on the index set $\mathcal{X}$. One special case is obtained by $\mathcal{X} = \{1\}$ in which the class of affine operators are obtained. The other extreme is letting $\mathcal{X} = \mathcal{Z}$,

in which any operator can be represented as piecewise affine. In this article intermediate groups are considered, by letting $\mathcal{X} = \{1, ..., m\}$ and $\mathcal{X} \subset \mathbf{R}^d$. The two examples above being one of each type. It will always be assumed that the operator is measurable with respect to the prior measure, $p\{z\}$, on $\mathcal{Z}$.

Define also the affine operators of a piecewise affine operator.

Definition 2 (The affine operators of a piecewise affine operator) *Let $\boldsymbol{K}$ be a piecewise affine operator with parameters $\{\mathcal{A}_x, \boldsymbol{K}_x, \boldsymbol{k}_x\}_{x\in\mathcal{X}}$. Then the affine operators $\{\boldsymbol{K}_x\}_{x\in\mathcal{X}}$ of $\boldsymbol{K}$ is defined by*

$$\boldsymbol{K}_x(z) = \boldsymbol{K}_x z + \boldsymbol{k}_x \ , \ z \in \mathcal{Z} \ ; \quad x \in \mathcal{X}$$

The affine operators of a given piecewise affine operator is hence defined by $\{\boldsymbol{K}_x, \boldsymbol{k}_x\}_{x\in\mathcal{X}}$, hence extending the affine pieces to the entire $\mathcal{Z}$.

The random variable, $\boldsymbol{Y}$, actually observed is defined through its conditional distribution,

$$p\{\boldsymbol{Y}|Z = z, \, z \in \mathcal{A}_x\} = N_r(\boldsymbol{K}_x z, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^x), \tag{2}$$

with $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^x$ being the covariance matrix of the observation error, possible dependent on the index. The marginal distribution of $\boldsymbol{Y}$ is not obvious due to the nonlinearity of $\boldsymbol{K}$. Define also the random variables $\{\boldsymbol{Y}_x\}_{x\in\mathcal{X}}$, that correspond to observe the affine operators of $\boldsymbol{K}$,

$$p\{\boldsymbol{Y}_x|Z = z\} = N_r(\boldsymbol{K}_x(z), \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^x) \ , \tag{3}$$

with $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^x$ being as for the likelihood in Expression (2). Note that the marginal distribution of $\boldsymbol{Y}_x$, is Gaussian, with parameters:

$$\begin{aligned} \boldsymbol{\mu}_{\boldsymbol{Y}_x} &= \mathrm{E}\{\boldsymbol{K}_x Z\} + \boldsymbol{k}_x \\ \boldsymbol{\Sigma}_{\boldsymbol{Y}_x} &= \mathrm{Cov}\{\boldsymbol{K}_x Z\} + \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^x \end{aligned}$$

In general $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^x$ may be singular, but $\boldsymbol{\Sigma}_{\boldsymbol{Y}_x}$ should be of full rank.

Example 1 The L^1 norm continued
The piecewise affine operator of this problem is the L^1 norm $\boldsymbol{K}(\boldsymbol{z}) = \|\boldsymbol{z}\|_1$, defined as:

$$\boldsymbol{K}(\boldsymbol{z}) = \begin{cases} z_1 + z_2 & z_1 \geq 0 \ ; \ z_2 \geq 0 \\ z_1 - z_2 & z_1 \geq 0 \ ; \ z_2 < 0 \\ -z_1 + z_2 & z_1 < 0 \ ; \ z_2 \geq 0 \\ -z_1 - z_2 & z_1 < 0 \ ; \ z_2 < 0 \end{cases} .$$

This problem has a discrete index $\mathcal{X} = \{1, 2, 3, 4\}$. In Figure 1, the contour $\|\boldsymbol{z}\|_1 = 2$ is plotted as a solid line, the extensions of the affine operators at the same level are plotted as dotted lines. Assume that $y = 2$ is observed and that the observation error is distributed as $N(0, 0.1^2)$. □

Example 2 The meteorologist continued
The meteorologist has observed the temperature each morning and the global extremes in between. To define this operator, consider its action on the function $z(t)$. The piecewise affine operator is:

$$\boldsymbol{K}(z) = \begin{cases} z(0) \\ z(t_{\max}) \\ z(t_{\min}) \\ z(24) \end{cases} \qquad z \in \mathcal{A}_{t_{\max},t_{\min}} \ ,$$

with

$$t_{\max} = \arg\max_{t\in[0,24]} z(t) \quad , \qquad t_{\min} = \arg\min_{t\in[0,24]} z(t) \ ,$$

and $\mathcal{A}_{t_{\max},t_{\min}} = \{z \in C^2([0,24]) : z(t_{\max}) \geq z(t) \geq z(t_{\min}) \ ; \ \forall t \in [0,24]\}$, where $C^2([0,24])$ denotes the functions on $[0,24]$ being two times continuously differentiable. Hence the operator is indexed by a continuous index $\mathcal{T} = [0,24] \times [0,24]$. The index correspond to the location where the maximum and the minimum occur, this is of course not observed. Assume $z(0) = 19.78$, $z(24) = 19.74$, $\max_{0\leq t\leq 24} z(t) = 26.04$ and $\min_{0\leq t\leq 24} z(t) = 16.61$ is observed, and that the observations are exact. The observations are indicated in Figure 2. □

3.2 Finite Index

The objective is to assess the posterior distribution. The full posterior will not be Gaussian due to the nonlinearity in $\boldsymbol{K}$, although within each $\mathcal{A}_x$ the posterior will be a truncated Gaussian distribution. The following theorem characterizes the posterior in terms of a mixing of these truncated Gaussians.

Theorem 1 (Finite partition) *Let Z be a Gaussian random field with distribution $p\{z\}$, such that $P\{Z \in \mathcal{Z}\} = 1$, and $\boldsymbol{K} : \mathcal{Z} \to \mathbf{R}^r$ be a $p\{z\}$-measurable piecewise affine operator with index set $\mathcal{X} \subset \boldsymbol{N}$ with $|\mathcal{X}| < \infty$ and parameters $\{\mathcal{A}_x, \boldsymbol{K}_x, \boldsymbol{k}_x\}_{x\in\mathcal{X}}$. Further let $\boldsymbol{Y}$ and $\{\boldsymbol{Y}_x\}_{x\in\mathcal{X}}$ be defined by Expressions (2) and (3) above. Assume further that:*

$$\begin{array}{ll} \forall x \in \mathcal{X} & \mathit{rank}\{\boldsymbol{\Sigma}_{\boldsymbol{Y}_x}\} = r \\ \exists x \in \mathcal{X} & P\{Z \in \mathcal{A}_x | \boldsymbol{Y}_x = \boldsymbol{y}\} > 0 \end{array}$$

Then,

$$p\{Z = z, Z \in \mathcal{A}_x | \boldsymbol{Y} = \boldsymbol{y}\} = \qquad (4)$$
$$p\{Z = z | \boldsymbol{Y}_x = \boldsymbol{y}, Z \in \mathcal{A}_x\} \times \frac{P\{Z \in \mathcal{A}_x | \boldsymbol{Y}_x = \boldsymbol{y}\} \cdot p\{\boldsymbol{Y}_x = \boldsymbol{y}\}}{p\{\boldsymbol{Y} = \boldsymbol{y}\}}$$

Proof :
A standard identity of conditional distributions is:

$$p\{Z = z, Z \in \mathcal{A}_x | \boldsymbol{Y} = \boldsymbol{y}\} = \frac{p\{Z = z | \boldsymbol{Y} = \boldsymbol{y}, Z \in \mathcal{A}_x\} p\{\boldsymbol{Y} = \boldsymbol{y}, Z \in \mathcal{A}_x\}}{p\{\boldsymbol{Y} = \boldsymbol{y}\}}.$$

By definition,

$$p\{Z = z | \boldsymbol{Y} = \boldsymbol{y}, Z \in \mathcal{A}_x\} = p\{Z = z | \boldsymbol{Y}_x = \boldsymbol{y}, Z \in \mathcal{A}_x\}$$

$$p\{Z \in \mathcal{A}_x, \boldsymbol{Y} = \boldsymbol{y}\} = p\{Z \in \mathcal{A}_x, \boldsymbol{Y}_x = \boldsymbol{y}\}$$

The result now follow by,

$$p\{Z \in \mathcal{A}_x, \boldsymbol{Y}_x = \boldsymbol{y}\} = p\{Z \in \mathcal{A}_x | \boldsymbol{Y}_x = \boldsymbol{y}\} p\{\boldsymbol{Y}_x = \boldsymbol{y}\} \qquad \text{QED.}$$

Note that the first term in Expression (4) is a truncated Gaussian distribution conditioned to the linear constraint $\boldsymbol{Y}_x = \boldsymbol{y}$. The second term gives the posterior probability of being in $\mathcal{A}_x$. The rank criterion is to assure that the mixing distribution on $\mathcal{X}$ do not have any singularities. The positivity criterion assures that there exists a solution to the problem.

The mixing distribution of Theorem 1, provides a sampling strategy to sample the posterior distribution.

Algorithm 1 *FTGM-algorithm (Finite Truncated Gaussian Mixing)*

1. *Sample* $x^* \sim q(x) \propto p\{\boldsymbol{Y}_x = \boldsymbol{y}\}$
2. *Sample* $Z^* \sim p\{Z | \boldsymbol{Y}_{x^*} = \boldsymbol{y}\}$
3. *If* $Z^* \in \mathcal{A}_{x^*}$ *stop.*

The algorithm is a variant of rejection sampling. The first objective is to obtain the posterior probability of being in $\mathcal{A}_x$. In Theorem 1 this probability is calculated as:

$$P\{Z \in \mathcal{A}_x | \boldsymbol{Y} = \boldsymbol{y}\} = \frac{P\{Z \in \mathcal{A}_x | \boldsymbol{Y}_x = \boldsymbol{y}\} \cdot p\{\boldsymbol{Y}_x = \boldsymbol{y}\}}{p\{\boldsymbol{Y} = \boldsymbol{y}\}} .$$

Note that $p\{\boldsymbol{Y} = \boldsymbol{y}\}$ is just the normalizing constant in the expression and $0 < \mathrm{P}\{Z \in \mathcal{A}_x | \boldsymbol{Y}_x = \boldsymbol{y}\} < 1$, hence proposing $\mathcal{A}_x$ according to $q(x) \propto p\{\boldsymbol{Y}_x = \boldsymbol{y}\}$ and accepting according to $\mathrm{P}\{Z \in \mathcal{A}_x | \boldsymbol{Y}_x = \boldsymbol{y}\}$, is a standard rejection algorithm for sampling $\mathcal{A}_x$. The algorithm does not give optimal acceptance rate for $q(x)$, since the acceptance probability in the algorithm $\alpha(x) = \mathrm{P}\{Z \in \mathcal{A}_x | \boldsymbol{Y}_x = \boldsymbol{y}\} < 1$, $\forall x \in \mathcal{X}$. The value $\mathrm{P}\{Z \in \mathcal{A}_x | \boldsymbol{Y}_x = \boldsymbol{y}\}$ is in

general not straight forward to calculate, however an event with this probability can be obtained by sampling Z^* from $p\{z|\boldsymbol{Y}_x = \boldsymbol{y}\}$ and accept if $Z^* \in \mathcal{A}_x$. The second objective is to sample $p\{z|\boldsymbol{Y}_x = \boldsymbol{y}, Z \in \mathcal{A}_x\}$, this is a truncated Gaussian distribution. By sampling $p\{z|\boldsymbol{Y}_x = \boldsymbol{y}\}$ and accepting if $Z \in \mathcal{A}_x$ this distribution is sampled correctly. The two acceptance criterions are the same for the two objectives of the algorithm, hence they are coupled in the final algorithm.

Due to the positivity constraint in Theorem 1, the algorithm produces a sample in finite time. The acceptance rate of the algorithm, can be calculated as:

$$p_{accept} = \frac{p\{\boldsymbol{Y} = \boldsymbol{y}\}}{\sum_{x \in \mathcal{X}} p\{\boldsymbol{Y}_x = \boldsymbol{y}\}}. \tag{5}$$

In comparison, the standard rejection algorithm with proposals drawn from the prior and accepted with probability proportional to the likelihood, have acceptance rate:

$$p_{accept} = p\{\boldsymbol{Y} = \boldsymbol{y}\}(2\pi)^{r/2}|\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}|^{1/2}, \tag{6}$$

for the case with fixed $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$. This is heavily dependent on the scale of the observation error. For the sampling algorithm proposed, the size of the observation error is of secondary importance. Resampling techniques and algorithms based on naive use of Markov Chains, also have a reduced performance when the observations are precise. Since $\mathcal{X}$ contains a finite number of states, $q(x)$ can be calculated exactly. The FTGM-algorithm is hence fully specified and provides exact independent samples from the posterior.

Example 1. The L^1 norm continued
The FTGM-algorithm is implemented for this example. Figure 1 shows a scatter plot of the samples obtained using the algorithm. The posterior have four modes which are visible in the scatter plot. Note that the algorithm is exact and the samples are independent. The observed acceptance rate is 88.5% in this example.

From Expression (5) it is clear that the acceptance probability is dependent on the observed value, y. Figure 3, shows this dependence for Example 1. For comparison the acceptance rate for naive rejection sampling is plotted as a dotted line in the same figure. □

The fact that the FTGM algorithm produces independent samples from the posterior is appealing, but in practical situations the acceptance rate $\mathrm{P}\{Z \in \mathcal{A}_x|\boldsymbol{Y}_x = \boldsymbol{y}\}$ may be very small. An example of this is found in Example 1 when y is small, see Figure 3. The decomposition of Theorem 1 imply that

$$p\{Z = z, Z \in \mathcal{A}_x|\boldsymbol{Y} = \boldsymbol{y}\} \propto p\{Z = z|\boldsymbol{Y}_x = \boldsymbol{y}\}\, p\{\boldsymbol{Y}_x = \boldsymbol{y}\}\, I\{z \in \mathcal{A}_x\},$$

with $I\{z \in \mathcal{A}_x\}$ being one if $z \in \mathcal{A}_x$, zero otherwise. This can be exploited to develop algorithms that are based on the generic principles, but specialized

to the inverse problem at hand. The idea being the same as for the FTGM - algorithm, first to sample the index, next exploit the affine structure to propose a sample.

3.3 Continuous Index

The objective is to assess the posterior distribution. In the case of a finite index one is free to choose the operators $\boldsymbol{K}_x$ and the partition $\{\mathcal{A}_x\}_{x\in\mathcal{X}}$, as long as the resulting operator $\boldsymbol{K}(z)$ is measurable. For a continuous partition further specifications are needed.

Definition 3 (Index continuity of operator) *An indexed set of bounded operators, $\boldsymbol{K}_x : \mathcal{Z} \to \mathbf{R}^r$ with $x \in \mathcal{X} \subset \mathbf{R}^d$ is continuous in the index with respect to a probability measure $p\{z\}$ on $\mathcal{Z}$, if:*

$$P\Big\{ \lim_{\Delta x \to 0} \|\boldsymbol{K}_x Z - \boldsymbol{K}_{x+\Delta x} Z\| = 0 \,;\, \forall x \in \mathcal{X} \Big\} = 1$$

That is, define a d-parameter, r-dimensional random field on $\mathcal{X}$, by for each $x \in \mathcal{X}$ associating the r-dimensional vector $\boldsymbol{K}_x Z$. Definition 3 implies sample path continuity for this field on $\mathcal{X}$.

Definition 4 (Restricted linear partition) *A partition $\{\mathcal{A}_x\}_{x\in\mathcal{X}}$ with $\mathcal{X} \subset \mathbf{R}^d$ of $\mathcal{Z}$ is a restricted linear partition if,*

$$\mathcal{A}_x = \{\boldsymbol{R}_x z + \boldsymbol{r}_x = \mathbf{0}\} \cap \mathcal{C}_x,$$

with $\boldsymbol{R}_x : \mathcal{Z} \to \mathbf{R}^d$ being a bounded linear operator; $\boldsymbol{r}_x$ being a d-dimensional vector function; and $\mathcal{C}_x$ is any subset of $\mathcal{Z}$. The set of triplets $\{\boldsymbol{R}_x, \boldsymbol{r}_x, \mathcal{C}_x\}_{x\in\mathcal{X}}$ are the parameters of the restricted linear partition.

Hence for a restricted linear partition, having parameters $\{\boldsymbol{R}_x, \boldsymbol{r}_x, \mathcal{C}_x\}_{x\in\mathcal{X}}$, the part $\mathcal{A}_{x^*}$ contain only those z's in the subset $\mathcal{C}_{x^*}$ for which $\boldsymbol{R}_{x^*} z + \boldsymbol{r}_{x^*} = \mathbf{0}$. Further regularity conditions must be assumed about the restricted linear partition with respect to the prior measure on $\mathcal{Z}$.

Definition 5 (Regular restricted linear partition) *Let $Leb(\cdot)$ denote the Lebesgue-measure on $\mathbf{R}^d$. A restricted linear partition with parameters $\{\boldsymbol{R}_x, \boldsymbol{r}_x, \mathcal{C}_x\}_{x\in\mathcal{X}}$ is regular with respect to a prior measure $p\{z\}$ on $\mathcal{Z}$ if:*

i) $rank(Cov\{\boldsymbol{R}_x Z\}) = d\,;\ \forall x \in \mathcal{X}$

ii) $\boldsymbol{\nabla}_x \boldsymbol{R}_x : \mathcal{Z} \to \mathbf{R}^{d\times d}$ *is a bounded linear operator,continuous in index.*

iii) $Leb(\{x : \det\{\boldsymbol{\nabla}_x \boldsymbol{R}_x Z\} = 0\}) = 0 \quad a.s.$

iv) $\boldsymbol{r}_x$ *is continuously differentiable in each component.*

v) $\mathcal{C}_x$ *is* $p\{z|\boldsymbol{R}_x Z + \boldsymbol{r}_x = \mathbf{0}\}$ *measurable*

vi) $\exists \mathcal{B} \in \mathcal{X}$ *with* $Leb(\mathcal{B}) > 0 : P\{Z \in \mathcal{C}_x|\boldsymbol{R}_x Z + \boldsymbol{r}_x = \mathbf{0}\} > 0\,;\ \forall \boldsymbol{x} \in \mathcal{B}$

For a regular restricted linear partition, the linear part, $\boldsymbol{R}_x Z + \boldsymbol{r}_x$, constitutes a non stationary d - dimensional Gaussian random field on $\mathcal{X}$. The restriction, $\boldsymbol{R}_x Z + \boldsymbol{r}_x = \mathbf{0}$, in the partition are the zero crossings of this random field. The level crossings of $\boldsymbol{R}_x Z + \boldsymbol{r}_x$ is a point process due to the match of the dimension of the parameter space and the value space of the process. For one specific realization z each point is also marked with $\boldsymbol{I}\{z \in \mathcal{C}_x\}$, the indicator of $\mathcal{C}_x$ being one if $z \in \mathcal{C}_x$, zero otherwise. This mark determines the value of x uniquely among the points of the point process, hence determines the part, $\mathcal{A}_x$, which z belongs to. Let this be illustrated by Example 2.

Example 2 The meteorologist continued
Assume the extreme values do not occur at the boundary. The global maximum and minimum during the 24 hour period are hence obtained in the interior of the region at critical points. Since z is twice continuously differentiable, the derivative of z is zero at critical points, hence the parameters of the partition are :

$$\begin{aligned} \boldsymbol{R}_{t_{\max},t_{\min}} z &= [z'(t_{\max}), z'(t_{\min})] \\ \boldsymbol{r}_{t_{\max},t_{\min}} &= \mathbf{0} \\ \mathcal{C}_{t_{\max},t_{\min}} &= \mathcal{A}_{t_{\max},t_{\min}}, \end{aligned}$$

with $\mathcal{A}_{t_{\max},t_{\min}}$ as previously defined. The partition is regular with probability one. Criterion *i)* is fulfilled if $t_{\max} \neq t_{\min}$, this possibility is ruled out since equality corresponds a constant temperature profile. Criterion *ii)* and *iii)* are fulfilled since:

$$\boldsymbol{\nabla}_{t_{\max},t_{\min}} \boldsymbol{R}_{t_{\max},t_{\min}} z = \begin{bmatrix} z''(t_{\max}) & 0 \\ 0 & z''(t_{\min}) \end{bmatrix}$$

Criterion *iv)* is fulfilled by the definition of $\boldsymbol{r}_{t_{\max},t_{\min}}$. Criterion *v)* is fulfilled since z is twice continuous differentiable. The number of critical points within the domain will be finite with probability one. By assumption one of these points are the global maximum and one the global minimum point hence criterion *vi)* is also fulfilled. □

In the continuous index problem, an extended piecewise affine operator is introduced.

Definition 6 (Extended piecewise affine operator) *For a piecewise affine operator defined on a continuous domain, $\mathcal{X}$, having the parameters $\{\mathcal{A}_x, \boldsymbol{K}_x, \boldsymbol{k}_x\}_{x\in\mathcal{X}}$ and a regular restricted linear partition with parameters $\{\boldsymbol{R}_x, \boldsymbol{r}_x, \mathcal{C}_x\}_{x\in\mathcal{X}}$ the extended piecewise affine operator is defined by:*

$$\tilde{\boldsymbol{K}}(z) = \tilde{\boldsymbol{K}}_x z + \tilde{\boldsymbol{k}}_x \quad for \;\; z \in \mathcal{A}_x \; ; \quad x \in \mathcal{X},$$

$$\tilde{\boldsymbol{K}}_x z = \begin{bmatrix} \boldsymbol{K}_x z \\ \boldsymbol{R}_x z \\ \boldsymbol{\nabla}_x \boldsymbol{R}_x z \end{bmatrix} \;\; ; \;\; \tilde{\boldsymbol{k}}_x = \begin{bmatrix} \boldsymbol{k}_x \\ \boldsymbol{r}_x \\ \boldsymbol{\nabla}_x \boldsymbol{r}_x \end{bmatrix}$$

The random variables that correspond to observe the affine operators of the extended piecewise affine operator, is denoted $\tilde{\boldsymbol{Y}}_x$ and are defined through the conditional distribution,

$$p\{\tilde{\boldsymbol{Y}}_x | Z = z, \, z \in \mathcal{A}_x\} = N_{r+d+d^2}(\tilde{\boldsymbol{K}}_x z, \tilde{\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}}^{x}), \tag{7}$$

with $\tilde{\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}}^{x}$ being a matrix consisting of $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{x}$ in the upper left corner and zeros otherwise.

Theorem 2 (Continuous partition) *Let Z be a Gaussian random field with distribution $p\{z\}$, such that $P\{Z \in \mathcal{Z}\} = 1$. Further let $\boldsymbol{K} : \mathcal{Z} \to \mathbf{R}^r$ be a $p\{z\}$-measurable piecewise affine operator with index set $\mathcal{X} \subset \mathbf{R}^d$, $0 < Leb(\mathcal{X}) < \infty$ and parameters $\{\mathcal{A}_x, \boldsymbol{K}_x, \boldsymbol{k}_x\}_{x\in\mathcal{X}}$, having a regular restricted linear partition, with parameters $\{\boldsymbol{R}_x, \boldsymbol{r}_x, \mathcal{C}_x\}_{x\in\mathcal{X}}$. Denote the parameters of the extended piecewise affine operator by $\{\mathcal{A}_x, \tilde{\boldsymbol{K}}_x, \tilde{\boldsymbol{k}}_x\}_{x\in\mathcal{X}}$. Let $\boldsymbol{Y}$, $\{\tilde{\boldsymbol{Y}}_x\}_{x\in\mathcal{X}}$ be as defined by Expressions (2) and (7) above. Assume $\tilde{\boldsymbol{K}}_x$ is continuous in index; $\tilde{\boldsymbol{k}}_x, \tilde{\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}}^{x}$ being continuously dependent on x; and that $rank\{Cov\{\boldsymbol{\nabla}_x \boldsymbol{R}_x Z\}\} = n$. Assume further:*

$$\begin{aligned} &rank\{Cov\{\tilde{\boldsymbol{Y}}_x\}\} = r + d + n \; ; \forall \; x \in \mathcal{X} \\ &\exists \mathcal{B} \times \mathcal{R} \subset \mathbf{R}^d \times \mathbf{R}^n \; with \; Leb(\mathcal{B} \times \mathcal{R}) > 0 : \quad \forall (x, \boldsymbol{r}) \in \mathcal{B} \times \mathcal{R} \\ &\qquad P\{Z \in \mathcal{C}_x | \, \tilde{\boldsymbol{Y}}_x = (\boldsymbol{y}, \boldsymbol{0}, \boldsymbol{r}) \, \} > 0 \end{aligned}$$

Then,

$$\begin{aligned} &p\{Z = z, \boldsymbol{\nabla}_x(\boldsymbol{R}_x Z + \boldsymbol{r}_x) = \boldsymbol{r}, Z \in \mathcal{C}_x, | \boldsymbol{Y} = \boldsymbol{y}\} \\ &= p\{Z = z | \, \tilde{\boldsymbol{Y}}_x = (\boldsymbol{y}, \boldsymbol{0}, \boldsymbol{r}), Z \in \mathcal{C}_x \, \} \\ &\quad \times \; P\{Z \in \mathcal{C}_x | \, \tilde{\boldsymbol{Y}}_x = (\boldsymbol{y}, \boldsymbol{0}, \boldsymbol{r}) \, \} \times \frac{|\det(\boldsymbol{r})| p\{ \, \tilde{\boldsymbol{Y}}_x = (\boldsymbol{y}, \boldsymbol{0}, \boldsymbol{r}) \, \}}{p\{\boldsymbol{Y} = \boldsymbol{y}\}} \end{aligned} \tag{8}$$

The first term in Expression (8) is a truncated Gaussian distribution. The second term is the acceptance criterion and the third term is the proposal distribution on $\mathcal{X} \times \mathbf{R}^n$. Note that there are only linear equality constraints in the

first term and that the second and third term is the posterior density for being in $\mathcal{A}_x$ with $\boldsymbol{\nabla}_x(\boldsymbol{R}_x Z + \boldsymbol{r}_x)$ having the value $\boldsymbol{r}$.

In comparison to the finite index case, Expression (8) contains the determinant of the Jacobian. Due to this factor, the mixing distribution must be extended. Instead of just drawing the region $\mathcal{A}_x$ containing z, the value of the Jacobian of the linear part of the restriction, $\boldsymbol{R}_x Z + \boldsymbol{r}_x$, must be sampled simultaneously. This is also the feature that makes the continuous index a nontrivial extension of the finite index case. The appearance of such a determinant when turning from the case of a discrete index, to the case of a continuous index, is similar to what appears for transforms of random variables. The proof of Theorem 2 is left to Appendix A.

The mixing distribution of Theorem 2, provides a sampling strategy for the posterior.

Algorithm 2 *CTGM-algorithm (Continuous Truncated Gaussian Mixing)*

1. *Sample* $x^*, \boldsymbol{r}^* \sim q(x, \boldsymbol{r})$ *with* $q(x, \boldsymbol{r}) \propto |\det(\boldsymbol{r})| p\{\tilde{\boldsymbol{Y}}_x = (\boldsymbol{y}, \boldsymbol{0}, \boldsymbol{r})\}$
2. *Sample* $Z^* \sim p\{Z | \tilde{\boldsymbol{Y}}_{x^*} = (\boldsymbol{y}, \boldsymbol{0}, \boldsymbol{r}^*)\}$
3. *If* $Z^* \in \mathcal{C}_{x^*}$ *stop.*

The algorithm is a variant of rejection sampling, and works exactly as for the case of a finite index. The only difference is that the index is extended. The acceptance rate of the algorithm can be calculated as

$$p_{accept} = \frac{p\{\boldsymbol{Y} = \boldsymbol{y}\}}{\int_{\mathcal{X}} \int_{\mathbf{R}^{d^2}} |\det(\boldsymbol{r})| p\{\tilde{\boldsymbol{Y}}_x = (\boldsymbol{y}, \boldsymbol{0}, \boldsymbol{r})\} \, d\boldsymbol{r} \, dx}$$

The difference from the finite index case, is that the normalizing constant of $q(x, \boldsymbol{r})$ is unknown, hence the proposal distribution for the mixing, must be sampled by the use of generic algorithms.

Example 2. The meteorologist continued
The CTGM-algorithm is implemented for this case. The proposal distribution q is sampled using a SIR-algorithm (Rubin 1988) using 500000 proposals. Figure 4 show 400 independent samples from the SIR approximation to the posterior. The true curve is plotted in white. The acceptance rate in step 3 of the CTGM-algorithm is observed to be 43.8% in this example. The correct distribution is sampled within the discretization effect of the SIR-algorithm, even for this case with exact observations. The naive use of any generic algorithm will fail for this case, due to the exactness of the observations. □

The acceptance rate in step three of the CTGM-algorithm may become small, the problem is however not as severe as for the finite index case since additional

information is provided, i.e. $\{\boldsymbol{R}_x Z + \boldsymbol{r}_x = \boldsymbol{0}\}$. The decomposition in Theorem 2 may be used to describe the following identity,

$$p\{Z = z, \boldsymbol{\nabla}_x(\boldsymbol{R}_x Z + \boldsymbol{r}_x) = \boldsymbol{r}, Z \in \mathcal{C}_x, |\boldsymbol{Y} = \boldsymbol{y}\}$$

$$\propto |\det(\boldsymbol{r})|\, p\{Z = z|\tilde{\boldsymbol{Y}}_x = (\boldsymbol{y}, \boldsymbol{0}, \boldsymbol{r})\} p\{\tilde{\boldsymbol{Y}}_x = (\boldsymbol{y}, \boldsymbol{0}, \boldsymbol{r})\}\, I\{z \in \mathcal{C}_x\}\,,$$

with $I\{z \in \mathcal{C}_x\}$ being one if $z \in \mathcal{C}_x$, zero otherwise This result can again be used to specialize a generic algorithm to the problem at hand.

Since the current approach requires the use of generic algorithms, the direct use of such algorithms to sample Z should be considered an alternative. The current approach compare favorable in two respects: firstly it reduces the number of parameters that have nonlinear relations to a minimum; secondly under the assumptions of Theorem 2, the distributions to be sampled are nonsingular even if the observations are exact, for this case the direct use of a generic algorithm fails and provides no alternative.

4 Conclusions and discussion

A Bayesian approach to solve piecewise affine inverse problems, is developed. Piecewise affine inverse problems have an intuitive definition and are easy to picture mentally. Although piecewise affine inverse problems only provide a small step into the world of nonlinearity, they possess genuine nonlinear features. Piecewise affine inverse problems constitutes a large class of problems, including travel time tomography and event migration of travel times from reflection seismic.

Both problems with a finite and a continuous index are considered. The general result is the decomposition of the posterior distribution as a mixture of truncated Gaussian distributions in both cases. The general formulation has to the authors knowledge not previously appeared, although results for a finite index are generally known. The results for a continuous index, is however a nontrivial extension of those for the finite index.

An algorithm that uses the decomposition and is based on rejection sampling is proposed. When tested on small example problems, the algorithm gives reasonable acceptance rates, and provides solutions to problems that can not be solved by direct use of generic sampling algorithms.

The decomposition can also be used to develop more sophisticated generic sampling algorithms, or to obtain a different goal than to sample the posterior distribution, for example to estimate expectations of functionals of Z. Contrary to direct use of generic algorithms, algorithms that makes use of the decomposition exploits the global structures of the inverse problem. Major benefit of using the proposed approach is expected to appear when the observations have high precision.

The algorithm has similarities to an auxiliary variable approach, but is not so in a strict sense since the index introduced is a function of the parameter.

The theory is developed for Gaussian random field priors, but can easily be extended to include mixtures of Gaussian random fields.

Acknowledgments

The work is supported by a PhD grant from the Research Council of Norway.

References

Abramovich, F., and Silverman B.W. (1998), "Wavelet decomposition approaches to statistical inverse problems," *Biometrika*, 85, 115-129.

Adler, R.J. (1981), "The geometry of random fields"; Wiley.

Berryman, J.G.(1997), "Variational structure of inverse problems in wave propagation and vibration," in *Inverse Problems in Wave Propagation*, ed. G. Chavent, G. Papanicolaou, P. Sacks, and W. W. Symes; New York; Springer; pp. 13-44.

Donoho, D.L. (1995), "Nonlinear solution of linear inverse problems by wavelet-vaguelet decomposition," *Applied and computational Harmonic analysis*, 2, 101-126.

Gelman, A., Carlin, J., Stern, H., and Rubin, D.B. (1995), "Bayesian data analysis"; London; Chapman & Hall.

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

Johnstone, I.M. and Silverman, B.W. (1990) "Speed of estimation in positron emission tomography and related inverse problems." Ann. Statist. Vol 18, 1, pp 251-280.

Johnstone, I. M. (1999), "Wavelet shrinkage for Correlated data and inverse problems: adaptivity results," *Statistica Sinca*, 1, 51- 83.

Kolessa A.E. (1986), "Recursive filtering algorithms for systems with piecewise linear non-linearities",*Avtomat.Telemekh.*, 4, 48-55.

Lehmann, E.L. (1999), "Elements of large-sample theory"; New York; Springer.

Liu J.S., and Chen,R. (1998), "Sequential Monte Carlo Methods for Dynamical Systems," *Journal of the American Statistical Association*, 93, 1032-1044.

O'Sullivan, F. (1986), "A statistical perspective on ill-posed inverse problems," *Statistical Science*, 1, 502-518.

von Neumann, J. (1951), "Various techniques in connection with random digits," *NBS Appl.Math Ser.*, 12, 36-38.

Pinsker, M. (1980), "Optimal filtering of square integrable signals in Gaussian white noise," *Problems of Information Transmission* 16, pp. 120-133. Originally in Russian in, *Problemy Peredatsii Informatsii*, 16, pp. 52-68.

Rubin, D.B. (1988), "Using SIR algorithm to simulate posterior distributions," in *Bayesian statistics 3* ed. J.M. Bernardo, M.H. DeGroot, D.V. Lindley, and A.F.M. Smith; New York: Oxford University Press; pp. 395-402.

Tarantola, A. (1987), "Inverse problem Theory"; Elsevier.

Tikhonov, A. (1963), "Solution to incorrectly formulated problems and the regularization method," *Soviet Math. Doklady*, 5, 1035-1038.

Vanmarcke, E. (1983), "Random fields"; The MIT press.

Wahba, G. (1990), "Spline models for Observational data", *NSF-CBMS Regional Conference in Mathematics*, (Vol. 59); Philadelphia; SIAM.

A Proof of Theorem 2

To simplify notation, define the random variable X implicitly by $Z \in \mathcal{A}_{X(Z)}$, and the random vector fields $\boldsymbol{R}(x) = \boldsymbol{R}_x Z + \boldsymbol{r}_x$, $\boldsymbol{Y}(x) = \boldsymbol{Y}_x$; $x \in \mathcal{X}$. In the current notation the content of Theorem 2 is restated as:

$$p\{X = x, \boldsymbol{Y} = \boldsymbol{y}\} = \int_{\mathbf{R}^{d^2}} |det(\boldsymbol{r})| P\{Z \in \mathcal{C}_x | \boldsymbol{R}(x) = \boldsymbol{0}, \boldsymbol{\nabla}_x \boldsymbol{R}(x) = \boldsymbol{r}, \boldsymbol{Y}(x) = \boldsymbol{y}\} \times p\{\boldsymbol{R}(x) = \boldsymbol{0}, \boldsymbol{\nabla}_x \boldsymbol{R}(x) = \boldsymbol{r}, \boldsymbol{Y}(x) = \boldsymbol{y}\}\, d\boldsymbol{r}\,.$$

Proof :
Consider the following identity

$$P\{X \in \mathcal{B} \cap \boldsymbol{Y} \in \Omega\} = P\{\exists x \in \mathcal{B} : \boldsymbol{R}(x) = \boldsymbol{0}, \boldsymbol{Y}(x) \in \Omega, Z \in \mathcal{C}_x\}$$

Define a random counting process by

$$N_\Omega(\mathcal{B}) = \#\{x \in \mathcal{B} : \boldsymbol{R}(x) = \boldsymbol{0}, \boldsymbol{Y}(x) \in \Omega, Z \in \mathcal{C}_x\}$$

Note that $N_\Omega(\mathcal{B}) \in \{0, 1\}$ since $\{\boldsymbol{R}(x) = \boldsymbol{0}\} \cap \mathcal{C}_x$ constitutes a partition, hence:

$$\mathrm{E}\{N_\Omega(\mathcal{B})\} = P\{\exists x \in \mathcal{B} : \boldsymbol{R}(x) = \boldsymbol{0}, \boldsymbol{Y}(x) \in \Omega, Z \in \mathcal{C}_x\}$$

This expectation can be obtained under suitable regularity conditions for the random fields involved (Adler 1981). First note that

$$N_\Omega(\mathcal{B}) = \lim_{\varepsilon \to 0} \int_{\mathcal{B}} \delta_\varepsilon(\boldsymbol{R}(x)) I(Z \in \mathcal{C}_x \cap \boldsymbol{Y}(x) \in \Omega) |\det(\boldsymbol{\nabla}_x \boldsymbol{R}(x))|\, dx\,, \qquad (9)$$

with δ_ε being a delta sequence; $I(\cdot)$ being the indicator function and $\boldsymbol{\nabla}_x \boldsymbol{R}(x)$ being the Jacobian of $\boldsymbol{R}(x)$. The result is obtained by integrating the delta sequence in the value set of $\boldsymbol{R}(x)$ and then flip the variable of integration to the definition set, resulting in a local Jacobian. The zero crossings of $\boldsymbol{R}(x)$ is a point process on $\mathcal{X}$. Each point in the point process, is marked with $I(Z \in \mathcal{C}_x \cap \boldsymbol{Y}(x) \in \Omega)$, $N_\Omega(\mathcal{B})$ count only those points for which this mark have the value 1, hence $N_\Omega(\mathcal{B})$ is a thinning of the point process being the zero crossings of $\boldsymbol{R}(x)$. The expectation of Expression (9) is obtained by integrating the limit of the expected value of the integrand. The interchange of limit, integration and expectation is valid under the regularity conditions of Theorem 2 according to Adler (1981).

$P\{N_\Omega(\mathcal{B}) = 1\} =$

$$\int_{\mathcal{B}} E\left\{ |\det(\boldsymbol{\nabla}_x \boldsymbol{R}(x))| \cdot I(Z \in \mathcal{C}_x \cap \boldsymbol{Y}(x) \in \Omega) \,|\boldsymbol{R}(x) = \boldsymbol{0} \right\} \times p\{\boldsymbol{R}(x) = \boldsymbol{0}\}\, dx$$

The final result is obtained by using the sentence of double expectation twice to get

$P\{N_\Omega(\mathcal{B}) = 1\} =$

$$\int_{\mathcal{B}} \int_{\Omega} \int_{\mathbf{R}^{d^2}} |det(\boldsymbol{r})| \cdot E\left\{ I(Z \in \mathcal{C}_x) \,|\boldsymbol{R}(x) = \boldsymbol{0}, \boldsymbol{Y}(x) = \boldsymbol{y}, \boldsymbol{\nabla}_x \boldsymbol{R}(x) = \boldsymbol{r} \right\} \times p\{\boldsymbol{R}(x) = \boldsymbol{0}, \boldsymbol{Y}(x) = \boldsymbol{y}, \boldsymbol{\nabla}_x \boldsymbol{R}(x) = \boldsymbol{r}\}\, d\boldsymbol{r}\, d\boldsymbol{y}\, dx$$

QED.

Note that the dimension of the mixing distribution in the general case is $d^2 + d$, but for most problems there will be many zeros off the diagonal in the Jacobian. The effective dimension of the mixing distribution will therefore usually be of order d.

B Tables and figures

Table 1: Parameters of the piecewise affine operator in Example 1.

x	$\mathcal{A}_x$	$\boldsymbol{K}_x$	k_x
1	$\{z_1 \geq 0\ ;\ z_2 \geq 0\}$	$[\ 1,\ \ 1]'$	0
2	$\{z_1 \geq 0\ ;\ z_2 < 0\}$	$[\ 1, -1]'$	0
3	$\{z_1 < 0\ ;\ z_2 \geq 0\}$	$[-1,\ \ 1]'$	0
4	$\{z_1 < 0\ ;\ z_2 < 0\}$	$[-1, -1]'$	0

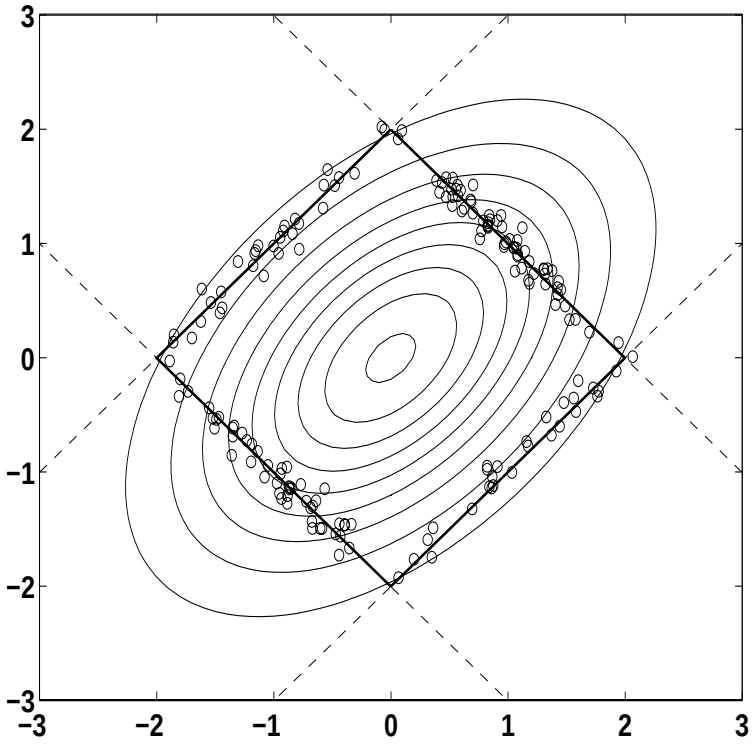


Figure 1: Conditioning to the L^1 norm. A scatter plot containing 177 values sampled from the posterior by the use of FTGM-algorithm. Superimposed on this is the contour lines from prior distribution; a solid square showing the contour of the piecewise affine operator at the observed level, $\|\boldsymbol{z}\|_1 = 2$; and dashed lines showing the extension of the affine operators at the same level.

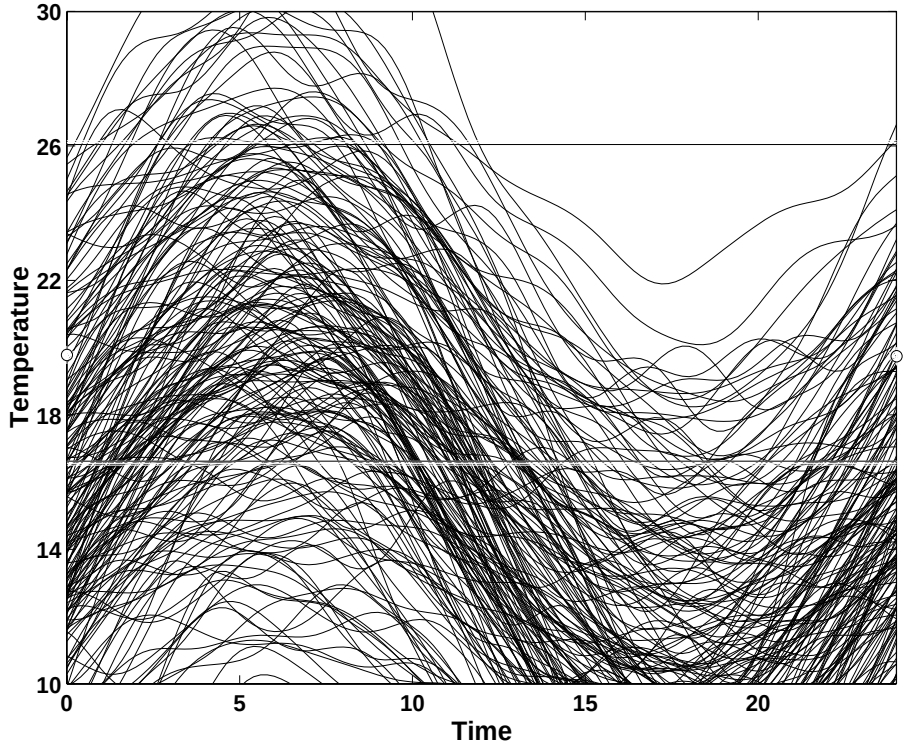


Figure 2: Conditioning to extreme value observations; observations and prior. The horizontal lines shows the observed level of maximum and minimum values. The white circles at each end, show the values observed at the boundaries, in addition 200 samples from the prior distribution used in Example 2, is displayed, some of the samples are partially or completely outside the scaling of the figure.

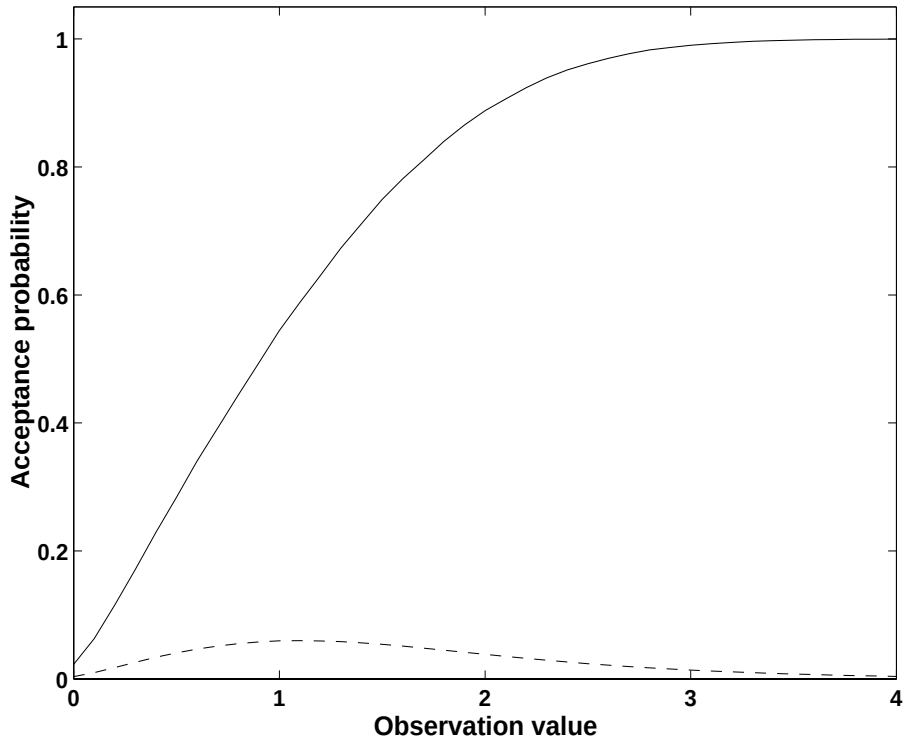


Figure 3: Comparison to naive rejection sampling. The acceptance rate is plotted as a function of the observed value in Example 1. The solid line is the accept probability for the FTGM-algorithm, the dotted line is the corresponding for naive rejection sampling.

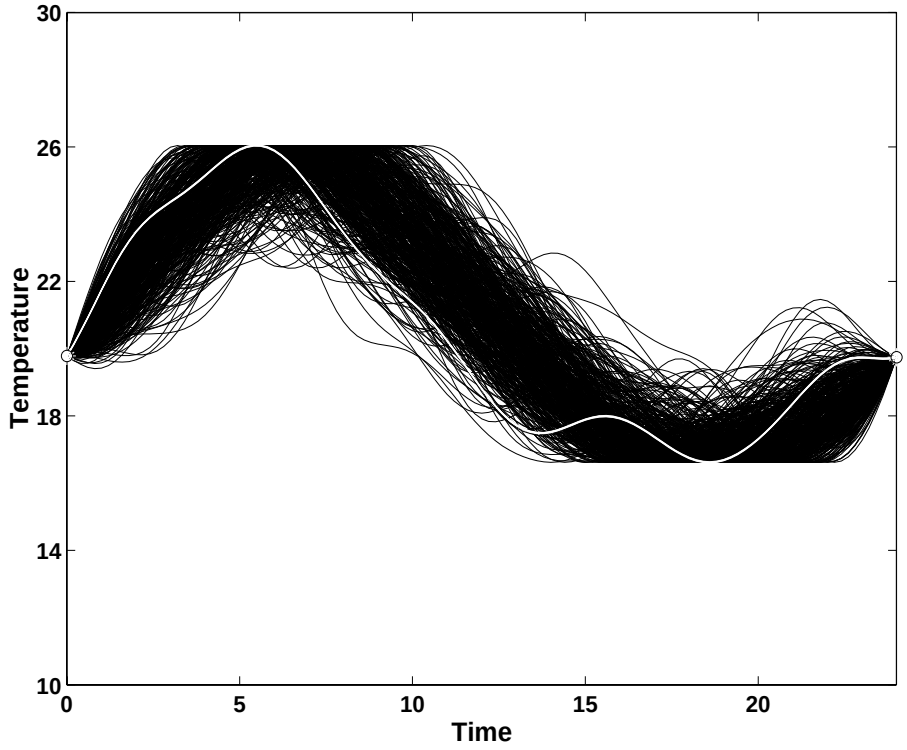


Figure 4: Conditioning to extreme value observations; the posterior. The actual temperature is displayed as a white curve, together with 400 realizations from the posterior distribution, sampled by the CTGM-algorithm in Example 2.

II

Geostatistical approach to event migration of seismic reflection times

Geostatistical approach to event migration of seismic reflection times

Odd Kolbjørnsen

Department of mathematical sciences
Norwegian university of science and technology
Norway

Abstract

Zero offset traveltimes from reflection seismic, are integrated in a Bayesian framework to localize a geological subsurface. The first arrival is fitted into the framework of piecewise affine operators. A partially analytical expression for the posterior distribution is used to develop an algorithm to sample from the posterior distribution. An efficient approximate algorithm to sample from the posterior distribution is also proposed. A synthetic example illustrate the results.

1 Introduction

Inverse problems arising in geophysics are frequently solved by using geostatistical methodology. Working with reflection time inversion, Delprat-Jannaud and Lailly (1993), recognize the benefit of defining a continuous model independent of discretization. They pose the solution as being the argument that minimize an objective function. The objective function is a trade off between the residual sum of squares and a regularization term. The uncertainty is assessed by evaluating the Hessian of the residual sum of squares for a linearized problem (Delprat-Jannaud and Lailly 1992).

When the regularization term is a norm in a reproducing kernel Hilbert space, there exists a dual probability measure for random functions, (Tarantola 1987; Wahba 1990). Using this measure as a prior in a Bayesian model, the maximum posterior estimate, is identical to the solution found by regularization. The maximum posteriori estimate is frequently used for Bayesian models (O'Sullivan 1986), although the classical Bayes estimate is the conditional expectation (Robert and Casella 1999). In general the posterior distribution may be regarded as the Bayesian answer to an inverse problem. The Bayesian methodology hence provide a stable estimate and assess the associated uncertainty.

The objective is to localize a geological subsurface, by using zero offset non-migrated reflection times. The subsurface is the borderline between two layers, it is assumed that the velocity is constant in the top layer. If the exact traveltimes

are known in a region, there is an one to one relation between the traveltimes and the subsurface, see for example Kleyn (1977). The traveltimes are however discretely sampled with errors. A standard solution is to fit a spline function through the traveltimes, see for example Gjøystdal and Ursin (1981). Event migration as used in this article differs from Kirchhoff migration (Bleistein, 1987). For a background model having constant velocity, Kirchhoff migration takes events in data space and back projects them along elliptic curves. Event migration localize the recorded event along the same elliptic curve. In the current work the problem is solved by a Bayesian methodology in a Gaussian framework. The traveltime observations are fit into the framework of piecewise affine operators (Kolbjørnsen and Omre 2002). This formulation gives an expression for the posterior distribution that is partially analytical tractable. Using this expression, an algorithm to sample the posterior distribution is proposed, as well as an efficient sampling algorithm that samples an approximation to the posterior distribution. The approximate approach still honor the nonlinearities in the observations. The method is illustrated in a synthetic example.

The constant velocity model is not sufficiently complex to describe a realistic earth model. The problem is however related to reflection tomography (Farra and Madariaga 1988; Delprat-Jannaud and Lailly 1993). In reflection tomography both the subsurface and the velocity field above the subsurface is unknown. A linearized analysis of reflection tomography show that large components of the velocity field and reflector position remains undetermined by traveltime observations (Delprat-Jannaud and Lailly 1992; Bube and Meadows 1999). A Bayesian solution to the inverse problem in reflection tomography, would give a contribution by its ability represent uncertainty in such nonlinear problems.

2 Model

Consider a two layer model, with constant seismic velocity, v, in the top layer. The twice continuously differentiable Gaussian random function $\{Z(x)\,,\, x \in \mathcal{R}\}$, represents the geological subsurface being the depth to the boundary between the two layers.

The Gaussian random function is defined by its mean and its covariance,

$$\begin{aligned} \mu_Z(x) &= \mathrm{E}\{Z(x)\}\,,\; x \in \mathcal{R} \\ \mathrm{C}_Z(x_1, x_2) &= \mathrm{Cov}\{Z(x_1), Z(x_2)\}\,,\; x_1, x_2 \in \mathcal{R}. \end{aligned}$$

For the Gaussian random function to be twice differentiable with probability one, the expectation is required to be twice differentiable, and regularity conditions must be enforced to the covariance function, see Stein (1999) for details. There is however no stationarity assumption, hence a model formulation as in Bayesian kriging (Omre and Halvorsen 1989), is admissible. The space of twice continuous differentiable functions on $\mathcal{R}$ is denoted $C^2(\mathcal{R})$. To keep notation

short, Z will be used to denote the random function it self, $Z(x)$ will denote the random function evaluated at the location x. Lower case letters will denote deterministic functions such as specific sample paths. Bold letters are used to denote vectors and vector valued functions.

3 Observations

The current section describes the experimental set up that produces the traveltime observations, and phrases the problem in the setting of a piecewise affine inverse problem.

Geometrical aspects of the data collection, is illustrated in Figure 1. A pulse is generated on the surface in the shot location, x_s. The pulse propagates into the ground and is reflected by the subsurface. The part of the pulse that is reflected in the location x propagates back to the surface and is the first to arrive in the receiver location x_r. In the point x, where this reflection occurs, the elliptic curve with foci at x_s and x_r is tangent to the subsurface, see Figure 1. The time from the pulse is generated in x_s until it is received at x_r is denoted the traveltime. The offset is half of the horizontal distance between the shot and receiver location. In the current presentation only zero offset traveltimes are considered hence $x_s = x_r$, and the ellipse degenerate to a circle.

For a given shot/receiver location the two way reflection time is a nonlinear multi valued functional of the geological subsurface $\{z(x)\,,\, x \in \mathcal{R}\}$. In the geological model described above there is at least one reflection for each shot. This reflection occur at minimum distance from the shot location to the subsurface. Only this first arrival is considered in this article, but a similar approach can be used to deal with multiple arrivals. For a subsurface z and a shot/receiver location x_s the traveltime of the first arrival can be expressed as:

$$t(z, x_s) = \min_{x \in \mathcal{R}} \frac{2}{v} \sqrt{z(x)^2 + (x - x_s)^2},$$

with v being the velocity in the top layer, and the value of x that obtains the minimum is the reflection location. Observations of first arrival will be fitted into the framework of piecewise affine operators as defined in Kolbjørnsen and Omre (2002).

Definition 1 (Piecewise affine operator) *An operator $\boldsymbol{K} : \mathcal{Z} \rightarrow \mathbf{R}^n$, is said to be piecewise affine, if it can be represented as:*

$$\boldsymbol{K}(z) = \boldsymbol{K}_x z + \boldsymbol{k}_x \quad \textit{for} \quad z \in \mathcal{A}_x \;; \quad x \in \mathcal{X}$$

with $\mathcal{X}$ being an index set, $\{\mathcal{A}_x\}_{x \in \mathcal{X}}$ being a partition of $\mathcal{Z}$; $\boldsymbol{K}_x : \mathcal{Z} \rightarrow \mathbf{R}^n$ being bounded linear operators on $\mathcal{Z}$ and $\boldsymbol{k}_x$ being n dimensional vectors. The indexed set of triplets $\{\mathcal{A}_x, \boldsymbol{K}_x, \boldsymbol{k}_x\}_{x \in \mathcal{X}}$ are the parameters of the piecewise affine operator.

Kolbjørnsen and Omre (2002) considers both finite and continuous index sets, $\mathcal{X}$. The problem of event migration is in the current presentation adapted to the framework of piecewise affine operators having a continuous index set. For the case of a continuous index set, the partition is assumed to have a special form.

Definition 2 (Restricted linear partition) *A partition* $\{\mathcal{A}_x\}_{x\in\mathcal{X}}$ *with* $\mathcal{X} \subset \mathbf{R}^d$ *of* $\mathcal{Z}$ *is a restricted linear partition if,*

$$\mathcal{A}_x = \mathcal{C}_x \cap \{\boldsymbol{R}_x(z) = \mathbf{0}\},$$

with $\mathcal{C}_x$ *being any subset of* $\mathcal{Z}$*;* $\boldsymbol{R}_x(z) = \boldsymbol{R}_x z + \boldsymbol{r}_x$*; with* $\boldsymbol{R}_x : \mathcal{Z} \to \mathbf{R}^d$ *being a bounded linear operator; and* $\boldsymbol{r}_x$ *being a d-dimensional vector function. The set of triplets* $\{\mathcal{C}_x, \boldsymbol{R}_x, \boldsymbol{r}_x\}_{x\in\mathcal{X}}$ *are the parameters of the restricted linear partition.*

For zero offset traveltimes, measured for the shot location, x_s, define the indexed functionals,

$$K_x(z) = \frac{2}{v}\sqrt{z(x)^2 + (x - x_s)^2}\ , \tag{1}$$

$$R_x(z) = z(x) \cdot z'(x) + x - x_s\ , \tag{2}$$

with v being the velocity in the top layer; z being the subsurface; z' being the derivative of the subsurface; and x being the spatial reference. In addition define the function sets

$$\mathcal{C}_x = \left\{z \in C^2(\mathcal{R}) : z(x)^2 + (x - x_s)^2 < z(u)^2 + (u - x_s)^2;\ u \in \mathbf{R}^1 \setminus \{x\}\right\} \tag{3}$$

The functionals K_x, R_x and the function sets $\mathcal{C}_x$ correspond to the parameters of the piecewise affine operator, and the restricted linear partition. Any reflection from the subsurface, z, back to the shot location, x_s, will occur in a location x with $R_x(z) = 0$, that is the place where the subsurface is tangent to the circle with center x_s. A reflection from this point will have the traveltime $K_x(z)$. In addition only the first arrival is considered, hence the reflection from x must have the minimum time, that is $z \in \mathcal{C}_x$. The parameters of the traveltime operator are hence identified by the expressions above, but does not automatically conform to the framework of the piecewise affine operators since K_x and R_x are nonlinear functionals. However since $Z(x)$ and $Z'(x)$ can be solved exactly when $K_x(Z)$ and $R_x(Z)$ are known only a minor adjustments is needed. The exact details for this is in Appendix A.

Having observed the minimum traveltime at n shot locations, denoted $\boldsymbol{x_s} = [x_{s\,1}, ..., x_{s\,n}]$, an indexed operator is made by stacking the operators corresponding to each shot location into a vector. For each shot location a new dimension is added to the index, $\boldsymbol{x} = [x_1, ..., x_n]$. This gives

$$\boldsymbol{K}(z) = \boldsymbol{K}_{\boldsymbol{x}}(z) \ \text{ for } \ z \in \mathcal{A}_{\boldsymbol{x}}\ ; \quad \boldsymbol{x} \in \mathcal{X}\ , \tag{4}$$

with

$$\begin{aligned}
\boldsymbol{K}_{\boldsymbol{x}}(z) &= [\, K_{x_1}(z), \ldots, K_{x_n}(z) \,]^T \ , \\
\boldsymbol{R}_{\boldsymbol{x}}(z) &= [\, R_{x_1}(z), \ldots, R_{x_n}(z) \,]^T \ , \\
\mathcal{A}_{\boldsymbol{x}} &= \bigcap_{i=1}^{n} \mathcal{C}_{x_i} \cap \{ \boldsymbol{R}_{\boldsymbol{x}}(z) = \boldsymbol{0} \} \ .
\end{aligned}$$

The index, $\boldsymbol{x}$, of the resulting piecewise affine operator have an intuitive interpretation. When $z \in \mathcal{A}_{\boldsymbol{x}}$, z have the minimum distance to the shot locations $\boldsymbol{x}_s$ in the reflection locations $\boldsymbol{x}$. Since only the minimum distance is considered there will be a monotone relation between the shot locations and the reflection locations.

The exact relation between the traveltimes and the subsurface is described above, but the traveltimes are observed with errors. Let the observed random variable, $\boldsymbol{Y}$, be defined by its conditional distribution,

$$p(\boldsymbol{y}|Z = z) = N_n(\boldsymbol{K}(z), \boldsymbol{\Sigma}_\varepsilon)$$

with $\boldsymbol{K}(z)$ being defined by Expression (4); and $\boldsymbol{\Sigma}_\varepsilon$ being the the covariance matrix for the observation error. The marginal distribution of $\boldsymbol{Y}$ is not Gaussian due to the nonlinearity of $\boldsymbol{K}$.

4 Sampling the posterior

The main result in this section is an algorithm for sampling the posterior when conditioning to traveltimes. This algorithm is based on a decomposition of the posterior obtained in Kolbjørnsen and Omre (2002). The Jacobian of the restrictions in the partition, $\boldsymbol{\nabla}_{\boldsymbol{x}} \boldsymbol{R}_{\boldsymbol{x}}(z)$, is needed to obtain the decomposition. It is further convenient to introduce the extended operators

$$\tilde{\boldsymbol{K}}_{\boldsymbol{x}}(z) = \tilde{\boldsymbol{K}}_{\boldsymbol{x}}(z) \ \text{ for } \quad \boldsymbol{x} \in \mathcal{X},$$

with

$$\tilde{\boldsymbol{K}}_{\boldsymbol{x}} z = \begin{bmatrix} \boldsymbol{K}_{\boldsymbol{x}}(z) \\ \boldsymbol{R}_{\boldsymbol{x}}(z) \\ \boldsymbol{\nabla}_{\boldsymbol{x}} \boldsymbol{R}_{\boldsymbol{x}}(z) \end{bmatrix} .$$

Note that this is a collection of operators indexed by the same index as the original problem. Define the random variables $\tilde{\boldsymbol{Y}}_{\boldsymbol{x}}$ corresponding to observations of each of the operators $\tilde{\boldsymbol{K}}_{\boldsymbol{x}}$,

$$p(\tilde{\boldsymbol{y}}_{\boldsymbol{x}}|Z = z) = N_{n+d+d^2}(\tilde{\boldsymbol{K}}_{\boldsymbol{x}}(z), \tilde{\boldsymbol{\Sigma}}_\varepsilon) \ ,$$

with $\tilde{\boldsymbol{K}}_{\boldsymbol{x}}$ being the extended operator for index $\boldsymbol{x}$; and $\tilde{\boldsymbol{\Sigma}}_{\varepsilon}$ being an extension of $\boldsymbol{\Sigma}_{\varepsilon}$. The values of $\tilde{\boldsymbol{\Sigma}}_{\varepsilon}$ are like $\boldsymbol{\Sigma}_{\varepsilon}$ in the upper left $n \times n$ corner and zero otherwise.

The objective is to compute the posterior distribution of Z given $\boldsymbol{Y} = \boldsymbol{y}$. In Kolbjørnsen and Omre (2002) this distribution is derived as a mixing of truncated Gaussian distributions, under given regularity conditions. The result for the traveltime observations reads,

$$
\begin{aligned}
& p\{Z = z, Z \in \mathcal{C}_{\boldsymbol{x}}, \boldsymbol{\nabla}_{\boldsymbol{x}} \boldsymbol{R}_{\boldsymbol{x}}(Z) = \boldsymbol{r} | \boldsymbol{Y} = \boldsymbol{y}\} \\
& \quad = p\{Z = z |\, Z \in \mathcal{C}_{\boldsymbol{x}}, \tilde{\boldsymbol{Y}}_{\boldsymbol{x}} = (\boldsymbol{y}, \boldsymbol{0}, \boldsymbol{r})\,\} \\
& \qquad \times\ \mathrm{P}\{Z \in \mathcal{C}_{\boldsymbol{x}} |\, \tilde{\boldsymbol{Y}}_{\boldsymbol{x}} = (\boldsymbol{y}, \boldsymbol{0}, \boldsymbol{r})\,\} \times \frac{|\det(\boldsymbol{r})| p\{\, \tilde{\boldsymbol{Y}}_{\boldsymbol{x}} = (\boldsymbol{y}, \boldsymbol{0}, \boldsymbol{r})\,\}}{p\{\boldsymbol{Y} = \boldsymbol{y}\}}
\end{aligned}
\tag{5}
$$

The distribution on the left hand side is the distribution of Z when Z is in $\mathcal{C}_{\boldsymbol{x}}$ and $\boldsymbol{\nabla}_{\boldsymbol{x}} \boldsymbol{R}_{\boldsymbol{x}}(Z)$ have the value $\boldsymbol{r}$. The posterior distribution of Z is obtained by integrating out $\boldsymbol{x}$ and $\boldsymbol{r}$. In Appendix A the expressions above are given in terms of $Z(x)$, $Z'(x)$ and $Z''(x)$, for a traveltime observation with reflection in the location x. The first term on the right hand side of Expression (5) is a truncated Gaussian distribution when the observations are without error, the case with errors being treated in Appendix A. The second term is a probability and the third term is a measure on $\mathcal{X} \times \mathbf{R}^{d^2}$. The product of the second and third term is the joint posterior density for Z being in $\mathcal{C}_{\boldsymbol{x}}$ and $\boldsymbol{\nabla}_{\boldsymbol{x}} \boldsymbol{R}_{\boldsymbol{x}}(Z)$ having the value $\boldsymbol{r}$. Note that $\boldsymbol{\nabla}_{\boldsymbol{x}} \boldsymbol{R}_{\boldsymbol{x}}(z)$ is diagonal and that the elements on the diagonal are positive for $z \in \mathcal{C}_{\boldsymbol{x}}$ due to the second order criterion for minima.

The mixing distribution of Expression (5), provides a sampling algorithm for the posterior.

Algorithm 1 *CTGM-algorithm (Continuous Truncated Gaussian Mixing)*

1. *Sample* $\boldsymbol{x}^*, \boldsymbol{r}^* \sim q(\boldsymbol{x}, \boldsymbol{r})$ *with* $q(\boldsymbol{x}, \boldsymbol{r}) \propto |\det(\boldsymbol{r})| p\{\, \tilde{\boldsymbol{Y}}_{\boldsymbol{x}} = (\boldsymbol{y}, \boldsymbol{0}, \boldsymbol{r})\,\}$
2. *Sample* $Z^* \sim p\{Z | \tilde{\boldsymbol{Y}}_{\boldsymbol{x}^*} = (\boldsymbol{y}, \boldsymbol{0}, \boldsymbol{r}^*)\}$
3. *If* $Z^* \in \mathcal{C}_{\boldsymbol{x}^*}$ *stop.*

The algorithm is a variant of rejection sampling. The acceptance rate of the algorithm is given by

$$
p_{accept} = \frac{p\{\boldsymbol{Y} = \boldsymbol{y}\}}{\int_{\mathcal{X}} \int_{\mathbf{R}^{d^2}} |\det(\boldsymbol{r})| p\{\, \tilde{\boldsymbol{Y}}_{\boldsymbol{x}} = (\boldsymbol{y}, \boldsymbol{0}, \boldsymbol{r})\,\}\, d\boldsymbol{r}\, d\boldsymbol{x}}
$$

The algorithm yields samples from the posterior when conditioning to a piecewise affine operator. The posterior density for Z being in $\mathcal{C}_{\boldsymbol{x}}$ with $\boldsymbol{\nabla}_{\boldsymbol{x}}\boldsymbol{R}_{\boldsymbol{x}}(Z)$ having the value $\boldsymbol{r}$ is sampled correctly since the proposal distribution in Step 1 is proportional to the third factor of Expression(5) and the acceptance part in Step 3 is exactly the second factor of Expression (5). When a pair $(\boldsymbol{x}^*, \boldsymbol{r}^*)$ is accepted, the sampled value z^* is a valid sample. Note that the reflection locations, $\boldsymbol{x}$, are monotone as a function of the shot locations, $\boldsymbol{x}_s$ and that the elements of $\boldsymbol{r}$ are positive. These inequality constraints should be imposed when sampling $q(\boldsymbol{x}, \boldsymbol{r})$.

The challenging part of the algorithm is to sample the proposal distribution,

$$q(\boldsymbol{x}, \boldsymbol{r}) \propto |\det(\boldsymbol{r})| p\{\, \tilde{\boldsymbol{Y}}_{\boldsymbol{x}} = (\boldsymbol{y}, \boldsymbol{0}, \boldsymbol{r})\,\}, \tag{6}$$

which is known apart from the normalizing constant, see Appendix A for details. There are several ways to sample this distribution. One approach is to sample the distribution using a McMC algorithm. As an alternative, an algorithm that samples an approximation to $q(\boldsymbol{x}, \boldsymbol{r})$ is proposed. The approximate approach still honor the nonlinear structure of the problem since only Step 1 of the CTGM-algorithm is approximated.

The distribution $q(\boldsymbol{x}, \boldsymbol{r})$ is approximated by a sequence of truncated Gaussian distributions. Firstly $q(\boldsymbol{x}, \boldsymbol{r})$ is approximated by a Gaussian distribution, $q_G(\boldsymbol{x}, \boldsymbol{r})$. The distribution $q_G(\boldsymbol{x}, \boldsymbol{r})$ is sampled sequentially, but for each variable in the sequence the relevant constraint is imposed on the conditional distribution. The resulting approximation is neither a Gaussian nor a truncated Gaussian distribution, but the absolute magnitude of the proposal density can be assessed directly in the sampling approach. Since the exact distribution is known to a multiplicative factor, importance weights for the sampled values can be calculated. These weights can be used to remove bias in the sample. This is not done in the current study, however.

5 Example

The methodology is tested in a synthetic example. Uncertainty in the shot/receiver location is included in addition to the observation error. To correctly account for these effects, shot locations, $\boldsymbol{x}_{\boldsymbol{s}}$, and observation errors, $\boldsymbol{\varepsilon}$, should be sampled simultaneously with $\boldsymbol{x}$ and $\boldsymbol{r}$, see Appendix A.

The prior model for the reflector is defined to be a stationary Gaussian random field,

$$Z(x) = Z_0 + \widetilde{Z}(x) \quad ; \quad x \in [0,\, 10];$$

with $p\{Z_0\} = N(2, 1)$; and $\widetilde{Z}(x)$ being a zero mean Gaussian random field with covariance $\mathrm{Cov}\{\widetilde{Z}(x), \widetilde{Z}(x+h)\} = 0.15^2 \exp\{-3 \cdot (h/2)^2\}$. Several samples from the prior distribution are shown in Figure 2. Note the extreme uncertainty

in the level of the samples, that make several of the samples lie partially or completely outside the scales of the figure. The realization used in this example is plotted with a thick line.

Figure 3 shows the lines that connect the shot/receiver locations with the reflecting point at the subsurface to be recovered. The uncertainty in shot location is assigned standard deviation $\sigma_{x_s} = 10$ m. The observation error of the traveltime is assigned standard deviation $\sigma_t = 0.05$ sec. The interval velocity is assumed to be $v = 1$ km/sec. In Figure 4 the observed traveltimes and assumed shot locations are plotted together with the true traveltimes, which of course can be exactly computed in this synthetic example. The CTGM-algorithm is implemented in two simulation studies. In the first study only the five traveltimes that is marked in Figure 4 is used, in the second all 81 traveltime observations are used.

Figure 5 display the sample values from the study when only five traveltime observations are used. The proposal distribution $q(\boldsymbol{x}, \boldsymbol{r})$ is sampled by a McMC-algorithm. The acceptance rate in the final step of the CTGM-algorithm, is 98%, leaving the sampling of q as the most time consuming part. Compared with the samples from the prior distribution, see Figure 2, there is a dramatic reduction in uncertainty. The level is well determined and, the true subsurface is within the ensemble of samples from the posterior distribution.

Figure 6 display the sample values from the study when all 81 traveltime observations are used. The approximate sampling approach is used. The observed acceptance rate is 4.95% in the final step of the CTGM-algorithm. The relatively low acceptance rate is partially due to numerical instability in this particular problem since extreme smoothness is imposed by the second order exponential correlation function. In this case the uncertainty is very low within the region containing observations. At both ends there is larger uncertainty. Some of the samples have extreme values at the end of the interval. This is an artifact that is caused by the interaction of the approximate sampling algorithm and the extreme smoothness of the second order exponential correlation function. Apart from the valley between 6 and 7 km the true subsurface is well within the ensemble of samples from the posterior distribution. The deviation in this valley can partially be explained by the actual observation errors in this region, see Figure 4.

The ensemble of samples can be combined to a single estimate. To reduce the influence of the outliers, see Figure 6, the pointwise median is used as an estimator. For comparison the standard estimate based on inverting the smoothed traveltimes is computed. In Figure 7 (a)-(c) the estimates are compared with each other and to the true subsurface. The standard estimate have higher, sharper tops and more shallow, broader valleys, than both the true subsurface and the proposed estimate. The standard estimate miss out both the valley around 4 km and the valley between 6 and 7 km. In the flat regions such as the interval between 1 and 3 km the standard estimate is as good as the pro-

posed. The cause of the difference in the estimates is that the smoothing of the traveltimes in the standard approach, implicitly impose a spatial assumption of stationarity for the traveltime as a function of the shot location. In the proposed approach the stationarity is imposed directly on the subsurface.

6 Conclusions

Zero offset traveltimes from reflection seismic are used to localize a geological subsurface. A Bayesian approach is developed by first defining the prior and likelihood and next condition to the observations. The posterior distribution is explored by sampling. Further an approximate sampling algorithm that honor the nonlinearities of the problem is proposed.

The methodology yields satisfactory results when evaluated in an example. The uncertainty can be represented when few observations are present, and the function is well recovered within the shot section when a realistic amount of data is used. When compared to a standard estimate, the proposed approach is computationally more expensive, but gives a better estimate in curved sections.

Observations being linear operators of the random field, such as well observations can be included, by using the conditional distribution as input to the algorithm, or by extending the affine operator. Seismic observations with offset, can be treated in the same frame work and the method extend to traveltimes in 3-D. A small inhomogeneous deviation from the constant velocity can also be accounted for by using a perturbation argument, this will produce a colored error term that is dependent on the reflection point, and is a first step in the direction of surface reflection tomography.

Acknowledgment

I am grateful to my advisor Henning Omre for valuable comments. Attending the mathematical geophysical summer school at Stanford university, august 1999, and several discussions with professor Paul Switzer at Stanford university have been beneficial. Funding for the research is provided by a PhD grant from the the Research Council of Norway.

References

Bleistein, N. (1987) "On the imaging of reflectors in the earth." Geophysics, Vol. 52, pp 931-942.

Bube K.P. and Meadows M.A. (1999) "The null space of a generally anisotropic medium in linearized surface reflection tomography", Geophys.J.Int., Vol. 139,

pp 9-50.

Delprat-Jannaud, F. & Lailly, P. (1992) "What information on the Earth Model Do the Reflection traveltimes Provide?" Journal of Geophysical research, Vol. 97, no. B13, pp 19827- 19844.

Delprat-Jannaud, F. & Lailly, P. (1993) "Ill-Posed and well posed Formulations of the Reflection traveltime Tomography Problem." Journal of Geophysical research, Vol. 98, no. B4, pp 6589 - 6605.

Farra, V. and Madariaga, R. (1988) "Non-linear reflection tomography." Geophysical Journal Vol. 95, pp 135-147.

Gjøystdal, G. and Ursin, B. "Inversion of reflection times in three dimensions" Geophysics, Vol. 46, no. 7, pp 972-983.

Kleyn, A. H. (1977) On the migration of reflection time contour maps: Geophys.Prosp, Vol. 25, pp 125-140.

Kolbjørnsen, O. & Omre, H. (2002) "Bayesian inversion of piecewise affine operators in a Gaussian framework."

Omre, H. & Halvorsen, K.B. (1989) "The Bayesian bridge between Simple and Universal Kriging." Mathematical Geology, Vol.21, no. 7, pp 767- 786.

O'Sullivan, F. (1986), "A statistical perspective on ill-posed inverse problems," *Statistical Science*, Vol 1,no. 4, pp 502-518.

Robert C.P. and Casella G. (1999) "Monte Carlo statistical methods", Springer.

Tarantola A. (1987) "Inverse problem theory : methods for data fitting and model parameter estimation",Elsevier.

Stein, M.L. (1999) "Interpolation of spatial data : some theory for kriging." Springer, New York.

Wahba, G. (1990) "Spline models for observational data." Society for Industrial and Applied Mathematics.

A Adapting traveltime information to the CTGM algorithm

This appendix contain detailed calculations to fit traveltime observations from reflection seismic into the frame work piecewise affine operators. The calculations below are for the case of one observation, but the extension to several observations is obvious, since the transforms apply locally to each set of variables corresponding to each single traveltime.

Firstly assume that the observed traveltimes do not have any observation error. The mixing distribution to be sampled in the CTGM-algorithm is then,

$$q(x,r) = \text{const} \times |r| p\{K_x Z = t, R_x(Z) = 0, \nabla_x R_x(Z) = r\},$$

with $K_x Z$, $R_x(Z)$ and $\nabla_x R_x(Z)$ referring to Expressions (1) and (2). For a given x rename the random variables $(K_x Z,\ R_x(Z),\ \nabla_x R_x(Z))$ by (T, R_0, R_1) and $(\ Z(x),\ Z'(x),\ Z''(x)\ )$ by (Z_0, Z_1, Z_2). The relation between (T, R_0, R_1) and (Z_0, Z_1, Z_2) is given by:

$$\begin{aligned} T &= \frac{2}{v}\sqrt{Z_0^2 + (x - x_s)^2}\,, \\ R_0 &= Z_0 \cdot Z_1 + x - x_s\,, \\ R_1 &= Z_0 \cdot Z_2 + Z_1^2 + 1\,. \end{aligned} \tag{7}$$

This relation can be inverted to find :

$$\begin{aligned} Z_0 &= \sqrt{\left(\frac{Tv}{2}\right)^2 - (x - x_s)^2}\,, \\ Z_1 &= \frac{x_s - x - R_0}{Z_0}\,, \\ Z_2 &= \frac{R_1 - Z_1^2 - 1}{Z_0}\,. \end{aligned} \tag{8}$$

The variables (Z_0, Z_1, Z_2), have a known Gaussian distribution, hence the distribution of (T, R_0, R_1) may be calculated by the usual transformation rule. Assuming $(T, R_0, R_1) = (t, 0, r)$, the outcome of (Z_0, Z_1, Z_2) is given by the following expressions:

$$\begin{aligned} z_0(t, x, x_s) &= \sqrt{\left(\frac{tv}{2}\right)^2 - (x - x_s)^2}\,, \\ z_1(t, x, x_s) &= \frac{x_s - x}{z_0(t, x, x_s)}\,, \\ z_2(t, x, x_s, r) &= \frac{r - z_1(t, x, x_s)^2 - 1}{z_0(t, x, x_s)}\,. \end{aligned}$$

Using the transformation rule yields the result

$$q(x, r; x_s, t) = \text{const} \times \left| \frac{r\ \frac{tv}{2}}{z_0(t, x, x_s)^3} \right|$$

$$\times\ p\{Z(x) = z_0(t, x, x_s),\ Z'(x) = z_1(t, x, x_s),\ Z''(x) = z_2(t, x, x_s, r)\},$$

with x_s and t regarded as given parameters; and the joint probability distribution of $(Z(x),\ Z'(x),\ Z''(x)\ )$ being known from the prior.

When observation error is included in traveltime and shot location, the mixing distribution must be extended such that it include the true traveltime and the true shot location as well. Hence by denoting the observed traveltime y, the task is to sample the distribution

$$\tilde{q}(x, r, t, x_s; y) = q(x, r; x_s, t) p\{\varepsilon = y - t\} p\{x_s\},$$

with; $p\{\varepsilon = y - t\}$ being the likelihood of the observation; and $p\{x_s\}$ being a prior distribution for the shot location.

In general two conditional probabilities are not identical even if the events behind the conditioning bar match. Conditional statements are relative to the $\sigma-$algebra generating the events. The equivalence between sampling the distribution $p\{Z|Z(x), Z'(x), Z''(x)\}$ in place of $p\{Z|K_x(Z), R_x(Z), \nabla_x R_x(Z)\}$, is due to the one to one relation between $Z(x)$, $Z'(x)$ and $Z''(x)$; and $K_x Z$, $R_x(Z)$ and $\nabla_x R_x(Z)$), that is given by Expressions (7) and (8).

B Figures

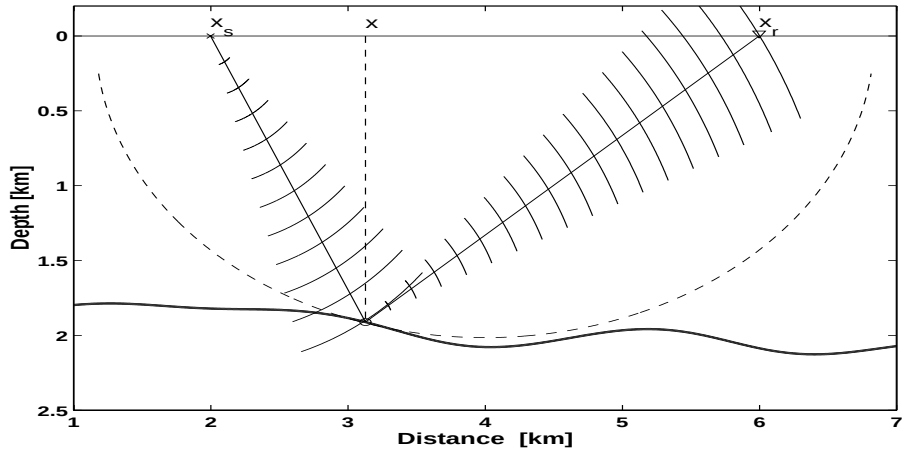


Figure 1: Traveltime geometry. The pulse generated in x_s is reflected in the point $(x, z(x))$ at the subsurface and is detected by the receiver in x_r. In the point, x, where the reflection occur the ellipse with foci in x_s and x_r, (dashed line) is tangent to the surface.

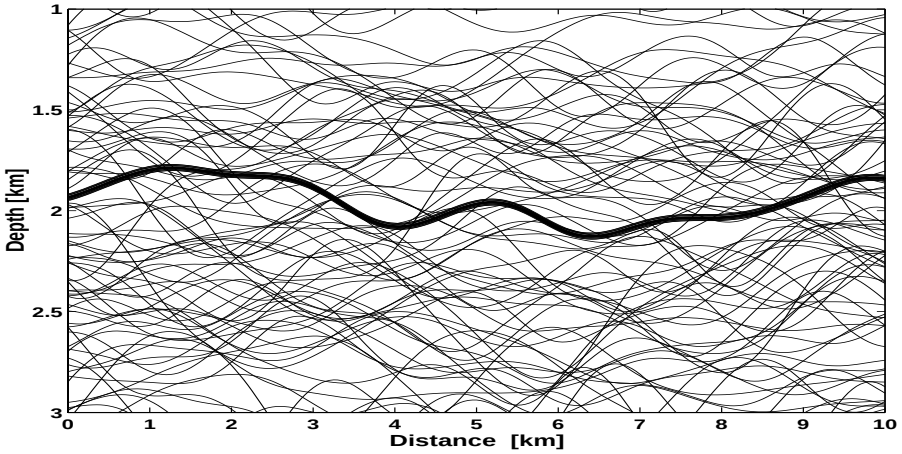


Figure 2: Traveltime migration; the prior. The actual curve is shown with a thick black line together with 100 samples from the prior model for the geological horizon, several of the samples are partially or completely outside the scales of the figure.

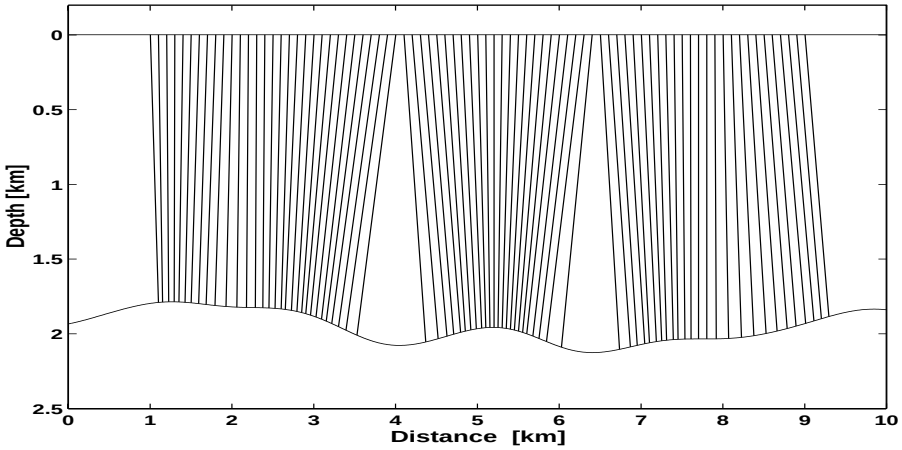


Figure 3: Traveltime geometry. The upper straight horizontal line is the surface, the geological horizon to be recovered is the lower curved horizontal line. The vertical lines are the ray paths connecting the shot location at the surface to to the nearest point at the geological horizon.

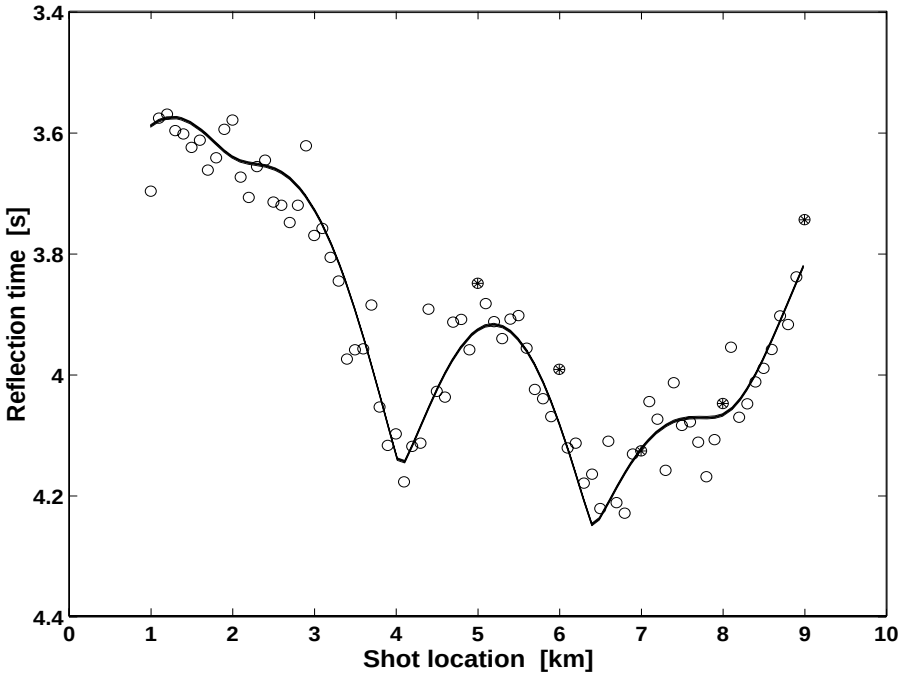


Figure 4: Traveltime migration; the observations. The observed traveltimes are plotted as circles together with the true traveltimes as a solid line. The observations have errors in both x_s and t. The filled circles is the subset of observations used when only five observations are considered.

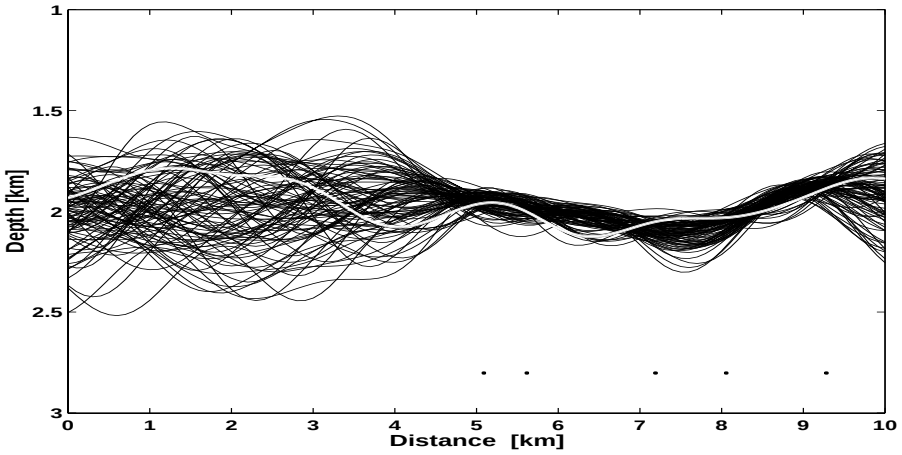


Figure 5: Traveltime migration; the posterior. The actual horizon is displayed in white together with 99 samples from the posterior, using only the five observations marked in Figure 4. The dots in the bottom of the figure indicates the actual reflection locations for the 5 observations.

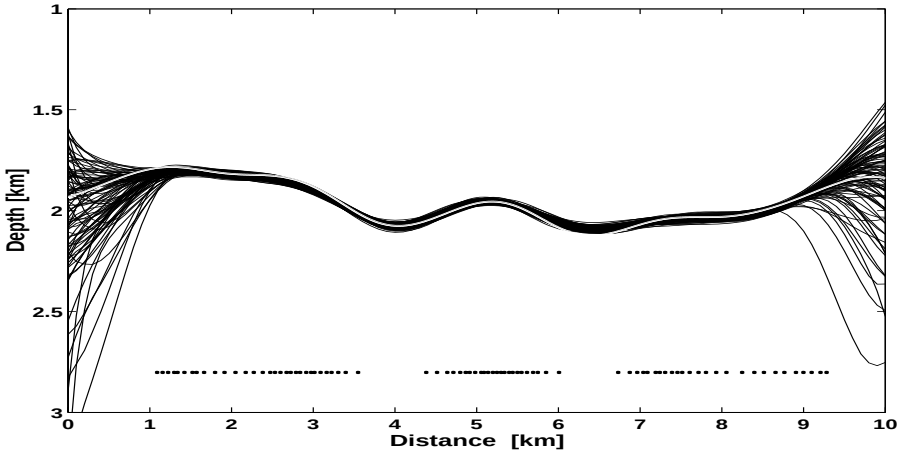
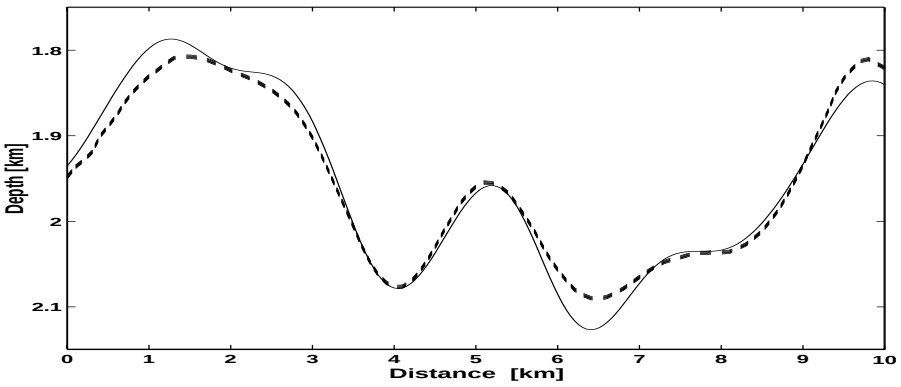
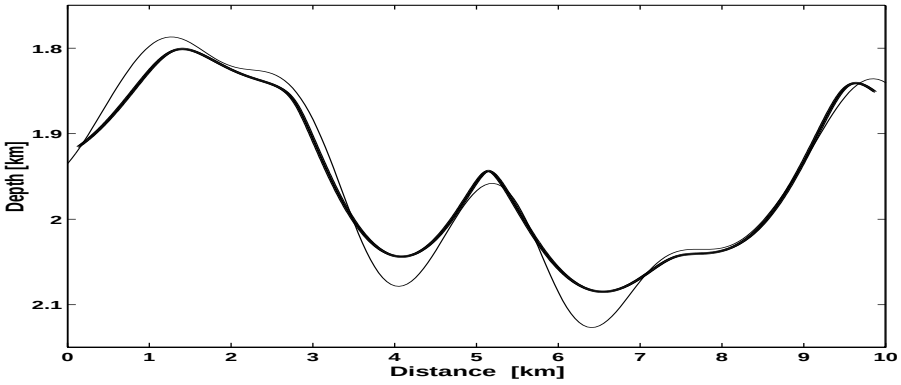


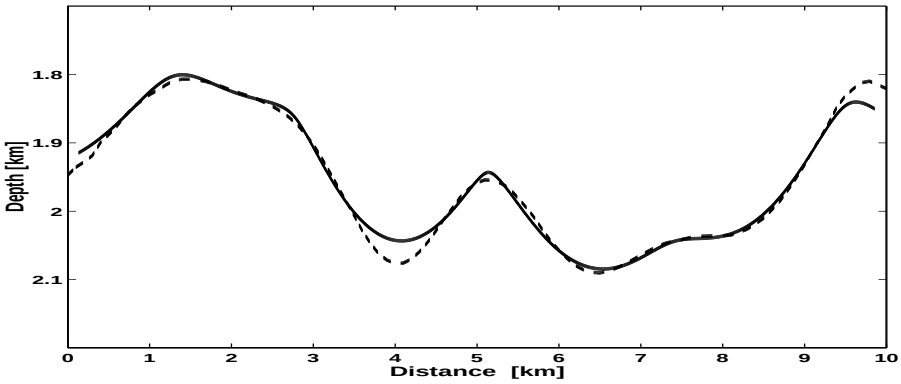
Figure 6: Traveltime migration; the posterior. The actual horizon is displayed in white together with 99 samples from the posterior, using all 81 observations in Figure 4. The dots in the bottom of the figure indicates the actual reflection locations for all 81 observations.



(a)



(b)



(c)

Figure 7: Comparison of estimates and true subsurface; The truth is plotted with a thin full line; the standard estimate with thick full line; and the proposed estimate with a thick dashed line. Comparison of truth and proposed estimate (a); truth and standard estimate (b); proposed and standard estimate (c).

III

Case specific uncertainty assessment in cross well tomography

Case specific uncertainty assessment in cross well tomography

Odd Kolbjørnsen

Department of mathematical sciences,
Norwegian university of science and technology
Trondheim, Norway

Abstract

The inverse problem in cross well tomography is solved by a Bayesian methodology in a Gaussian framework. A finite element approach is used to resolve the variational structure given by Fermat's principle, as a result the approximate forward map is piecewise affine. In the Gaussian framework the posterior distribution can be calculated as a mixture of truncated Gaussian distributions. A sampling algorithm that exploit this structure is proposed. The methodology is tested in a small synthetic example.

KEY WORDS: *Bayesian statistics, Sampling based inference, piecewise affine inverse problem, nonlinear traveltime tomography, Fermat's principle.*

1 Introduction

Cross well tomography is an important source of information about elastic parameters of the earth. Both the direct problem of wave propagation (Langan, Lerche and Cutler 1985; Vidale 1988; Auld 1990) and the inverse problem in cross well tomography (Menke 1984; Berryman 1990; Langan and Bube 1998) are subject to substantial research interest.

The direct problem in cross well tomography is nonlinear, the solution is given by the minimum of a set of linear functionals. In linearized cross well tomography the solution to the direct problem is approximated by picking one of the linear functionals in the set.

The primary goal in cross well tomography is to stably estimate material parameters of the earth based on traveltime observations. Linearized cross well tomography gives a qualitative understanding of the problem. Menke (1984) show that linearized cross well tomography resolves the material parameters of an isotropic earth poorly, especially in the horizontal direction. The case for anisotropic case is even worse (Bube and Meadows, 1998). Further, related operators such as X-ray and Radon transforms have unbounded inverses (Faridani,

1997). The inverse problem of cross well tomography is hence ill-posed, since it is both underdetermined and unstable.

Ill-posed inverse problems is commonly solved by regularization or equivalently by introducing rigid boundaries on the parameter space (Bertero, 1989). The solution is then obtained by minimizing an objective function. For nonlinear problems such as cross well tomography, iterative solvers are frequently used. Berryman (1990) notice that this formulation does not fully appreciate the variational structure that is present in the problem of cross well tomography. The first arrival time obeys Fermat's principle, i.e. it is the shortest traveltime that is physically possible (Aronsson 1970). Berryman (1990) uses Fermat's principle to construct feasibility constraints for the solution. When he uses the feasibility constraints to determine the step size in his solver he obtain a stable reconstruction.

A secondary and frequently equally important objective in cross well tomography is to assess the uncertainty of the estimate. Common approaches are resolution theory (Menke, 1984) and a singular value decomposition (Michelena 1993), in either case the operator is linearized. In nonlinear problems such as cross well tomography, it is hard to describe the underdetermined and badly determined features exactly, since they do not span a linear space.

The current work apply a Bayesian approach to the inverse problem of cross well tomography. In Bayesian analysis a likelihood is defined according to the statistical link between the parameter of interest and the observations and a prior distribution is defined for the parameter of interest. The prior distribution is frequently criticized by non-Bayesians. However for ill-posed inverse problems, such as cross well tomography, the prior distribution plays an essential role. The prior distribution stabilizes the solution and resolves the problem of underdetermination. The prior hence serve the purpose of regularization and define soft boundaries on the parameter space. In a specific case there is often available information about the scales of the slowness, either based on general geological knowledge or analog reservoirs. This information can be included through the prior distribution. The effect of the assumptions can be visualized by random samples from the prior distribution.

The Bayesian solution to the inverse problem, is the posterior distribution which is formally proportional to the product of the likelihood and the prior. For most practical problems it is beneficial to approximate the posterior distribution by a finite representation. In the current work the posterior distribution is approximated by random samples assigned equal weight. This approach apply to both linear and nonlinear problems. The Bayesian approach achieve both goals in cross well tomography. The posterior mean is a stable estimate. The posterior distribution itself describes the uncertainty of the estimate. Bayesian uncertainty assessment is hence case specific. The posterior distribution is relative to the observation at hand and depend on the prior distribution and the likelihood which are defined such that their characteristics are adapted to the case under study.

Bayesian approaches to problems in tomography is developed by several authors, (Natterer 1980; Carfantan and Mohammad-Djafari 1997), most authors only consider the maximum posterior estimate and does not use the full power of the Bayesian analysis. In the current work the Bayesian inversion, i.e. an algorithm to sample the posterior, is worked out in a Gaussian framework, taking account of the nonlinear features. In Kolbjørnsen and Omre (2002) the theory of piecewise affine inverse problems in a Gaussian framework is presented. The posterior distribution is a mixture of truncated Gaussian distributions in this case. The contribution in the current work is to use the Fermat's principle to phrase cross well tomography as a piecewise affine inverse problem and develop the methodology of Kolbjørnsen and Omre (2002) for this problem.

Section 2 describes the problem of cross well tomography. In section 3 the problem of cross well tomography is formulated as an piecewise affine inverse problem, by using a finite element approach to approximate Fermat's principle. Section 4 describes the statistical models that are used, and section 5 contains the posterior distribution together with the sampling approach. Section 6 discuss a generalization of the approach. In section 7 two small examples are investigated. Section 8 contain a discussion of the results.

2 Problem description

The current section gives a brief introduction to the problem of cross well tomography. The slowness, the inverse of the velocity, is the material parameter of relevance. In the current presentation the medium is assumed to be isotropic, but the approach can easily be extended to media with elliptical anisotropy (Bube and Meadows, 1998).

The objectives in cross well tomography is to reconstruct the slowness field in a region, $\mathcal{R}$, between two wells based on imperfect observations of traveltimes from sources in one well to receivers in the other well, and to assess the uncertainty of the reconstruction. Figure 1 illustrates the situation. A source is placed in one well at the location (x_s, z_s), a receiver is placed in the other well at the location (x_r, z_r). The traveltime is the time it takes for a wave to propagate from the source to the receiver.

For simplicity the earth is considered to vary only with depth, z, and the lateral component describing the inter distance between the two wells, x, i.e. $s(x, y, z) = s(x, z)$. Further the slowness is assumed to be twice continuously differentiable, i.e. $s \in C^2(\mathcal{R})$.

The traveltime between a source and a receiver is denoted the Fermat time because it obeys Fermat's principle. That is, it is the minimum traveltime from the source to the receiver. To make this precise Berryman (1997) introduce two types of functionals for traveltime. Let Γ be the set of continuous paths connecting the source and the receiver. For a given $\gamma \in \Gamma$ define the traveltime

functional, $\tau(\gamma, \cdot)$, associated with this path by its action on a slowness field, s,

$$\tau(\gamma, s) = \int_{\gamma} s(x, z)\, dl^{\gamma}$$

with dl^{γ} being the infinitesimal distance along γ. Define now the traveltime functional, τ^*, corresponding to the Fermat time. For given slowness field, s, this is defined as,

$$\tau^*(s) = \min_{\gamma \in \Gamma} \tau(\gamma, s)\ . \tag{1}$$

The Fermat time is the minimum path integral of the slowness along any continuous path connecting the source and the receiver. The Fermat path, γ^*, is defined as the path where this minimum occur,

$$\gamma^*(s) = \arg\min_{\gamma \in \Gamma} \tau(\gamma, s).$$

The Fermat path need not be unique, but for a given source/receiver pair it almost surely is so. The Fermat time can be expressed as

$$\tau^*(s) = \int_{\gamma^*(s)} s(x, z)\, dl^{\gamma^*(s)}\ ,$$

that is, if the Fermat path is known the traveltime is a linear functional of s.

In a medium of constant slowness, the Fermat paths are straight lines connecting the source and the receiver. A perturbation argument (Boyse and Keller 1995) show that the bending of the Fermat path is a second order effect, hence the traveltime can be approximated to the first order by the line integral along the straight line connecting the source and the receiver. This is the argument used in linearized cross well tomography to pick a particular path. Figure 2 show a slowness field where the perturbation argument is not valid due to large deviations from a constant background. Figure 2(a) show the linear paths for 16 source/receiver pairs. Figure 2(b) show the Fermat paths for the same slowness field. For such cases other approximations are needed.

3 Cross well tomography as a piecewise affine inverse problem

To phrase cross well tomography as a piecewise affine inverse problem, each traveltime is approximated by a piecewise affine functional. In the current work the Fermat time, $\tau^*(s)$ in Expression (1), is approximated by a finite element approach.

$$\tau_0^*(s) = \min_{\gamma \in \Gamma_0} \tau(\gamma, s)\ , \tag{2}$$

with $\tau_0^*(s)$ being the approximate Fermat time; and Γ_0 being the set of finite elements. The set Γ_0 consist of piecewise linear paths, parameterized with d internal nodes. The nodes are equispaced in the lateral directions and free to move in the vertical direction, see Figure 3. Each path is hence parameterized by a d dimensional parameter, $\gamma = (\gamma_1, \gamma_2, ..., \gamma_d)$, being the vertical coordinate of each node. In what follows there will not be made any notationally distinction between the parameter γ and the piecewise linear path that is associated with it. The path parameter is a vector with d components but it is denoted by a normal type letter to avoid confusion when several traveltimes are considered. Further let $\gamma_0^*(s) \in \Gamma_0$ denote the path where the minimum in Expression (2) occur. The path $\gamma_0^*(s)$ is hence the approximate Fermat path.

Figure 4 and 5 visualize the finite element approximation for the slowness field in Figure 2. Figure 4 show how the traveltime approximation improve with an increasing number of internal nodes for the 16 traveltimes indicated in Figure 2. Figure 5 show how one Fermat path change as the number of internal nodes increase. In this particular case the approximation is good even with a low number of internal nodes.

Note that the finite parameterization of the path does not force any particular parameterization of the slowness, this is in contrast to approaches that use block models and Snell's law for ray bending at the block boundaries. The accuracy of the approximation will of course depend on the slowness field. In the continuous formulation of the problem, paths between different source/receiver pairs can cross one time at most. This ordering is forced also in the discrete problem even if several crossings could occur for this case.

According to Kolbjørnsen and Omre (2002) a piecewise affine operator is defined as

Definition 1 (Piecewise affine operator) *An operator $\boldsymbol{K} : \mathcal{Z} \rightarrow \mathbf{R}^r$, is said to be piecewise affine, if it can be represented in the following way:*

$$\boldsymbol{K}(z) = \boldsymbol{K}_x z + \boldsymbol{k}_x \quad for \quad z \in \mathcal{A}_x \; ; \quad x \in \mathcal{X}$$

with $\mathcal{X}$ being an index set, $\{\mathcal{A}_x\}_{x \in \mathcal{X}}$ being a partition of $\mathcal{Z}$; $\boldsymbol{K}_x : \mathcal{Z} \rightarrow \mathbf{R}^r$ being bounded linear operators on $\mathcal{Z}$ and $\boldsymbol{k}_x$ being r dimensional vectors. The indexed set of triplets $\{\mathcal{A}_x, \boldsymbol{K}_x, \boldsymbol{k}_x\}_{x \in \mathcal{X}}$ are the parameters of the piecewise affine operator.

The finite element approximation to the Fermat times is a piecewise affine operator with a continuous index set, Γ_0. For operators having a continuous index set a special type of partition is treated in Kolbjørnsen and Omre (2002).

Definition 2 (Restricted linear partition) *A partition $\{\mathcal{A}_x\}_{x \in \mathcal{X}}$ with $\mathcal{X} \subset \mathbf{R}^d$ of $\mathcal{Z}$ is a restricted linear partition if,*

$$\mathcal{A}_x = \{\boldsymbol{R}_x z + \boldsymbol{r}_x = \boldsymbol{0}\} \cap \mathcal{C}_x,$$

with $\boldsymbol{R}_x : \mathcal{Z} \to \mathbf{R}^d$ being a bounded linear operator; $\boldsymbol{r}_x$ being a d-dimensional vector function; and $\mathcal{C}_x$ is any subset of $\mathcal{Z}$. The set of triplets $\{\boldsymbol{R}_x, \boldsymbol{r}_x, \mathcal{C}_x\}_{x \in \mathcal{X}}$ are the parameters of the restricted linear partition.

The approximate traveltime, $\tau_0^*(s)$ in Expression (2), can be represented as

$$\tau_0^*(s) = \tau(\gamma, s) \quad \textit{for} \quad s \in \mathcal{A}_\gamma \; ; \quad \gamma \in \Gamma_0 \tag{3}$$

with

$$\mathcal{A}_\gamma = \{s \in C^2(\mathcal{R}) : \tau(\gamma, s) \leq \tau(\tilde{\gamma}, s) \;\; \textit{for} \;\; \tilde{\gamma} \in \Gamma_0\} \;\; ,$$

hence $s \in \mathcal{A}_\gamma \Leftrightarrow \gamma = \gamma_0^*(s)$. That is, γ is the approximate Fermat path of s, using the predefined resolution given by Γ_0. Note further $\mathcal{A}_\gamma \subset \{\nabla_\gamma \tau(\gamma, s) = 0\}$ with $\nabla_\gamma \tau(\gamma, s)$ being the gradient of $\tau(\gamma, s)$ with respect to the path, evaluated for the Fermat path, γ. The operator $\nabla_\gamma \tau : \Gamma_0 \times C^2(\mathcal{R}) \to \mathbf{R}^d$ is linear in the second argument, i.e. slowness, for any value of the first, i.e. path. The partition $\{\mathcal{A}_\gamma\}_{\gamma \in \Gamma_0}$ in Expression (3) is hence a restricted linear partition according to Definition 2. Further the Hessian of the traveltime with respect to the path, $\nabla_\gamma \nabla_\gamma \tau$, is of importance. Note that due to the parameterization, the Hessian is tridiagonal.

For each traveltime there are two functionals, $\tau(\gamma, s)$ and $\tau_0^*(s)$, and two operators $\boldsymbol{g}(\gamma, s) = \nabla_\gamma \tau(\gamma, s)$ and $\boldsymbol{h}(\gamma, s) = \nabla_\gamma \nabla_\gamma \tau(\gamma, s)$, that are of importance. The operators $\boldsymbol{g}(\gamma, s)$ and $\boldsymbol{h}(\gamma, s)$ are both linear in the second argument and produce row vectors and matrices respectively. When r traveltimes are considered, the paths corresponding to each of the traveltimes are collected to form one large index, $\boldsymbol{\gamma} = [\gamma_1, \gamma_2, ..., \gamma_r]$, this should not be confused with the parameterization of the individual paths; i.e. $\gamma_i = (\gamma_{i1}, ..., \gamma_{id})$ for $i = 1, .., r$. The traveltime functionals are stacked to form vector valued operators, $\boldsymbol{\tau}(\boldsymbol{\gamma}, s)$ and $\boldsymbol{\tau}_0^*(s)$, and the relevant operators are joined,

$$\boldsymbol{g}(\boldsymbol{\gamma}, s) \;\; = \;\; \left[\begin{array}{cccc} \boldsymbol{g}(\gamma_1, s) & \boldsymbol{g}(\gamma_2, s) & \ldots & \boldsymbol{g}(\gamma_r, s) \end{array} \right] \tag{4}$$

$$\boldsymbol{h}(\boldsymbol{\gamma}, s) \;\; = \;\; \left[\begin{array}{cccc} \boldsymbol{h}(\gamma_1, s) & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{h}(\gamma_2, s) & & \\ \vdots & & \ddots & \mathbf{0} \\ \mathbf{0} & \ldots & \mathbf{0} & \boldsymbol{h}(\gamma_r, s) \end{array} \right] \;\; ,$$

further

$$\mathcal{A}_{\boldsymbol{\gamma}} = \bigcap_{i=1}^{r} \mathcal{A}_{\gamma_i} \; .$$

The traveltimes are hence approximated by a piecewise linear operator with a linear restricted partition, and can thereby be solved in a Gaussian framework by the methodology of Kolbjørnsen and Omre (2002). The Gaussian framework is defined next.

4 Statistical models

In Bayesian analysis knowledge and uncertainty is quantified by probability distributions. A generic distribution and a generic probability is denoted by p an P respectively. The relevant random variable will occasionally be displayed in the argument of p to clarify which distribution that is referred.

The likelihood is the statistical link between the parameter of interest and the observation. In Bayesian analysis it is given the interpretation of being the conditional distribution of traveltimes for a given slowness. In the current Gaussian framework, the observations are assigned additive Gaussian errors. Let $\boldsymbol{T}^*$ denote the random variable that is observed. The conditional distribution of $\boldsymbol{T}^*$ for a given s is then

$$p(\boldsymbol{t}^*|s) = N_r(\boldsymbol{\tau}_0^*(s), \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}) \ , \tag{5}$$

with $\boldsymbol{t}^*$ being the outcome of $\boldsymbol{T}^*$; s being a slowness field; N_r denoting the r dimensional multinormal distribution; $\boldsymbol{\tau}_0^*(s)$ being the approximate Fermat times for the slowness field s; and $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$ being the covariance for the observation error. Define also the indexed set of random variables $\boldsymbol{T}(\boldsymbol{\gamma})$, that is defined for each $\boldsymbol{\gamma}$ by the conditional distribution that correspond to observation of the path integrals, $\boldsymbol{\tau}(\boldsymbol{\gamma}, s)$,

$$p(\boldsymbol{t}_{\boldsymbol{\gamma}}|s) = N_r(\boldsymbol{\tau}(\boldsymbol{\gamma}, s), \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}) \ ,$$

with $\boldsymbol{t}_{\boldsymbol{\gamma}}$ denoting the outcome of $\boldsymbol{T}(\boldsymbol{\gamma})$; s being a slowness field; $\boldsymbol{\tau}(\boldsymbol{\gamma}, s)$ being the traveltimes in s along $\boldsymbol{\gamma}$; and $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$ being as in Expression (5). The marginal distribution of $\boldsymbol{T}^*$ is dependent on the distribution of the slowness and does not have an explicit representation in the current analysis. The marginal distribution of each of the random variables $\boldsymbol{T}(\boldsymbol{\gamma})$ will however be Gaussian if the slowness is so. The additive error term is modeled by a random error, it includes both observation errors and model errors.

In the current Gaussian framework the slowness, S, is assumed to be a Gaussian random field (Vanmarcke, 1983). The prior distribution is formally denoted $p(s)$, but is symbolic and not a density since S is a random field. The slowness to be reconstructed is assumed to be two times continuously differentiable, see Section 2. Gaussian random fields are well suited to represent different degrees of smoothness. In the presentation below it is assumed that the slowness field is almost surely two times continuously differentiable, i.e. $P(S \in C^2(\mathcal{R})) = 1$, see for example Stein (1999) for details about how to define such a Gaussian random field. This smoothness criterion is somewhat relaxed in Section 6, however.

The random variables that are defined by randomizing $\boldsymbol{g}(\boldsymbol{\gamma}, s)$ and $\boldsymbol{h}(\boldsymbol{\gamma}, s)$, see Expression (4), over the prior distribution of S for a fixed selection of paths, are denoted by capital letters, i.e. $\boldsymbol{G}(\boldsymbol{\gamma})$ and $\boldsymbol{H}(\boldsymbol{\gamma})$. Because the operators, $\boldsymbol{g}(\boldsymbol{\gamma}, s)$ and $\boldsymbol{h}(\boldsymbol{\gamma}, s)$, are linear in the second argument the corresponding random variables are Gaussian. These random variables are used to decompose the posterior distribution.

5 Representing the posterior distribution

The posterior distribution in the inverse problem of cross well tomography is decomposed as a mixture of truncated Gaussian distributions. This representation is in turn used to define an algorithm to sample the posterior distribution. The samples from the posterior distribution yields an approximation of the posterior distribution.

The theory of piecewise affine inverse problems, is developed in a Gaussian framework in Kolbjørnsen and Omre (2002). Using the notation introduced above, the posterior distribution of S can be represented as a mixture distribution

$$\begin{aligned} & p\{S = s, S \in \mathcal{A}_{\boldsymbol{\gamma}}, \boldsymbol{H}(\boldsymbol{\gamma}) = \boldsymbol{h} | \boldsymbol{T}^* = \boldsymbol{t}^*\} \\ & = p\{S = s | S \in \mathcal{A}_{\boldsymbol{\gamma}}, (\boldsymbol{T}(\boldsymbol{\gamma}), \boldsymbol{G}(\boldsymbol{\gamma}), \boldsymbol{H}(\boldsymbol{\gamma})) = (\boldsymbol{t}^*, \boldsymbol{0}, \boldsymbol{h})\} \\ & \times \mathrm{P}\{S \in \mathcal{A}_{\boldsymbol{\gamma}} | (\boldsymbol{T}(\boldsymbol{\gamma}), \boldsymbol{G}(\boldsymbol{\gamma}), \boldsymbol{H}(\boldsymbol{\gamma})) = (\boldsymbol{t}^*, \boldsymbol{0}, \boldsymbol{h})\} \\ & \times \frac{|det(\boldsymbol{h})| p\{(\boldsymbol{T}(\boldsymbol{\gamma}), \boldsymbol{G}(\boldsymbol{\gamma}), \boldsymbol{H}(\boldsymbol{\gamma})) = (\boldsymbol{t}^*, \boldsymbol{0}, \boldsymbol{h})\}}{p\{\boldsymbol{T}^* = \boldsymbol{t}^*\}} \quad . \end{aligned} \tag{6}$$

The distribution on the left hand side is the posterior distributions of S when $S \in \mathcal{A}_{\boldsymbol{\gamma}}$ and $\boldsymbol{H}(\boldsymbol{\gamma})$ have the value $\boldsymbol{h}$. The marginal posterior distribution of S is obtained by randomizing Expression (6) over $\boldsymbol{\gamma}$ and $\boldsymbol{h}$. The first term in Expression (6) is a truncated Gaussian distribution, since the equality constraint is linear. The second term is a probability and the third term is a non-negative measure on $\Gamma_0 \times \mathbf{R}^{(2d-1)r}$, with d being the number of internal nodes and r being the number of observations. The product of the second and third term is the posterior density for S being in $\mathcal{A}_{\boldsymbol{\gamma}}$ with $\boldsymbol{H}(\boldsymbol{\gamma})$ having the value $\boldsymbol{h}$. Note that $\boldsymbol{h}$ is fully described by $(2d-1)r$ values, see Section 3. In addition a necessary condition for $S \in \mathcal{A}_{\gamma}$ is that $\boldsymbol{H}(\boldsymbol{\gamma})$ is positive definite. The mixing distribution of Expression (6), provides a sampling strategy for the posterior.

Algorithm 1 *CTGM-algorithm (Continuous Truncated Gaussian Mixing)*

1. *Sample* $\boldsymbol{\gamma}^{\#}, \boldsymbol{h}^{\#} \sim q(\boldsymbol{\gamma}, \boldsymbol{h})$ *with*

$$q(\boldsymbol{\gamma}, \boldsymbol{h}) \propto |det(\boldsymbol{h})| p\{(\boldsymbol{T}(\boldsymbol{\gamma}), \boldsymbol{G}(\boldsymbol{\gamma}), \boldsymbol{H}(\boldsymbol{\gamma})) = (\boldsymbol{t}^*, \boldsymbol{0}, \boldsymbol{h})\}$$

2. *Sample* $s^{\#} \sim p\left\{S = s | (\boldsymbol{T}(\boldsymbol{\gamma}^{\#}), \boldsymbol{G}(\boldsymbol{\gamma}^{\#}), \boldsymbol{H}(\boldsymbol{\gamma}^{\#})) = (\boldsymbol{t}^*, \boldsymbol{0}, \boldsymbol{h}^{\#})\right\}$
3. *If* $S^{\#} \in \mathcal{A}_{\boldsymbol{\gamma}^{\#}}$ *stop.*

The algorithm splits the sampling into a nonlinear step, a linear step and an acceptance step. In the nonlinear step a value for the the Fermat paths and

the Hessian of the traveltimes along the Fermat paths is proposed. The matrix $\boldsymbol{h}^{\#}$ is restricted to be positive definite, hence the paths $\boldsymbol{\gamma}^{\#}$ are local minima. In the second step a slowness field, $s^{\#}$, that have local minima along the paths $\boldsymbol{\gamma}^{\#}$, with $\boldsymbol{h}(\boldsymbol{\gamma}^{\#}, s^{\#}) = \boldsymbol{h}^{\#}$ is drawn. In the third step it is controlled that $\boldsymbol{\gamma}^{\#}$ in fact is the Fermat paths, if not the sampled slowness is rejected and a new pair of $(\boldsymbol{\gamma}, \boldsymbol{h})$ must be drawn. Since the proposed paths are guaranteed to be local minima, there is usually a high acceptance rate in the third step. The nonlinear step in the algorithm is the challenge. To sample the distribution $q(\boldsymbol{\gamma}, \boldsymbol{h})$ a McMC algorithm is used. The decomposition given in Expression (6), can also be exploited in other types of algorithms. The benefit of using the decomposition is that it uses the global structure of the inverse problem.

6 Generalization to a non-smooth slowness

The smoothness assumption regarding the slowness is common in a continuous formulation of cross well tomography. In the current work it is however imposed by the solution method and is hence undesirable. In this section the theory is extended to account for small perturbations from a smooth background, let

$$s(x, z) = s_L(x, z) + \epsilon\, s_H(x, z) \ ,$$

with s_L being a lowfrequent background model; and $\epsilon\, s_H$ being a highfrequent perturbation with ϵ being a small number. By a standard perturbation argument, similar to the one used in Boyse and Keller (1995), the traveltime can be expanded in an asymptotic series in powers of ϵ. Including only the first order, this reads

$$\boldsymbol{\tau}^*(s_L + \epsilon\, s_H) = \boldsymbol{\tau}^*(s_L) + \epsilon\, \boldsymbol{\tau}(\boldsymbol{\gamma}^*(s_L), s_H) + \mathcal{O}(\epsilon^2)$$

with $\boldsymbol{\tau}^*(s_L)$ being the Fermat times in the lowfrequent part of the slowness; $\boldsymbol{\gamma}^*(s_L)$ being the Fermat path in the lowfrequent part; $\boldsymbol{\tau}(\boldsymbol{\gamma}^*(s_L), s_H)$ being the line integral of s_H along $\boldsymbol{\gamma}^*(s_L)$; and $\mathcal{O}(\epsilon^2)$ being higher order terms which are neglected in what follows. The likelihood in Expression (5) is now replaced by

$$p(\boldsymbol{t}^*|s_L, s_H) = N_r\left(\boldsymbol{\tau}_0^*(s_L) + \epsilon\, \boldsymbol{\tau}(\boldsymbol{\gamma}_0^*(s_L), s_H), \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\right) \ ,$$

with $\boldsymbol{\tau}_0^*(s_L)$ and $\boldsymbol{\gamma}_0^*(s_L)$ being the approximate Fermat times and paths in the lowfrequent part of the slowness respectively.

The sampling of the the lowfrequent and highfrequent part is done sequentially. For a fixed low frequent part the problem of sampling the highfrequent part is the linearized problem for a non-constant back ground. The challenge is hence to sample the lowfrequent part in the presence of the highfrequent part. This can be done by computing the marginal likelihood of s_L. Assuming S_H to be a Gaussian random field independent of s_L this can be done analytically. If S_H is centered the marginal likelihood for $s_L \in \mathcal{A}_{\boldsymbol{\gamma}}$ is,

$$p(\boldsymbol{t}^*|s_L) = N_r(\boldsymbol{\tau}_0^*(s_L), \boldsymbol{\Sigma}(\boldsymbol{\gamma}) + \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}) \ ,$$

with $\mathbf{\Sigma}(\boldsymbol{\gamma})$ being the covariance of line integrals of ϵS_H along the paths, $\boldsymbol{\gamma}$. Assuming that S_L is a Gaussian random field, this formulation is still within the scope of the theory of piecewise affine inverse problems developed by Kolbjørnsen and Omre (2002).

The characteristic that allows for the generalization is that the highfrequent part have an additive effect for which the statistical properties only depend on the index of the piecewise affine inverse problem, i.e. the Fermat paths. Sølna and Papanicolaou (2000) find a similar result for a different type of deviation from a smooth background.

7 Example

In the current section a synthetic example is investigated to highlight some of the differences between the current approach and a linearized problem. The traveltimes investigated relates to the slowness in Figure 2.

The slowness is a stationary Gaussian random field and is defined by its mean, variance and spectral density, these are denoted by μ_S, σ_S^2 and $\phi_S(k_x, k_z)$ respectively. It is convenient to specify the correlation in terms of the spectral density since this makes it easier to control the differentiability of the random field. The spectral density is assumed to have the form

$$\phi_S(k_x, k_z) \propto \left(1 + (k_z\, L_z)^2 + (k_x\, L_x)^2\right)^{-(\nu+2)/2},$$

with k_z and k_x being spatial frequencies; L_z and L_x being scales in depth and lateral direction respectively; and ν being the parameter that controls the smoothness. In the subsequent examples the prior distribution is defined by $\mu_S = 0.5\,\text{ms/m}$, $\sigma_S = 0.06\,\text{ms/m}$, $L_z = 225\,\text{m}$, $L_x = 130\,\text{m}$ and $\nu = 18$. Figure 6 show the resulting covariance function for the depth, z, and the lateral component, x. The slowness field in Figure 2 is a random sample from this prior distribution.

The observations have variance $\mathbf{\Sigma}_{\boldsymbol{\varepsilon}} = \sigma^2 \boldsymbol{I}_{r\times r}$, with $\sigma = 0.1\,\text{ms}$ being the standard deviation of the error; $\boldsymbol{I}_{r\times r}$ being the $r \times r$ identity matrix; and r being the number of observations. The observations are hence recorded with a high precision since the travel times are ranging from 48.1 ms to 76.6 ms

7.1 One observation

In this paragraph only one observation is considered. The source is in the left well at the depth 150 m and the receiver is in the right well at the depth 50 m. The approximation of the Fermat path in the true slowness field is displayed in Figure 5 for a variable number of internal nodes. The approximation of the Fermat time as a function of the number of internal nodes is displayed in the

top right corner in Figure 4. The observed traveltime is 76.6 ms. One traveltime observation hardly provide any information regarding the slowness field, hence no features of the true slowness can be expected to be retrieved. The example highlight differences between linear and nonlinear cross well tomography, however.

The inversion procedure is carried out for zero, one and seven internal nodes. The case with no internal nodes correspond to the linear case and is not discussed in any further detail. When only one internal node is considered, the mixing distribution $q(\gamma, h)$ is two dimensional, the density for the current case is displayed in Figure 7. To sample the mixing distribution for the case of seven internal nodes, a Markov chain is constructed. The algorithm use a diffusion step to sample the path. For a given path, γ, the distribution $q(\gamma, \boldsymbol{h})$ is approximated by a Gaussian distribution, this distribution is sampled sequentially to assure $\boldsymbol{h}$ to be positive definite. One sample is extracted for every 200 iteration, extracting a total of 3000 samples. Figure 8(a) show the value of the 2end, 4th and 6th internal node, and Figure 8(b) show the corresponding diagonal elements of $\boldsymbol{h}$. The plots show that the algorithm is slowly mixing. Figure 9(a)-(c) show the paths used in the inversion when zero, one and seven internal nodes are used respectively. For the case of one and seven internal nodes these are the samples from the corresponding mixing distribution in Step 1 of the CTGM-algorithm. For comparison the true path is plotted in the same figures. The uncertainty of the path is clearly illustrated by the figures. The acceptance rate in the third step of the CTGM-algorithm is 96% and 92% for the case with one and seven nodes respectively.

Figure 10(a)-(c) show the final estimates using the three strategies. The estimate for zero internal nodes is obtained analytically. Visually the estimates appear to be similar. All estimates increase the slowness along the line connecting the source and the receiver. The main effect of the internal nodes are better seen in cross sections of the estimates. Figure 11(a)-(c) show cross sections of the estimates at $x = 10\,\mathrm{m}$, $x = 50\,\mathrm{m}$ and $x = 75\,\mathrm{m}$ respectively. The nonlinear estimates are consistently larger, and have a larger region of influence. The deviation from the background is 20% larger for the case with seven internal nodes than it is for the linear estimate. Much of the nonlinear effect on the estimate is present in the case with only one internal node.

The main effect of the nonlinearity is however hidden by the averaging that is done in the estimation. The nonlinearity is present in the individual samples. To illustrate the differences, 500 samples from the three conditional distributions are used. Let γ_0 denote the direct line from the source to the receiver. For each sample, $s^{\#}$, the two traveltime functionals $\tau(\gamma_0, s^{\#})$ and $\tau^*(s^{\#})$ are computed. That is the line integral of the slowness along γ_0 and the Fermat time. Figure 12(a)-(c) display the scatter plot of these two functionals evaluated for each sample using three approaches. In linear tomography the line integral remains stable and the Fermat time fluctuates, whereas in the case with seven internal

nodes the opposite effect is observed. For the case with one internal node the Fermat times are quite stable, but some large deviations are present. Compare also some of the conditional probabilities that is illustrated in Figure 12. The percentage of the samples having Fermat time less than 75 ms is 29%, 5% and 0%, and the percentage of the samples having have line integral larger than 78 ms is 0%, 16% and 24% for the case with zero, one and seven internal nodes respectively. Much of the nonlinear effect is hence gained by including just one internal node.

In the algorithm, the Fermat path is drawn conditioned to the observed traveltime. The mixing distribution of the Fermat path, $q(\gamma, h)$, is hence dependent on the observed value of the traveltime. Figure 13(a) and (b) visualize this effect in the case with one internal node. The figures show $q(\gamma, h)$ for $t = 50$ ms and $t = 100$ ms. Note that $q(\gamma, h)$ is not the posterior distribution of (γ, h) since the acceptance probability is factored out, but $q(\gamma, h)$ still indicate the general shape of the distribution since the acceptrate in the third step of the CTGM-algorithm is large. When the observed value of t is small, i.e. $t = 50$ ms, it is likely that the path has followed the direct line from source to the receiver. This is illustrated in Figure 14(a) where 1000 paths sampled from $q(\gamma, h)$ are displayed. Notice the low spread of the samples that, indicate a channel of high velocity connecting the source and receiver. When the observed value of of t is large, i.e. $t = 100$ ms, it is likely that the Fermat path is bent either up or down as is indicated by the bi-modality in Figure 13(b). This is illustrated in Figure 14(b) where 1000 paths sampled from $q(\gamma, h)$ are displayed. Notice how most paths avoid the middle of the figure. This indicate a bump of low velocity located on the direct line connecting the source and the receiver.

7.2 Several observations

In this paragraph all 16 traveltimes, see Figure 2, are considered. Compared to the results of the previous paragraph, more of the structure of the slowness field is expected to be recovered.

The inversion procedure is carried out for zero and one internal node. The results for zero internal nodes are obtained analytically. The results for one internal node is obtained using the CTGM-algorithm. To sample the mixing distribution in Step 1 of the CTGM-algorithm a Markov chain is constructed in the same manner as in the previous paragraph. In each step a change is proposed in all the paths simultaneously. For the given path proposal the distribution $q(\gamma, \boldsymbol{h})$ is approximated by a Gaussian distribution and sampled sequentially to assure $\boldsymbol{h}$ to be positive definite. A sample is extracted after every 400 iteration, extracting a total of 2000 samples. Figure 15 show the mixing plot of the 32 random variables that are sampled. In general the mixing plots are satisfactory, but the internal node in the path that start in the left well at depth 150 m and arrive in the right well at depth 50 m to is however mixing slightly slower than

the other parameters. The mixing plot for this parameter is in the top right corner in Figure 15(a). The acceptance rate in the third step of the CTGM-algorithm is 98%.

Figure 16(a) and (b), show the estimates from the two models. Visually the estimates appear to be similar and have captured some of the features of the slowness. A high slowness region in the true slowness is located is located from depth 60 m to 120 m and at lateral position 30 m to 100 m. This is also present in the estimates, but the shape is slightly wrong. At the depth of 150 meters the estimates have a high value at the left and a low value at the right. This is also so for the true slowness. Comparing the two different estimates closely, the features are more diffuse in the nonlinear estimate than in the linear estimate. When the deviation from the true surface is measured, the nonlinear estimate improve the quadratic loss by 10%. It is however substantially more time consuming to compute the nonlinear estimate.

As a measure of the variability, the pointwise variance is integrated. The posterior in the linear case has 30 % lower integrated variance than the posterior in the nonlinear case. This does not necessarily mean that the full posterior is better determined in the linear case, since the integrated variance only respond to the marginal posterior distributions. In the linear case each observation will reduce the posterior integrated variance. This is generally not true for nonlinear observations.

8 Discussion

The inverse problem in nonlinear cross well tomography is solved by a Bayesian methodology in a Gaussian framework. The traveltimes obey Fermat's principle. This variational structure is approximated by a finite element method. Under the finite element approximation the forward map of nonlinear cross well tomography is piecewise affine. For a test example the approximation is reasonable even for a coarse resolution of the finite elements.

The estimate is taken to be the posterior expectation which is optimal under quadratic loss. The posterior distribution is explored by sampling and the expectation is approximated by the sample average. When the conditional expectation is used as estimator, the estimated slowness field will not reproduce the Fermat times in the case of exact observations. This is due to the convexity of the problem. Each individual sample will have a Fermat time corresponding to the observed time, but the Fermat path will differ between samples. When all the samples are averaged the Fermat time will be a lower bound for the path integral along any path, hence the Fermat time in the average medium will be larger. In general it is difficult to preserve nonlinear properties in an estimator. In nonlinear cross well tomography this can however be done by estimating the Fermat paths and then average the slowness for the given Fermat times under

the given selection of Fermat paths. This will however raise issues on estimating the Fermat paths. This is further complicated by the fact that the posterior distribution of the Fermat paths can be multi modal, see Figure 13.

The ray paths are flexible in the current nonlinear approach whereas in a linear approach, they are fixed. When studied in an example with one observation the nonlinear estimate have a deviation from the background that is 20% larger than the linear estimate. When studied in an example with 16 observations, the nonlinear estimate perform 10% better in terms of quadratic loss compared to the linear estimate. In both cases however the estimates look similar and only a small amount is gained by using the methodology in this respect. The major impact of the nonlinearity is however regarding typical deviations from the estimate, i.e. in the uncertainty.

The challenge in the methodology is to sample the mixing distribution $q(\boldsymbol{\gamma}, \boldsymbol{h})$, see Algorithm 1. In the current work this is done by a naive implementation of a McMC algorithm, the resulting chain is slowly mixing. Efficient exploration of $q(\boldsymbol{\gamma}, \boldsymbol{h})$ is of high importance for further development of the methodology.

The prior distribution of the slowness field is Gaussian. Gaussian random fields constitute a large class of prior distributions and is in particular well suited for modeling of smoothness. The methodology can also be extended to priors being mixtures of Gaussian distributions.

Acknowledgments

This work was supported by the Research Council of Norway. The author thanks Henning Omre for helpful comments.

References

Aronsson, G. (1970) "Axiomatic derivation of Fermat's principle", SIAM Journal on applied Mathematics, Vol. 18, no. 3, pp 675-681.

Auld, B.A (1990) "Acoustic Fields and Waves in solids",Vol. 1, Robert E. Krieger, Melbourn, F1.

Berryman J.G. (1990) "Stable iterative reconstruction algorithm for nonlinear traveltime tomography", Inverse problems 6, pp 21-42.

Berryman, J.G. (1997) "Variational structure of inverse problems in wave propagation and vibration", Inverse problems in wave propagation (Minneapolis, MN, 1995), 13–44,IMA Vol. Math. Appl., 90, Springer, New York.

Bertero, M. (1989), "Linear inverse and ill-posed Problems", Advances in Electronics and Electron Physics, Academic Press, New York.

Boyse, W. and Keller, J.B. (1995) "Short acoustic, electromagnetic, and elastic waves in random media", J. Opt. Soc. Amer. A., Vol 12 , no. 2, pp 380-389.

Bube K.P. and Meadows M.A. (1998) "Characterization of the null space of a generally anisotropic medium in linearized cross well tomography", Geophys.J.Int, 133 pp 65-84.

Carfantan, H. and Mohammad-Djafari, A. (1997) "An overview of nonlinear diffraction tomography within the Bayesian estimation framework", Inverse problems of wave propagation and diffraction, 107–124, Lecture Notes in Phys., 486, Springer, Berlin.

Faridani A. (1997) "Results, old and new, in computed tomography, in inverse problems in wave propagation", G. Chevalent et al. (editors), The IMA Volumes in mathematics and its Applications, vol 90, Springer Verlag New York, pp 167-193.

Kolbjørnsen, O and Omre H. (2002) "Bayesian inversion of piecewise affine operators in a Gaussian framework"

Langan R.T., Lerche I. and Cutler R.T. (1985) "Tracing rays through heterogeneous media an accurate and efficient procedure", Geophysics, 50 pp 1456-1465.

Langan R.T. and Bube, K.P. (1998) "A resolution analysis of cross well seismic tomography and its implications for the formulation of a traveltime inversion algorithm", in Mathematics of Reflection Seismology, ed Symes W.W.SIAM Frontiers in Applied Mathematics.

Menke, W. (1984) "The resolving power of cross-borehole tomography", Geophys. Res.Lett., 11, pp 105-108.

Michelena, R.J (1993) "Singular value decomposition for cross well tomography", Geophysics 58, 1655-1661.

Natterer, F. (1980) "Efficient implementation of "optimal" algorithms in computerized tomography", Math. Methods Appl. Sci. 2, no. 4, 545–555.

Stein, M. L. (1999) "Interpolation of spatial data. Some theory for Kriging", Springer Series in Statistics. Springer-Verlag, New York.

Sølna, K. and Papanicolaou, G. (2000) "Ray Theory for a Locally Layered Random Medium", Waves in Random Media, vol 10, pp. 155-202.

Vanmarcke, E. (1983), "Random fields", The MIT press.

Vidale J.E. (1988) "Finite difference traveltime calculation", Bull.seism. Soc. Am, 78 pp 2062-2076.

Tables and figures

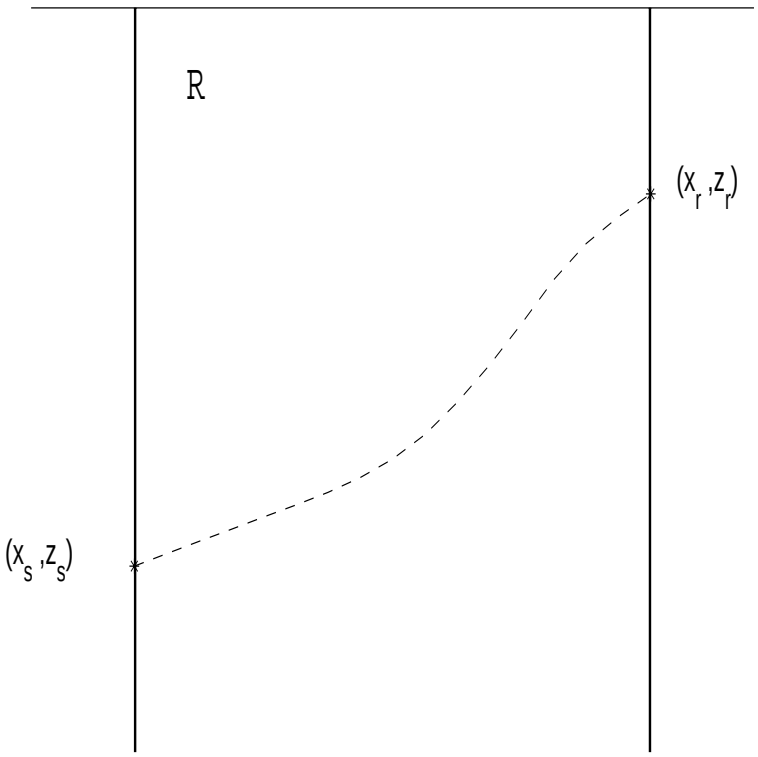
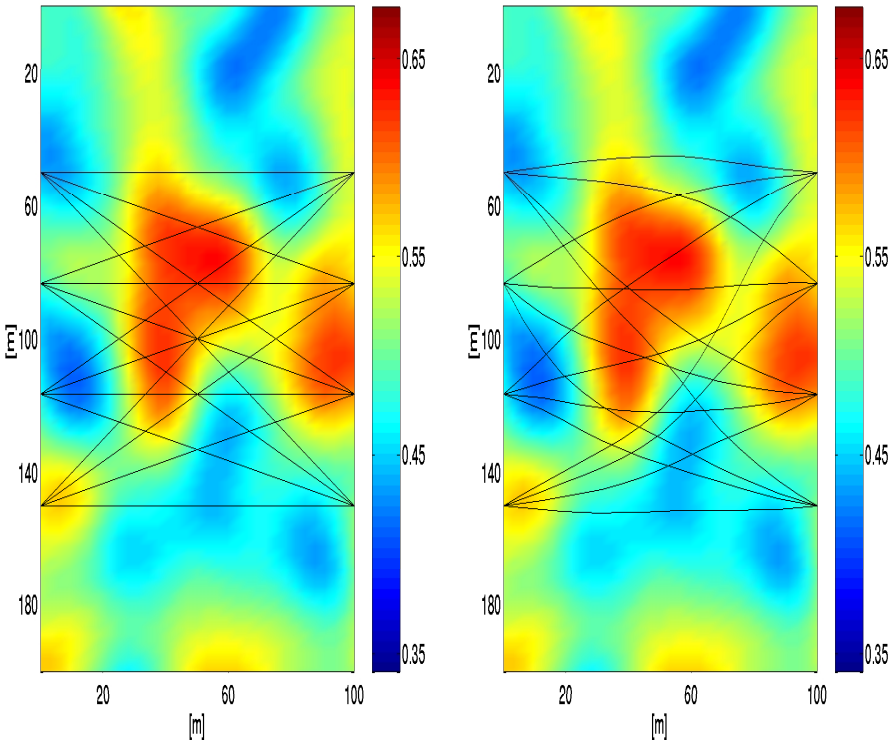


Figure 1: Cross well tomography. The two vertical lines are boreholes, the region between the two boreholes is $\mathcal{R}$. A source is situated in one well at the the location (x_s, z_s), a receiver is situated in the other well at the the location (x_r, z_r). The time it takes for a pulse to propagate from the source to the receiver is recorded.



(a) (b)

Figure 2: Linear paths and Fermat paths. The paths used for linear tomography (a); The Fermat paths for the superimposed slowness field (b).

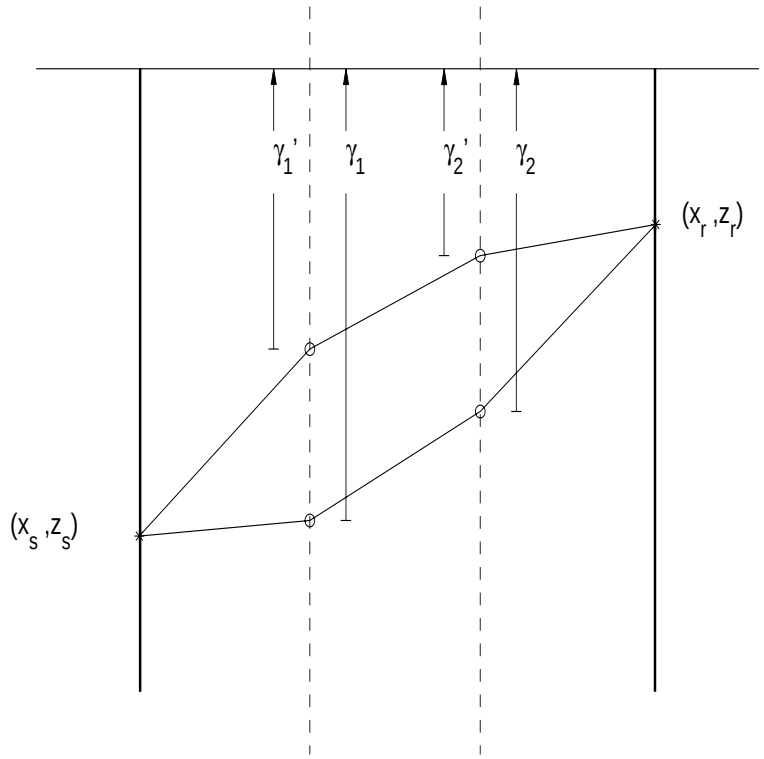


Figure 3: Finite element approximation to Fermat's principle. The path between the source and the receiver is restricted to be piecewise linear between internal nodes. Two different paths are displayed for the case with two internal nodes. The path is parameterized by the vertical distance to the knot point,(γ_1, γ_2) and (γ_1', γ_2') for the two paths respectively

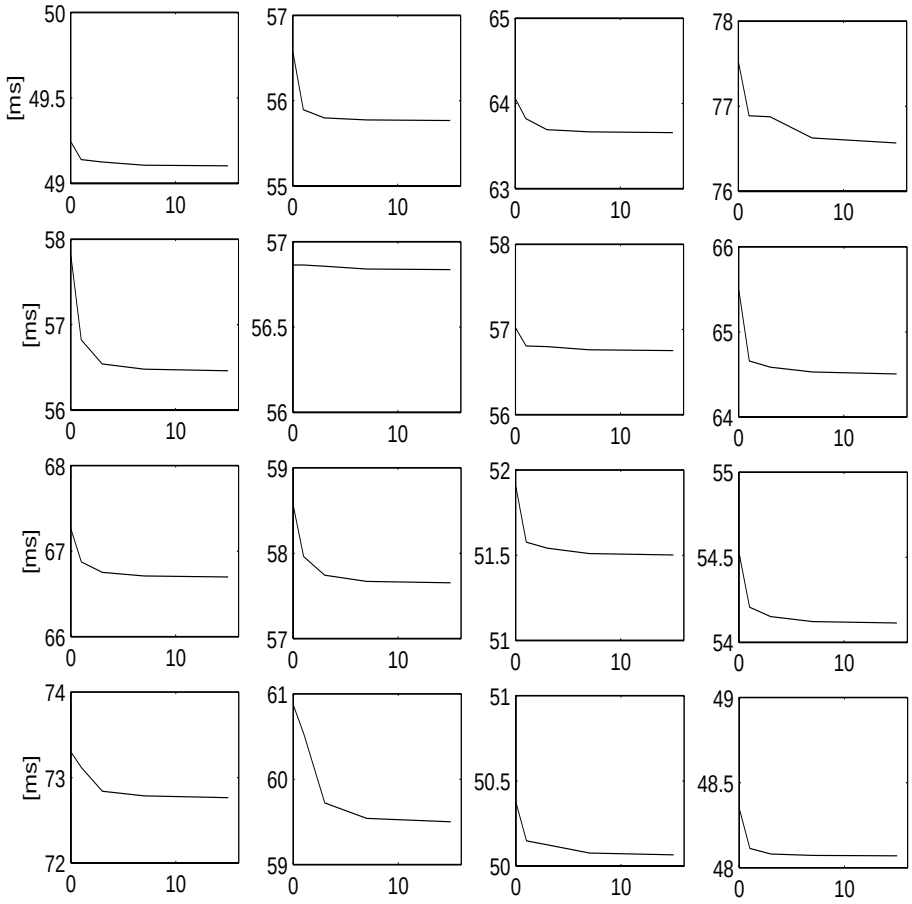


Figure 4: Minimum traveltime, dependence of number of internal nodes. For each of the 16 traveltimes, the minimum traveltime is plotted as a function of the number of internal nodes. Increasing column number correspond to increasing depth of starting point. Increasing row number correspond to increasing depth of end point. The values where computed for zero, one, three, seven and 15 internal nodes to have a monotone decay.

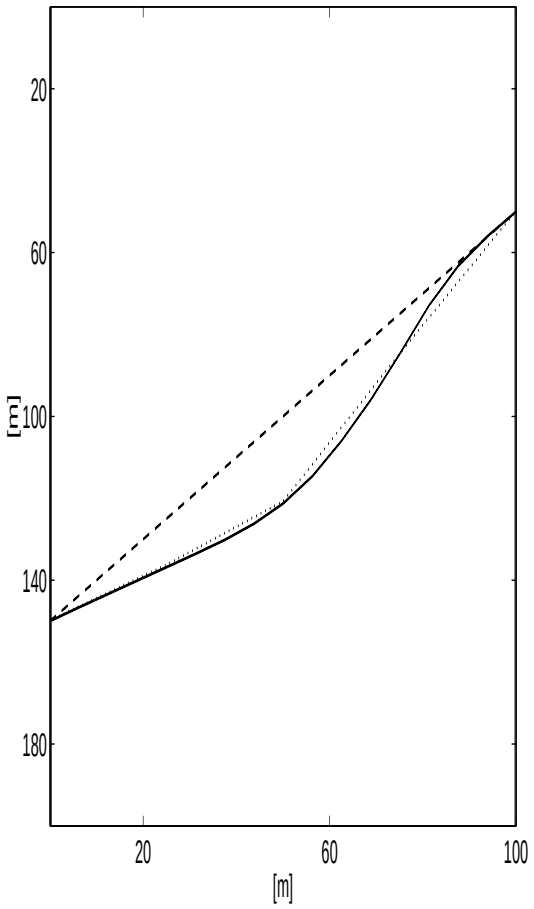


Figure 5: Minimum path, dependence of number of internal nodes. For one set of endpoints, the minimum path is plotted for zero (dashed line), three (dotted line) and 15 (full line) internal nodes.

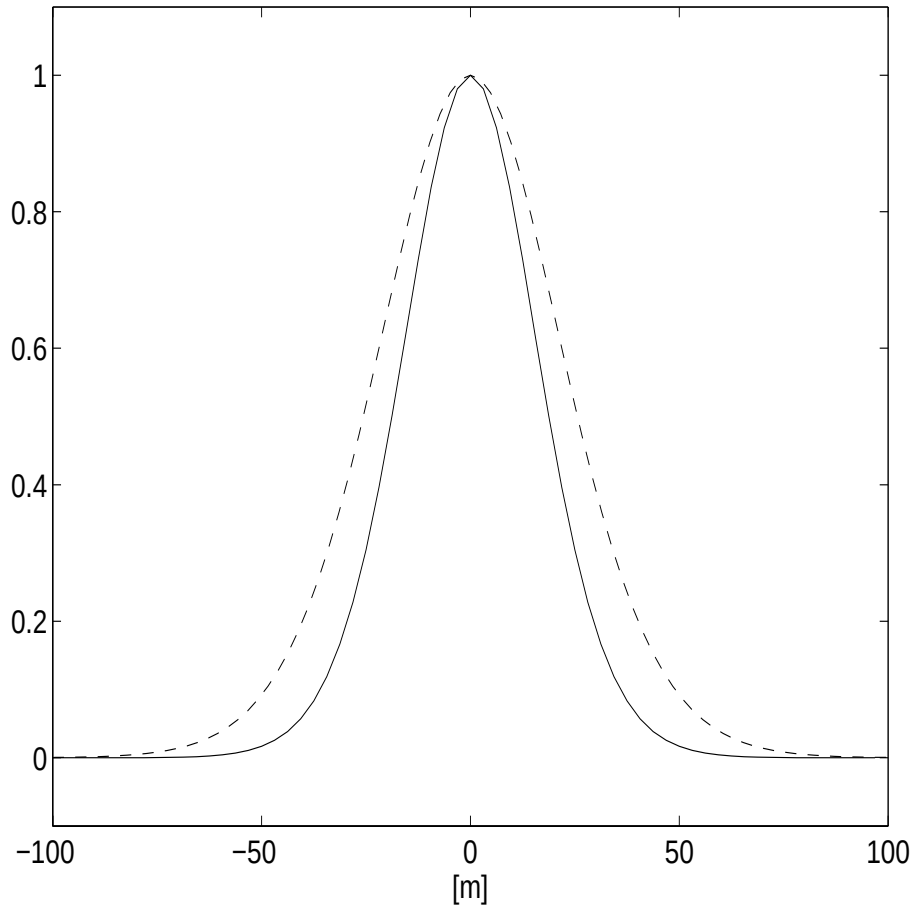


Figure 6: The correlation function used in the example of Section 7. The dashed line being for the depth, z, and the full line being for the lateral direction, x.

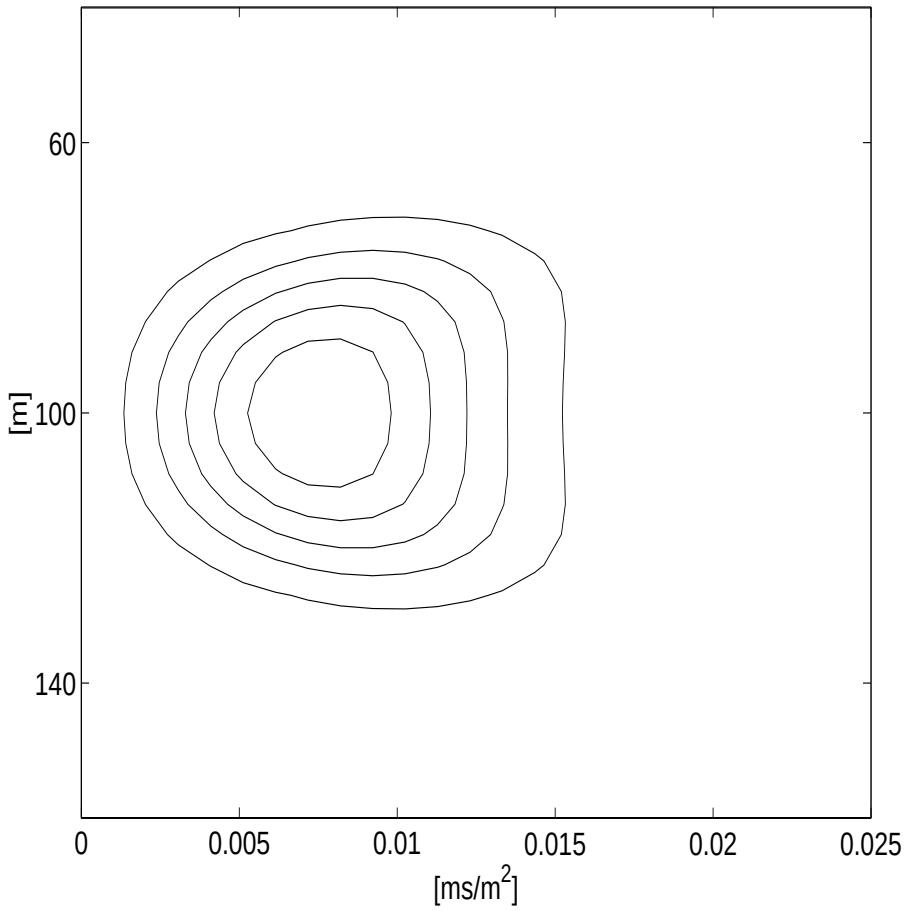
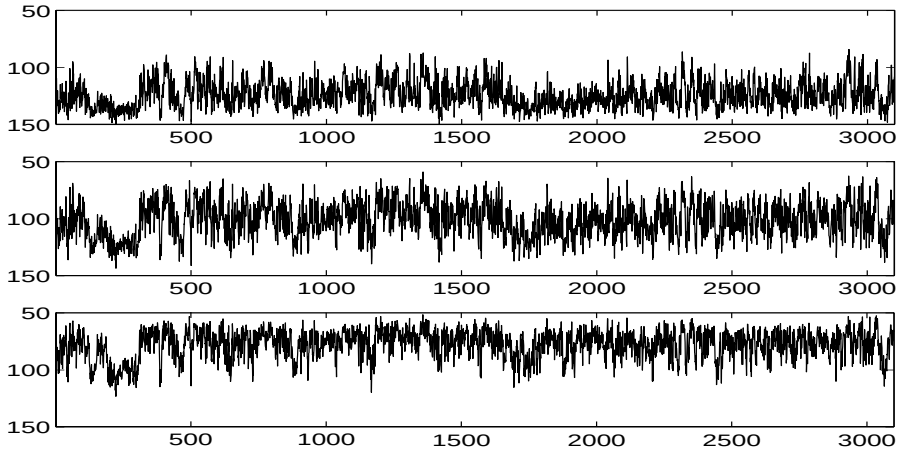
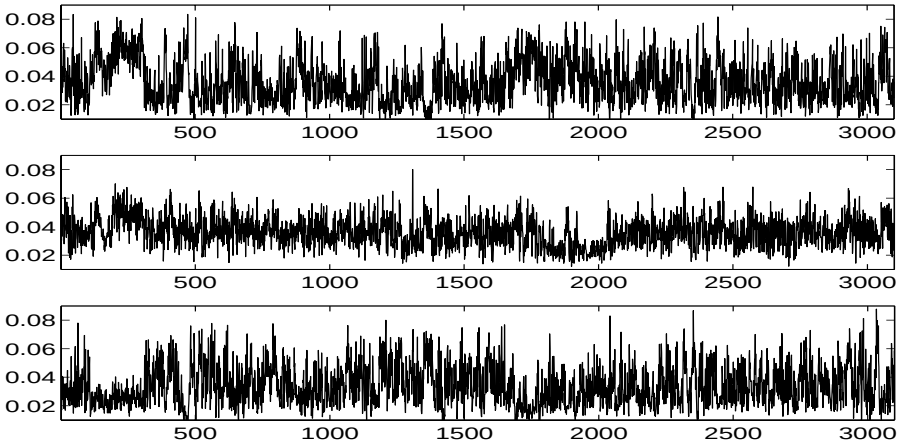


Figure 7: Proposal distribution for one internal node. The proposal distribution in the nonlinear step in the CTGM algorithm for the actual observation; i.e. $t = 76.6\,\text{ms}$.

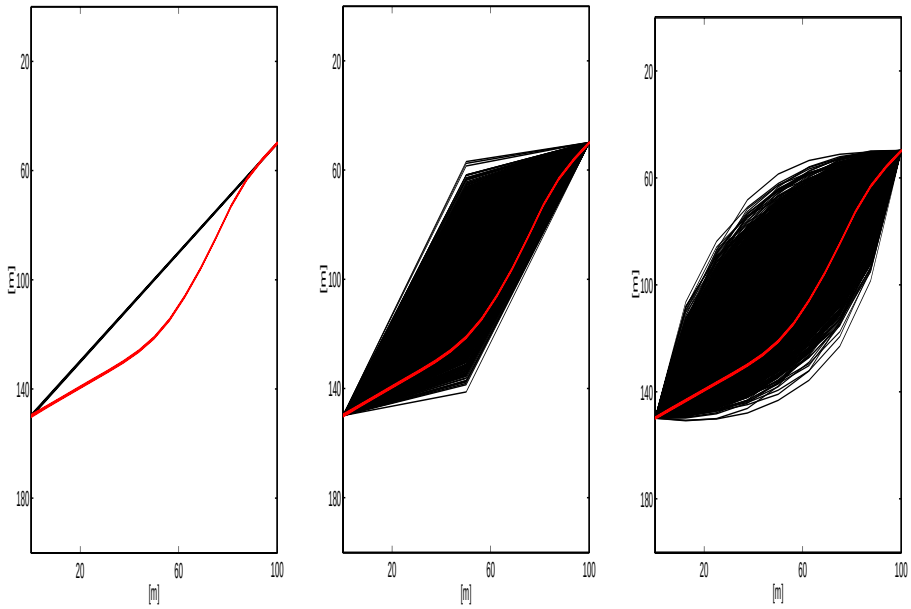


(a)



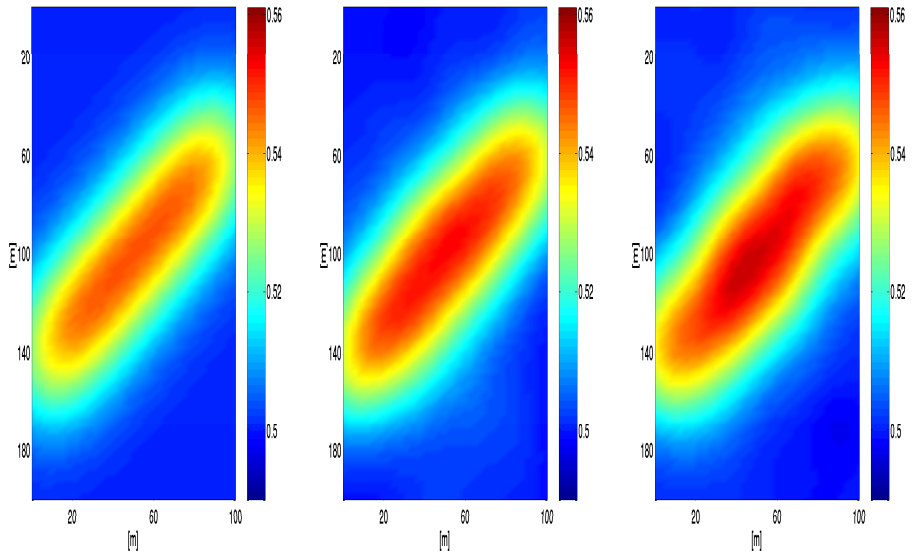
(b)

Figure 8: Mixing plot for the Markov chain used in nonlinear inversion for seven internal nodes. Three path parameters (a); three parameters for the gradient of the constraint (b). In both (a) and (b) the top is for the parameter corresponding to $x = 25\,\mathrm{m}$ the middle corresponding to $x = 50\,\mathrm{m}$ and the bottom corresponding to $x = 75\,\mathrm{m}$.



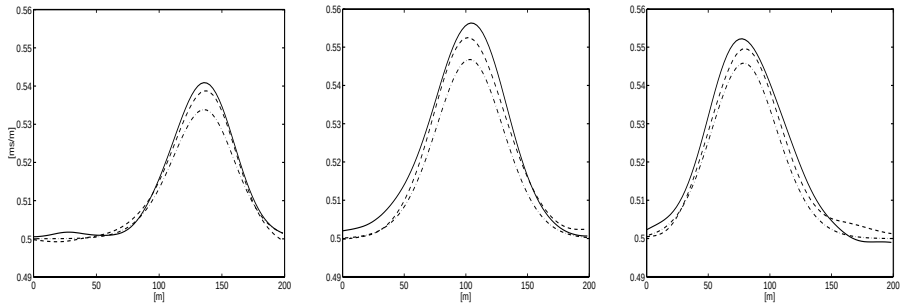
(a) (b) (c)

Figure 9: Comparison of paths used for reconstruction. The red line is the actual Fermat path in the problem, the black lines are the paths used in (a) linear inversion; (b) nonlinear inversion with one internal node;(c) nonlinear inversion with seven internal nodes. In (b) and (c) 3000 paths are displayed.



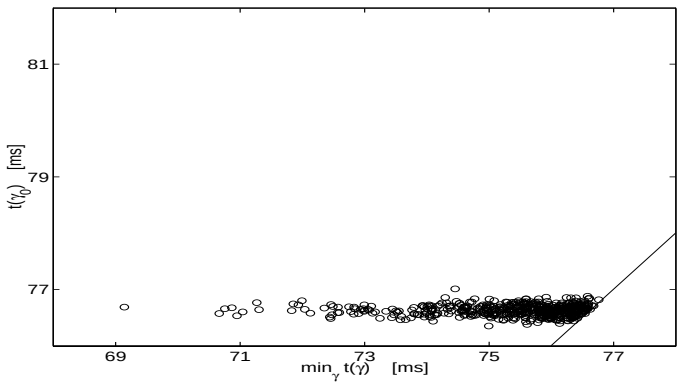
(a) (b) (c)

Figure 10: Comparison of estimates. The conditional expectation using (a) linear tomography; (b) nonlinear tomography one internal node; (c) nonlinear tomography seven internal nodes.

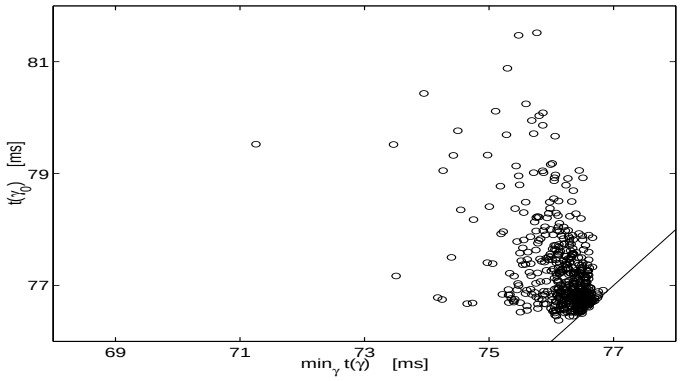


(a) (b) (c)

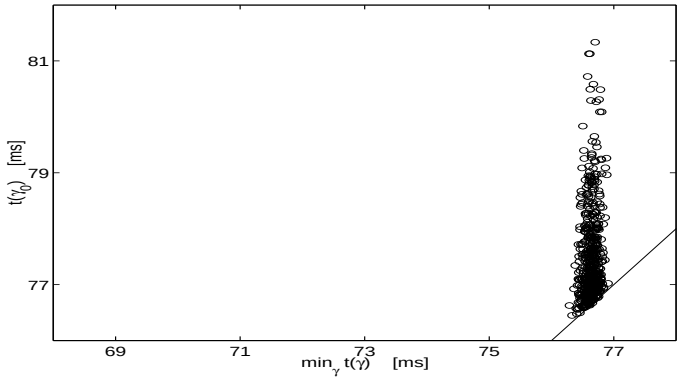
Figure 11: Cross sections of estimates. Dash/dot line - linear tomography; dashed line - nonlinear tomography with one internal node; full line - nonlinear tomography with seven internal nodes. The cross sections show vertical slices for lateral components (a) $x = 10\,\mathrm{m}$;(b) $x = 50\,\mathrm{m}$; (c) $x = 75\,\mathrm{m}$.



(a)



(b)



(c)

Figure 12: Comparison of results for linear and nonlinear inversion. The scatter plot of the posterior distribution of the traveltime along the linear path and the Fermat path for (a) linear tomography; (b) nonlinear tomography with one internal node; (c) nonlinear tomography with seven internal nodes.

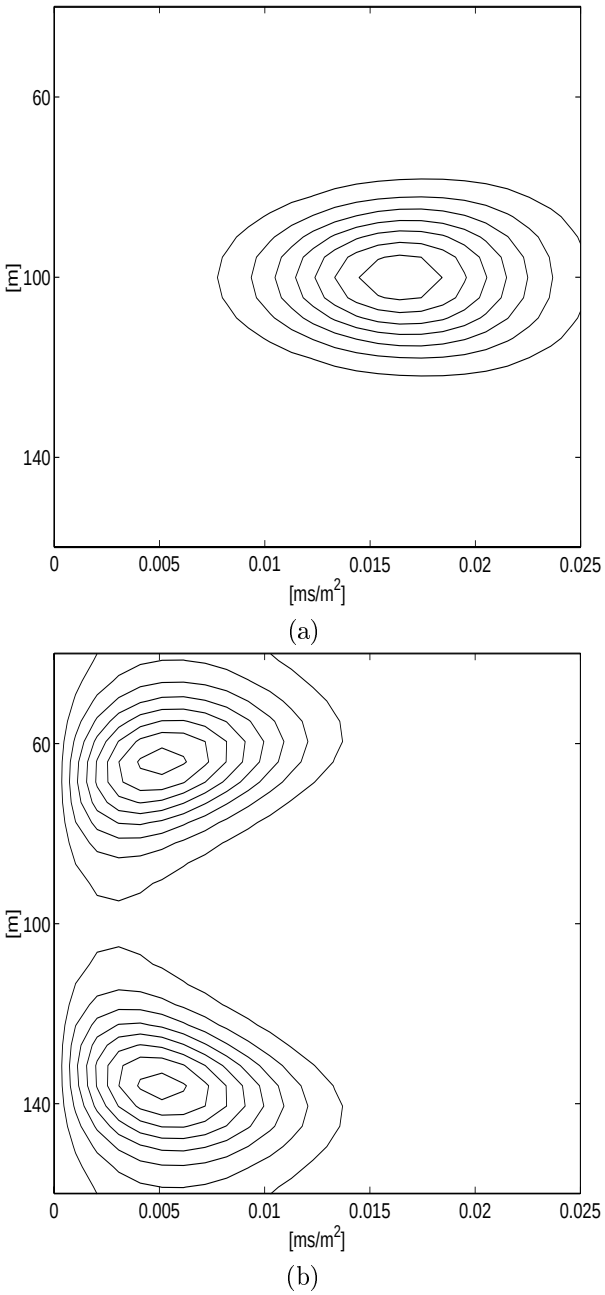
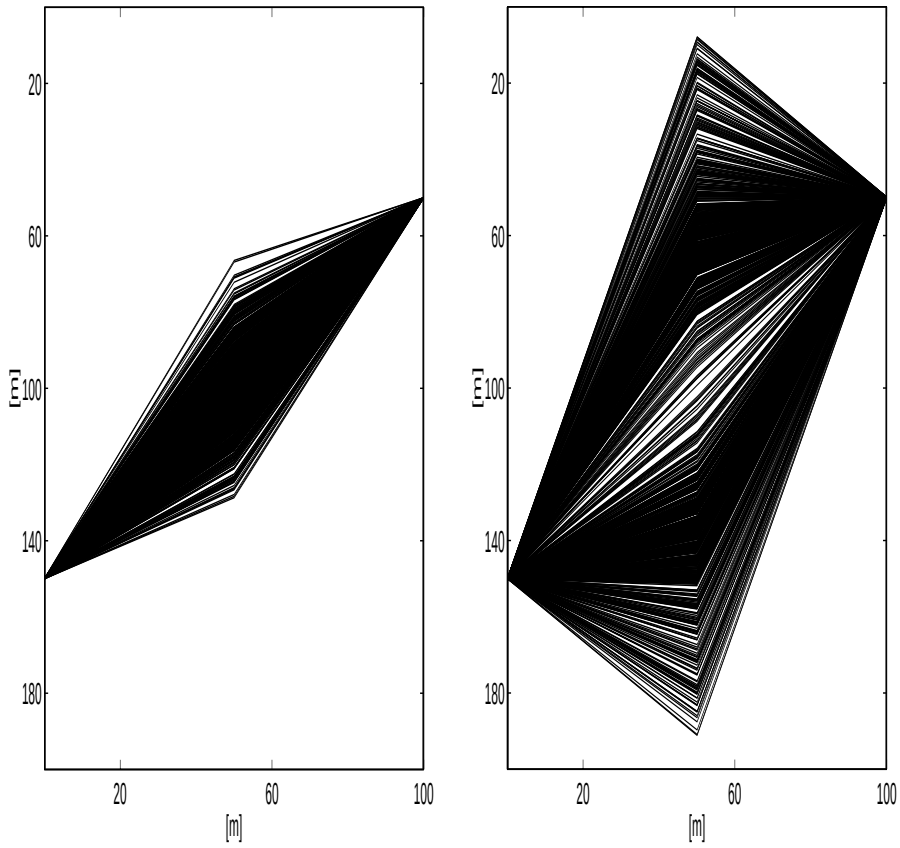
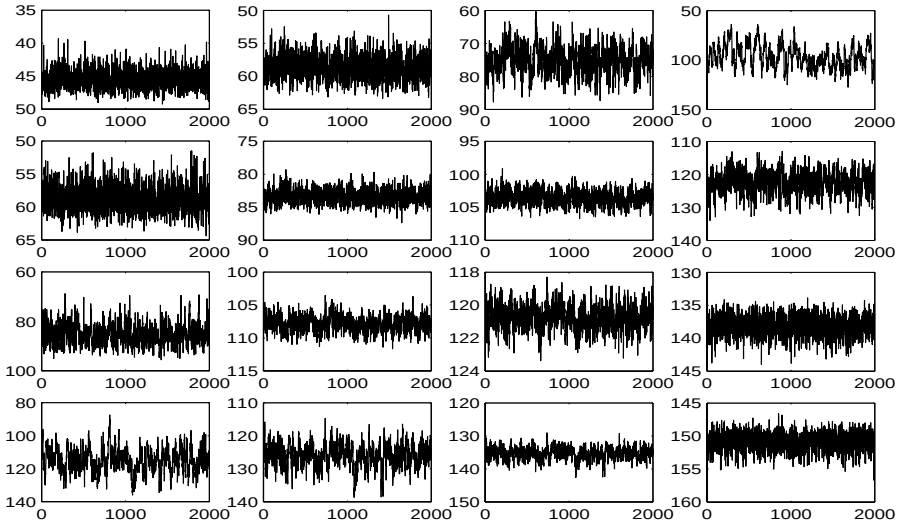


Figure 13: Proposal distribution for one internal node. The proposal distribution in the nonlinear step in the CTGM algorithm for extreme observations (a) $t = 50\,\text{ms}$; (b) $t = 100\,\text{ms}$.

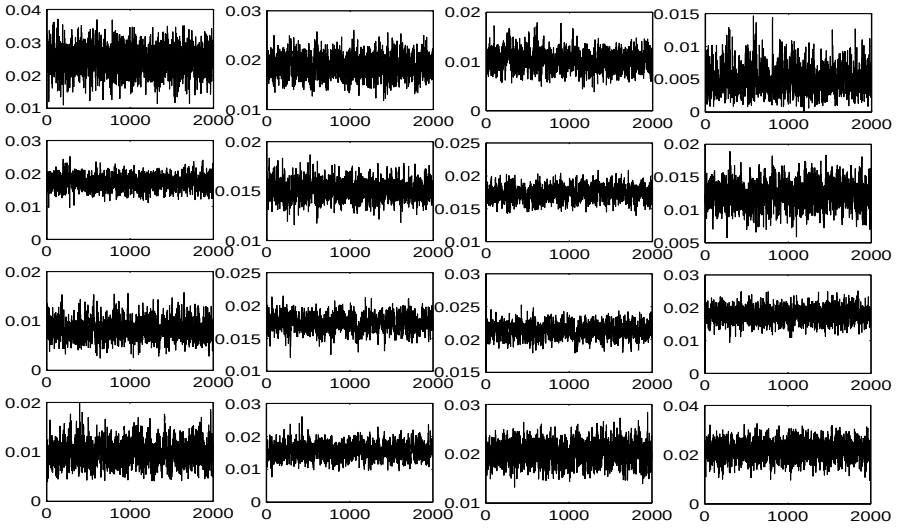


(a) (b)

Figure 14: Proposed paths with one internal node and extreme observations. 1000 proposed paths sampled from the proposal distributions plotted in Figure 13 in the nonlinear step in the CTGM algorithm for extreme observations (a) $t = 50.0\,\text{ms}$; (b) $t = 100\,\text{ms}$.

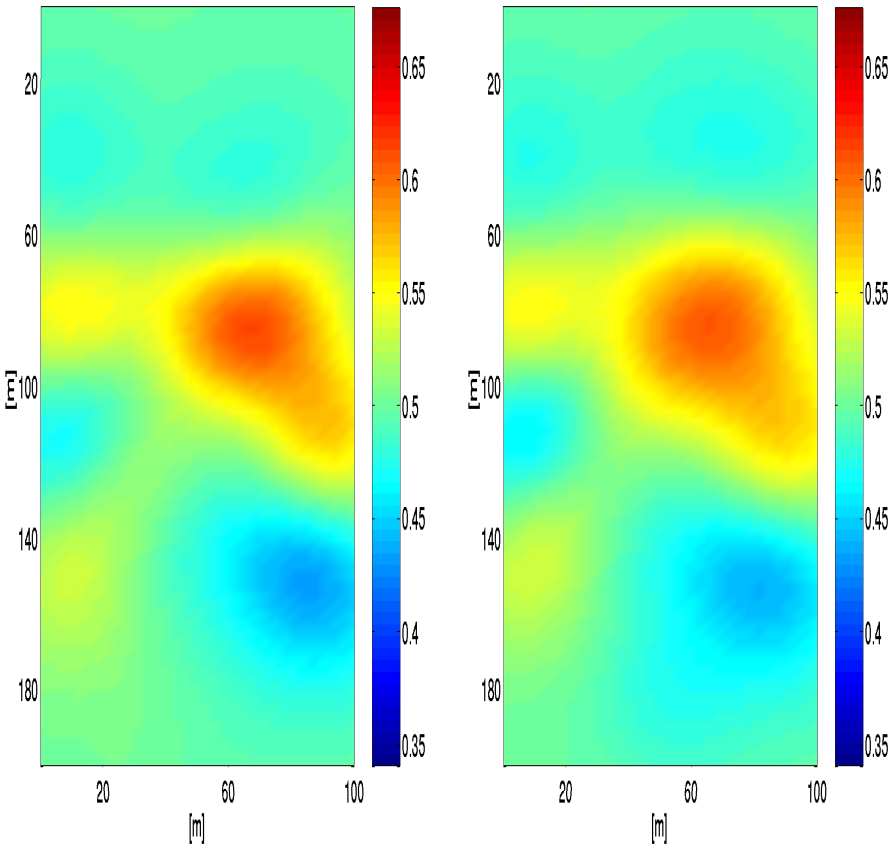


(a)



(b)

Figure 15: Mixing plot for the Markov chain used in nonlinear inversion for one internal node and 16 observations. The path parameters i.e. $\boldsymbol{\gamma}$ (a); the second derivative i.e. $\boldsymbol{h}$ (b). In both (a) and (b) the ordering of the figures is such that an increasing column number correspond to increasing depth of starting point of the path. Increasing row number correspond to increasing depth of end point of the path.



(a) (b)

Figure 16: Comparison of estimates 16 observations. The conditional expectation using (a) linear tomography; (b) nonlinear tomography one internal node.

IV

Cauchy prior for Bayesian linearized seismic AVO inversion

Cauchy prior for Bayesian linearized seismic AVO inversion

Odd Kolbjørnsen

Department of mathematical sciences,
Norwegian university of science and technology
Trondheim, Norway

Abstract

A Bayesian approach is used to estimate material parameters of the underground. The parameters to be estimated are pressure wave velocity, shear wave velocity and density. The data analyzed are angle gathers. The underground usually have a layered structure. A stationary log Gaussian prior model is frequently used, but is not adequate to describe a layered structure. In the current approach the prior is modeled by a superposition of a Cauchy and Gaussian processes on a logarithmic scale. The Cauchy process yields a model for the layering whereas the Gaussian processes describe fluctuations within a layer. The physics of the likelihood is approximated by a linear operator between the logarithm of the material parameters and the observed seismic traces, the error structure is assumed to be Gaussian. The final estimate is optimal under the loss criterion of absolute deviation and is evaluated by Monte Carlo integration.

The current methodology is compared to a pure log Gaussian model. The material parameters observed in a well at the Sleipner Øst field is used as a test case and synthetic seismic observations are generated. Over all the velocity estimates in the current model reduces the risk by 7%, and the average length of the 90% credibility interval is reduced by 7%. In a region where the true velocity have large fluctuations, the velocity estimate in the current model improve by 14% and 10%. In a region where the true velocity is slowly varying, the 90% credibility interval is reduced by 10%. There are only minor effects of the model concerning the estimate of the rock density.

The current model is tested for real seismic observations collected in a marine seismic survey above the Sleipner Øst field. The inversion results are satisfactory, but the information content in the observations is small due to large errors in the data.

KEY WORDS: *Bayesian statistic, Independently scattered random measures, Deconvolution, Seismic inversion*

1 Introduction

The objective of seismic inversion, is to estimate material parameters of the underground. The observations are obtained by generating an acoustic wave above the target area and record the signal reflected from the underground.

A simplified model for the wave propagation, is obtained by regarding the reflected signal as the response of a locally vertical 1D-earth model, see for example Sheriff and Geldart (1995). In this model the reflected signal can be approximated by a convolution of a wavelet and the seismic reflectivity. The seismic reflectivity is again connected to the material parameters through the Zoeppritz equation.

The problem of seismic inversion is inherently ill-posed. The high frequency components of the wave that respond to high frequency changes in the rock are dampened due to intrinsic absorption, hence an exact stable inversion is beyond reach. A stable reconstruction of the material parameters, can only be obtained by providing, directly or indirectly, information about the preferred solution. The Bayesian formalism is well suited for this task. The Bayesian choice of prior distribution is a direct way of introducing preferences in the solution space.

Stationary Gaussian random fields are frequently used to construct prior distributions, this choice is particularly successful for solving linear inverse problems, due to the simplicity of the solution, (Tarantola 1987). Buland and Omre (2002a) treats the current problem in the Gaussian framework. A stationary Gaussian random field prior, give preference towards smooth solutions. At the geological scales considered in seismic exploration, the earth frequently have slow variations within a layered structure. Hence there appears to be a conflict between a stationary Gaussian random field prior, and the phenomenon under study.

In the current work, a Bayesian methodology is used to solve the problem of seismic inversion. This requires a prior distribution for the material parameters and a likelihood for the observations. To account for the layered structure, the prior distribution is described by a superposition of Gaussian and Cauchy random fields. Both these types of random fields can be constructed by the theory of independently scattered random measures (Rajput and Rosinski 1989). The likelihood model in the current paper is identical to the one used in Buland and Omre (2002a).

The objective of modern Bayesian inference is to explore the posterior distribution, which is formally proportional to the product of the prior and the likelihood. In the current approach this is done by sampling. The samples represent the space of uncertainty with respect to the inversion. In the current presentation the samples are combined to a single estimate using the loss criterion of absolute deviation.

There are alternative Bayesian approaches to modeling of layered structures. Common alternatives are to use a Bernoulli Gauss prior (Mendel 1983), or to model the layering as a point process and mark each point with a random leap size (Malinverno and Leaney 2000). There are also non-Bayesian approaches to inverse problems. Traditional approaches such as quadratic regularization (Tikhonov 1963) and filtering of singular values (Bertero 1989; Hansen 1998) are formally equivalent to Gaussian random field priors (Tarantola 1987; Whaba 1990), and suffer the same deficiencies. However, resent development in computational harmonic analysis, allows for reconstruction by wavelet-vaguelette decomposition (Donoho 1995; Abramovich and Silverman 1998). For a general class of function spaces, the reconstruction adaptively obtain the minimax rate of convergence in the zero noise limit within a log term. This approach is not pursued in the current paper, this is partially because the inverse problem that arise in the current setting does not have sufficient regularity, i.e. the singular values does not have a power law decay.

The data collection procedure and geophysical aspects of the the likelihood model is discussed in Section 2. The statistical construction of the prior and the likelihood is presented in Section 3. The posterior distribution is developed in Section 4, together with the sampling algorithm. In Section 5 the methodology is compared to the standard Gaussian theory in an example where synthetic seismic is generated based on material parameters observed in a well at the Sleipner Øst Field. Section 6 presents inversion of a seismic inline from the Sleipner Øst Field. Lastly, a discussion of the results is included in Section 7.

2 Data collection and geophysical model

The current section gives a brief introduction to practical aspects regarding the data collection and the geophysical assumptions used in the current work. This is all standard methodology in the geophysical community. A more detailed discussion on an introductory level can be found in Sheriff and Geldart (1995).

The material parameters of the earth considered in this work are α being the pressure wave velocity; β being the shear wave velocity; and ρ being the density. These parameters are sufficient to characterize an isotropic elastic medium. Other sets of three parameters are frequently used by geophysicists, but there is an one to one relation between different choices. Below each point at the surface, the material parameters are described by the time profile relative to the reflected wave, $\{\alpha(t), \beta(t), \rho(t)\}$. Figure 1 display the material parameters observed in a well at the Sleipner Øst Field. The observed depth profile is converted to a time profile. The conversion between time and depth is a standard problem in seismic inversion, but is not discussed here.

The seismic data that will be considered in the current presentation is recorded in the North Sea above the Sleipner Øst Field. In a marine seismic survey an

air gun attached to a ship, generates an acoustic wave. The wave propagates through the earth, and is reflected due to contrasts in the sub sea rock formations. The reflected signal propagates back to the surface and is recorded in several hydrophones located in a cable being towed by the ship. The hydrophones record the amplitude of the reflected pressure field as a function of time. A seismic trace is a strain or a pressure amplitude as a function of time. A collection of seismic traces is denoted a gather. The collection of seismic traces recorded in the hydrophones after one shot with the air gun is denoted a common shot point gather, for obvious reasons. In a seismic survey several common shot point gathers are collected, these gathers are further processed. Buland and Omre (2002a) lists 24 different steps of the processing sequence. The processed data are in the form of common depth point gathers, CDP gathers for short. In a CDP gather, the seismic signal correspond to the amplitudes in reflections occurring below one location at the surface, the distance from this location to the shot location is denoted the offset. The processing objective is to make the traces in the CDP gather correspond to primary reflections from a locally vertical 1D-earth model. In a vertical 1D-earth model, the material parameters of the earth is assumed to vary only with depth. The assumption of a vertical 1D-earth model is local, hence material parameters below different locations at the surface may vary at a larger scale.

The seismic signal in CDP gathers used in the inversion, is indexed by angle, θ, in addition to time, t, and are hence denoted angle gathers. The ray path of a wave is defined as the normal vector of the wave front. The angle reference, θ, indicate the angle between the ray path and the vertical line in the point where the reflection occur. Figure 2 show three ray paths that have a common angle of incidence to the vertical line. The amplitudes from these reflections will have the same same angle reference in the angle gather. As the time increases in the angle gather data must be collected at larger offset to keep the angle θ fixed. The time reference in the angle gather defines the depth to the reflecting point in terms of the zero offset reflection time. Figure 3 show three ray paths that have the same depth to the reflection. The amplitudes from these reflections will have the same same time reference in the angle gather. The time reference in the angle gather does not correspond to the physical reflection time. The physical reflection time increase with an increasing offset since the length of the ray path increase, see Figure 3.

A seismic inline of 176 angle gathers, each containing seismic traces for nine angles, were recorded and processed as a part of a seismic survey above the Sleipner Øst Field. Figure 4 shows a time window of the average trace for each angle gather in this inline. The averaging of traces in a gather is denoted stacking in the geophysical terminology. In Figure 5 one of the collected angle gathers is displayed. This gather has approximately the same surface coordinates as the well for which the material parameters are observed, see Figure 1.

The seismic signal in an angle gather can for each angle, θ, be modeled as a

convolution of a wavelet and seismic reflectivity corresponding to this angle,

$$d(\theta,t) = s_\theta * c_{PP}(\theta,t) + e_m(\theta,t)\ , \tag{1}$$

with t being zero offset reflection time; s_θ being a wavelet specific to the angle θ; $c_{PP}(\theta,t)$ being the seismic reflectivity for reflections occurring at the angle θ; and $e_m(\theta,t)$ being model error. For each time coordinate, the seismic reflectivity measure the strength of the reflection from this particular point. The subscript of c_{PP}, indicates that this is the reflectivity of a propagated pressure wave to a reflected pressure wave. There are also reflection coefficients involving shear waves. Only the pure pressure wave reflections are considered here, since only pressure measurements are recorded in the hydrophones.

The Zoeppritz equation describes the dependence between the local material parameters, $\{\alpha(t),\beta(t),\rho(t)\}$, and the seismic reflectivity, $c_{PP}(\theta,t)$, for any angle of the ray path in the 1D-earth model. Stolt and Weglein (1985) introduce a time continuous weak contrast approximation to the Zoeppritz equation. By additional assumptions defined below the dependencies of the angle and the material parameter separates for the seismic reflectivity, yielding the relation,

$$c_{PP}(\theta,t) = a_\alpha(\theta)\frac{d}{dt}\ln\alpha(t) + a_\beta(\theta)\frac{d}{dt}\ln\beta(t) + a_\rho(\theta)\frac{d}{dt}\ln\rho(t) + e_c(\theta,t) \tag{2}$$

with $e_c(\theta,t)$ being model error; and

$$\begin{aligned} a_\alpha(\theta) &= \frac{1}{2}(1+\tan^2\theta)\,, \\ a_\beta(\theta) &= -4\frac{\beta_0^2}{\alpha_0^2}\sin^2\theta\,, \\ a_\rho(\theta) &= \frac{1}{2}\left(1-4\frac{\beta_0^2}{\alpha_0^2}\sin^2\theta\right)\,, \end{aligned} \tag{3}$$

with β_0 and α_0 being constants. The assumption made in Expression (2) is that the ratio of $\beta(t)/\alpha(t)$ can be approximated by β_0/α_0. Buland and Omre (2002a) demonstrate that the modeling error due to this assumption is small compared with the typical noise level in real seismic data. Note that the constant ratio only is assumed in order to approximate Expression (2) and will not carry through to the final estimates. Buland and Omre (2002a) formulate a slightly more general approximation by allowing the ratio to have a preselected time dependence. This type of time dependence could also be included in the current formulation.

The convolutional model and the weak contrast approximation, see Expression (1) and (2), has a limited range of validity. The crucial point being that the seismic preprocessing achieve its goal within a reasonable error. The largest contrast in the material parameters that is observed in the well at the Sleipner Øst Field occur about 2380 ms and have the magnitude 0.3. This is close to

the limit of both the convolutional model and the weak contrast approximation. Below this contrast the angle and time index in the angle gather may be disoriented since the rays bend at the boundary and this effect is not fully accounted for in the preprocessing. Also for large angles surface waves may form in the boundary layer and create additional noise.

The objective is now to reconstruct the material parameters, $\alpha(t)$, $\beta(t)$ and $\rho(t)$ based on the seismic traces in the corresponding angle gather. The material parameters below each location is inverted independently.

3 Statistical model

Bayesian inference requires a statistical model for both the prior and the likelihood. The material parameters are defined on a continuous domain and hence the prior should be modeled by random functions. In addition the prior model should be sufficiently flexible to capture essential characteristics of the material parameters. The likelihood includes the physical link between the observations and the material parameters, and the statistical model for the errors.

3.1 The prior model

The prior model is defined for the logarithm of the material parameters:

$$[\,\alpha_L(t),\,\beta_L(t),\,\rho_L(t)\,] = [\,\ln\alpha(t),\,\ln\beta(t),\,\ln\rho(t)\,]\,.$$

This parameterization guarantees the the material parameters to be positive and is convenient due to the linear relation to the likelihood, see Expression (2). The prior model is described by the locationwise relations;

$$\begin{aligned}
\alpha_L(t) &= \alpha_L^0 + b_\alpha^C \varepsilon_C(t) + b_\alpha^1\,\varepsilon_1(t) + b_\alpha^2\,\varepsilon_2(t) + b_\alpha^3\,\varepsilon_3(t)\,,\\
\beta_L(t) &= \beta_L^0 + b_\beta^C\,\varepsilon_C(t) + b_\beta^1\,\varepsilon_1(t) + b_\beta^2\,\varepsilon_2(t) + b_\beta^3\,\varepsilon_3(t)\,,\\
\rho_L(t) &= \rho_L^0 + b_\rho^C\,\varepsilon_C(t) + b_\rho^1\,\varepsilon_1(t) + b_\rho^2\,\varepsilon_2(t) + b_\rho^3\,\varepsilon_3(t)\,,
\end{aligned} \tag{4}$$

with $\varepsilon_C(t)$ being a centered Cauchy random process; $\varepsilon_1(t), \varepsilon_2(t)$ and $\varepsilon_3(t)$ being centered Gaussian random processes; the b's being scale parameters describing the locationwise dependencies between the material parameters; and α_L^0, β_L^0, and ρ_L^0 being Gaussian random variables centered at the median values for $\alpha_L(t)$, $\beta_L(t)$ and $\rho_L(t)$ respectively. All random components on the right hand side of Expression (4) are assumed to be independent. The Cauchy process $\varepsilon_C(t)$ primarily model the abrupt changes in material parameters, while the Gaussian processes primarily model the smooth variations.

The random processes involved are defined by smoothing of independently scattered random measures (ISRM),

$$\begin{aligned} \varepsilon_C(t) &= \int \phi_C(t-h)\, dC(h) \\ \varepsilon_j(t) &= \int \phi_j(t-h)\, dW_j(h)\,; \quad j \in \{1, 2, 3\} \end{aligned} \tag{5}$$

with $dC(h)$ being the Cauchy measure; $dW_j(h)$; $j \in \{1, 2, 3\}$ being independent Wiener measures; and ϕ_j; $j \in \{C, 1, 2, 3\}$ being kernel functions. A short introduction to ISRM is included in Appendix A, a more rigorous presentation is found in Rajput and Rosinski (1989). It is assumed that $\|\phi'_C\|_1 < \infty$ and $\|\phi'_j\|_2 < \infty$; $j \in \{1, 2, 3\}$, with $'$ denoting differentiation with respect to time. This imply that the derivative fields, $\varepsilon'_j(t)$; $j \in \{C, 1, 2, 3\}$, are stationary. It is further assumed that $\int \varepsilon_j(t)\, dt = 0$ for $j \in \{C, 1, 2, 3\}$, the integral being over the region under consideration, hence all the variability regarding the global level is represented by α^0_L, β^0_L, ρ^0_L.

The prior distribution as defined above, is stationary for the derivatives of the logarithm of the material parameters, $\alpha'_L(t)$, $\beta'_L(t)$ and $\rho'_L(t)$. By integrating these and center the integrated field, the logarithms of the material parameters will generally not have a stationary distribution, but have higher variability at both ends of the interval. To prevent this effect, information regarding the increase of level in the material parameters could be supplied. For the Sleipner Øst Field this is done by extracting low frequency information from a well and enforce this as an additional constraint to define the prior.

3.2 The likelihood

The profile of the material parameters, is reconstructed independently below each location, using the seismic traces in the corresponding angle gather. The physical link between the observations and the material parameters is defined by combining Expression (1) and (2). As a result the likelihood is defined by,

$$d_{obs}(\theta, t) = a_\alpha(\theta) s_\theta * \alpha'_L(t) + a_\beta(\theta) s_\theta * \beta'_L(t) + a_\rho(\theta) s_\theta * \rho'_L(t) + e(\theta, t)\,, \tag{6}$$

with t being the zero offset reflection time; θ being the angle reference in the angle gather; $d_{obs}(\theta, t)$ being the observed seismic signal; s_θ and the a's being as in Expression (1) and (3); and $e(\theta, t)$ being the error term. The error term represent both observation errors and model errors, and is modeled as a centered Gaussian random field on $[0,\, \pi/2\,] \times \boldsymbol{R}$, being independent of the material parameters. The errors are correlated both in time and angles, but to maintain simplicity it is assumed that the dependencies separates in the covariance function,

$$\mathrm{Cov}\{e(\theta_k, t_i), e(\theta_l, t_j)\} = \sigma_\theta(\theta_k)\, \sigma_\theta(\theta_l)\, \nu_\theta(\theta_k, \theta_l)\, \nu_t(t_i, t_j)\,, \tag{7}$$

with $\sigma_\theta(\theta)$ being angle dependent standard deviation; $\nu_\theta(\theta_k, \theta_l)$ being a correlation function describing the dependencies in angular direction; and $\nu_t(t_i, t_j)$ being a correlation function describing the dependencies in time direction. Note that the likelihood only involve the derivative processes, hence it is invariant to changes in α_L^0, β_L^0, and ρ_L^0.

The likelihood is linear with respect to the random measures $dC(h)$ and $dW_j(h)$; $j \in \{1, 2, 3\}$ in Expression (5). Focusing these random measures in Expression (6), one may write,

$$d_{obs}(\theta, t) = \int K_C(\theta, t-h)\, dC(h) + \sum_{j=1}^{3} \int K_j(\theta, t-h)\, dW_j(h) + e(\theta, t) \quad (8)$$

with $d_{obs}(\theta, t)$ being the observed seismic signal; $e(\theta, t)$ being the error; and

$$K_j(\theta, t) = \Big(b_\alpha^j\, a_\alpha(\theta) + b_\beta^j\, a_\beta(\theta) + b_\rho^j\, a_\rho(\theta) \Big)\, \frac{\partial}{\partial t}(s_\theta * \phi_j); \quad j \in \{C, 1, 2, 3\},$$

with s_θ being the seismic wavelet, see Expression (1); ϕ_j being kernel functions, see Expression (5); and the a's and the b's being as in Expression (3) and (4).

4 The posterior

The joint posterior distribution of the random measures, $dC(h)$ and $dW_j(h)$; $j \in \{1, 2, 3\}$, see Expression (5) and (8) are explored in this section. The logarithm of the material parameters can then be found by a linear transform of the measures according to Expression (4) and (5). The likelihood is linear, but the Cauchy measure disturbs the traditional Gaussian-linear machinery.

Formally, let C, W and D denote the Cauchy measure, the Wiener measures and the seismic observations respectively. The posterior distribution is explored by splitting it according to the identity:

$$p(w, c|d) = p(c|d)p(w|d, c). \quad (9)$$

The factor $p(c|d)$ is the distribution of a Cauchy measure under a linear constraint with Gaussian errors. The factor $p(w|d, c)$ is the distribution of the Wiener measure under a linear constraint with Gaussian errors. The first factor can be sampled by McMC algorithms, while the second factor can be evaluated analytically.

4.1 Discretization of the problem

In order to implement the Bayesian methodology on a computer, the problem is discretized. This is done by creating discrete equivalents of the relations

above. Having the prior and the likelihood defined on a continuous domain enables control of discretization and assure consistency if the discretization is refined. The random measures are discretized into independent random seeds, see Appendix A. The random processes involved, are then modeled by the discrete equivalent of Expression (5),

$$\begin{aligned} \varepsilon_C &= \boldsymbol{\Phi}_C \boldsymbol{C} \\ \varepsilon_j &= \boldsymbol{\Phi}_j \boldsymbol{W}_j \,; \quad j \in \{1, 2, 3\} \,, \end{aligned} \tag{10}$$

with $\boldsymbol{\Phi}_j$; $j \in \{C, 1, 2, 3\}$ being a matrix representing the convolving kernels, ϕ_j, in Expression (5); $\boldsymbol{C}$ being iid Cauchy seeds; and $\boldsymbol{W}_j$; $j \in \{1, 2, 3\}$ being three independent sets of iid Gaussian seeds. For later reference let τ and σ denote the scale of Cauchy seeds and Gaussian seeds respectively.

The discrete equivalent of Expression (8), may be written as

$$\boldsymbol{d} = \boldsymbol{K}_{\mathrm{C}} \boldsymbol{C} + \boldsymbol{K}\, \boldsymbol{W} + \boldsymbol{e} \tag{11}$$

with $\boldsymbol{d}$ being the discretized seismic signal in the angle gather; $\boldsymbol{C}$ and $\boldsymbol{W}^T = [\boldsymbol{W}_1^T, \boldsymbol{W}_2^T, \boldsymbol{W}_3^T]$ being the random seeds, see Expression (10); $\boldsymbol{K}_{\mathrm{C}}$ and $\boldsymbol{K}^T = [\boldsymbol{K}_1^T, \boldsymbol{K}_2^T, \boldsymbol{K}_3^T]$ being matrices representing the kernels in Expression (8); and $\boldsymbol{e}$ being the discretized error term. The error vector have a multi Gaussian distribution centered at zero and covariance $\boldsymbol{\Sigma}_E$. Exact definitions of the random variables $\boldsymbol{d}$ and $\boldsymbol{e}$ and the matrices $\boldsymbol{K}_{\mathrm{C}}$ and $\boldsymbol{K}$ are given in Appendix B.

The objective is now to sample the posterior distribution of the random seeds $\boldsymbol{C}$ and $\boldsymbol{W}$. Given a sample from this distribution, a sample of the material parameters are obtained by using Expression (4) and (10).

4.2 The posterior Cauchy seed

When the focus is on the Cauchy seed, Expression (11) can be restated as,

$$\boldsymbol{d} = \boldsymbol{K}_{\mathrm{C}} \boldsymbol{C} + \boldsymbol{e}_N$$

with $\boldsymbol{d}$, $\boldsymbol{K}_{\mathrm{C}}$, $\boldsymbol{C}$ being as in Expression (11); and $\boldsymbol{e}_N = \boldsymbol{K}\, \boldsymbol{W} + \boldsymbol{e}$, hence the error structure is altered,

$$p(\boldsymbol{e}_N) = N_{mn}\left(\boldsymbol{0}, \boldsymbol{\Sigma}_N\right),$$

with $\boldsymbol{\Sigma}_N = \sigma^2 \boldsymbol{K}\boldsymbol{K}^T + \boldsymbol{\Sigma}_E$ being the covariance matrix. The posterior can hence be written as:

$$\begin{aligned} p(\boldsymbol{c}|\boldsymbol{d}) = \quad & const \times \left[\prod_{i=1}^{n} \left(\tfrac{1}{\pi\tau(1+(c_i/\tau)^2)}\right)\right] \\ & \times \exp\left\{-\tfrac{1}{2}(\boldsymbol{K}_{\mathrm{C}}\boldsymbol{c} - \boldsymbol{d})^T \boldsymbol{\Sigma}_N^{-1} (\boldsymbol{K}_{\mathrm{C}}\boldsymbol{c} - \boldsymbol{d})\right\} \,. \end{aligned} \tag{12}$$

The posterior distribution of $\boldsymbol{C}$ is multi modal. By differentiation of Expression (12) it is easily found that any local mode $\boldsymbol{c}^{\text{LM}}$ satisfy the relation,

$$\boldsymbol{c}^{\text{LM}} = \left(\boldsymbol{K}_{\text{C}}^T \boldsymbol{\Sigma}_N^{-1} \boldsymbol{K}_{\text{C}} + \boldsymbol{D}(\boldsymbol{c}^{\text{LM}}) \right)^{-1} \boldsymbol{K}_{\text{C}}^T \boldsymbol{\Sigma}_N^{-1} \boldsymbol{d} \,, \tag{13}$$

with $\boldsymbol{D}(\boldsymbol{c}^{\text{LM}})$ being a diagonal matrix with $\boldsymbol{D}(\boldsymbol{c}^{\text{LM}})_{ii} = \frac{2}{\tau^2 + (c_i^{\text{LM}})^2}$. The expression for $\boldsymbol{c}^{\text{LM}}$ above should be compared with the corresponding for a Gaussian prior, for which $\boldsymbol{D}(\boldsymbol{c}^{\text{LM}})_{ii} = \frac{1}{\tau_G^2}$. The maximum posterior estimator correspond to linear shrinkage in the Gaussian model, whereas in the current model the maximum posterior correspond to nonlinear shrinkage.

The modes of the posterior distribution, $p(\boldsymbol{c}|\boldsymbol{d})$, are located close to subspaces constructed by setting most of the coefficients of $\boldsymbol{C}$ to zero. This is due to the sparse structure of the Cauchy seed, as is pointed out in Appendix A. This fact allows for efficient sampling based on the multi directional Gibbs sampler (Liu and Sabatti 2000). Details of the algorithm is in Appendix C and E.

4.3 The posterior Gaussian seed

When the value of the Cauchy seed is given, $\boldsymbol{C} = \boldsymbol{c}$, Expression (11) can be expressed as,

$$\boldsymbol{d} - \boldsymbol{K}_{\text{C}} \boldsymbol{c} = \boldsymbol{K}\boldsymbol{W} + \boldsymbol{e} \,.$$

The conditional posterior can be calculated explicitly ,

$$p(\boldsymbol{w}|\boldsymbol{d}, \boldsymbol{c}) = N_{3n} \left(\boldsymbol{\mu}_{\text{w}}, \boldsymbol{\Sigma}_{\text{w}} \right),$$

with

$$\boldsymbol{\mu}_{\text{w}} = \sigma^2 \boldsymbol{K}^T \left(\sigma^2 \boldsymbol{K}\boldsymbol{K}^T + \boldsymbol{\Sigma}_E \right)^{-1} (\boldsymbol{d} - \boldsymbol{K}_{\text{C}} \boldsymbol{c})$$

$$\boldsymbol{\Sigma}_{\text{w}} = \sigma^2 \boldsymbol{I} - \sigma^4 \boldsymbol{K}^T \left(\sigma^2 \boldsymbol{K}\boldsymbol{K}^T + \boldsymbol{\Sigma}_E \right)^{-1} \boldsymbol{K}$$

Note in particular that the posterior covariance does not depend on the value of the Cauchy seed, hence if the maximum posterior solution for the seeds is sought, the search may be done sequentially in Expression (9).

Since the full conditional distribution is Gaussian, there are several standard ways to sample the posterior distribution in this case. The approach used here is direct simulations, details of the algorithm is in Appendix E.

Because the Gaussian seeds are easy to sample when the Cauchy seed is given, several Gaussian seeds are generated for each Cauchy seed.

5 Comparison to a pure Gaussian model

The model defined in the current work is compared with a model having a purely Gaussian prior. The test case is based on material parameters observed in the well at the Sleipner Øst Field, see Figure 1, synthetic seismic observations are generated based on this profile. In Figure 1 the depth profile is converted to a time profile. The region under consideration is the time interval from $2000 - 2400$ ms. The drilling stopped after 2390 ms below this depth the value of the material parameters is fixed at a constant level.

The prior parameter values are estimated from the well observations. The estimation procedure is described in the next section and Appendix D. The level and scale parameters, see Expression (4), are listed in Table 1 and 2, the derivative of the kernels, ϕ_j'; $j \in \{C, 1, 2, 3\}$, see Expression (5), are displayed in Figure 6 and 7 for the Cauchy model and the pure Gaussian model respectively. A random effect could be added to the level parameters, but this serves no purpose in this example since the seismic observations are unaffected by a change in these parameters.

The likelihood is modeled by the linearized 1D-earth model, see Expression (1), (2) and (3), using $\beta_0/\alpha_0 = 0.5$. The data is collected for nine equispaced angles from 5° to 37°. The smoothing wavelet of Expression (1) is assumed to have the functional form

$$s_\theta(t) = a(\nu)\,(1.5 - 2\pi^2\nu^2t^2)\exp\{-\pi^2\nu^2t^2\}, \tag{14}$$

with $a(\nu)$ being an amplitude scale; and $\nu = 25$ Hz being the peak frequency. The wavelet is independent of angle and the amplitude is selected such that $\|s_\theta\|_2 = 1$. This is a Ricker like wavelet, see Sheriff and Geldart (1995), modified such that the low frequency components are larger. This is to avoid large fluctuations in the low frequent components in the solution, and is done instead of redefining the prior to contain lowfrequent information.

The error structure have a white and a colored component. The white noise component contribute 1% of the variance, and the colored component contribute the remaining 99%. The colored component has the form given in Expression (7). The random error in time direction is obtained by convolving the seismic wavelet, see Expression (14), with the Wiener measure, this defines $\nu_t(t_i, t_j)$. The error correlation in the direction of angles is given by,

$$\nu_\theta(\theta_k, \theta_l) = \exp\{-3|\theta_k - \theta_l|/\eta_\theta\} \tag{15}$$

with $\eta_\theta = 60°$ being the length scale for the correlation. The standard deviation, $\sigma_\theta(\theta)$, is assumed constant as a function of angle and is 0.002. A discussion of the correlation structure is included in the next section.

The synthetic observations are obtained by applying the linearized forward model to the parameters from the well logs, and add errors according to the

likelihood. The synthetic observations are displayed in Figure 8. The signal to noise ratio, measured as the ratio of the squared L^2 norms, is about 64. Under the loss criterion of absolute deviation the estimate is the median of the marginal posterior distribution for the relevant parameter.

The estimates in the Cauchy model are evaluated using 800 sampled Cauchy seeds, sampling five Gaussian seeds for each Cauchy seed. The results from the Gaussian model can be obtained analytically. The estimates and the true parameter values are displayed in Figure 9 and 10 for the Cauchy model and the pure Gaussian model respectively. Figure 11 and 12 show the error in the estimates together with a 90% credibility interval.

Evaluating the estimates by eye, both methods keep the true value reasonable well within the error margin and captures the main features in the velocities, $\alpha(t)$ and $\beta(t)$, whereas neither model resolve the density, $\rho(t)$, satisfactory. The biggest visual difference appears at the leap in the parameter values at 2380 ms. For the Cauchy model the leap is present in the estimate, whereas in the Gaussian model it is smoothed over a region, compare Figure 9 and 10. The error for the Cauchy model seems unaffected by the leap, but the smoothing in the pure Gaussian model results in large fluctuations in the error around 2380 ms, compare Figure 11 and 12. A similar effect is observed around the peak at 2310 ms.

When the estimates are evaluated by the criterion of absolute deviation the Cauchy model produces a better fit for the velocity estimates but there is no gain for the density. For the material parameters in Figure 1 the risk is estimated by averaging the loss for 100 errors simulated according to the likelihood. The overall risk improvement is 7% for the velocity estimates. No improvement is noted for the rock density. If only the interval from 2000 ms to 2250 ms is considered there is essentially no difference between the two models. The main advantage of the Cauchy model is observed in the region 2250 ms to 2400 ms, where the velocities have large fluctuations. In this region the estimates are improved by 14 % and 10 % for $\alpha(t)$ and $\beta(t)$ respectively.

The Cauchy model reduces the average error margin in the 90% credibility interval by 7% for the velocities $\alpha(t)$ and $\beta(t)$. However in the regions with large fluctuations, the error margins for the Cauchy model is larger than for the pure Gaussian model. Note in particular the region around 2310 ms where the credibility interval increases in the Cauchy model. Note also the characteristic peaks in the credibility intervals in the regions where the leaps are, i.e. around 2310 ms and 2380 ms. This indicate uncertainty in the jump location. In the top region between 2000 ms to 2250 ms the error margin in the 90% credibility interval is about 10% shorter in the Cauchy model compared with the pure Gaussian model.

6 Sleipner Øst Field

In this section the seismic inline of 176 angle gathers collected above the Sleipner Øst Field, see Figure 4, is inverted independently in each location. The inversion of the Sleipner data is based on the procedure defined above, using the observed part of the well log in Figure 1 to fit parameters in the prior model, and the colocated angle gather, see Figure 5, to estimate the parameters in the likelihood. Normal plots of the derivative of the logarithmic material parameters from the well log are displayed in Figure 13. The figure clearly illustrate the heavy tailed nature of the phenomenon, hence the need for a non-Gaussian model in this case study.

By assuming the processes defined as the derivative of the logarithmic material parameters to be ergodic, averages under the marginal distribution can be evaluated by time averages of these processes. The scale parameters in Expression (4) can hence be estimated by a modified method of moments, keeping the scale of the Cauchy field at fixed ratios. The final estimates are listed in Table 1. Details about the estimation procedure are left to Appendix D. The Cauchy process is primarily modeling the layered structure, hence the smoothing kernel for the derivative process is fixed as a Dirac at the level of grid resolution. The kernel is displayed in Figure 6 (a). The Gaussian processes, $\varepsilon_1'(t)$, $\varepsilon_2'(t)$ and $\varepsilon_3'(t)$, are then estimated by subtracting an estimate of the Cauchy field, and decouple the processes by the inverse of the estimated scale matrix for the Gaussian processes. The correlation function for each of the Gaussian processes are then fit by eye to an empirical estimate. A more advanced technique combining the methodology of Appendix D with tapering (Fodor and Stark 2000) could be developed, but is not believed to give any significant advantage in the current application. The estimated kernels, being the symmetric square root of the correlation operator, are displayed in Figure 6 (b)-(d). The kernels are those used in the example of the previous section.

In the likelihood the seismic wavelet and the noise covariance must be estimated. This is an inverse problem by itself. Buland and Omre (2002b) uses a Bayesian approach with vague priors to estimate the parameters. A slightly modified method is used here. Before the estimation is performed the well log and the seismic traces are aligned in time such that the highest peak of the seismic traces match the leap in parameter values at 2380 ms. The estimated wavelets are displayed in Figure 14.

The residuals after the wavelet estimation have two components of error, being due to observation errors in the well logs and in the seismic data. Only the sum of the errors is identifiable. Hence a subjective choice must be made regarding the seismic error structure. The variance of the error is modeled by a colored component of 99 % and a white component of 1 %. The colored component of the error is mainly due to effects related to the approximate model and imperfections of the seismic processing, hence the error is on the same scales as

the data. The time correlation of the error is hence obtained by convolving an average wavelet with white noise, this defines $\nu_t(t_i, t_j)$. The angle correlation of the error is chosen to be a first order exponential correlation function, see Expression (15), with length scale $\eta_\theta = 30°$. The standard deviation of the error $\sigma_\theta(\theta)$ is chosen to be constant and have the value 0.35. This correspond to a signal to noise ratio ranging from five to two since the energy in the wavelets vary with angle.

The seismic wavelet contain only intermediate frequencies. This gives large fluctuations in the sampled values due to uncertainty in the low frequencies. In Gaussian models it is common to extract the low frequency components from the well and center the samples and estimates around this mean value. This approach is not optimal for the Cauchy model since it interfere with the structure of the prior distribution. In the current approach the low frequency content of the well is included by extracting the information below 10 Hz from the well log. The prior distribution is now redefined. The new prior distribution is the distribution previously defined conditioned to observations of the low frequency components according to the well observations. Figure 15 show the pointwise median and 90% credibility interval for the material parameters in the well location when the low frequency information is included.

The estimates are based on 200 samples of the Cauchy seeds, sampling 10 Gaussian seeds for each Cauchy seed. The mixing of the sampling algorithm is briefly commented in Appendix C and appears to be satisfactory. A larger number of samples would have been desirable to reduce the Monte Carlo variation in the estimates. The parameters are reconstructed in the region between 2050 ms and 2450 ms to avoid boundary effects due to missing observations.

Figure 16 displays the estimates below the well location together with 90% credibility interval. Comparing this figure with Figure 15 it is seen that the seismic only carry a moderate amount of information regarding the parameters, however the estimates contain more details than the prior and the credibility intervals are generally shorter than in the prior. In Figure 17 the current estimates are compared with the well logs. The estimates have a reasonable good correspondence, except from the peak at about 2310 ms that bears no effect in the estimate of $\alpha(t)$. Contrary to the synthetic example, the estimate is smoothed in the region containing the leap. This is due to uncertainty in the jump location. Most individual samples have one jump at about the right position, but this location varies slightly between samples, this reflects the multi modality of the posterior distribution. When the posterior distribution is multi modal, such as in the current case, it is in general impossible to find one estimate that both have a characteristic shape and represent average properties of the posterior. The global mode of the posterior distribution is an alternative estimator. This estimator will reproduce a leap in the parameter value, and hence be visually attractive. It would however have a worse performance if measured by average quantities.

Figure 18 show the the 90% credibility interval for the error and the actual error for the well. The smoothing effect of the estimate around the leap produces a large error in this location, but even this large error is within the credibility interval due to the characteristic peak in the credibility interval around the leap. The peak in the credibility interval indicate the uncertainty in jump location.

The final estimates of α, β and ρ for all the 176 gathers are displayed in Figure 19(a)-(c). The leap value between 2350 ms and 2450 ms dominates the picture, but other details are also present. The Monte Carlo variation of the estimates causes some disturbance to the pictures in particular for the density estimate, see Figure 19(c). For the estimate of β, see Figure 19(b), the the estimate is threshold at 3500 m/s in order to represent the contrasts in the estimate better, the estimate exceed this value in some areas below the leap these areas are colored red. the maximum value of β is about 4250 m/s and occur in the read area at the leap boundary for gather 43. Below the leap especially in the lower left corner of the figure the estimates fluctuate. The basis of these fluctuations are present i the data, but may be due to imperfections of the preprocessing. There are several sources of errors. If time axis is shrunk to much in the preprocessing energy is migrated to higher frequencies, if the effect of geometrical spreading and absorption is over compensated the signal is amplified, also the velocity dependence of the references in the angle gather may be problematic. In general the estimates are reasonable.

7 Discussion with conclusions

The prior model is defined in a consistent way by the use of independent scattered random measures. A superposition of Cauchy and Gaussian processes models a layered structure with slow variations within each layer. Compared to the more common Bernoulli Gauss model, the Cauchy model introduces the layering without introducing dichotomy explicitly, and is defined independent of grid. The current model is an alternative to modeling the layers by point processes. The likelihood model is based on well founded geophysical principles.

The estimator is evaluated by stochastic simulation. The loss criterion of absolute deviation account for all the generated samples in a robust way. In an example the Cauchy model is compared to a pure Gaussian model. The test example is based on real material parameters observed in the Sleipner Øst Field and synthetic seismic data is generated. For the velocities the Cauchy model is found to reduce the over all risk by 7%. In regions where the material parameters have large fluctuations the estimates of the velocities improve by 14 % and 10 %. Over all the error margins in the 90% credibility interval is reduced by 7%. In regions where the material parameters are slowly varying the 90% credibility intervals for the velocities are reduced by 10%. The model only have a minor impact on estimation of the rock density. In general the uncertainty of the estimates are well represented in the Cauchy model in particular the

uncertainty in leap locations.

For the Sleipner Øst Field, the model define a prior distribution that account for the layered structure. To restore the material parameters, with reasonable error margins, low frequency information must be included from the well log. In the current approach this is done by redefining the prior distribution to be the conditional distribution when given low frequency information from the well log. The inversion results are satisfactory and the uncertainty is well represented. The uncertainty is however large due to large errors in the data.

Ideally a model that accounts for lateral dependencies should be developed. In such a model the well information could be included in a consistent way. The current model does not immediately generalize to include lateral dependencies. It is however possible to define a spatial model by superposition of stationary Cauchy fields in higher dimensions, such a model would however substantially increase the computational cost and raise additional questions regarding estimation of prior parameters and sampling algorithm.

The Cauchy model might be considered to have too heavy tails, hence unrealistically large leaps in the parameter values may occur according to the prior distribution. The random fields defined above are related to random fields of type $\mathcal{G}$ (Barndorff-Nielsen and Prez-Abreu 2002). Random fields of type $\mathcal{G}$ are very general and offer a broad specter of prior distributions that can be investigated. In particular multivariate normal inverse Gaussian distributions is a class of flexible distributions that bridges the gap between multivariate Cauchy distributions and Gaussian distributions. To utilize such random fields as priors is a topic for further research.

Acknowledgments

This work was supported by the Research Council of Norway. The author thanks Henning Omre, Arild Buland and Håkon Tjelmeland for helpful comments, and Statoil and the Sleipner licence (Statoil, Exxon/Mobil, Norsk Hydro and TotalFinaElf) for permission to publish this paper

References

Abramovich F. and Silverman B.W. (1998), "Wavelet decomposition approaches to statistical inverse problems," *Biometrika*, 85, 115-129.

Barndorff-Nielsen, O.E. and Prez-Abreu, V. (2002) "Multivariate Type $\mathcal{G}$ Distributions " To appear in Theory Prob. Its Appl

Bertero, M. (1989), "Linear inverse and ill-posed problems" Advances in electronics and electron physics, Vol. 75, pp 1-120.

Buland, A. and Omre, H. (2002a), "Bayesian linearized AVO inversion" To appear in Geophysics.

Buland, A. and Omre, H. (2002b), "Bayesian wavelet estimation from seismic and well data " To appear in Geophysics.

Donoho, D.L. (1995), "Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition," *Applied and computational harmonic analysis*, 2, 101-126.

Fodor, I.K., Stark, P.B. (2000) "Multitaper Spectrum Estimation for time series with gaps" IEEE Transactions on signal processing Vol. 48, 3472-3483.

Hansen, P.C. (1998) "Rank-deficient and discrete ill-posed problems," SIAM Monographs on Mathematical Modeling and Computation. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA

Liu, J.S. and Sabatti, C. (2000) "Generalized Gibbs Sampler and Multigrid Monte Carlo for Bayesian Computation", *Biometrika*, 87, 353-369.

Malinverno, A. and Leaney, S. (2000) A Monte Carlo method to quantify uncertainty in the inversion of Zero-Offset VPS data.

Mendel, J.M. (1983) "Optimal Seismic Deconvolution: An Estimation-Based Approach", Academic Press, New York.

Rajput, B.S. and Rosinski, J. (1989) "Spectral Representation of Infinitely Divisible Processes." Probability Theory and related fields. Vol 82, 451-487.

Sheriff, R.E. and Geldart,L.P. (1995) "Exploration Seismology" , 2 ed., Cambridge University Press.

Stolt, R. H. and Weglein, A.B., (1985), "Migration and inversion of seismic data"; Geophysics Vol. 50, pp 2458-2472.

Tarantola, A. (1987), "Inverse problem Theory"; Elsevier.

Tikhonov, A. (1963), "Solution to incorrectly formulated problems and the regularization method," *Soviet Math. Doklady*, 5, 1035-1038.

Vanmarcke, E. (1983), "Random fields"; The MIT press.

Wahba, G. (1990), "Spline models for Observational data", *NSF-CBMS Regional Conference in Mathematics*, (Vol. 59); Philadelphia; SIAM.

A Independently scattered random measures

Independently scattered random measures offers a way of constructing prior distributions for functions defined on a continuous domain in $\mathbf{R}^d$. The presentation below is motivated from the point of application, a more rigorous presentation is found in Rajput and Rosinski (1989).

Definition 1 (Independently scattered random measures,(ISRM)) *An independently scattered random measure, $Z \in \mathbf{R}^d$ is a set of random variables indexed by the Borel sets in $\mathbf{R}^d$, such that for any sequence of disjoint Borel sets $\{\mathcal{A}_i\}_{i=1}^{\infty}$ in $\mathbf{R}^d$ the two properties hold.*

$$\begin{aligned} &i) \quad Z(\mathcal{A}_i)\ ;\ i = 1, 2, \ldots\ ;\ \textit{are independent,} \\ &ii) \quad Z(\textstyle\bigcup_i \mathcal{A}_i) = \sum_i Z(\mathcal{A}_i) \qquad \textit{a.s.,} \end{aligned}$$

with a.s. denoting almost sure convergence of the series.

The Wiener measure and the Cauchy measure will be defined next. Let $\stackrel{\mathcal{D}}{=}$ denote equality in distribution, and $|\cdot|$ denote the Lebesgue measure on $\boldsymbol{R}^d$.

The Wiener measure, can be defined by $i)$ and $ii)$ in addition to

$$iii)_w \quad Z(\mathcal{A}_i) \stackrel{\mathcal{D}}{=} \text{Gauss}(0, |\mathcal{A}_i|),$$

with $\text{Gauss}(0, \sigma^2)$ being a Gaussian distributed random variable with zero mean and variance σ^2. The Wiener measure is a signed random measure on $\mathbf{R}^d$. In what follows W is used to refer to this measure.

The Cauchy measure, can be defined by $i)$ and $ii)$ in addition to

$$iii)_c \quad Z(\mathcal{A}_i) \stackrel{\mathcal{D}}{=} \text{Cauchy}(|\mathcal{A}_i|),$$

with $\text{Cauchy}(\tau)$ being a centered Cauchy distributed random variable with the scale factor τ. The density of a $\text{Cauchy}(\tau)$ random variable is:

$$p_C(x;\tau) = \frac{1}{\pi\tau\left(1 + \left(\frac{x}{\tau}\right)^2\right)} \quad , \qquad x \in \boldsymbol{R}\,, \tau > 0\,.$$

The Cauchy measure is a signed random measure on $\mathbf{R}^d$. In what follows C is used to refer to this measure.

A.1 Stationary random fields defined by ISRM

The stationary fields obtained by convolving a kernel ϕ with the ISRM,

$$\varepsilon(t) = \int_{\mathbf{R}^d} \phi(t-h)\, dZ(h), \tag{16}$$

is of particular interest in the current application.

The random fields generated by the Wiener measure and the Cauchy measure will be used for constructing the prior distribution. Let $\|\phi\|_1$ and $\|\phi\|_2$ denote the L^1 and L^2-norms respectively and assume $\|\phi\|_1 + \|\phi\|_2 < \infty$.

A stationary Gaussian random field is defined by

$$\varepsilon_G(t) = \int_{\mathbf{R}^d} \phi(t-h)\, dW(h).$$

According to the standard theory $\varepsilon_G(t)$ have a covariance function uniquely defined by ϕ (Vanmarcke 1983).

A stationary Cauchy random field is defined by

$$\varepsilon_C(t) = \int_{\mathbf{R}^d} \phi(t-h)\, dC(h).$$

The term Cauchy field is twofold deserved. Firstly, it is constructed based on the Cauchy measure, secondly all marginal distributions are Cauchy distributed,

$$\varepsilon_C(t) \stackrel{\mathcal{D}}{=} \text{Cauchy}(\|\phi\|_1).$$

Linear transforms of stationary fields defined by Expression (16), are given by transforming the kernel ϕ correspondingly. Let K denote the linear transform and apply this transform to the random field, then

$$K\varepsilon(t) = \int_{\mathbf{R}^d} K\phi(t-h)\, dZ(h),$$

with $K\varepsilon$ being the transformed field; and $K\phi$ being the linear transform applied to the kernel ϕ. Appropriate regularity conditions must apply to ϕ and K in order to make $K\varepsilon$ well defined. In the current application the linear transforms of interest are integration, convolution, and differentiation.

A.2 Discretization of ISRM

In the current article independent scattered random measures are used to define prior distributions for functions on a continuous domain. Having a continuously defined prior, enables control of the discretization error, and guarantees stability of the discretization as the resolution increases. An independently scattered random measure is discretized into independent random seeds by integrating over small volumes. For $t \in \mathbf{R}^d$ use the multi index notation to denote $t_\alpha = \alpha \cdot \Delta t = (\alpha_1 \Delta t_1, \alpha_2 \Delta t_2, ..., \alpha_d \Delta t_d)$ for $\alpha \in \boldsymbol{Z}^d$, and let $|\Delta t|$ denote the Lebesgue measure of the volume element Δt.

Discretization of the Wiener measure results in,

$$W_\alpha = \int_{t_\alpha}^{t_\alpha + \Delta t} dW(h),$$

with W_α being an iid sequence of random variables such that

$$W_\alpha \stackrel{\mathcal{D}}{=} \text{Gauss}(0, |\Delta t|).$$

For the Wiener measure the standard normal seed is scaled by $\sqrt{|\Delta t|}$.

Discretization of the Cauchy measure results in,

$$C_\alpha = \int_{t_\alpha}^{t_\alpha + \Delta t} dC(h),$$

with C_α being an iid sequence of random variables such that

$$C_\alpha \stackrel{\mathcal{D}}{=} \text{Cauchy}(|\Delta t|).$$

For the Cauchy measure the standard Cauchy seed is scaled by $|\Delta t|$.

In Figure 20 the Wiener measure and the Cauchy measure are visualized by discretized realizations. The figure clearly reveal the underlying structure of the random measures. The Wiener measure distributes the energy equally along the region, while the Cauchy measure concentrates most of the energy in a few locations. The focusing of energy is one of the properties that motivated the use of Cauchy fields in the current application.

B Details regarding discretization of the problem

The observations are collected for the angles θ_k; $k = 1, ..., m$; for each angle the functions are discretized into vectors of length n. For a given angle θ_k Expression (6) then translates into

$$\boldsymbol{d}_k = a_\alpha(\theta_k)\boldsymbol{S}_k\boldsymbol{D}\boldsymbol{\alpha}_L + a_\beta(\theta_k)\boldsymbol{S}_k\boldsymbol{D}\boldsymbol{\beta}_L + a_\rho(\theta_k)\boldsymbol{S}_k\boldsymbol{D}\boldsymbol{\rho}_L + \boldsymbol{e}_k, \qquad (17)$$

with $\boldsymbol{d}_k$ being the discretized seismic traces; the a's being as for Expression (3); $\boldsymbol{S}_k$ being a matrix representing convolution with the wavelet s_{θ_k}; $\boldsymbol{D}$ being a matrix representing differentiation; $\boldsymbol{\alpha}_L$, $\boldsymbol{\beta}_L$ and $\boldsymbol{\rho}_L$ being discretization of the material parameters; and $\boldsymbol{e}_k$ being the discrete error at angle θ_k. The full error vector $\boldsymbol{e}^T = [\boldsymbol{e}_1^T, ..., \boldsymbol{e}_m^T]$ have a multivariate Gaussian distribution,

$$p(\boldsymbol{e}) = N_{nm}\left(\boldsymbol{0}, \boldsymbol{\Sigma}_E\right)$$

with $\boldsymbol{\Sigma}_E$ being the covariance matrix. Due to the separability, see Expression (7), the covariance matrix have the form $\boldsymbol{\Sigma}_E = \boldsymbol{\Sigma}_\theta \otimes \boldsymbol{\Sigma}_t$, with $\boldsymbol{\Sigma}_\theta$ and $\boldsymbol{\Sigma}_t$ being $m \times m$ and $n \times n$ matrices describing the error covariance and correlation in direction of angles and time respectively; and $\otimes$ being the Kronecker tensor product.

Focusing on the random seeds $\boldsymbol{C}$ and $\boldsymbol{W}_j$; $j \in \{1, 2, 3\}$ in Expression (10), the likelihood is linear,

$$\boldsymbol{d} = \boldsymbol{K}_{\mathrm{C}}\boldsymbol{C} + \boldsymbol{K}\,\boldsymbol{W} + \boldsymbol{e} \tag{18}$$

with $\boldsymbol{d}^T = [\boldsymbol{d}_1^T, ..., \boldsymbol{d}_m^T]$ being the discretized seismic traces; $\boldsymbol{C}$ and $\boldsymbol{W}^T = [\boldsymbol{W}_1^T, \boldsymbol{W}_2^T, \boldsymbol{W}_3^T]$ being the random seeds; $\boldsymbol{K}_{\mathrm{C}}$ being a $nm \times n$ matrix; $\boldsymbol{K}$ being a $mn \times 3n$ matrix; and $\boldsymbol{e}$ being the error term as defined above. $\boldsymbol{K}^T = [\boldsymbol{K}_1^T, \boldsymbol{K}_2^T, \boldsymbol{K}_3^T]$. The matrices $\boldsymbol{K}_j$; $j \in \{C, 1, 2, 3\}$ all have the same form, the relation is

$$\boldsymbol{K}_j = \begin{bmatrix} a_1^j \boldsymbol{S}_1 \boldsymbol{D} \boldsymbol{\Phi}_j \\ \vdots \\ a_m^j \boldsymbol{S}_m \boldsymbol{D} \boldsymbol{\Phi}_j \end{bmatrix},$$

with $\boldsymbol{S}_k$ being the matrix representing convolution with the wavelet at at angle θ_k;$\boldsymbol{D}$ being a matrix representing differentiation; $\boldsymbol{\Phi}_j$ being matrices representing the kernel functions, see Expression (10); and

$$a_k^j = a_\alpha(\theta_k) b_\alpha^j + a_\beta(\theta_k) b_\beta^j + a_\rho(\theta_k) b_\rho^j\,, \quad j \in \{C, 1, 2, 3\}\,,$$

with the a's on the right hand side being as in Expression (3); and the b's being as in Expression (4).

C The multi directional Gibbs sampler

The multi directional Gibbs sampler is a particular case of the generalized Gibbs sampler (Liu and Sabatti 2000). The generalized Gibbs sampler is a Markov chain based method for sampling a distribution. A Markov chain having the target distribution as stationary distribution is constructed by defining the updating rules.

Let $\boldsymbol{c}^{\mathrm{old}}$ be the current state of the chain; $\{\boldsymbol{v}^l\}_{l=1}^L$ be an over complete pool of basis vectors; and s be a scalar. For the purpose of describing the algorithm, let $p(\cdot)$ be the target density. The multi directional Gibbs sampler is described by the updating rule:

1. Draw l^* uniformly from $\{1, ..., L\}$
2. Draw s^* from $q(s) \propto p(\boldsymbol{c}^{\mathrm{old}} + s \cdot \boldsymbol{v}^{l^*})$
3. Let $\boldsymbol{c}^{\mathrm{new}} = \boldsymbol{c}^{\mathrm{old}} + s^* \cdot \boldsymbol{v}^{l^*}$

In the current application of the multi directional Gibbs sampler, the unit vectors are chosen as translations of the vectors in Figure 21. The density values

of $q(s)$ is known to proportionality by Expression (12),

$$q(s) \propto \left[\prod_{\{i\,:\,v_i^{l^*} \neq 0\}} \frac{1}{\pi\tau \left(1 + \left(\frac{c_i^{\text{old}} + s\, v_i^{l^*}}{\tau}\right)^2\right)} \right] \exp\{-\tfrac{1}{2\sigma_{l^*}^2}[s - \boldsymbol{\mu}_{l^*}^T(\boldsymbol{d} - \boldsymbol{K}_{\text{C}}\boldsymbol{c}^{\text{old}})]^2\},$$

with

$$\sigma_{l^*}^2 = \left[(\boldsymbol{v}^{l^*})^T \boldsymbol{K}_{\text{C}}^T \boldsymbol{\Sigma}_N^{-1} \boldsymbol{K}_{\text{C}} \boldsymbol{v}^{l^*}\right]^{-1},$$

and

$$\boldsymbol{\mu}_{l^*} = \sigma_{l^*}^2 \boldsymbol{\Sigma}_N^{-1} \boldsymbol{K}_{\text{C}} \boldsymbol{v}^{l^*}.$$

The likelihood give essential bounds for the posterior range. The univariate density $q(s)$ is calculated on a dense grid centered at $\boldsymbol{\mu}_{l^*}^T(\boldsymbol{d} - \boldsymbol{K}_{\text{C}}\boldsymbol{c}^{\text{old}})$ and stretching $5\sigma_{l^*}$ to each side. Note that σ_l and $\boldsymbol{\mu}_l$ are specific to each unit vector $\boldsymbol{v}^l$, and that their expressions only depend on the angles for which the traces are observed. Since the same set of angles are used for all gathers, σ_{l^*} and $\boldsymbol{\mu}_{l^*}$ only need to be calculated once for the total study.

In the current work the simulation is initiated in a local mode. The local mode is found by iterating Expression (13). After a burn in of 20 random scans through the pool of basis vectors, a sample is extracted at the end of every 2end random scan through the pool of basis vectors. The mixing for the Cauchy seed in the Sleipner Øst Field is displayed in Figure 22. The Cauchy seed is visualized in the figure. The middle plot show the sampled seed value that correspond to the leap value at 2380 ms. The seeds form the two nearest neighbors at both sides are in the plots above and below, respectively. In the figure three samples are extracted in each random scan through the pool of basis vectors. The figure clearly show how the sampling algorithm move between different modes corresponding to different locations for the leap in the parameter values.

D Estimation of prior scale parameters

In this appendix a methodology for estimating the scale parameters of Expression (4) is supplied. The scaling of the Gaussian processes is the square root of the pointwise covariance matrix for the logarithm of the material parameters given the Cauchy process. This part requires six parameters to be estimated. Let $\boldsymbol{B}_G$ denote a symmetric matrix with the values of the scaling of the Gaussian field. For the Cauchy process three parameters must be estimated.

If all nine parameters are to be determined, the estimates are unstable. This is seen by the fact that if all random processes are Gaussian, only six parameters can be identified. To further reduce the number of parameters, the ratios of the parameters for the Cauchy process are held fixed; $[\,b_\alpha^C,\, b_\beta^C,\, b_\rho^C\,] = \gamma \boldsymbol{v}^T$ with

$\boldsymbol{v}^T = [1.0,\, 1.0,\, 0.15]$, only the global scale γ is estimated. According to geophysical literature it is reasonable to assume that the first two components are of approximately of the same size, the third component is selected to be about size of $(\rho_L(2500) - \rho_L(2000))/(\alpha_L(2500) - \alpha_L(2000))$. The scale parameters are estimated by a modified method of moments, by computing the sample averages and tune the scale parameters so that the population averages match the sample averages for a given set of functions.

For a stationary random process, $\{\varepsilon(t),\ t \in \boldsymbol{R}\}$, define $\varepsilon_i = \int_{t_i}^{t_i+\Delta t} \varepsilon(s)ds$. The random variables ε_i are then identically distributed. Let ε denote a generic random variable being distributed according to the law of ε_i . If $\{\varepsilon(t),\ t \in \boldsymbol{R}\}$, is ergodic, the sample averages of ε_i approaches the marginal distribution of ε as the number of observations increases and the step length, Δt, is kept fixed. Using this notation for the random fields involved in the current problem, the functions used in the method of moments are,

$$\boldsymbol{I}(|\gamma\, \varepsilon'_C + \varepsilon'_G| > K)$$

and

$$\begin{bmatrix} \alpha'_L \\ \beta'_L \\ \rho'_L \end{bmatrix} [\, \alpha'_L,\, \beta'_L,\, \rho'_L\,] \boldsymbol{I}(|\gamma\varepsilon'_C + \varepsilon'_G| < K),$$

with γ being the global scale for the Cauchy process; K being a fixed constant; $\boldsymbol{I}(\cdot)$ being an indicator function for an event; ε'_C being as for Expression (4); and ε'_G being a linear combination of the Gaussian processes, ε'_1, ε'_2 and ε'_3, see Expression (4). In the current setting, $K = 0.0495$ and

$$\gamma\, \varepsilon'_C + \varepsilon'_G = 0.60\, \alpha'_L + 0.32\, \beta'_L + 0.54\, \rho'_L$$

The estimates must be solved numerically. Since the main focus is not on these parameters, approximate values for the population averages are computed, by using the approximations

$$P(|\gamma\, \varepsilon'_C + \varepsilon'_G| > K) \approx P(|\varepsilon'_C| > K/\gamma)$$

and

$$P(|\varepsilon'_C| > K/\gamma) \approx \frac{2\gamma}{\pi K}.$$

The approximations improves as K increases.

E Simplifications by using the Fourier transform

The expressions stated in the previous sections are valid for more general models than the particular that is specified. In the specified model, several expressions

simplify since the Fourier transform simultaneously diagonalize the stationary operators considered. To use the simplified formulas the convolutions must be defined cyclic. This is done by tapering the data and extend the vectors by zero padding to avoid boundary problems.

Let $\boldsymbol{X}^T$ denote the orthogonal Fourier transform, and $\boldsymbol{X}$ denote its inverse. Now introduce the relations,

$$\begin{aligned}
\boldsymbol{S}_k &= \boldsymbol{X}\boldsymbol{\Lambda}_k^s\boldsymbol{X}^T\,, \quad k \in \{1,2,...,m\}\,,\\
\boldsymbol{\Phi}_j &= \boldsymbol{X}\boldsymbol{\Lambda}^j\boldsymbol{X}^T\,, \quad j \in \{C,\,1,\,2,\,3\,\}\,,\\
\boldsymbol{\Sigma}_E &= \boldsymbol{\Sigma}^\theta \otimes \boldsymbol{\Sigma}^t\,,\\
\boldsymbol{\Sigma}^t &= \boldsymbol{X}\boldsymbol{\Lambda}^t\boldsymbol{X}^T\,,\\
\boldsymbol{X}^{\otimes m} &= \boldsymbol{I}_{m\times m} \otimes \boldsymbol{X}\,,\\
\boldsymbol{X}^{T\otimes m} &= \boldsymbol{I}_{m\times m} \otimes \boldsymbol{X}^T\,,
\end{aligned}$$

with the $\boldsymbol{\Lambda}$'s being diagonal matrices; $\otimes$ being the Kronecker tensor product; and

$$\boldsymbol{\Sigma}^\theta = \begin{bmatrix} \sigma^\theta_{11} & \dots & \sigma^\theta_{1m} \\ \vdots & \ddots & \vdots \\ \sigma^\theta_{m1} & \dots & \sigma^\theta_{mm} \end{bmatrix}\,.$$

According to these identities;

$$\boldsymbol{\Sigma}_N = \sigma^2\boldsymbol{K}\boldsymbol{K}^T + \boldsymbol{\Sigma}_E = \boldsymbol{X}^{\otimes m}\boldsymbol{D}_N\boldsymbol{X}^{T\otimes m}$$

with

$$\boldsymbol{D}_N = \begin{bmatrix} \boldsymbol{D}_{11} & \dots & \boldsymbol{D}_{1m} \\ \vdots & \ddots & \vdots \\ \boldsymbol{D}_{m1} & \dots & \boldsymbol{D}_{mm} \end{bmatrix}\,,$$

with $\boldsymbol{D}_{kl}$ being diagonal matrices having the form:

$$\boldsymbol{D}_{kl} = \sigma^\theta_{kl}\boldsymbol{\Lambda}^t + \sigma^2\sum_{j=1}^{3} a^j_k a^j_l \boldsymbol{\Lambda}^s_k\boldsymbol{\Lambda}^s_l(\boldsymbol{\Lambda}^j)^2.$$

The inverse of $\boldsymbol{\Sigma}_N$ can be calculated by solving a $m\times m$ system for each Fourier component separately. When $\boldsymbol{\Sigma}_E$ have a small white noise component the inversion is stable. The inverse of $\boldsymbol{D}_N$ have the same structure as $\boldsymbol{D}_N$,

$${\boldsymbol{D}_N}^{-1} = \begin{bmatrix} \boldsymbol{D}^{11} & \dots & \boldsymbol{D}^{1m} \\ \vdots & \ddots & \vdots \\ \boldsymbol{D}^{m1} & \dots & \boldsymbol{D}^{mm} \end{bmatrix}\,.$$

with $\boldsymbol{D}^{kl}$ being diagonal matrices.

For the sampling the Cauchy seed, σ_{l^*} and $\boldsymbol{\mu}_{l^*}$, need to be computed. First simplify the matrix $\boldsymbol{K}_{\mathrm{C}}^T \boldsymbol{\Sigma}_N^{-1} \boldsymbol{K}_{\mathrm{C}}$ using the notation above,

$$\boldsymbol{K}_{\mathrm{C}}^T \boldsymbol{\Sigma}_N^{-1} \boldsymbol{K}_{\mathrm{C}} = \boldsymbol{X}^T \boldsymbol{\Lambda}_H^2 \boldsymbol{X}^T$$

with

$$\boldsymbol{\Lambda}_H^2 = \sum_{k=1}^{m} \sum_{l=1}^{m} a_k^C a_l^C \boldsymbol{\Lambda}_k^s \boldsymbol{\Lambda}_l^s \boldsymbol{D}^{kl} (\boldsymbol{\Lambda}^C)^2,$$

hence:

$$\sigma_{l^*} = \| \boldsymbol{\Lambda}_H \boldsymbol{X}^T \boldsymbol{v}^{l^*} \|_2^{-1},$$

$$\boldsymbol{\mu}_{l^*}^T \boldsymbol{K}_{\mathrm{C}} \boldsymbol{c}^{\mathrm{old}} = \sigma_{l^*}^2 (\boldsymbol{X}^T \boldsymbol{v}^{l^*})^T \boldsymbol{\Lambda}_H^2 \boldsymbol{X}^T \boldsymbol{c}^{\mathrm{old}},$$

$$\boldsymbol{\mu}_{l^*}^T \boldsymbol{d} = \sigma_{l^*}^2 (\boldsymbol{X}^T \boldsymbol{v}^{l^*})^T \left(\sum_{k=1}^{m} \sum_{l=1}^{m} \frac{a_k^C \boldsymbol{\Lambda}_k^s \boldsymbol{D}^{kl} \boldsymbol{\Lambda}^C}{\boldsymbol{\Lambda}_H} \boldsymbol{X}^T \boldsymbol{d}_l \right),$$

with $\boldsymbol{d}_l$ being the discretized traces at angle θ_l.

For the normal seed computations, the important quantities are $\boldsymbol{\mu}_{\mathrm{w}}$ and $\boldsymbol{\Sigma}_{\mathrm{w}}$. These are split into the components corresponding to each of the normal seeds, $\boldsymbol{W}_j;\ j \in \{1, 2, 3\}$,

$$\boldsymbol{\mu}_{\mathrm{w}} = \begin{bmatrix} \boldsymbol{\mu}_1^{\mathrm{w}} \\ \boldsymbol{\mu}_2^{\mathrm{w}} \\ \boldsymbol{\mu}_3^{\mathrm{w}} \end{bmatrix}$$

and,

$$\boldsymbol{\Sigma}_{\mathrm{w}} = \begin{bmatrix} \boldsymbol{\Sigma}_{11}^{\mathrm{w}} & \boldsymbol{\Sigma}_{12}^{\mathrm{w}} & \boldsymbol{\Sigma}_{13}^{\mathrm{w}} \\ \boldsymbol{\Sigma}_{21}^{\mathrm{w}} & \boldsymbol{\Sigma}_{22}^{\mathrm{w}} & \boldsymbol{\Sigma}_{23}^{\mathrm{w}} \\ \boldsymbol{\Sigma}_{31}^{\mathrm{w}} & \boldsymbol{\Sigma}_{32}^{\mathrm{w}} & \boldsymbol{\Sigma}_{33}^{\mathrm{w}} \end{bmatrix}.$$

By the above notation,

$$\boldsymbol{\mu}_j^{\mathrm{w}} = \sigma^2 \boldsymbol{X} \sum_{k=1}^{m} \sum_{l=1}^{m} a_k^j \boldsymbol{D}^{kl} \boldsymbol{\Lambda}^j \boldsymbol{\Lambda}_k^s \boldsymbol{X}^T (\boldsymbol{d}_l - (\boldsymbol{K}_{\mathrm{C}} \boldsymbol{c}^{\mathrm{old}})_l),$$

with $(\boldsymbol{K}_{\mathrm{C}} \boldsymbol{c}^{\mathrm{old}})_l$ being the vector containing the, $(l-1)n+1$, $(l-1)n+2, ..., ln$, components of $\boldsymbol{K}_{\mathrm{C}} \boldsymbol{c}^{\mathrm{old}}$, further

$$\boldsymbol{\Sigma}_{ij}^{\mathrm{w}} = \sigma^2 \delta_{ij} \cdot \boldsymbol{I} - \sigma^4 \boldsymbol{X} \left(\sum_{k=1}^{m} \sum_{l=1}^{m} a_k^j a_l^i \boldsymbol{D}^{kl} \boldsymbol{\Lambda}^j \boldsymbol{\Lambda}^i \boldsymbol{\Lambda}_k^s \boldsymbol{\Lambda}_l^s \right) \boldsymbol{X}^T,$$

with $\delta_{ij} = 1$ if $i = j$, and $\delta_{ij} = 0$ if $i \neq j$,

The random seeds can hence be sampled by sampling the frequencies of $\boldsymbol{W}_1, \boldsymbol{W}_2, \boldsymbol{W}_3$ simultaneously. For each of the n frequencies, a 3×3 matrix must be factored.

Tables and figures

material parameter	level	scale ε_C	scale ε_1	scale ε_1	scale ε_3
α	7.9941	0.0059	0.0183	−0.0014	0.0004
β	7.2651	0.0059	−0.0014	0.0303	−0.0014
ρ	7.7629	0.0009	0.0004	−0.0014	0.0106

Table 1: Level and scale parameters for the Cauchy model. The table relates to Expression (4).

material parameter	level	scale ε_C	scale ε_1	scale ε_1	scale ε_3
α	7.9941	0.0000	0.0319	0.0102	0.0032
β	7.2651	0.0000	0.0102	0.0449	0.0007
ρ	7.7629	0.0000	0.0032	0.0007	0.0120

Table 2: Level and scale parameters for the pure Gaussian model. The table relates to Expression (4).

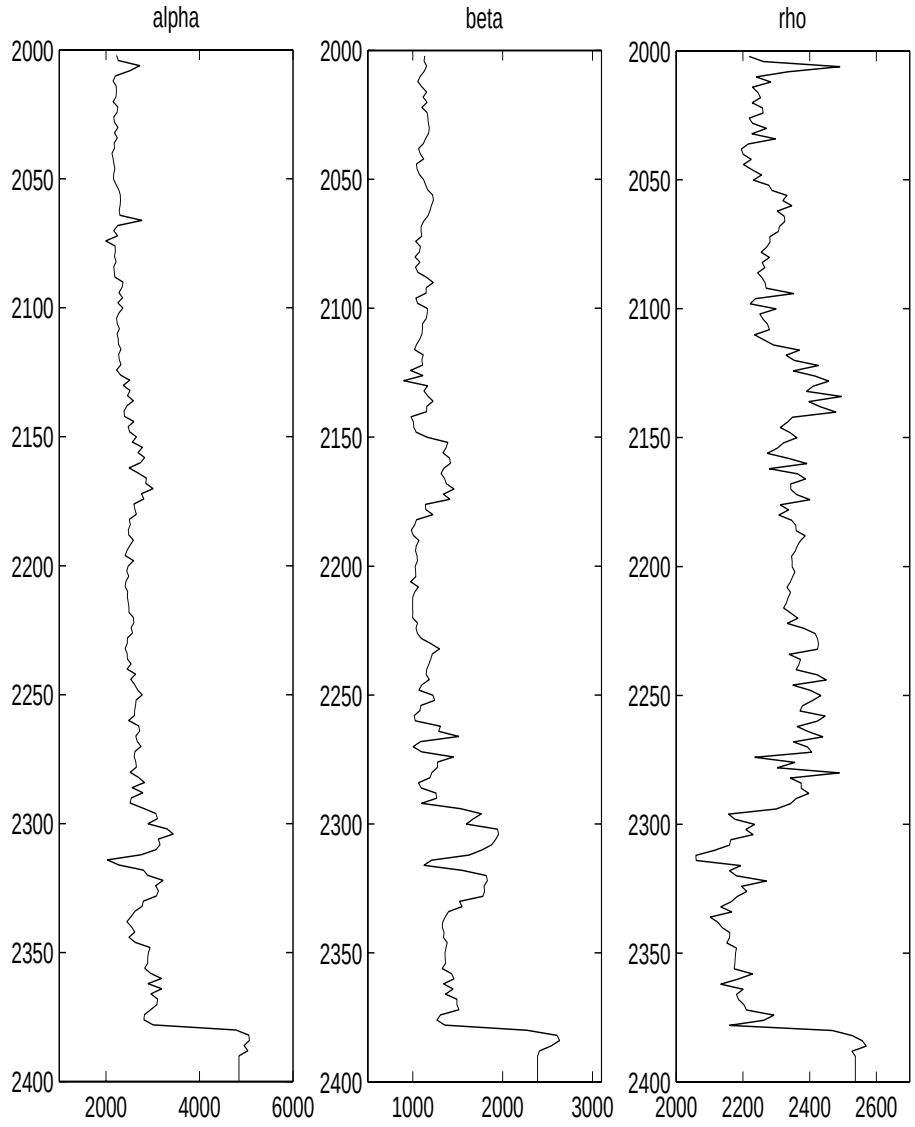


Figure 1: Well logs. The material parameters observed in a well at the Sleipner Øst Field. The observed depth profile is converted to a time profile. The drilling stopped at 2390 ms below this depth the value is fixed at a constant level.

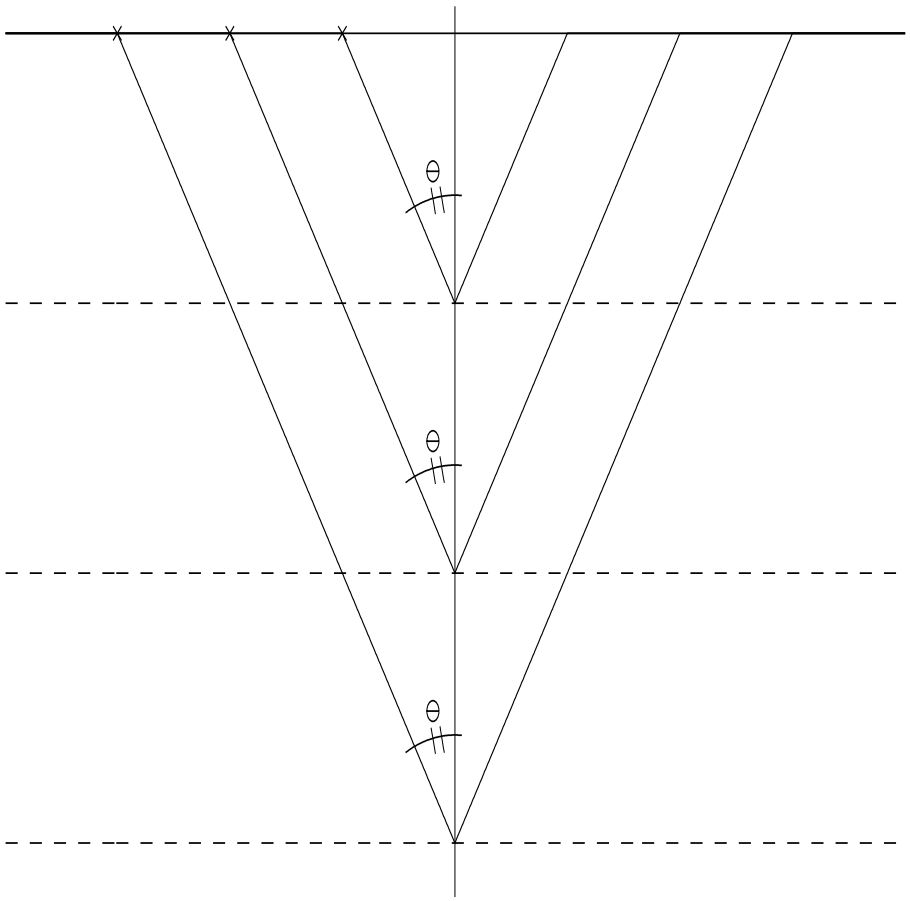


Figure 2: Angle gather-common angle. The ray paths all have a common angle to the vertical line, and hence a common angle in the angle gather

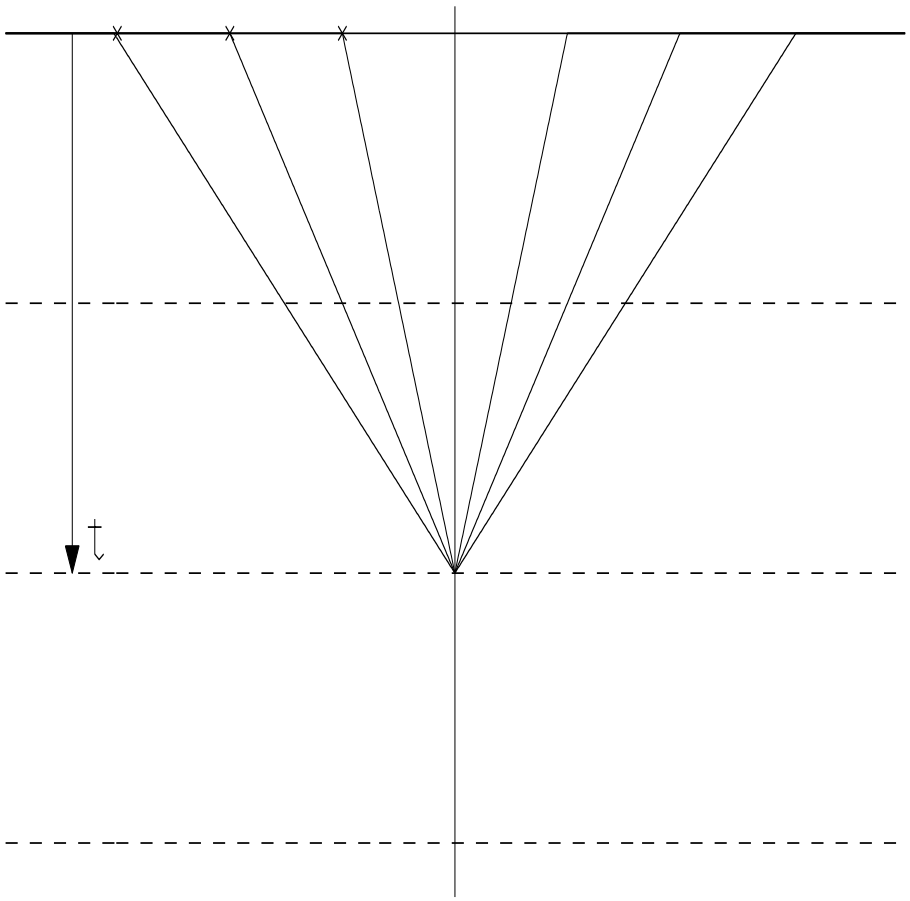


Figure 3: Angle gather-common time. The ray paths all have a common depth of reflection, and hence the same time reference in the angle gather.

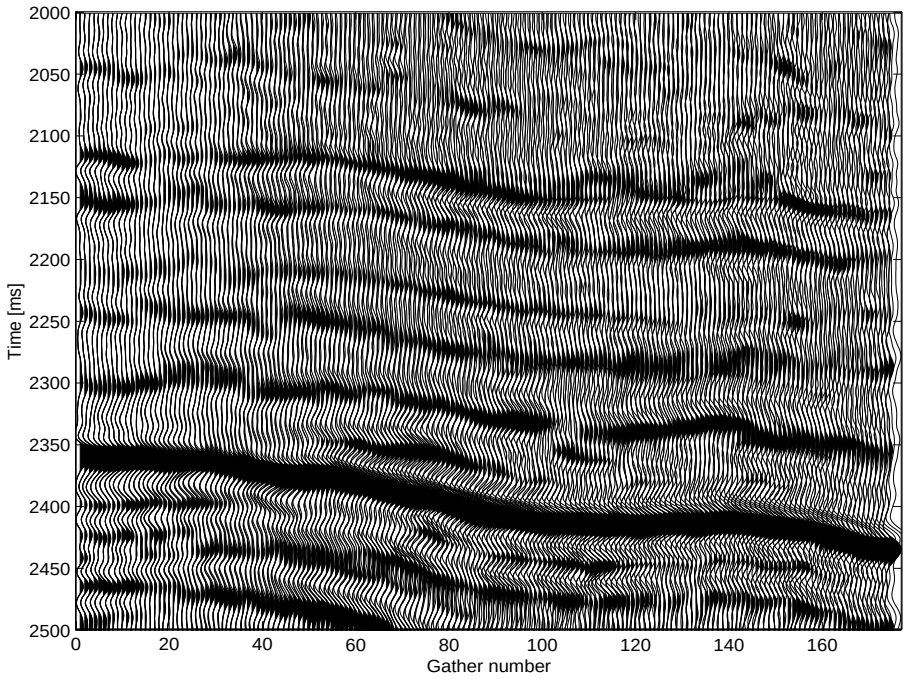


Figure 4: Stack section for the seismic inline. The seismic inline is observed in a marine seismic survey above the Sleipner Øst Field.

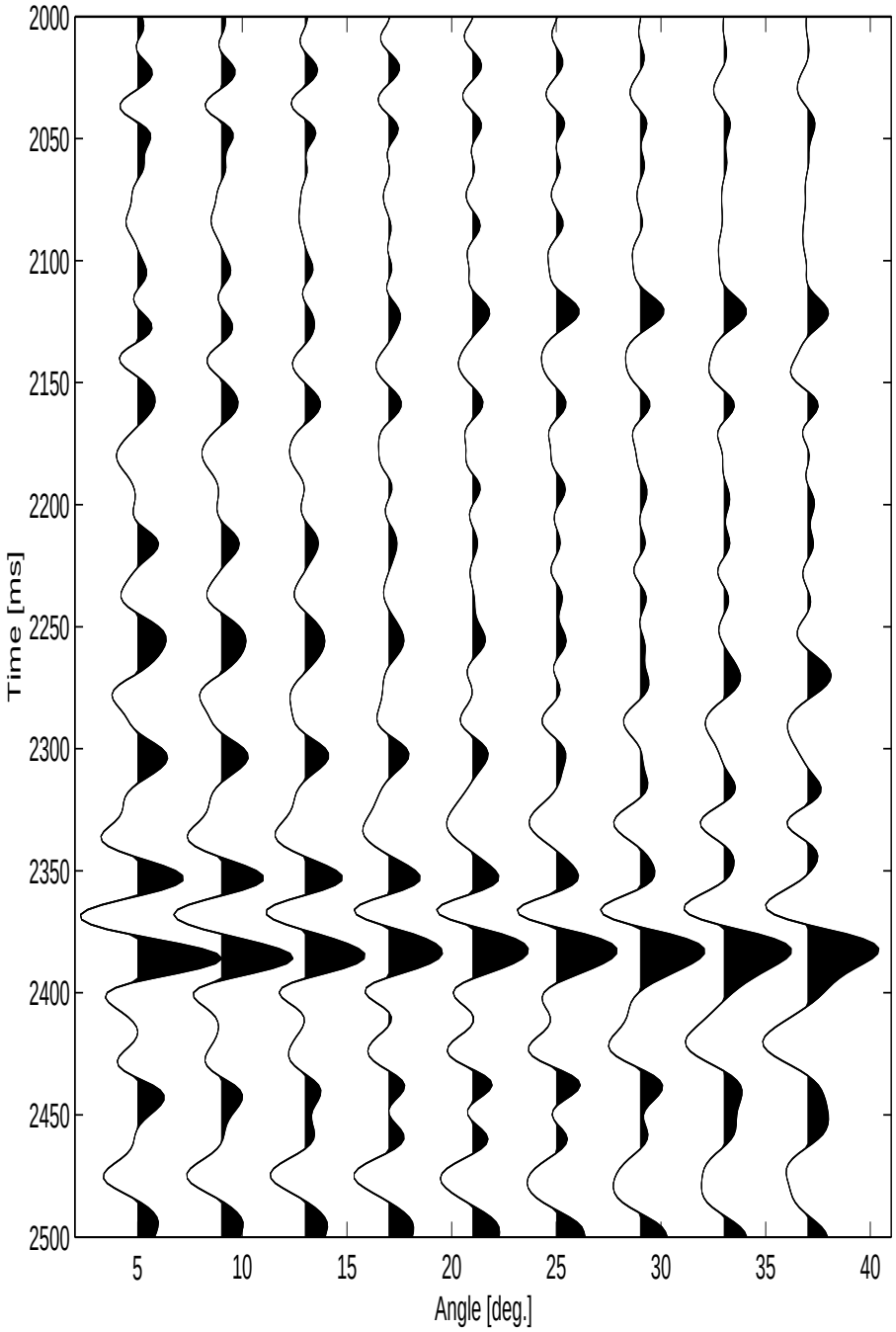


Figure 5: Angle gather number 67. The seismic traces in this angle gather is recorded in the well location.

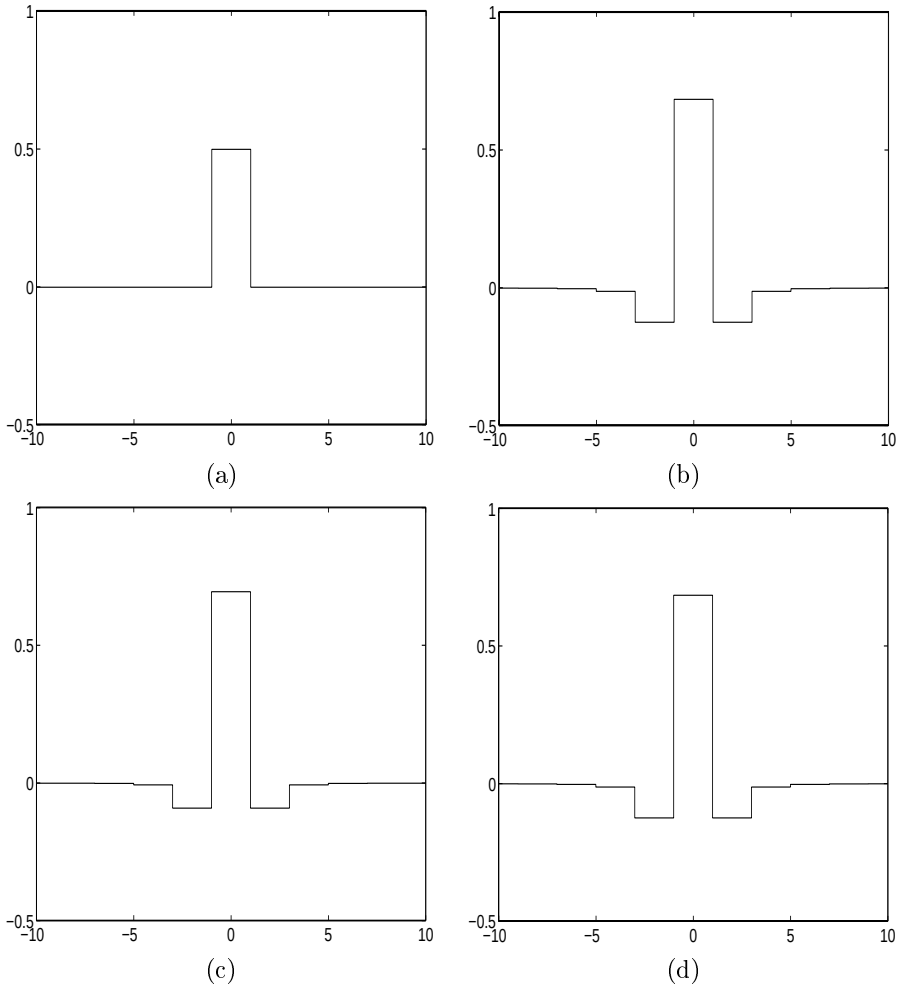


Figure 6: Smoothing kernels for the Cauchy model. The kernels are relates to the random components of Expression (4). The functions shown are, $\phi'_C(t)$ (a); $\phi'_1(t)$ (b); $\phi'_2(t)$ (c) and $\phi'_3(t)$ (d), respectively.

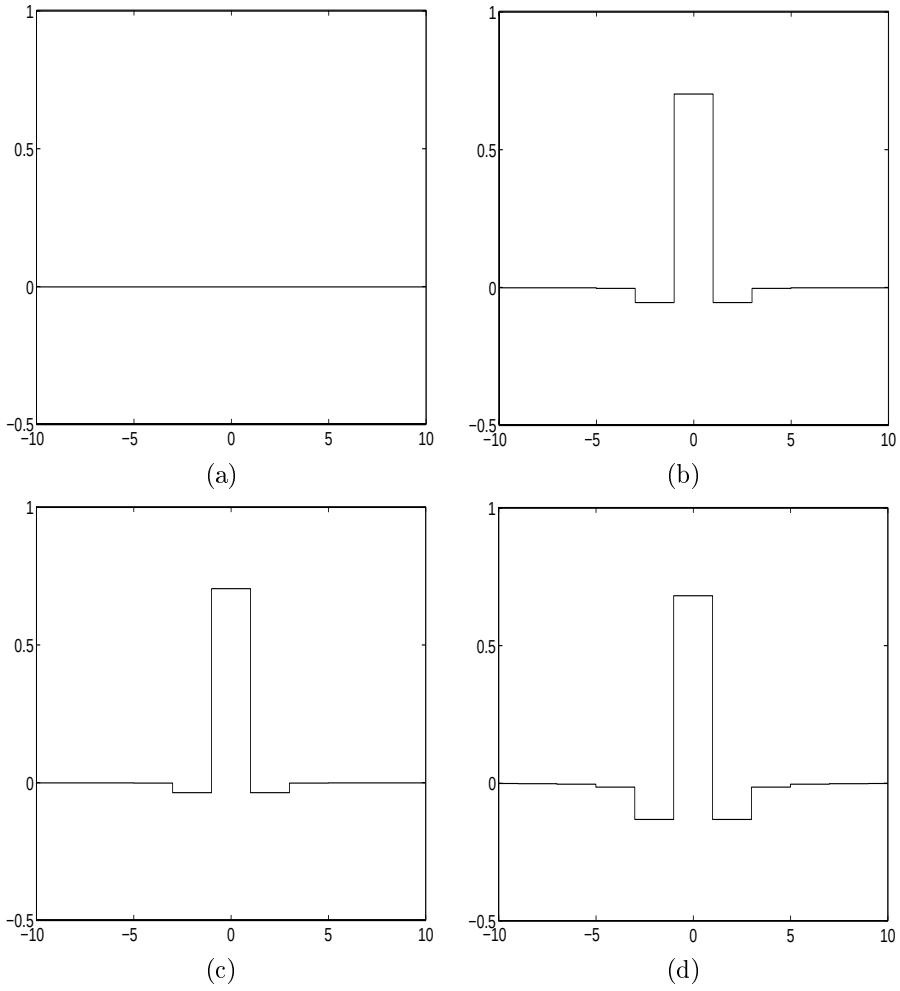


Figure 7: Smoothing kernels for the pure Gaussian model. The kernels are relates to the random components of Expression (4). The functions shown are, $\phi'_C(t)$ (a); $\phi'_1(t)$ (b); $\phi'_2(t)$ (c) and $\phi'_3(t)$ (d), respectively.

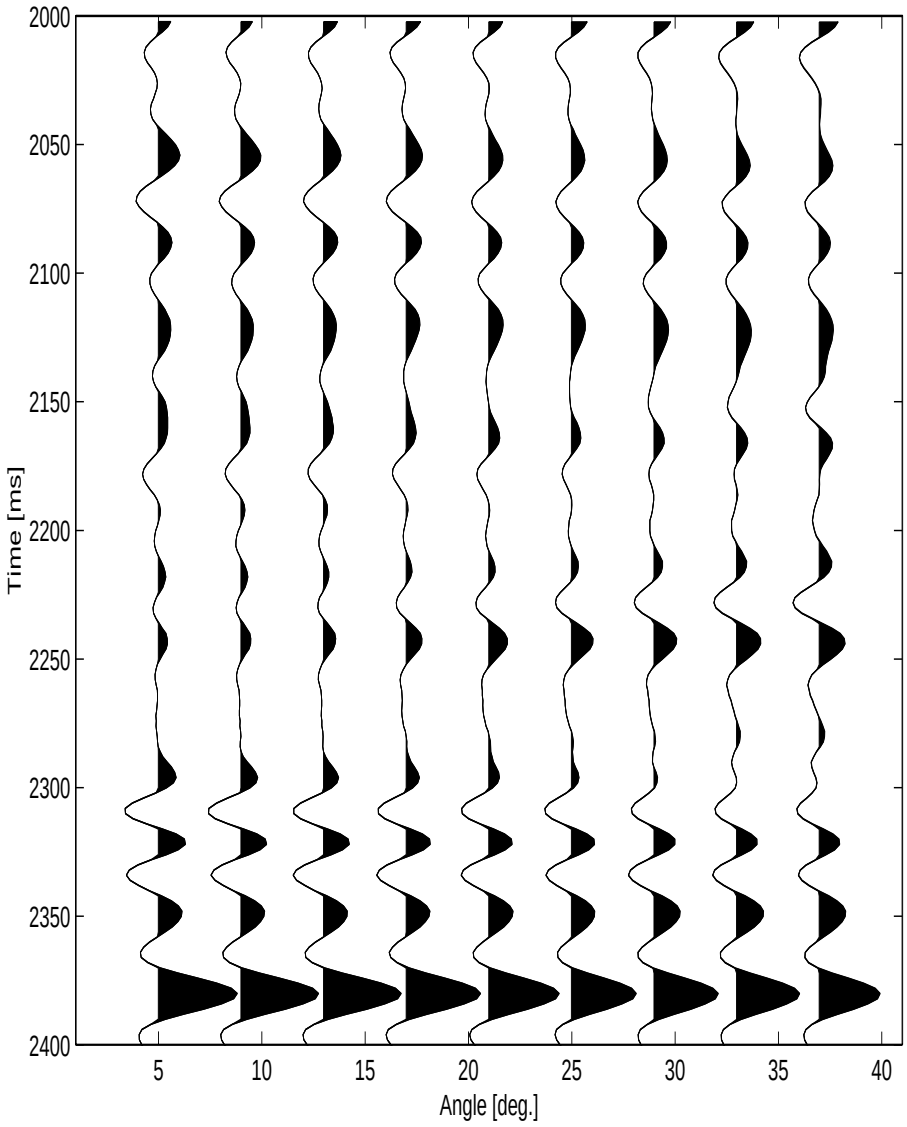


Figure 8: Synthetic observations. The CDP gather inverted in the example in Section 6. The observations are generated by the linearized model and errors where added according to the likelihood.

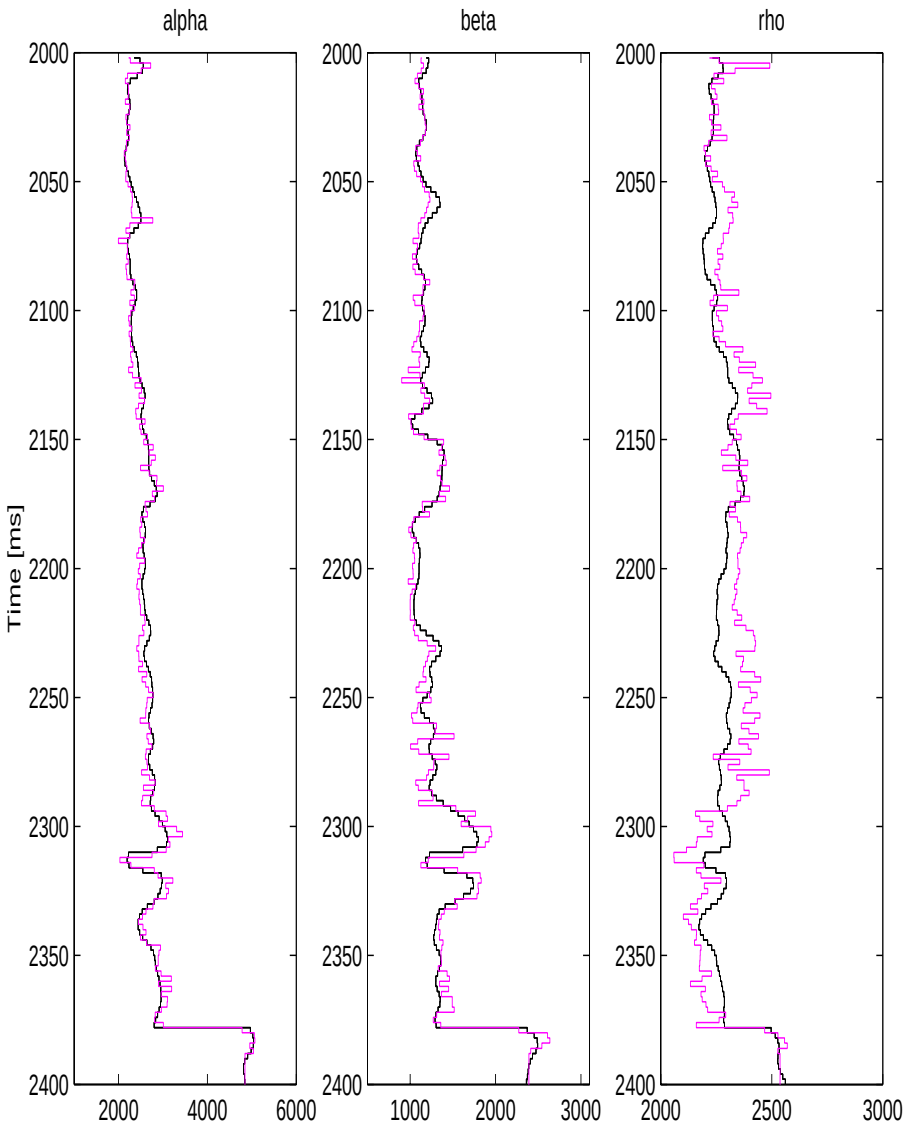


Figure 9: Inversion results for the Cauchy model. The median and is plotted together with the true parameter values.

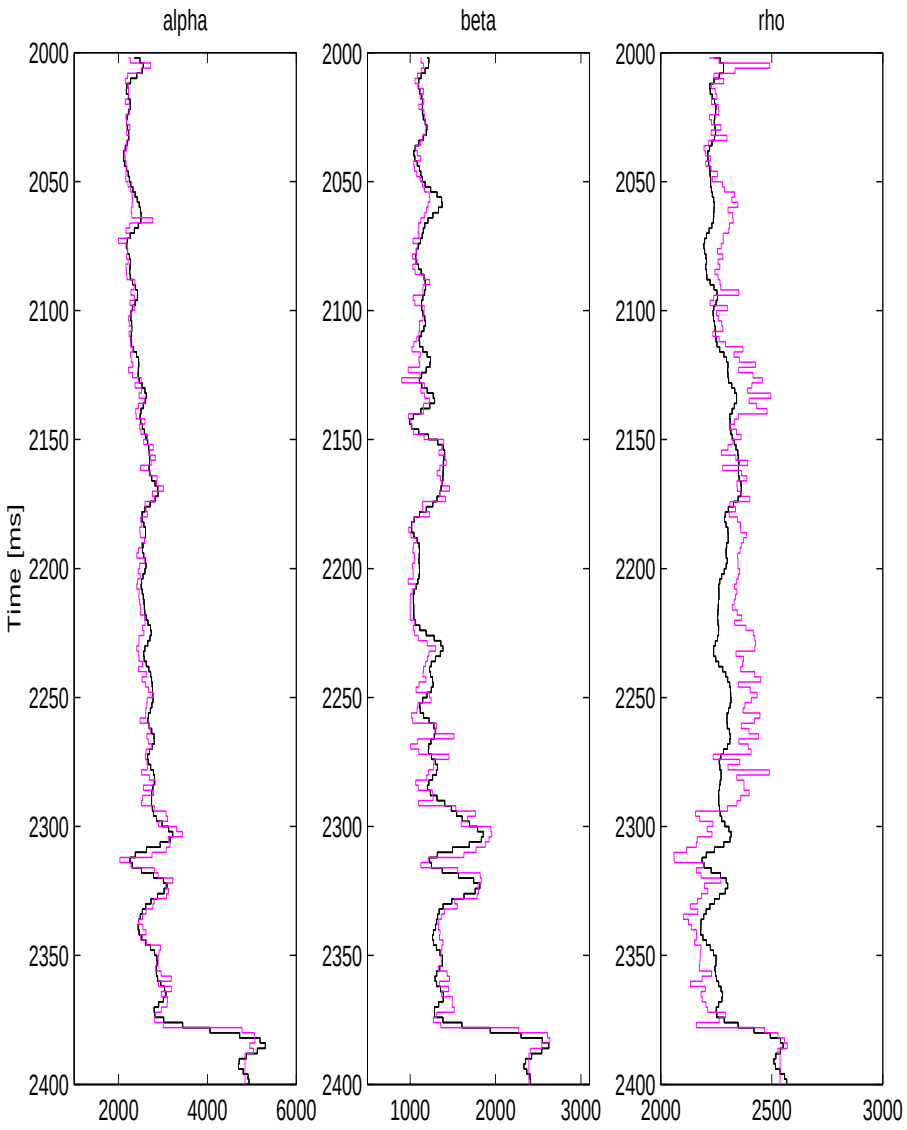


Figure 10: Inversion results for the pure Gaussian model. The median is plotted together with the true parameter values.

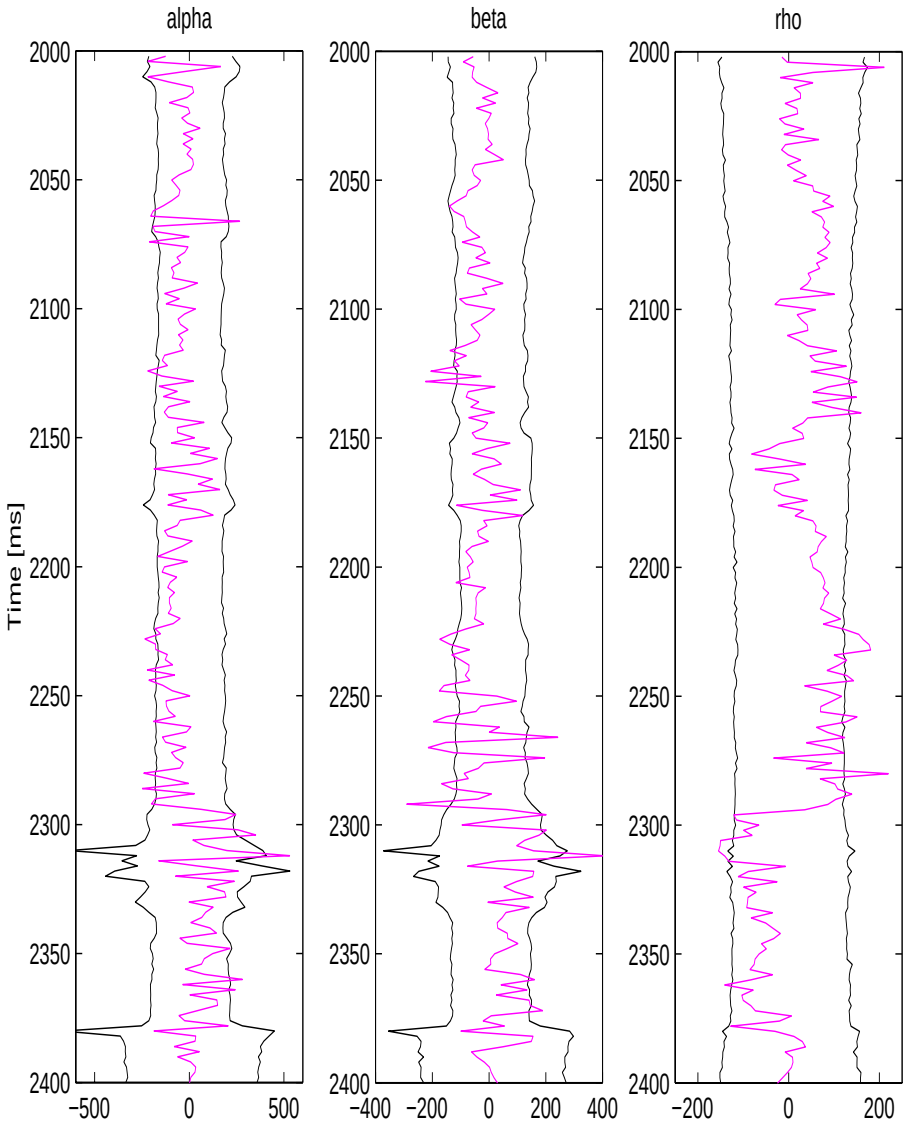


Figure 11: Errors for the Cauchy model. The 90% credibility interval for the error predicted by the simulations displayed together with the actual error.

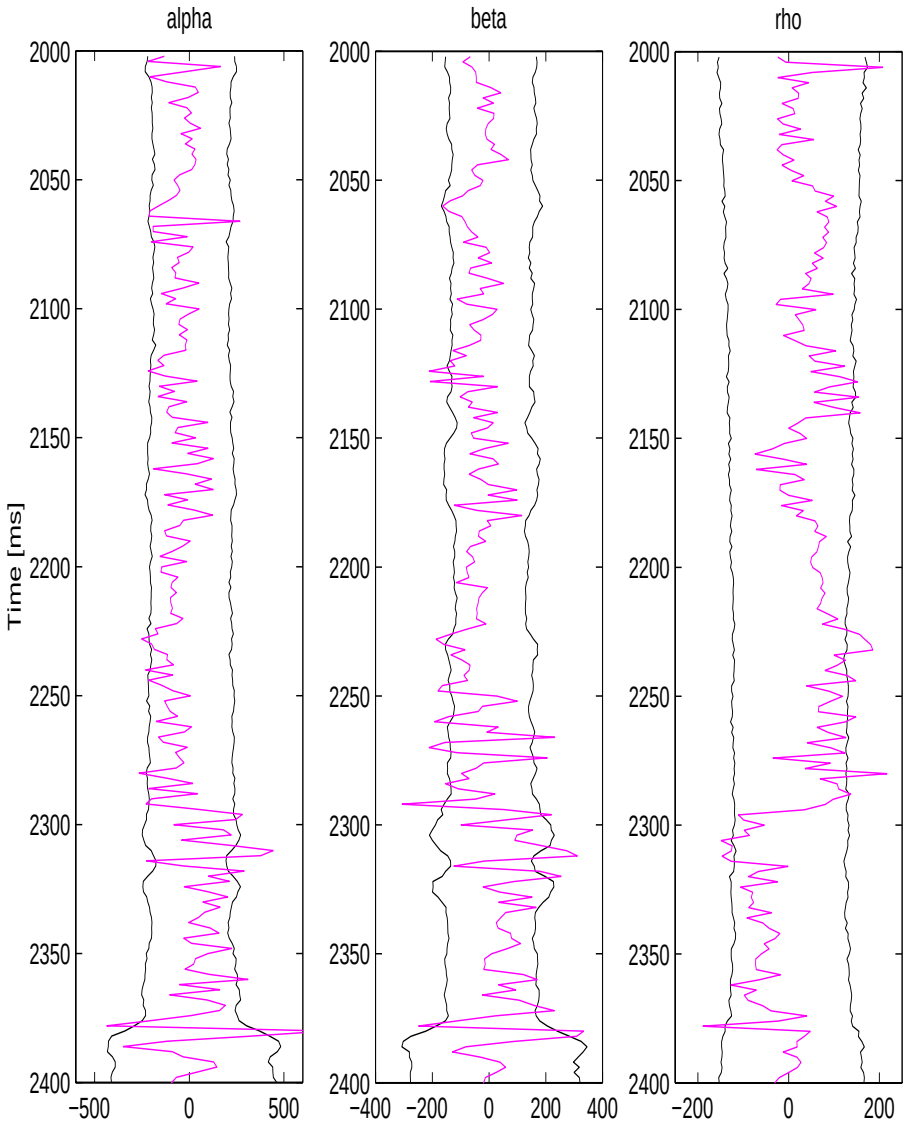


Figure 12: Errors for the pure Gaussian model. The 90% credibility interval for the error predicted by the simulations displayed together with the actual error.

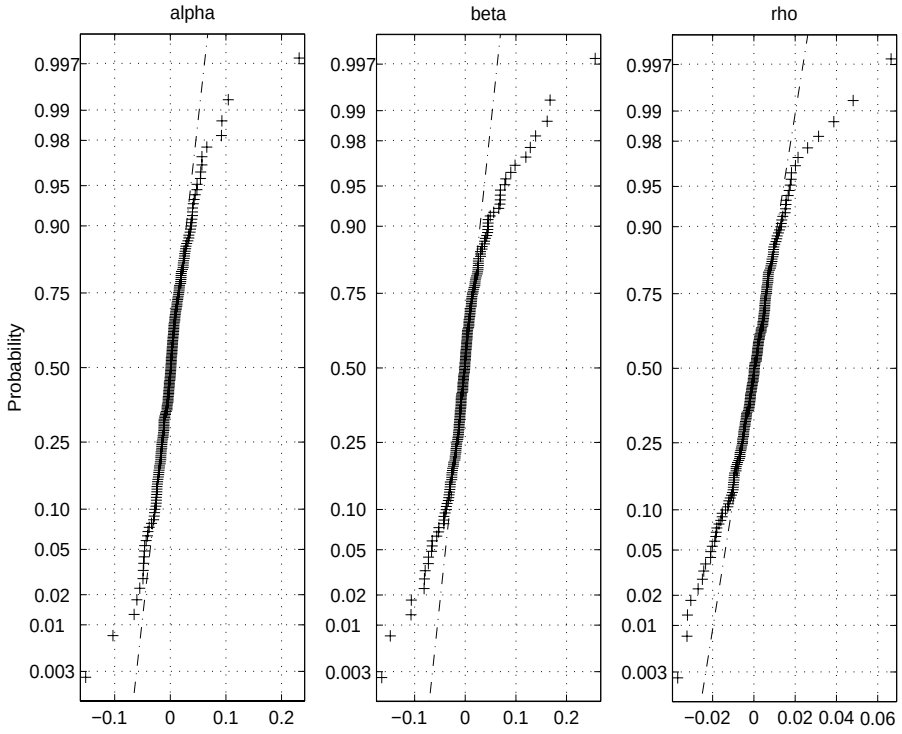


Figure 13: Normal plots for the derivative of the logarithm of the material parameters.

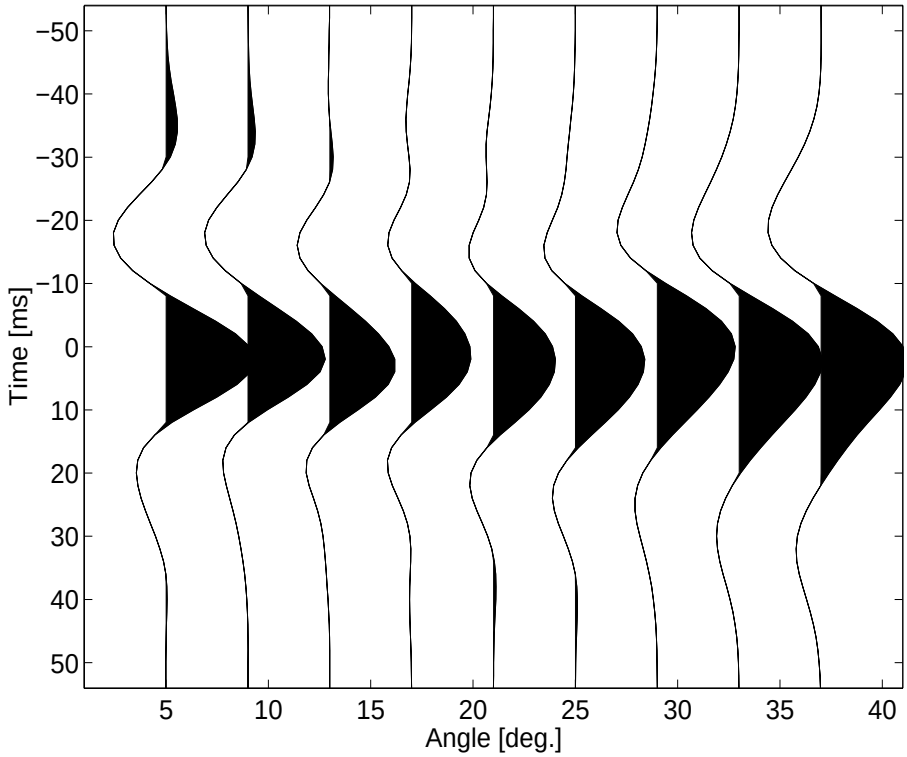


Figure 14: The wavelets for the inversion. The wavelets where estimated from the well log displayed in Figure 1 and the angle gather displayed in Figure 5.

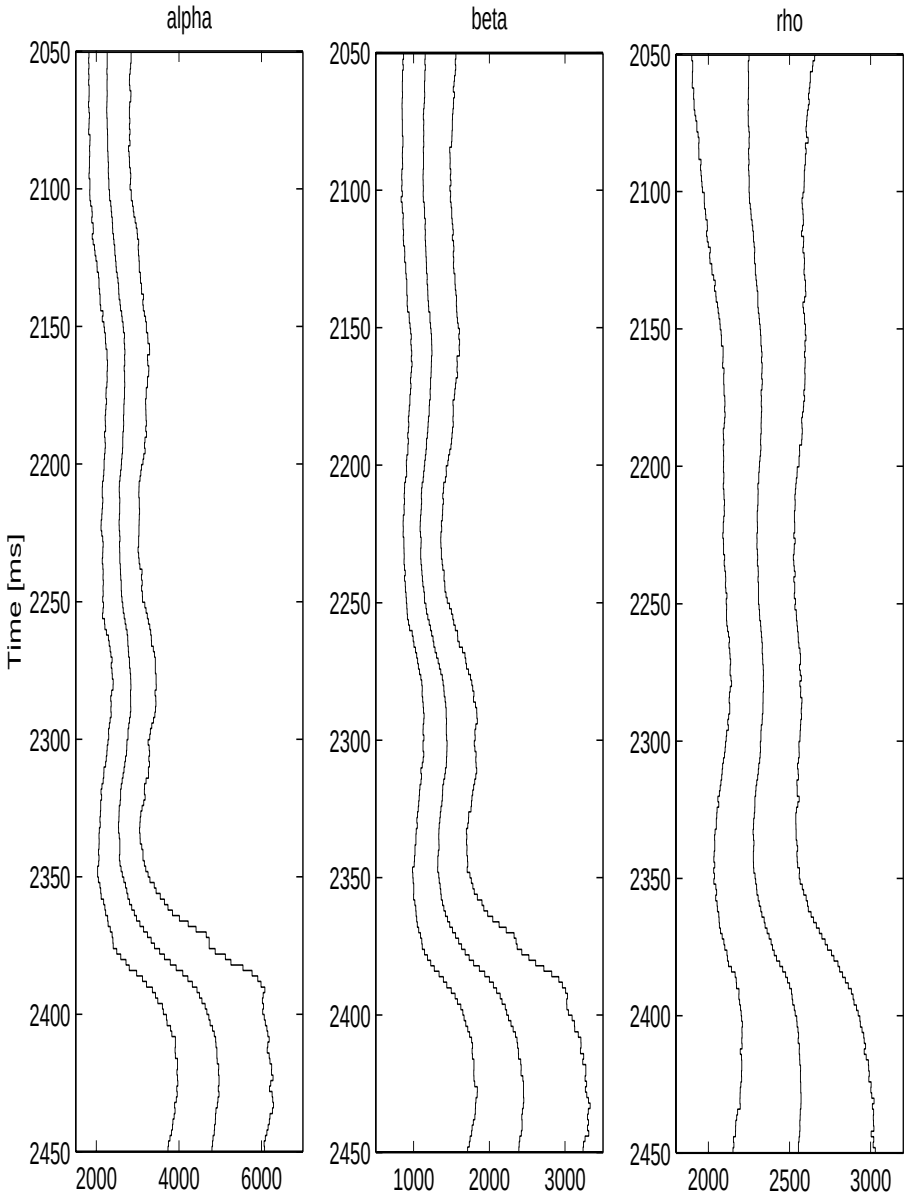


Figure 15: The prior distribution gather 67. The prior median and pointwise 90% credibility interval.

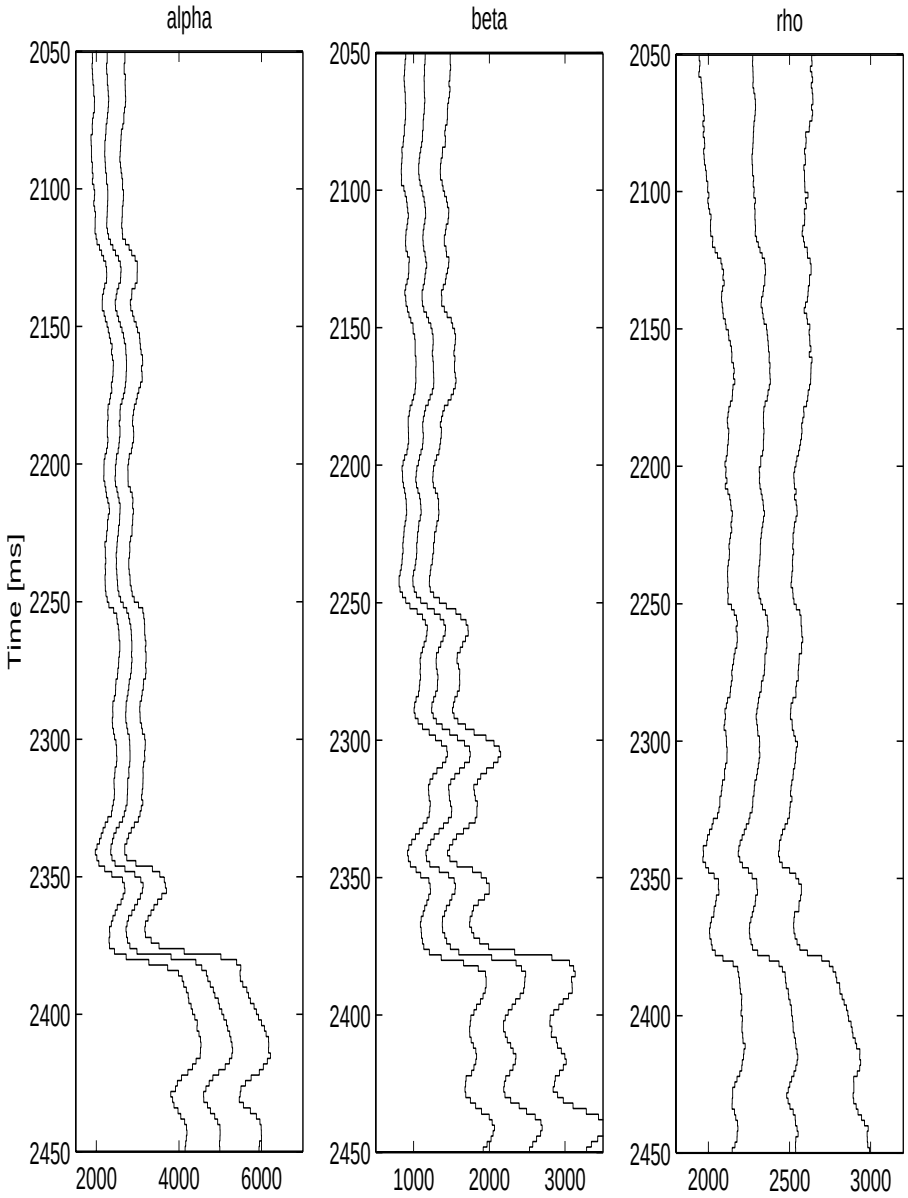


Figure 16: The posterior distribution gather 67. The posterior median and pointwise 90% credibility interval.

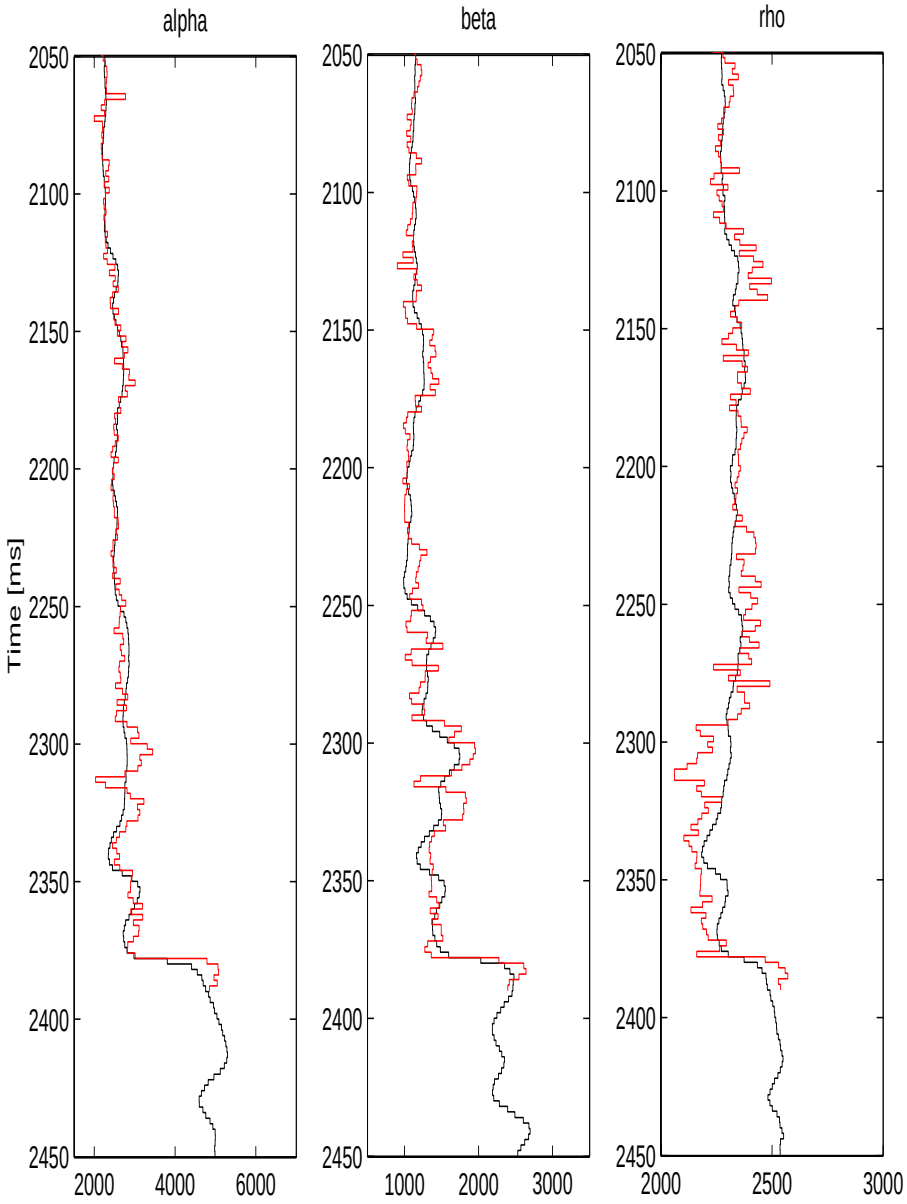


Figure 17: The inversion results for Sleipner Øst gather number 67. The figure display the inverted path together with the values observed in the well in red.

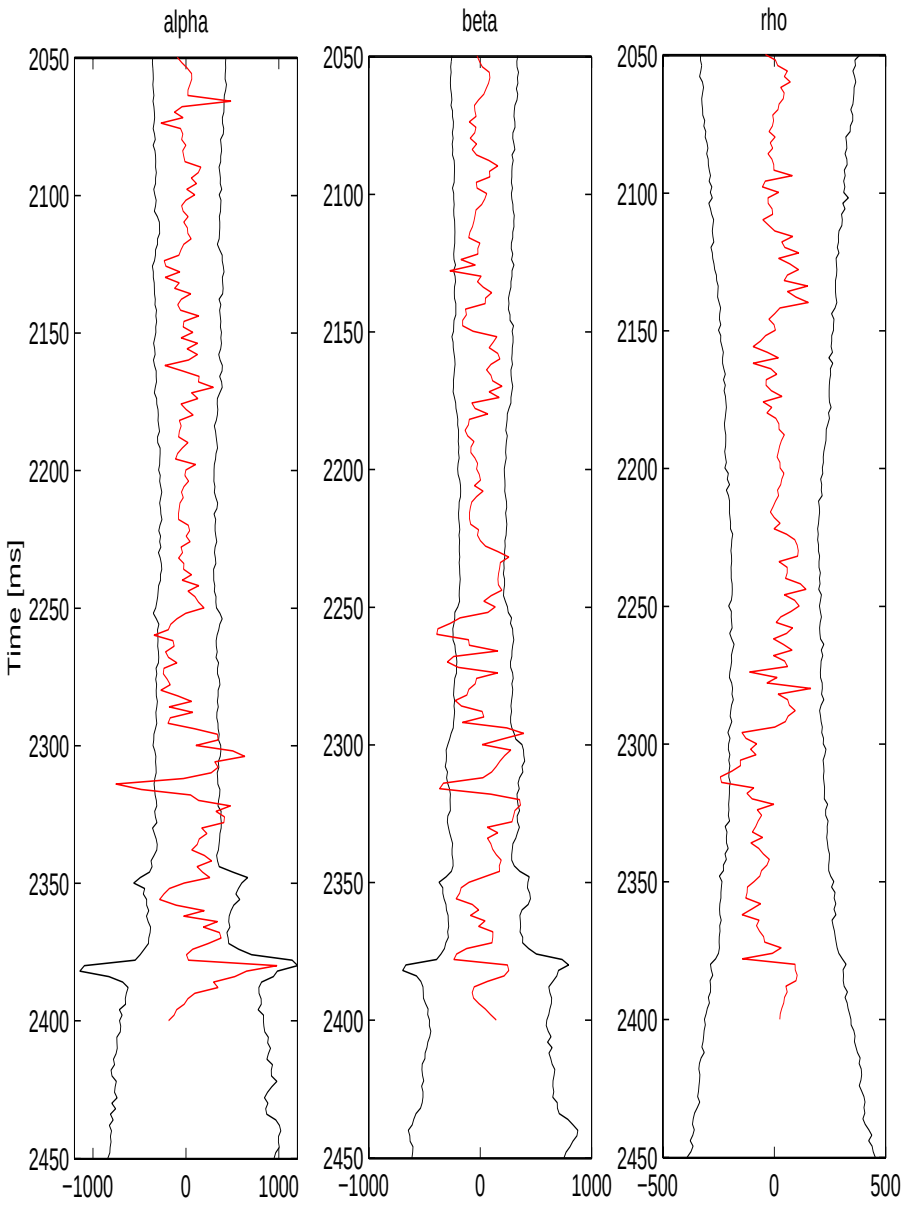


Figure 18: Errors for for Sleipner Øst gather number 67. The 90% credibility interval for the error predicted by the simulations displayed together with the actual error.

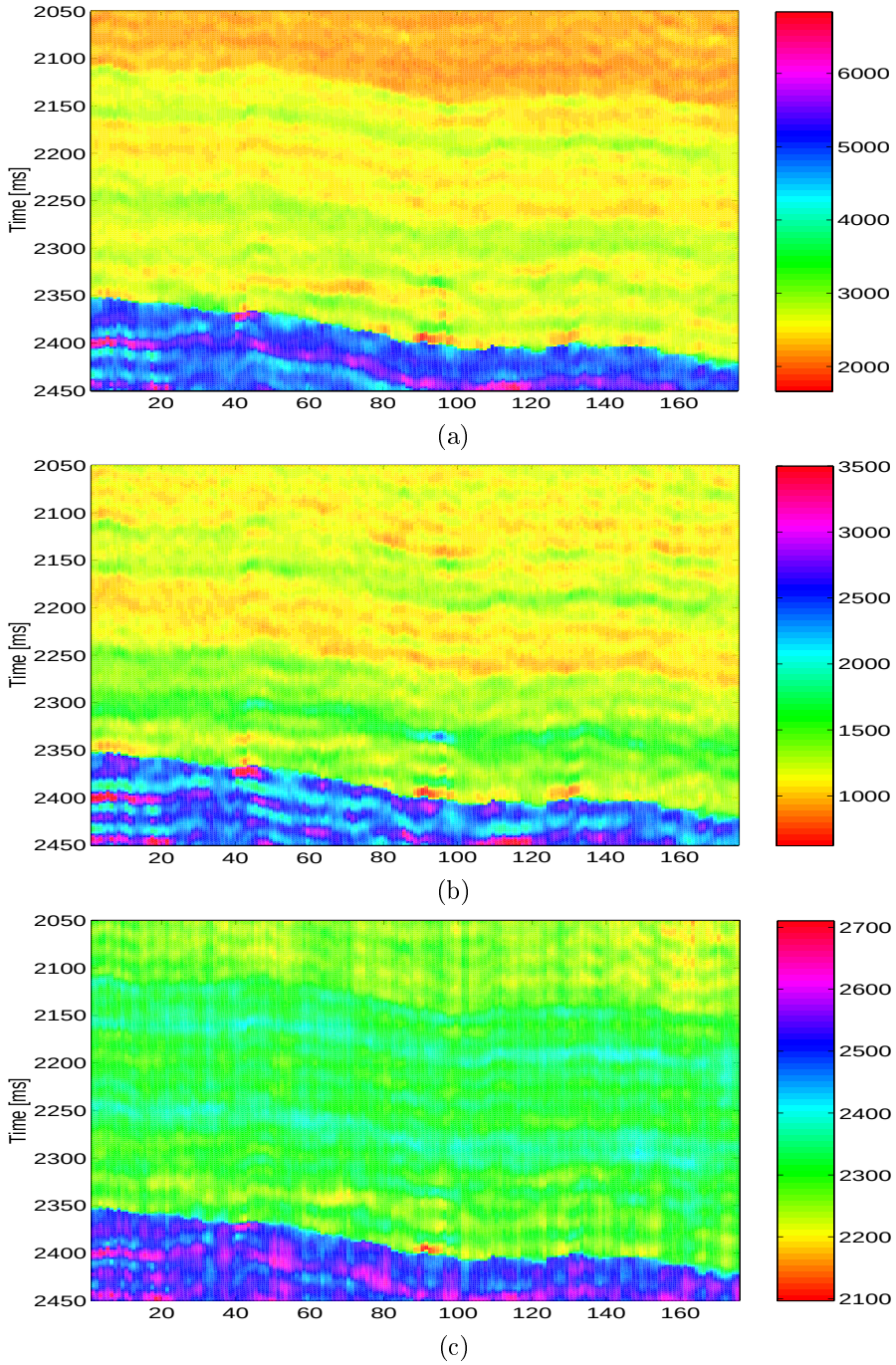


Figure 19: The inversion results for Sleipner Øst data. The estimated values being the posterior medians for (a) α; (b) β; and (c) ρ, respectively.

URN:NBN:no-333

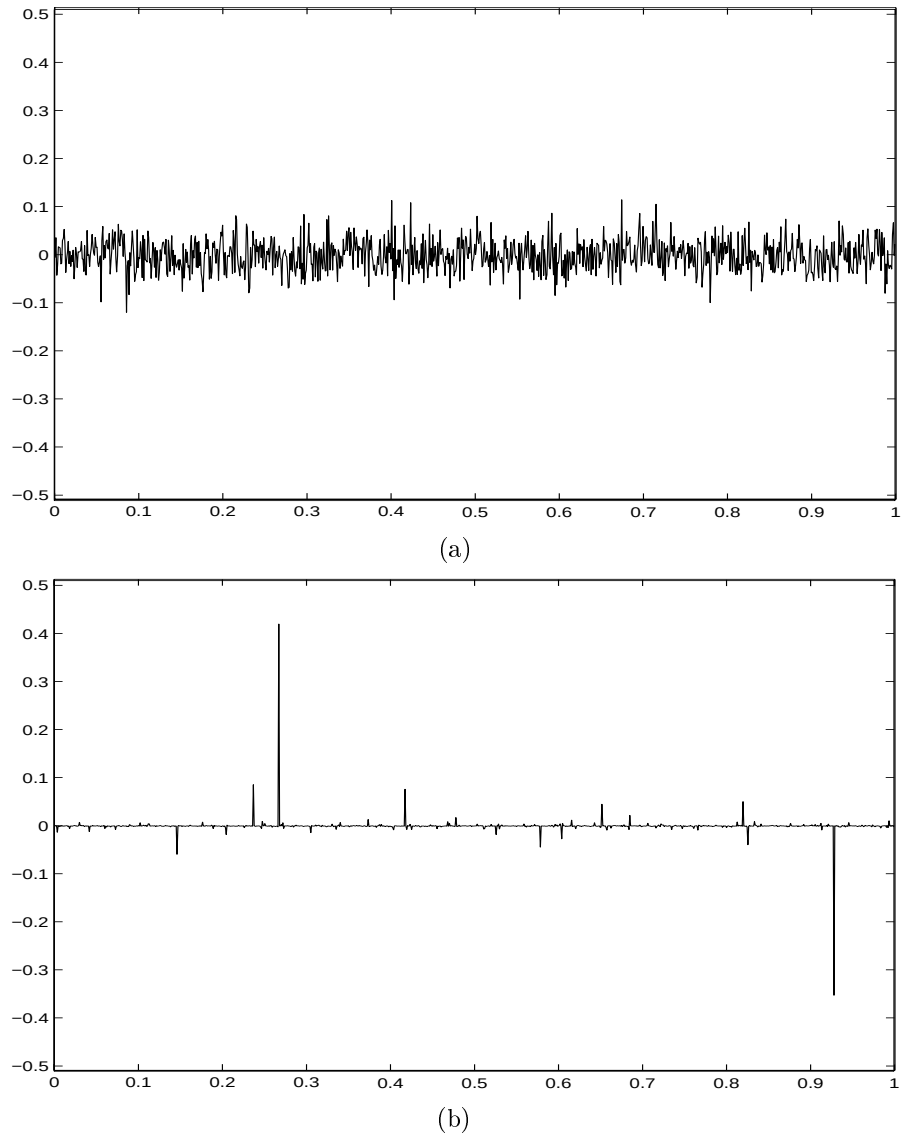


Figure 20: Comparison of the Winer measure and the Cauchy measure. The measures are discretized by integrating over intervals of length 1/1024. (a) One random sample of the discretized Wiener measure. The Gaussian seed have the scaling factor $1/\sqrt{1024}$. (b) One random sample of the discretized Cauchy measure. The Cauchy seed have the scaling factor $\tau_c/1024$. The factor $\tau_c \approx 0.416$ adjust the scale so that the the measure of the total region $[0, 1]$ have the same 90'th percentile for the two measures.

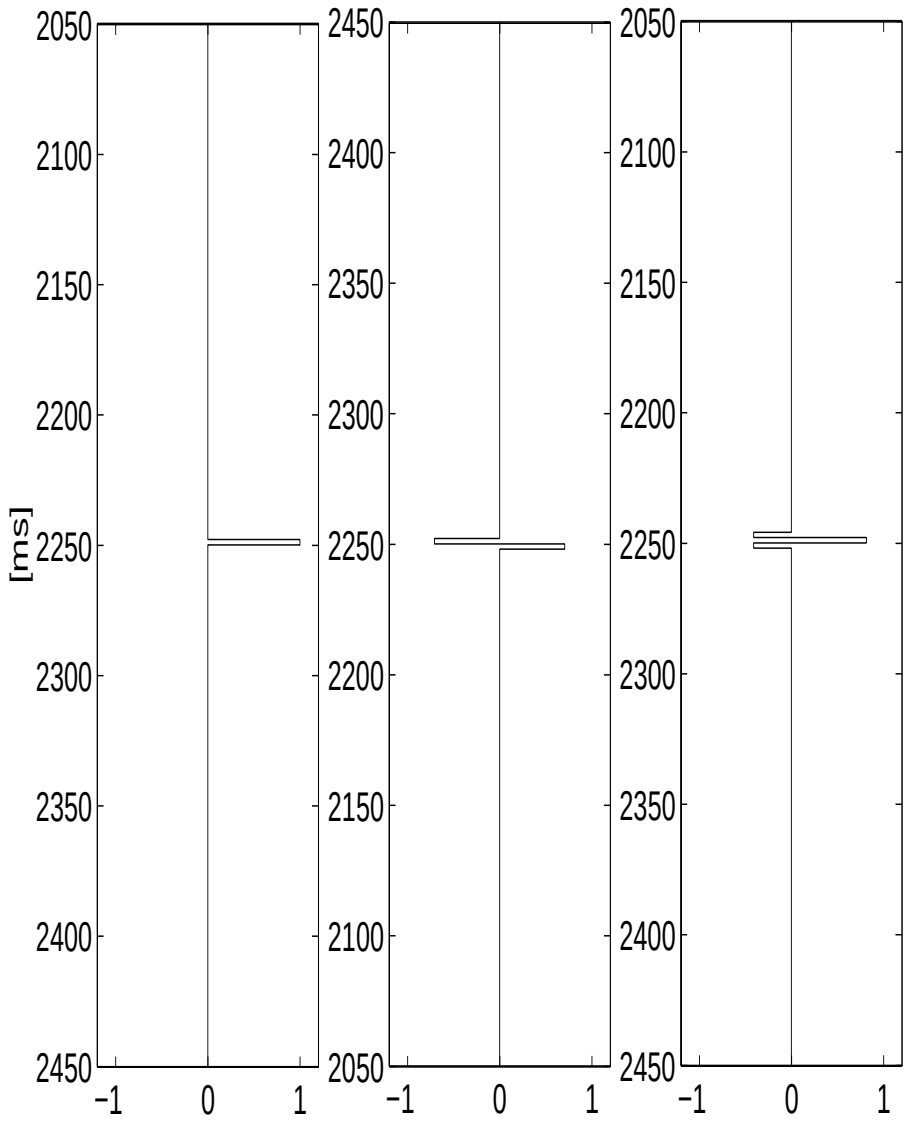


Figure 21: Directions for multi directional Gibbs sampler. The pool of basis vectors used in the multi directional Gibbs sampler, are translations of the functions above. The resolution of the figure is such that each plateau correspond to the size of one vector component.

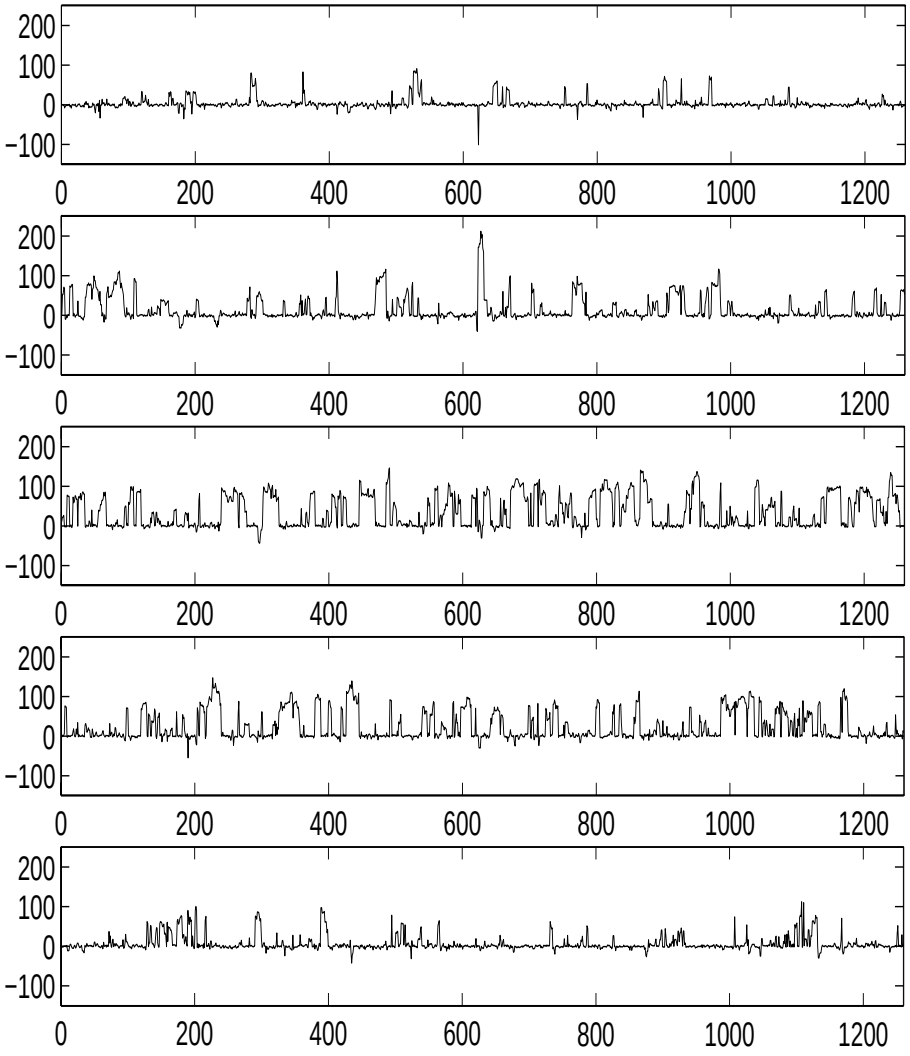


Figure 22: Mixing of multi directional Gibbs sampling, for the Sleipner Øst Field. Time traces for the the Cauchy seed. The sampled seed value that correspond to the leap value at 2380 ms is in the middle plot, the two nearest neighbors at both sides are above and below. Note that the scale of the figure is for the Cauchy seed. To get the scale that apply to the derivative of the logarithm of the material parameters, the seed should be divided by 2 due to the kernel in Figure 6(a) and multiplied by the second column in Table 1, due to the relation in Expression (4).

V

Rapid spatially coupled AVO inversion in the Fourier domain

Rapid spatially coupled AVO inversion in the Fourier domain

Arild Buland[1], Odd Kolbjørnsen & Henning Omre

Department of mathematical sciences
Norwegian University of Science and Technology
Trondheim, Norway

Abstract

Spatial coupling of the model parameters in an inversion problem provides lateral consistence and robust solutions. We have defined the inversion problem in a Bayesian framework, where the solution is represented by a posterior distribution obtained from a prior distribution and a likelihood model for the recorded data. The spatial coupling of the model parameters is imposed via the prior distribution by a spatial correlation function. In the Fourier domain, the spatially correlated model parameters can be decoupled, and the inversion problem can be solved independently for each frequency component.

For a spatial model parameter represented on n grid nodes, the computing time for the inversion in the Fourier domain follows a linear function of the number of grid nodes, while the computing time for the fast Fourier transform follows an $n \log n$ function. We have developed a 3-D linearized AVO inversion method with spatially coupled model parameters, where the objective is to obtain posterior distributions for P-wave velocity, S-wave velocity, and density.

The inversion algorithm has been tested on a 3-D dataset from the Sleipner Field with 4 million grid nodes, each with three unknown model parameters. The computing time was less than 3 minutes on the inversion in the Fourier domain, while each 3-D Fourier transform used about 30 seconds on a single 400 MHz Mips R12000 CPU.

KEY WORDS: *Bayesian statistic, Seismic inversion, 3-D inversion, Sampling algorithm, Merging observations*

[1] Arild Buland also work at Statoil Research Centre,Trondheim, Norway.

1 Introduction

Many geophysical inverse problems can naturally be cast in a Bayesian framework, where it is possible to combine available prior knowledge with the information contained in the measured data, see e.g., Tarantola and Valette (1982); Tarantola (1987); Duijndam (1988a,b). The solution of a Bayesian inverse problem is represented by the posterior distribution. From the posterior distribution, the best estimate of the solution and the corresponding uncertainty can be extracted. A set of plausible solutions can also be drawn directly from the posterior distribution.

Amplitude versus offset (AVO) inversion can be used to extract information about the elastic subsurface parameters from the angle dependency in the reflectivity, see e.g., Hampson and Russell (1990); Lörtzer and Berkhout (1993); Pan et al. (1994); Buland et al. (1996); Gouveia and Scales (1998). In practice, and especially for 3-D surveys, linearized AVO inversion is attractive since it can be performed with use of moderate computer resources. Prior to a linearized AVO inversion, the seismic data must be processed to remove nonlinear relations between the model parameters and the seismic response. Important steps in the processing are the removal of the moveout, multiples, and the effects of geometrical spreading and absorption. The seismic data should be prestack migrated, such that dip related effects are removed. After prestack migration, it is reasonable to assume that each single bin-gather can be regarded as the response of a local 1-D earth model. The benefits of prestack migration before AVO analysis is discussed in Brown (1992); Mosher et al. (1996); Buland and Landrø (2001). We further assume that wave mode conversions, interbed multiples and anisotropy effects can be neglected after processing. Finally, the prestack gathers must be transformed from offsets to reflection angles.

Under Gaussian model assumptions, an explicit analytical solution of a Bayesian linearized AVO inversion problem is worked out for a single angle gather, see Buland and Omre (2001a). The objective of this method is to obtain posterior distributions for the P-wave velocity, S-wave velocity, and density. The solution is fast to compute and the method is therefore suitable for inversion of seismic 3-D data. However, the model parameters are not laterally coupled, so each CDP gather is inverted independently of the neighbor CDPs.

In the current paper, a spatially coupled model is defined to obtain a spatial consistent and robust solution of the linearized AVO inversion problem. The consequence of the spatial coupling is that the solution in each location depends on the solutions in all other locations. Even for small data sets, this results in an enormously system of equations. For example, a small 3-D inversion problem may have dimension $100 \times 100 \times 100$, that is $n = 10^6$ grid nodes. The computing time for inversion of the corresponding equation system is proportional to n^3, denoted $\mathcal{O}(n^3)$. An obvious approximate approach to this problem is to assume that the solution in a specific location only depends on the solutions at the

nearest neighbor locations, see e.g., Omre et al. (1993); Rue (2000). Domain decomposition constitutes another approximate technique, where the inversion area is divided into several subareas, each limited to a size which efficiently can be handled by the actual computer. The problems with this method are related to boundary effects and the final coupling of the inverted subareas. In this paper we present a Bayesian AVO inversion method where the spatial coupling can be handled exactly under certain assumptions. The method utilizes that the covariance matrix for a homogeneously correlated spatial variable sampled on a regular grid can be diagonalized by a Fourier transform, see Wood (1995). In the Fourier domain, the inversion problem can be solved independently for each frequency component. The computing time for the inversion in the Fourier domain is then $\mathcal{O}(n)$, which is the optimal scaling property for an inversion algorithm. However, the computing time for a fast Fourier transform is according to $\mathcal{O}(n \log n)$, such that the Fourier and inverse Fourier transforms will dominate the computing time asymptotically.

2 Methodology

The seismic reflection coefficients depend on the material properties of the subsurface. An isotropic, elastic medium is completely described by the three material parameters $\{\alpha(\boldsymbol{x},t), \beta(\boldsymbol{x},t), \rho(\boldsymbol{x},t)\}$, where α, β, and ρ are P-wave velocity, S-wave velocity, and density, $\boldsymbol{x}$ is the lateral position, and t is the vertical seismic traveltime. A weak contrast approximation to the seismic reflectivity is (Aki and Richards, 1980; Stolt and Weglein, 1985)

$$\begin{aligned} c(\boldsymbol{x},t,\theta) = a_\alpha(\boldsymbol{x},t,\theta)\,\frac{\partial}{\partial t}\ln\alpha(\boldsymbol{x},t) \\ +a_\beta(\boldsymbol{x},t,\theta)\,\frac{\partial}{\partial t}\ln\beta(\boldsymbol{x},t) + a_\rho(\boldsymbol{x},t,\theta)\,\frac{\partial}{\partial t}\ln\rho(\boldsymbol{x},t), \end{aligned} \quad (1)$$

where θ is the reflection angle, and

$$\begin{aligned} a_\alpha(\boldsymbol{x},t,\theta) &= \frac{1}{2}\left(1+\tan^2\theta\right), \\ a_\beta(\boldsymbol{x},t,\theta) &= -4\frac{\beta^2(\boldsymbol{x},t)}{\alpha^2(\boldsymbol{x},t)}\sin^2\theta, \\ a_\rho(\boldsymbol{x},t,\theta) &= \frac{1}{2}\left(1-4\frac{\beta^2(\boldsymbol{x},t)}{\alpha^2(\boldsymbol{x},t)}\sin^2\theta\right). \end{aligned} \quad (2)$$

The elastic subsurface parameters can be collected in a vector field. Motivated by the form of the reflectivity function in expression (1), let

$$\boldsymbol{m}(\boldsymbol{x},t) = \left[\ln\alpha(\boldsymbol{x},t), \ln\beta(\boldsymbol{x},t), \ln\rho(\boldsymbol{x},t)\right]^T, \quad (3)$$

where T denotes transpose such that $\boldsymbol{m}(\boldsymbol{x},t)$ is a column vector. Further, let $\boldsymbol{a}(\boldsymbol{x},t,\theta)$ be a row vector

$$\boldsymbol{a}(\boldsymbol{x},t,\theta) = [a_\alpha(\boldsymbol{x},t,\theta), a_\beta(\boldsymbol{x},t,\theta), a_\rho(\boldsymbol{x},t,\theta)]\,. \tag{4}$$

For zero incidence reflections, the reflectivity function $c(\boldsymbol{x},t,\theta)$ reduces to

$$c(\boldsymbol{x},t,0) = \frac{1}{2}\,\frac{\partial}{\partial t}\ln Z_P(\boldsymbol{x},t), \tag{5}$$

where $Z_P = \alpha\rho$ is the acoustic impedance. In this case, $\boldsymbol{a}(\boldsymbol{x},t,0)$ reduces to $1/2$, and $\boldsymbol{m}(\boldsymbol{x},t)$ reduces to $\ln Z_P(\boldsymbol{x},t)$. Inversion for acoustic impedance from zero incidence data can be defined by a simple reformulation of the AVO inversion problem, and is therefore not further discussed in this paper.

The seismic data are represented by the convolutional model

$$d_{obs}(\boldsymbol{x},t,\theta) = \int s(\tau,\theta)\; c(\boldsymbol{x},t-\tau,\theta)\; d\tau + e(\boldsymbol{x},t,\theta), \tag{6}$$

where s is the wavelet, and e is an error term. Note that the wavelet is allowed to be angle dependent, but independent of the lateral position $\boldsymbol{x}$. The wavelet is assumed to be stationary within a limited target window.

2.1 The Fourier transform

The spatial coupled inversion problem can be decoupled in the Fourier domain. The inversion problem can then be solved independently for each frequency component. Let the Fourier transform be defined as

$$\tilde{f}(\boldsymbol{k},\omega) = \iiint f(\boldsymbol{x},t)\exp[-i(\boldsymbol{k}\cdot\boldsymbol{x}+\omega t)]\; d\boldsymbol{x}\; dt, \tag{7}$$

with inverse transform

$$f(\boldsymbol{x},t) = \frac{1}{(2\pi)^3}\iiint \tilde{f}(\boldsymbol{k},\omega)\exp[i(\boldsymbol{k}\cdot\boldsymbol{x}+\omega t)]\; d\boldsymbol{k}\; d\omega, \tag{8}$$

where $i = \sqrt{-1}$, ω is the temporal frequency, and $\boldsymbol{k}$ is the spatial frequency vector with components k_x and k_y. In the following, frequency means a $(\boldsymbol{k},\omega)$ pair. Note that geophysicists often call k_x and k_y wavenumbers, and the sign of the temporal frequency ω is often defined opposite of the definition above.

The Fourier transform of the convolutional model in expression (6) is

$$\tilde{d}_{obs}(\boldsymbol{k},\omega,\theta) = \tilde{s}(\omega,\theta)\;\tilde{c}(\boldsymbol{k},\omega,\theta) + \tilde{e}(\boldsymbol{k},\omega,\theta). \tag{9}$$

We use a constant β/α ratio in expression (2), such that $\boldsymbol{a}(\boldsymbol{x},t,\theta) = \boldsymbol{a}(\theta)$, then the Fourier transform of the convolutional model can be written

$$\tilde{d}_{obs}(\boldsymbol{k},\omega,\theta) = \boldsymbol{g}(\omega,\theta)\cdot\tilde{\boldsymbol{m}}(\boldsymbol{k},\omega) + \tilde{e}(\boldsymbol{k},\omega,\theta), \tag{10}$$

where $\boldsymbol{g}$ is a row vector defined by

$$\boldsymbol{g}(\omega,\theta) = i\omega\ \tilde{s}(\omega,\theta)\ \boldsymbol{a}(\theta), \tag{11}$$

and $\tilde{\boldsymbol{m}}(\boldsymbol{k},\omega)$ is the Fourier transform of the elements in $\boldsymbol{m}(\boldsymbol{x},t)$, that is $\ln\alpha(\boldsymbol{x},t)$, $\ln\beta(\boldsymbol{x},t)$, and $\ln\rho(\boldsymbol{x},t)$, see expression (3). Although a constant β/α ratio is used in an approximative expression for the reflection coefficient, the solution will in general have a varying β/α ratio. Further, note that the differentiations in equation (1) now appear as an $i\omega$ term in expression (11). In Buland and Omre (2001a), it was assumed that $\boldsymbol{m}(\boldsymbol{x},t)$ was differentiable with respect to time, but here it is sufficient to assume that the convolution of $s(t,\theta)$ and $\boldsymbol{a}(\theta)\cdot\boldsymbol{m}(\boldsymbol{x},t)$ is differentiable.

For a set of n_θ specified reflection angles, the Fourier transformed seismic data can be written in the vector form

$$\tilde{\boldsymbol{d}}_{obs}(\boldsymbol{k},\omega) = \boldsymbol{G}(\omega)\ \tilde{\boldsymbol{m}}(\boldsymbol{k},\omega) + \tilde{\boldsymbol{e}}(\boldsymbol{k},\omega), \tag{12}$$

where $\boldsymbol{G}(\omega)$ is an $n_\theta \times 3$ matrix defined by

$$\boldsymbol{G}(\omega) = \begin{bmatrix} \boldsymbol{g}(\omega,\theta_1) \\ \vdots \\ \boldsymbol{g}(\omega,\theta_{n_\theta}) \end{bmatrix}, \tag{13}$$

and $\tilde{\boldsymbol{d}}_{obs}(\boldsymbol{k},\omega)$ and $\tilde{\boldsymbol{e}}(\boldsymbol{k},\omega)$ are n_θ-dimensional vectors.

2.2 The prior model

The elastic parameters $\alpha(\boldsymbol{x},t)$, $\beta(\boldsymbol{x},t)$, and $\rho(\boldsymbol{x},t)$ are assumed to be log-Gaussian random fields, hence the vector field $\boldsymbol{m}(\boldsymbol{x},t)$, which contains the logarithm of these parameters, is Gaussian with expectation

$$\boldsymbol{\mu}_m(\boldsymbol{x},t) = [\mu_\alpha(\boldsymbol{x},t), \mu_\beta(\boldsymbol{x},t), \mu_\rho(\boldsymbol{x},t)]^T, \tag{14}$$

where the elements are the expectations of $\ln\alpha(\boldsymbol{x},t)$, $\ln\beta(\boldsymbol{x},t)$, and $\ln\rho(\boldsymbol{x},t)$, respectively, and with covariance

$$\boldsymbol{\Sigma}_m(\boldsymbol{x}_1,t_1;\boldsymbol{x}_2,t_2) = \mathrm{Cov}\{\boldsymbol{m}(\boldsymbol{x}_1,t_1), \boldsymbol{m}(\boldsymbol{x}_2,t_2)\}. \tag{15}$$

We assume that the covariance function is stationary and homogeneous, and can be factorized as

$$\boldsymbol{\Sigma}_m(\boldsymbol{x}_1,t_1;\boldsymbol{x}_2,t_2) = \boldsymbol{\Sigma}_{0,m}\ \nu_m(\boldsymbol{\xi},\tau), \tag{16}$$

where $\nu_m(\boldsymbol{\xi},\tau)$ is a spatial correlation function, $\boldsymbol{\xi} = [|x_2 - x_1|, |y_2 - y_1|]^T$, $\tau = |t_2 - t_1|$, and

$$\boldsymbol{\Sigma}_{0,m} = \begin{bmatrix} \sigma_\alpha^2 & \sigma_\alpha\sigma_\beta\nu_{\alpha\beta} & \sigma_\alpha\sigma_\rho\nu_{\alpha\rho} \\ \sigma_\alpha\sigma_\beta\nu_{\alpha\beta} & \sigma_\beta^2 & \sigma_\beta\sigma_\rho\nu_{\beta\rho} \\ \sigma_\alpha\sigma_\rho\nu_{\alpha\rho} & \sigma_\beta\sigma_\rho\nu_{\beta\rho} & \sigma_\rho^2 \end{bmatrix}. \tag{17}$$

The diagonal elements of $\mathbf{\Sigma}_{0,m}$ are the variances, and $\nu_{\alpha\beta}$, $\nu_{\alpha\rho}$, and $\nu_{\beta\rho}$, are the correlations between $\ln\alpha(\boldsymbol{x},t)$, $\ln\beta(\boldsymbol{x},t)$ and $\ln\rho(\boldsymbol{x},t)$, respectively. A more general covariance function is also allowed, where the covariance function is composed of a sum of terms with the form on the right-hand side of expression (16).

The model parameters and the seismic data are so far defined for continuous $\boldsymbol{x}$ and t. In practice, the seismic data are available in a discrete form. In the following we assume an identical sampling of the model parameters and the seismic data on a regular grid in space and time. This is a required assumption for this method. Let the discrete representation of the model parameter field $\boldsymbol{m}(\boldsymbol{x},t)$ in a time window and for a set of lateral positions be written $\boldsymbol{m}$. The discrete model parameter vector $\boldsymbol{m}$ is Gaussian with expectation vector $\boldsymbol{\mu}_m$ and covariance matrix $\mathbf{\Sigma}_m$, shortly denoted

$$\boldsymbol{m} \sim \mathcal{N}_{n_m}(\boldsymbol{\mu}_m, \mathbf{\Sigma}_m), \tag{18}$$

where n_m is the dimension of $\boldsymbol{m}$. For a 3-D problem on a regular grid with $n = n_x n_y n_t$ grid nodes, the dimensions of $\boldsymbol{m}$ and $\boldsymbol{\mu}_m$ are $n_m = 3n$, while the dimension of $\mathbf{\Sigma}_m$ is $n_m \times n_m$. Since the covariance function can be written as in expression (16), the complete covariance matrix for $\boldsymbol{m}$ can be written as a Kronecker product

$$\mathbf{\Sigma}_m = \mathbf{\Sigma}_{0,m} \otimes \mathbf{\Upsilon}_m, \tag{19}$$

where each of the elements in the 3×3 constant matrix $\mathbf{\Sigma}_{0,m}$, defined in expression (17), are multiplied with the $n \times n$ correlation matrix $\mathbf{\Upsilon}_m$, defined from $\nu_m(\boldsymbol{\xi},\tau)$.

The spatial dependency can be decoupled by Fourier transforming the problem. The Fourier transform of $\boldsymbol{m}$, denoted $\tilde{\boldsymbol{m}}$, is Gaussian with Fourier transformed expectation vector $\tilde{\boldsymbol{\mu}}_m$ and covariance matrix

$$\tilde{\mathbf{\Sigma}}_m = \mathbf{\Sigma}_{0,m} \otimes \tilde{\mathbf{\Lambda}}_m, \tag{20}$$

where $\tilde{\mathbf{\Lambda}}_m$ is the diagonal eigenvalue matrix of $\mathbf{\Upsilon}_m$ scaled by the dimension of the discrete Fourier transform, see Appendix A. The important consequence of this diagonalization is that the frequency components of $\tilde{\boldsymbol{m}}$ are independent, with each component being Gaussian

$$\tilde{\boldsymbol{m}}_k \sim \mathcal{N}_3(\tilde{\boldsymbol{\mu}}_{m,k}, \tilde{\mathbf{\Sigma}}_{m,k}), \tag{21}$$

with index k corresponding to a specific discrete $(\boldsymbol{k},\omega)$ pair. The covariance matrix for frequency component k is a 3×3 matrix defined by

$$\tilde{\mathbf{\Sigma}}_{m,k} = \tilde{\lambda}_{m,k}\mathbf{\Sigma}_{0,m}, \tag{22}$$

with $\tilde{\lambda}_{m,k}$ being the corresponding diagonal element in the scaled diagonal eigenvalue matrix $\tilde{\mathbf{\Lambda}}_m$. Further, and of crucial importance, is that the $n = n_x n_y n_t$

eigenvalues can be calculated efficiently by a 3-D Fourier transform of n discrete samples of the correlation function $\nu_m(\boldsymbol{\xi}, \tau)$ extended to a circulant form, see Appendix A. That means that the complete $n \times n$ correlation matrix $\boldsymbol{\Upsilon}_m$ and the even larger covariance matrix $\boldsymbol{\Sigma}_m$ are not involved in the computations.

2.3 The statistical model for the seismic data

We assume that the error term $e(\boldsymbol{x}, t, \theta)$, introduced in the convolutional model in expression (6), is zero mean colored Gaussian noise. The covariance of the error vector $\boldsymbol{e}(\boldsymbol{x}, t) = [e(\boldsymbol{x}, t, \theta_1), \ldots, e(\boldsymbol{x}, t, \theta_{n_\theta})]^T$ is

$$\boldsymbol{\Sigma}_e(\boldsymbol{x}_1, t_1; \boldsymbol{x}_2, t_2) = \boldsymbol{\Sigma}_{0,e}\, \nu_e(\boldsymbol{\xi}, \tau), \tag{23}$$

where $\boldsymbol{\Sigma}_{0,e}$ is an $n_\theta \times n_\theta$ covariance matrix containing the noise variances for the different reflection angles and the correlations between the angles, and $\nu_e(\boldsymbol{\xi}, \tau)$ is a spatial and temporal correlation function. Again we allow sums of terms with the form on the right-hand side of expression (23). Note that white noise is a special case, where $\boldsymbol{\Sigma}_{0,e}$ is diagonal, and $\nu_e(\boldsymbol{\xi}, \tau) = 0$ except for $\nu_e(\boldsymbol{0}, 0) = 1$.

As for the prior model above, the frequency components of the discrete Fourier transformed error vector $\tilde{\boldsymbol{e}}$ are now independent Gaussian

$$\tilde{\boldsymbol{e}}_k \sim \mathcal{N}_{n_\theta}(\boldsymbol{0}, \tilde{\boldsymbol{\Sigma}}_{e,k}). \tag{24}$$

From expressions (12), (21), and (24), each frequency component of the seismic data is then apriori Gaussian

$$\tilde{\boldsymbol{d}}_{obs,k} \sim \mathcal{N}_{n_\theta}(\tilde{\boldsymbol{\mu}}_{d,k}, \tilde{\boldsymbol{\Sigma}}_{d,k}), \tag{25}$$

where

$$\tilde{\boldsymbol{\mu}}_{d,k} = \boldsymbol{G}_k \tilde{\boldsymbol{\mu}}_{m,k}, \tag{26}$$

$$\tilde{\boldsymbol{\Sigma}}_{d,k} = \boldsymbol{G}_k \tilde{\boldsymbol{\Sigma}}_{m,k} \boldsymbol{G}_k^* + \tilde{\boldsymbol{\Sigma}}_{e,k}, \tag{27}$$

and $*$ denotes the conjugate transpose (adjoint).

The cross-covariance between the Fourier transform of seismic data and the model parameters is

$$\mathrm{Cov}\{\tilde{\boldsymbol{d}}_{obs,k}, \tilde{\boldsymbol{m}}_k\} = \boldsymbol{G}_k \tilde{\boldsymbol{\Sigma}}_{m,k}. \tag{28}$$

The cross-covariance is needed to compute the posterior distribution.

2.4 The posterior model

The posterior distribution is defined by a Gaussian conditional distribution. A general presentation of Gaussian and conditional Gaussian distributions can be

found in Anderson (1984). Using expressions (21) and (25)-(28), the posterior distribution for frequency component k is given by the Gaussian conditional distribution

$$\tilde{\boldsymbol{m}}_k|\tilde{\boldsymbol{d}}_{obs,k} \sim \mathcal{N}_3(\tilde{\boldsymbol{\mu}}_{m|d_{obs},k}, \tilde{\boldsymbol{\Sigma}}_{m|d_{obs},k}), \tag{29}$$

where

$$\tilde{\boldsymbol{\mu}}_{m|d_{obs},k} = \tilde{\boldsymbol{\mu}}_{m,k} + (\boldsymbol{G}_k\tilde{\boldsymbol{\Sigma}}_{m,k})^*\tilde{\boldsymbol{\Sigma}}_{d,k}^{-1}(\tilde{\boldsymbol{d}}_{obs,k} - \tilde{\boldsymbol{\mu}}_{d,k}) \tag{30}$$

$$\tilde{\boldsymbol{\Sigma}}_{m|d_{obs},k} = \tilde{\boldsymbol{\Sigma}}_{m,k} - (\boldsymbol{G}_k\tilde{\boldsymbol{\Sigma}}_{m,k})^*\tilde{\boldsymbol{\Sigma}}_{d,k}^{-1}\boldsymbol{G}_k\tilde{\boldsymbol{\Sigma}}_{m,k}. \tag{31}$$

The core part of the inversion is the calculation of the 3 elements in the posterior mean vector in expression (30) and the 3×3 posterior covariance matrix in expression (31) for all frequency components k, that is for all discrete $(\boldsymbol{k}, \omega)$ pairs. The solution is transformed back to the $(\boldsymbol{x}, t)$ domain by 3-D inverse Fourier transforms, 3 for the posterior mean, and 6 for the posterior covariance since the covariance is symmetrical. The posterior distribution of the model parameters is represented by the posterior mean $\boldsymbol{\mu}_{m|d_{obs}}$ and the posterior covariance $\boldsymbol{\Sigma}_{m|d_{obs}}$. The posterior covariance is stationary and homogeneous and hence can be represented by six cubes of size n.

A set of possible solutions can be generated by simulation from the posterior distribution. This can be done efficiently in the Fourier domain: For each frequency component k, draw $\tilde{\boldsymbol{m}}_k$ from the posterior distribution, and then transform $\tilde{\boldsymbol{m}}$ to $\boldsymbol{m}$ by an inverse 3-D Fourier transform for each of the three model parameters in $\boldsymbol{m}$. Since $\boldsymbol{m}$ represents the logarithm of the elastic material parameters, see expression (3), the corresponding set of simulated solutions of the P-wave velocity, S-wave velocity, and density are obtained by the inverse transform $\exp[\boldsymbol{m}]$. Since the posterior distribution for the model parameters can be represented explicitly by the posterior mean and covariance, the inversion results can be merged with a set of well logs to refine the solution around wells. This can be done both for the conditional mean and the conditional simulations using Kriging, see e.g., Cressie (1991).

2.5 The inversion procedure

The inversion procedure can shortly be summarized by the following steps :

1. Define the prior model for the model parameters based on the available knowledge, that is $\boldsymbol{\mu}_m(\boldsymbol{x}, t)$, $\boldsymbol{\Sigma}_{0,m}$, and $\nu_m(\boldsymbol{\xi}, \tau)$, see expressions (14)-(17).
2. Estimate the wavelet $\boldsymbol{s}(t, \theta)$.
3. Estimate the noise covariance, that is $\boldsymbol{\Sigma}_{0,e}$ and $\nu_e(\boldsymbol{\xi}, \tau)$, see expression (23).
4. Calculate the discrete 3-D Fourier transform of $\boldsymbol{\mu}_m(\boldsymbol{x}, t)$, $\nu_m(\boldsymbol{\xi}, \tau)$, $\nu_e(\boldsymbol{\xi}, \tau)$, and the 1-D Fourier transform of $\boldsymbol{s}(t, \theta)$. Sort the seismic data $d_{obs}(\boldsymbol{x}, t, \theta)$

into common angle cubes, and 3-D Fourier transform each of these angle cubes to $\tilde{d}_{obs}(\boldsymbol{k}, \omega, \theta)$.

5. For each frequency component k, calculate the posterior expectation $\tilde{\boldsymbol{\mu}}_{m|d_{obs},k}$ and the posterior covariance $\tilde{\boldsymbol{\Sigma}}_{m|d_{obs},k}$, see expressions (30) and (31).

6. Inverse Fourier transform the solution represented by the posterior mean and covariance.

3 Inversion example of Sleipner data

A rectangular portion of a seismic survey from the Sleipner Øst Field is used in this inversion example. This is the same dataset which was used in Buland and Omre (2001a), where a more detailed presentation of the dataset can be found, including seismic processing, prior model definition, wavelet estimation and estimation of the noise covariance. More on wavelet estimation and the estimation of the noise covariance can be found in Buland and Omre (2001b). The main focus in this paper is on the lateral coupling of the model parameters.

The inversion area is defined from inlines 1411 to 1751, and from crosslines 1225 to 1400, covering 9.3 km^2, or 12% of the total survey. Only each second line is used, such that $n_x = 176$ and $n_y = 171$. The seismic data set is reduced to three angle stacks, $n_\theta = 3$, representing 9°, 21°, and 33°. The thickness of the target area is 250 ms in two-ways traveltime, such that $n_t = 126$ with sampling interval 2 ms. The time window follows an interpretation of the main layering in this target zone. The corresponding number of frequency components are $n_\omega = 318$, $n_{k_x} = 350$, and $n_{k_y} = 340$, that is 38 million frequency components. Compared to the grid size, this increase is caused by extending the problem to a circulant form, see Appendix A. The grid size is not optimal with respect to fast Fourier transform, so the fast radix-2 Fourier transform can not be applied.

The spatial coupling of the model parameters is imposed through the spatial correlation function in expression (16). The lateral correlation of the model parameters is estimated from the seismic data and found to be adequately fitted by a first order exponential correlation function,

$$\nu_{m,\xi}(\boldsymbol{\xi}) = \exp\left[-\frac{\sqrt{\xi_x^2 + \xi_y^2}}{d_r}\right], \tag{32}$$

with lateral range $d_r = 250$ m. The temporal correlation of the model parameters can not be directly estimated from the seismic data since they are blurred by the seismic wavelet. However, a temporal correlation function can be estimated directly from the well logs, here modeled by the composite correlation

function

$$\nu_{m,\tau}(\tau) = \frac{1}{2}\left\{ \exp\left[-\left(\frac{\tau}{d_{t_1}}\right)^2\right] + \left(1 - \frac{2\tau^2}{d_{t_2}^2}\right) \exp\left[-\left(\frac{\tau}{d_{t_2}}\right)^2\right] \right\}, \qquad (33)$$

with temporal range parameters $d_{t_1} = 1.8\,\text{ms}$ and $d_{t_2} = 9\,\text{ms}$, see Buland and Omre (2001a). The complete spatial correlation function $\nu_m(\boldsymbol{\xi}, \tau)$ is the product of the lateral and the temporal correlation functions defined in expressions (32) and (33).

The error term $e(\boldsymbol{x}, t, \theta)$ in the convolutional model, expression (6), includes both seismic noise and errors related to the inversion methodology. We have assumed that the error is zero-mean Gaussian with covariance function on the form given in expression (23). The simplest form of the covariance function is obtained for white noise, that is noise with no spatial correlation. However, in seismic inversion, the most serious noise is usually source generated noise, where remaining multiples are an important example. Such noise components have a smooth waveform similar to the waveform of the primary events. The estimated temporal correlation of this noise can be modeled by a scaled second derivative of a second order exponential correlation function,

$$\nu_{e,\tau}(\tau) = \left(1 - \frac{2\tau^2}{d_t^2}\right) \exp\left[-\left(\frac{\tau}{d_t}\right)^2\right], \qquad (34)$$

where the temporal range is estimated to $d_t = 13\,\text{ms}$. Note that this correlation function can be recognized as a Ricker wavelet with center frequency $f_c = 25\,\text{Hz}$, using the relation $d_t = 1/(\pi f_t)$. Further, we model the lateral correlation of the seismic coherent noise with same correlation function as the model parameters, see expression (32). A first order exponential correlation function is also used to model the correlation between the different reflection angles, with range estimated to $d_\theta = 10°$. The variance of the coherent noise is estimated to 0.1, and the variance of the white noise component is estimated to 0.01. In the frequency domain, the coherent seismic noise colored by the seismic wavelet is band-limited, lacking the lowest and the highest frequencies. In contrast, the white noise distributes equally to all frequencies.

The complete solution of the inversion is represented by the posterior distribution, defined by the posterior mean and covariance, see expressions (30) and (31). In Figure 1, the P-wave velocity, S-wave velocity, and density corresponding to the exponent of the posterior mean of $\boldsymbol{m}$, $\exp[\boldsymbol{\mu}_{m|d_{obs}}]$, are shown for inline 1627. A well is located at crossline 1291, and the well logs are plotted for comparison, showing good agreement with the inversion results. A constant time slice of the P-wave velocity and the S-wave velocity at 2320 ms are shown in Figure 2. The real data of inline 1627, the synthetic data computed from the posterior mean solution, and the corresponding residual are shown in Figures 3, 4, and 5 for 9°, 21°, and 33°. It is important to realize that the objective of

the inversion is not only to minimize the data residual, but to estimate a solution which honor both the seismic data and our prior knowledge. The residual could have been reduced by erroneously altering the error covariance, e.g., by erroneously reducing the variance or the spatial correlation range.

A set of possible solutions for the P-wave velocity, S-wave velocity, and density can be found by drawing a set of vectors $\tilde{\boldsymbol{m}}$ from the posterior distribution, inverse Fourier transform them to $\boldsymbol{m}$, and then calculate $\exp[\boldsymbol{m}]$. One such simulated solution is shown in Figure 6 for inline 1627. The simulated solution differs significantly from the smoother posterior mean solution in Figure 1, but both give a good explaination of the real seismic data.

The prior model specifies the prior values for the variances of $\ln\alpha$, $\ln\beta$, and $\ln\rho$. These values are defined on the diagonal of $\boldsymbol{\Sigma}_{0,m}$, see expression (17). In this example, the prior variances are estimated from the well logs to be

$$\mathrm{Diag}(\boldsymbol{\Sigma}_{0,m}) = 10^{-4} \times [39, 123, 4]. \tag{35}$$

After inversion, the corresponding posterior variances are

$$\mathrm{Diag}(\boldsymbol{\Sigma}_{0,m|d_{obs}}) = 10^{-4} \times [22, 85, 4]. \tag{36}$$

The variance of $\ln\alpha$ has the relatively strongest decrease, followed by the variance of $\ln\beta$. The variance of $\ln\rho$ is hardly changed, that means that the inversion does not provide significant new information which reduces the uncertainty about this parameter.

In Figures 1 and 6, the well logs are plotted for comparison with the inversion result. However, the well logs can be included in a refined solution by Kriging, see e.g. Cressie (1991). This requires specification of a covariance matrix for the well log errors. For simplicity, this covariance matrix is here set to zero, that means that the well logs are defined to be exact. Including the well logs to the sections in Figures 1 and 6 by Kriging gives the corresponding updated solutions shown in Figure 7 and 8. Since the well logs are defined to be exact, the solutions updated by Kriging is equal to the well logs in the well position. The influence of the well logs decreases with increasing distance to the well position, and the uncertainty of the refined solution decreases near the well. The effect of the merge of the well log information with the seimic inversion results is most distinct for the posterior mean density in Figure 7. The reason is that seismic data provides little information about the density, resulting in a smooth posterior mean solution. After the merge with the density log, the solution is far more detailed near the well. Also the simulated solutions in Figure 8 are updated near the well and practically unchanged far from the well. The average of a large number of simulated solutions updated by Kriging will approach the solution in Figure 7.

4 Discussion and conclusions

We have developed an efficient AVO inversion technique where the spatially correlated model parameters are decoupled in the Fourier domain. The seismic data and the model parameters are assumed to be represented on an identical, regularly sampled grid. Further, the covariance functions for the model parameters and the data errors are assumed to be homogeneous and stationary, that is translationally invariant. When the range of the spatial dependency is shorter than the total spatial extension of the grid, the inversion technique is exact with respect to the spatial coupling.

The solution of the inversion problem is represented by a Gaussian posterior distribution with explicit matrix expressions for the posterior mean and covariance. The posterior mean can be interpreted as a smooth best estimate of the solution, while the posterior covariance contains the uncertainty and the correlation structures of the solution. The posterior covariance is homogeneous and stationary, such that the estimated uncertainty of the solution is equal for all positions $(\boldsymbol{x}, t)$. The uncertainty at the boundary of the inversion area is in general underestimated, most severely at the corners. This problem is related to the assumed symmetry in the spatial coupling of the model parameters. At the boundary of the inversion area, this symmetry is lacking. The thickness of the influenced boundary zone depends on the correlation range.

The computing time for the inversion in the Fourier domain follows a linear function of the total number of grid nodes, $\mathcal{O}(n)$, while the computing time for the fast Fourier transform follows an $\mathcal{O}(n \log n)$ function. A 3-D dataset from the Sleipner Field represented by 3 angle stacks on a grid with 4 million grid cells, each with 3 unknown model parameters, used less than 3 minutes on the inversion on a single 400 MHz Mips R12000 CPU. In addition, each Fourier transform used about 30 seconds, but asymptotically the Fourier transforms will dominate the computing time when n approaches infinity. The inversion method is suitable for parallellization, since the inversion problem can be solved independently for each frequency component. Utilizing that the seismic data are band limited, a further speedup can be obtained by inverting only the significant frequencies.

5 Acknowledgments

We thank Statoil for permission to publish this paper.

References

Aki, K., and Richards, P. G., 1980, Quantitative seismology: W.H. Freeman and Co.

Anderson, T. W., 1984, An introduction to multivariate statistical analysis: John Wiley and Sons Inc.

Brockwell, P. J., and Davis, R. A., 1987, Time series: Theory and methods: Springer Verlag.

Brown, R. L., 1992, Estimation of AVO in the presence of noise using prestack migration: 62nd Ann. Internat. Mtg., Soc. Expl. Geophys., Expanded Abstracts, 855–888.

Buland, A., and Landrø, M., 2001, The impact of common offset migration on porosity estimation by AVO inversion: Geophysics, **66**, 755–762.

Buland, A., and Omre, H., 2001a, Bayesian linearized AVO inversion: Submitted for publication in Geophysics, Oct. 2000, http://www.math.ntnu.no/-preprint/statistics/2001.

——— 2001b, Bayesian wavelet estimation from seismic and well data: Submitted for publication in Geophysics, March 2001, http://www.math.ntnu.no/-preprint/statistics/2001.

Buland, A., Landrø, M., Andersen, M., and Dahl, T., 1996, AVO inversion of Troll Field data: Geophysics, **61**, 1589–1602.

Cressie, N., 1991, Statistics for spatial data: John Wiley and Sons Inc.

Duijndam, A. J. W., 1988a, Bayesian estimation in seismic inversion. Part I: Principles: Geophys. Prosp., **36**, 878–898.

——— 1988b, Bayesian estimation in seismic inversion. Part II: Uncertainty analysis: Geophys. Prosp., **36**, 899–918.

Gouveia, W. P., and Scales, J. A., 1998, Bayesian seismic waveform inversion: parameter estimation and uncertainty analysis: J. Geophys. Res., **103**, 2759–2779.

Hampson, D., and Russell, B., 1990, AVO inversion: theory and practice: 60th Ann. Internat. Mtg., Soc. Expl. Geophys., Expanded Abstracts, 1456–1458.

Lörtzer, G. J. M., and Berkhout, A. J., 1993, Linearized AVO inversion of multicomponent seismic data, *in* Castagna, J., and Backus, M., Eds., Offset-dependent reflectivity - theory and practice of AVO analysis: Soc. Expl. Geophys., 317–332.

Mosher, C. C., Keho, T. H., Weglein, A. B., and Foster, D. J., 1996, The impact of migration on AVO: Geophysics, **61**, 1603–1615.

Omre, H., Sølna, K., and Tjelmeland, H., 1993, Simulation of random functions on large lattices, *in* Soares, A., Ed., Geostatistics Tróia '92: Kluwer Academic Publisher, 179–199.

Pan, G. S., Young, C. Y., and Castagna, J. P., 1994, An integrated target-oriented prestack elastic waveform inversion: Sensitivity, calibration, and application: Geophysics, **59**, 1392–1404.

Rue, H., 2000, Fast sampling of Gaussian Markov random fields with applications: Submitted for publication, http://www.math.ntnu.no/preprint/-statistics/2000.

Stolt, R. H., and Weglein, A. B., 1985, Migration and inversion of seismic data: Geophysics, **50**, 2458–2472.

Tarantola, A., and Valette, B., 1982, Inverse problems = quest for information: J. Geophys., **50**, 159–170.

Tarantola, A., 1987, Inverse problem theory: Elsevier Science Publ. Co. Inc.

Wood, A. T. A., and Chan, G., 1994, Simulation of stationary Gaussian processes in $[0,1]^d$: Jour. of Computational and Graphical Statistics, **3**, 409–432.

Wood, A. T. A., 1995, When is a truncated covariance function on the line a covariance function on the circle?: Statistics & probability letters, **24**, 157–164.

A Diagonalization of a covariance matrix by DFT

In the following, the relationship between the discrete Fourier transform (DFT) and the eigen-values and eigen-vectors of a circulant matrix is presented. Further, it is shown how this can be used to diagonalize a homogeneous covariance function sampled on a regular grid. For simplicity, the presentation is limited to 1-D, but the extension to higher dimensions is straightforward. More details on these topics can be found in Brockwell and Davis (1987); Wood and Chan (1994); Wood (1995).

A.1 The DFT

The 1-D discrete Fourier transform (DFT) of the sequence $f(k),\ k = 0, \ldots, n-1$, can be written

$$\tilde{f}(l) = \sum_{k=0}^{n-1} f(k) \exp\left[-2\pi i \frac{kl}{n}\right], \quad l = 0, \ldots, n-1, \tag{37}$$

with inverse transform (IDFT)

$$f(k) = \frac{1}{n} \sum_{l=0}^{n-1} \tilde{f}(l) \exp\left[2\pi i \frac{kl}{n}\right], \quad k = 0, \ldots, n-1. \tag{38}$$

The DFT can alternatively be written as a matrix-vector product

$$\tilde{\boldsymbol{f}} = \boldsymbol{F}\boldsymbol{f}, \tag{39}$$

where $\boldsymbol{f} = [f(0), \ldots, f(n-1)]^T$, and

$$\boldsymbol{F} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & w^1 & \cdots & w^{n-1} \\ \vdots & \vdots & & \vdots \\ 1 & w^{n-1} & \cdots & w^{(n-1)^2} \end{bmatrix}, \tag{40}$$

where $w = \exp[-2\pi i/n]$. The matrix corresponding to the IDFT is $\boldsymbol{F}^{-1} = n^{-1}\boldsymbol{F}^*$, where $*$ denotes the conjugate transpose. If the dimension n is a power of 2, the fast radix-2 Fourier transform (FFT) can be used.

A.2 Circulant matrices

An $n \times n$ matrix $\boldsymbol{M}$ is a circulant matrix if the elements m_{kl} are defined by a function $m(\cdot)$ with period n such that $m_{kl} = m_{l-k} = m(l-k)$, that is

$$\boldsymbol{M} = \begin{bmatrix} m_0 & m_1 & \cdots & m_{n-1} \\ m_{n-1} & m_0 & \cdots & m_{n-2} \\ \vdots & \vdots & & \vdots \\ m_1 & m_2 & \cdots & m_0 \end{bmatrix}, \tag{41}$$

see Brockwell and Davis (1987). Note that a circulant matrix is Toeplitz, but the opposite is generally not true. The eigenvalues of a circulant matrix $\boldsymbol{M}$ are

$$\lambda_l = \sum_{k=0}^{n-1} m(k) \exp\left[-2\pi i \frac{kl}{n}\right], \quad l = 0, \ldots, n-1, \tag{42}$$

with orthonormal eigenvectors

$$\boldsymbol{v}_l = n^{-1/2} \begin{bmatrix} 1 \\ w^l \\ \vdots \\ w^{(n-1)l} \end{bmatrix}. \tag{43}$$

The circulant matrix $\boldsymbol{M}$ can be diagonalized by

$$\boldsymbol{V}\boldsymbol{M}\boldsymbol{V}^* = \boldsymbol{\Lambda}_M, \tag{44}$$

where $\boldsymbol{\Lambda}_M = \mathrm{diag}\{\lambda_0, \lambda_1, \ldots, \lambda_{n-1}\}$, and $\boldsymbol{V}$ is the unitary eigenvector matrix

$$\boldsymbol{V} = [\boldsymbol{v}_0, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_{n-1}]. \tag{45}$$

For the following, it is important to recognize that the eigenvalues of a circulant matrix $\boldsymbol{M}$ are equal to the DFT of the first row, and that $\boldsymbol{F} = n^{1/2}\boldsymbol{V}$.

A.3 Diagonalization of a circulant covariance matrix

Let $r(x)$ be a zero mean Gaussian variable with homogeneous covariance function

$$\Sigma(x_1, x_2) = \sigma^2 \nu(\xi), \tag{46}$$

where $\xi = |x_2 - x_1|$. Let $\boldsymbol{r}$ be a discrete representation of $r(x)$ sampled on a regular grid, $x_k = k\Delta x$, where $k = 0, \ldots, n_x - 1$. The corresponding $n_x \times n_x$ covariance matrix is then symmetric Toeplitz,

$$\boldsymbol{\Sigma} = \sigma^2 \begin{bmatrix} \nu_0 & \nu_1 & \cdots & \nu_{n_x-1} \\ \nu_1 & \nu_0 & \cdots & \nu_{n_x-2} \\ \vdots & \vdots & & \vdots \\ \nu_{n_x-1} & \nu_{n_x-2} & \cdots & \nu_0 \end{bmatrix}, \tag{47}$$

where $\nu_k = \nu(k\Delta x)$. This covariance matrix is not circulant, but it can be embedded in a symmetric circulant $n \times n$ matrix

$$\boldsymbol{\Sigma}_c = \sigma^2 \begin{bmatrix} \nu_0 & \nu_1 & \cdots & \nu_{n/2} & \cdots & \nu_1 \\ \nu_1 & \nu_0 & \cdots & \nu_{n/2-1} & \cdots & \nu_2 \\ \vdots & \vdots & & \vdots & & \vdots \\ & & & \vdots & & \\ \nu_{n/2} & \nu_{n/2-1} & \cdots & & & \nu_{n/2-1} \\ \vdots & \vdots & & \vdots & & \vdots \\ \nu_1 & \nu_2 & \cdots & \nu_{n/2-1} & \cdots & \nu_0 \end{bmatrix}, \tag{48}$$

where $n \geq 2(n_x - 1)$, and such that the top left $n_x \times n_x$ sub matrix of $\boldsymbol{\Sigma}_c$ is equal to $\boldsymbol{\Sigma}$. The circulant matrix $\boldsymbol{\Sigma}_c$ is a legal covariance matrix if and only if it is positive definite. A sufficient, but not neccessary condition for positive definiteness is that $\nu_k = 0$ for all $k > k_0$, where $k_0 < n_x$, see Wood (1995). Strictly, this excludes many of the most common correlation functions, for example exponential correlation functions with order $(1, 2]$. However, the range of a correlation function will often be much shorter than the total spatial dimension, such that $\nu_k \approx 0$ for all $k > k_0$. In such cases a truncation of the correlation function may be adequate.

Let now $\boldsymbol{r}_c$ be an extension of $\boldsymbol{r}$ with dimension n and covariance matrix $\boldsymbol{\Sigma}_c$. While $\boldsymbol{r}$ is sampled on a line, $\boldsymbol{r}_c$ can be interpreted to be sampled on a circle. Then the Fourier transform of $\boldsymbol{r}_c$, $\tilde{\boldsymbol{r}}_c = \boldsymbol{F}\boldsymbol{r}_c$, has a diagonal covariance matrix

$$\tilde{\boldsymbol{\Sigma}}_c = \boldsymbol{F}\boldsymbol{\Sigma}_c\boldsymbol{F}^* = n\boldsymbol{V}\boldsymbol{\Sigma}_c\boldsymbol{V}^{-1} = n\boldsymbol{\Lambda}_\Sigma = \tilde{\boldsymbol{\Lambda}}_\Sigma, \tag{49}$$

where $\boldsymbol{\Lambda}_\Sigma$ is the eigenvalue matrix of $\boldsymbol{\Sigma}_c$ with real nonnegative eigenvalues. This means that the correlated variables in $\boldsymbol{r}$ are transformed to independent variables in the Fourier domain. From above, we know that $\boldsymbol{\Lambda}_\Sigma$ can simply be

calculated by a DFT of the first row of $\mathbf{\Sigma}_c$. This means that it is not necessary to compute the matrix products $\boldsymbol{F}\mathbf{\Sigma}_c\boldsymbol{F}^*$. In fact, the complete matrix $\mathbf{\Sigma}_c$ is not involved in the computations.

The extension to 2-D and 3-D problems is straightforward. Let $\boldsymbol{r}$ be a discrete representation of a zero mean Gaussian variable with homogeneous covariance function, sampled on a regular 2-D or 3-D grid. The corresponding covariance matrices are block Toeplitz in 2-D and nested block Toeplitz in 3-D, and they can be embedded in block or nested block circulant matrices. Similarly to the 1-D case, the n eigenvalues can be found by a 2-D or 3-D DFT of a circulant discrete representation of the correlation function.

B Figures

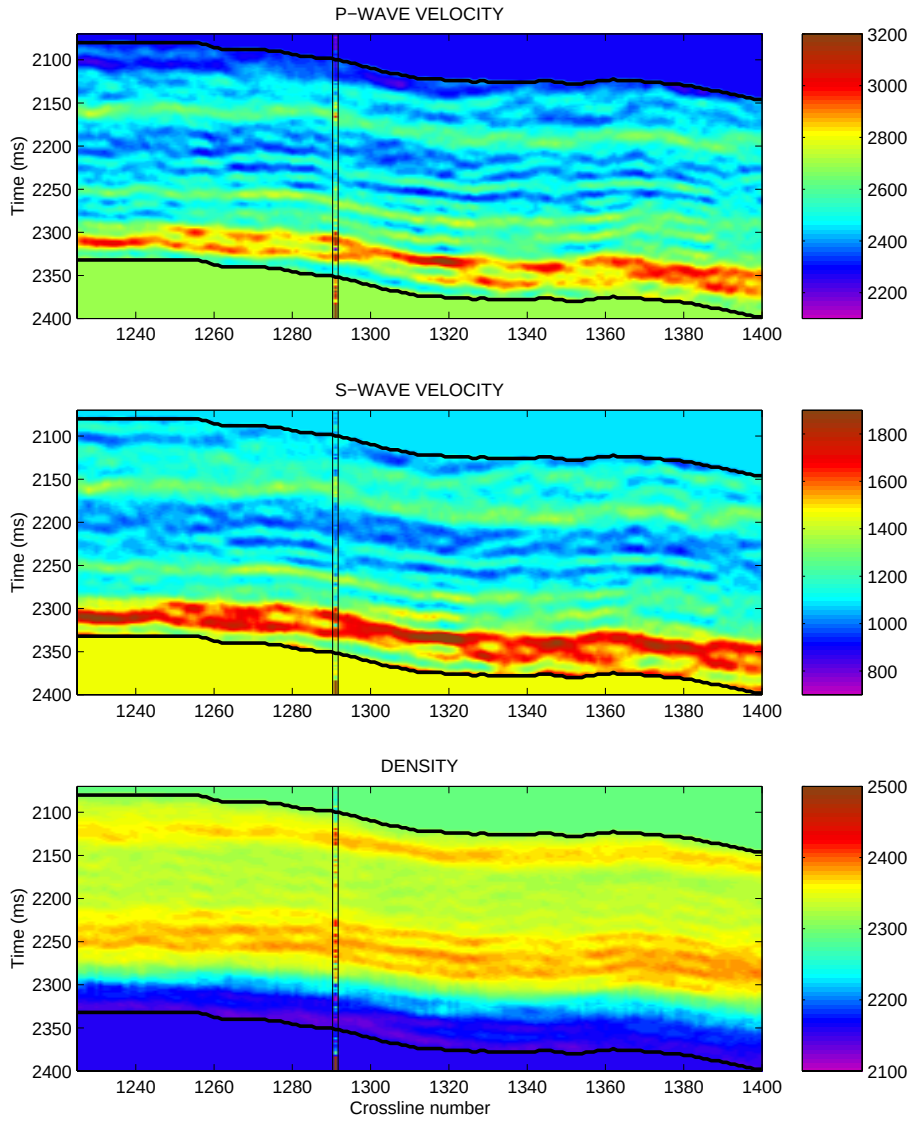


Figure 1: P-wave velocity (top), S-wave velocity (middle), and density (bottom) corresponding to the posterior mean.

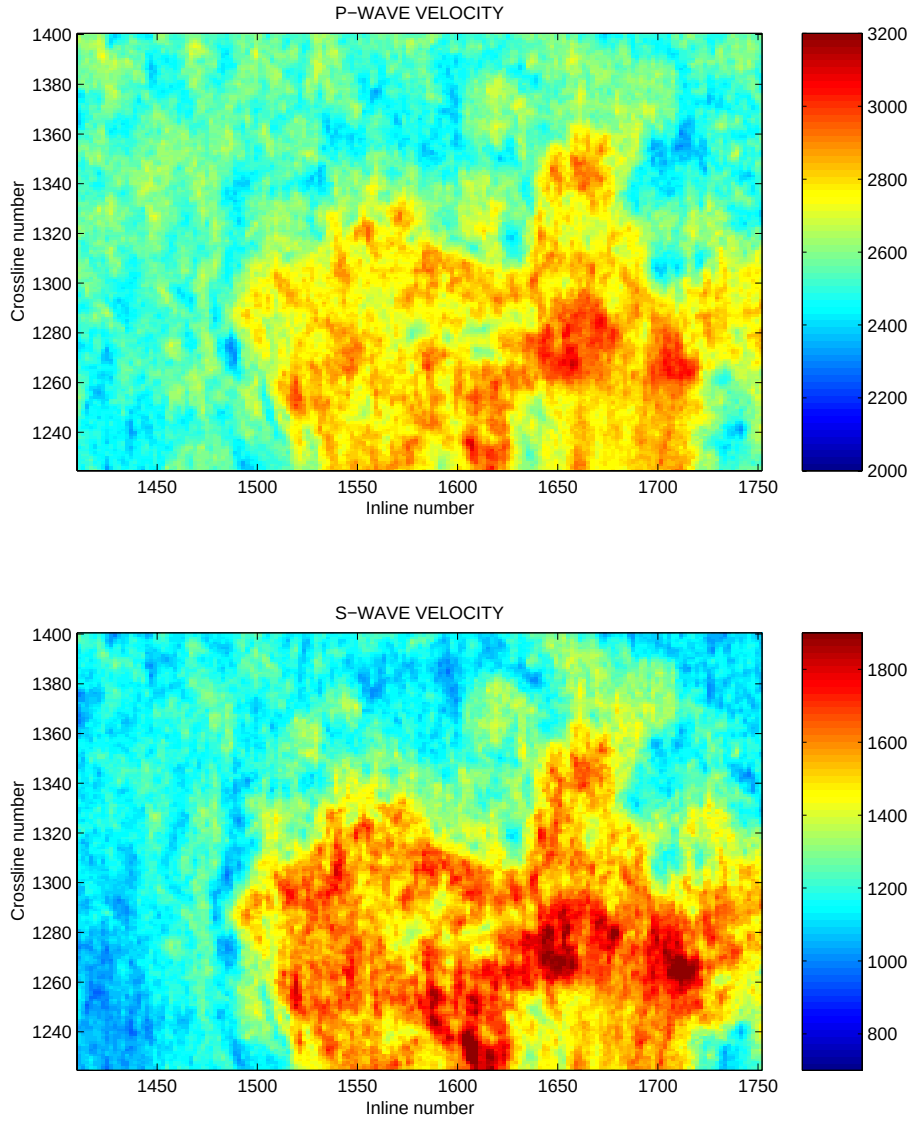


Figure 2: Time slice of the P-wave velocity (top) and the S-wave velocity (bottom) at 2320 ms.

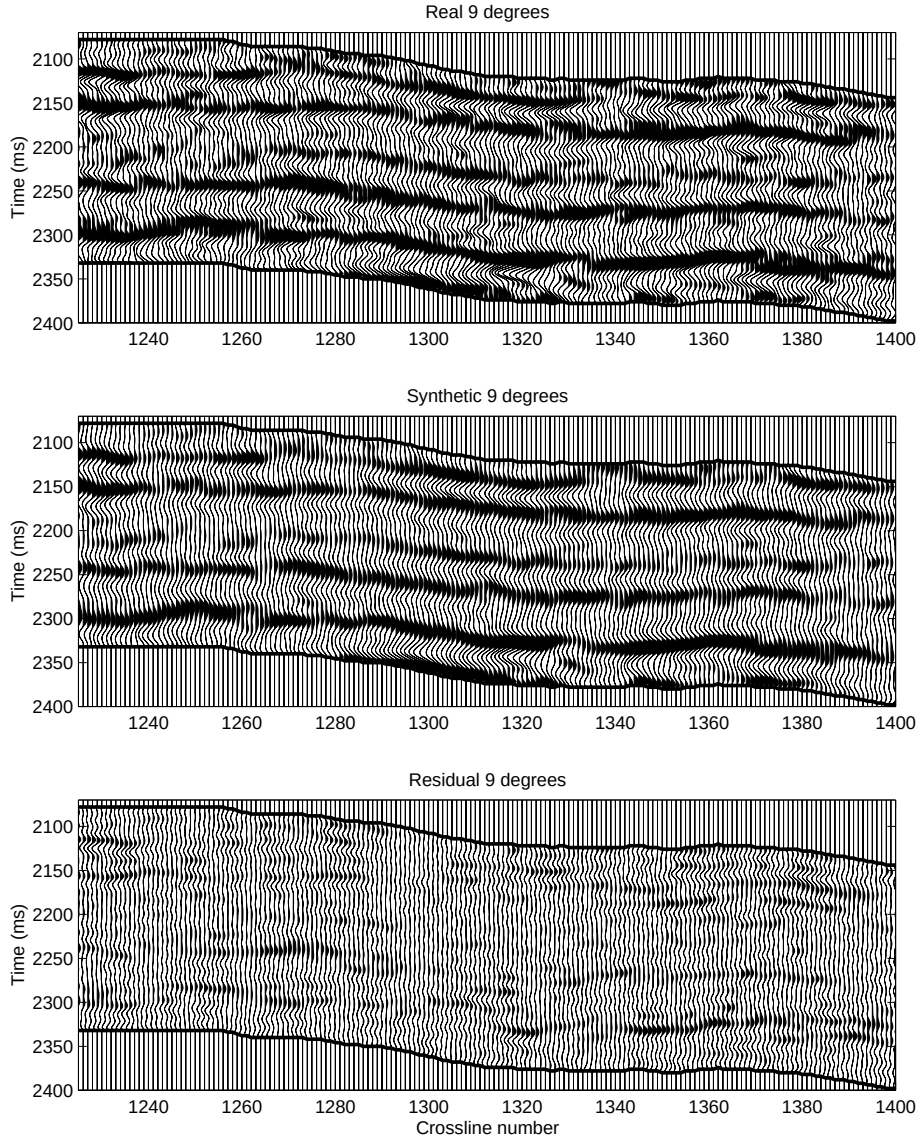


Figure 3: Real data (top), synthetic (middle), and residual (bottom) for 9°.

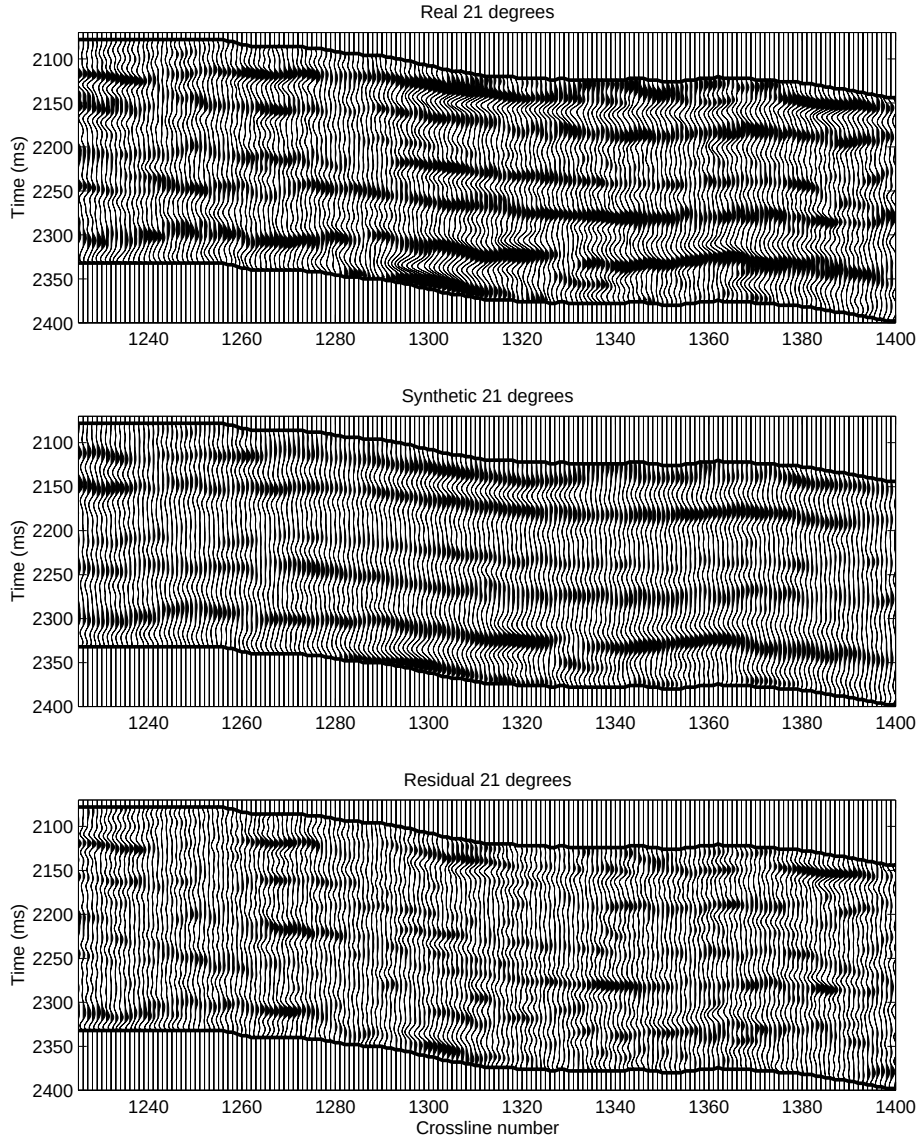


Figure 4: Real data (top), synthetic (middle), and residual (bottom) for 21°.

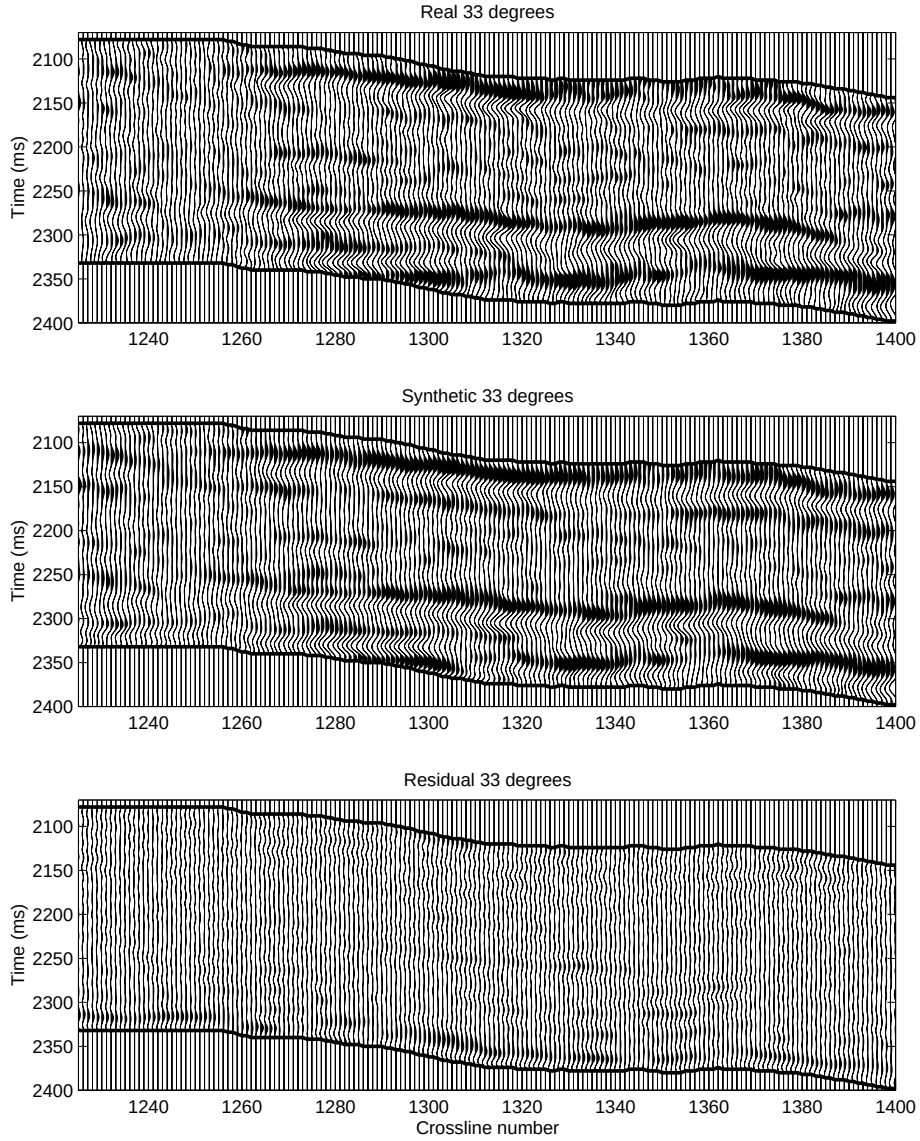


Figure 5: Real data (top), synthetic (middle), and residual (bottom) for 33°.

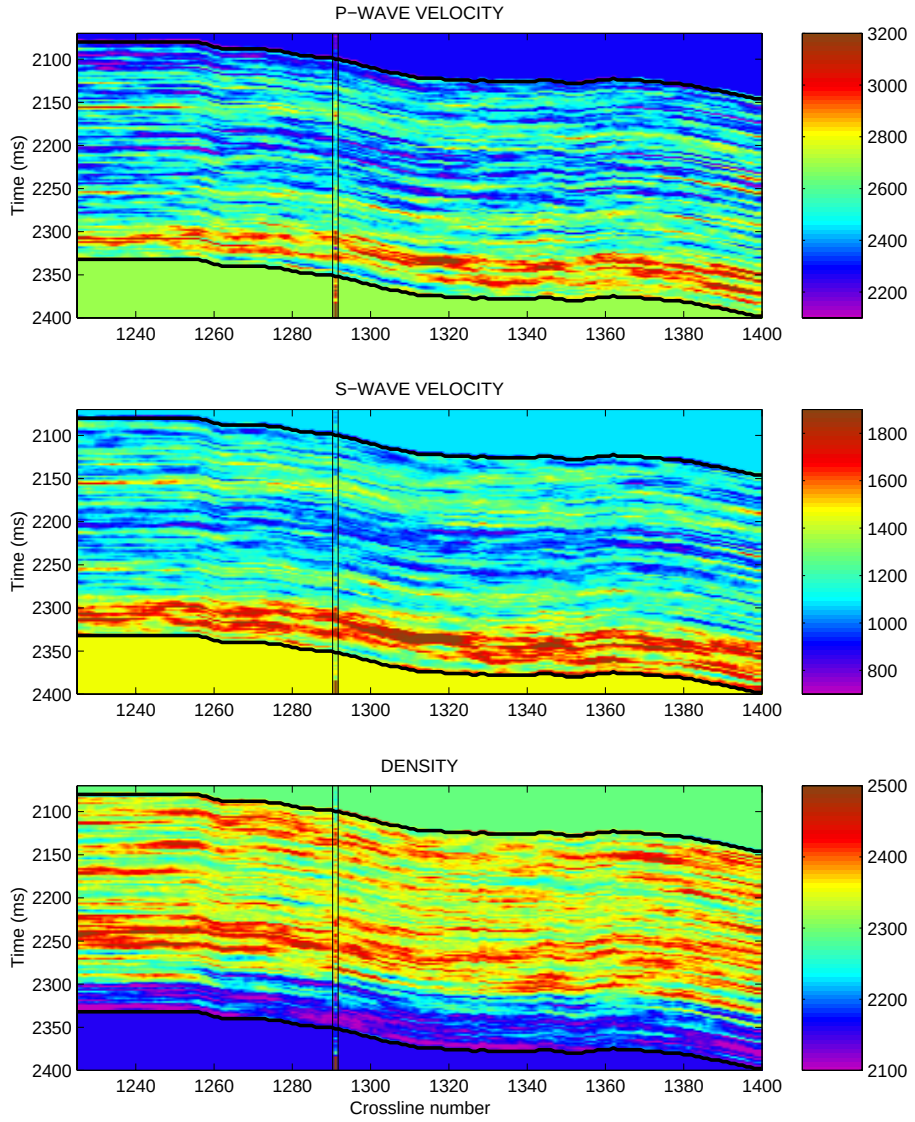


Figure 6: Simulated P-wave velocity (top), S-wave velocity (middle), and density (bottom).

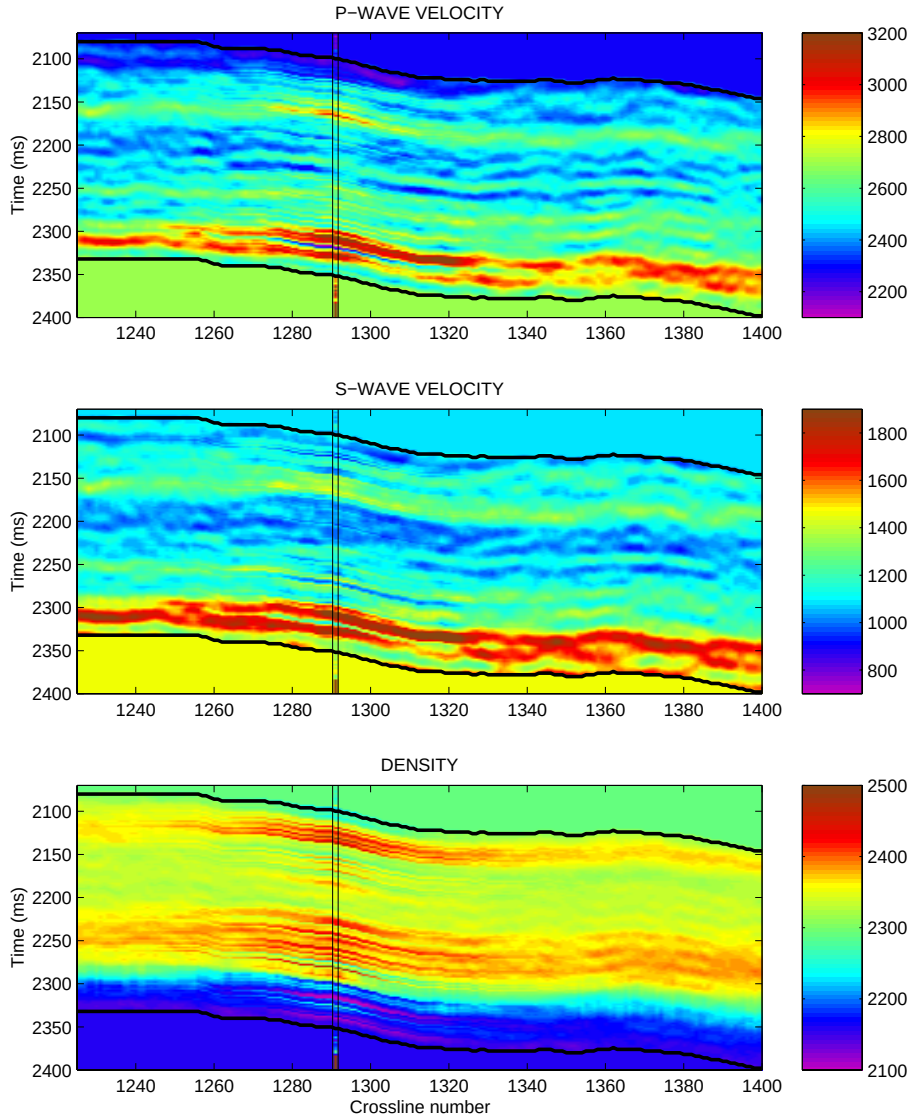


Figure 7: The posterior mean solution conditioned to the well logs.

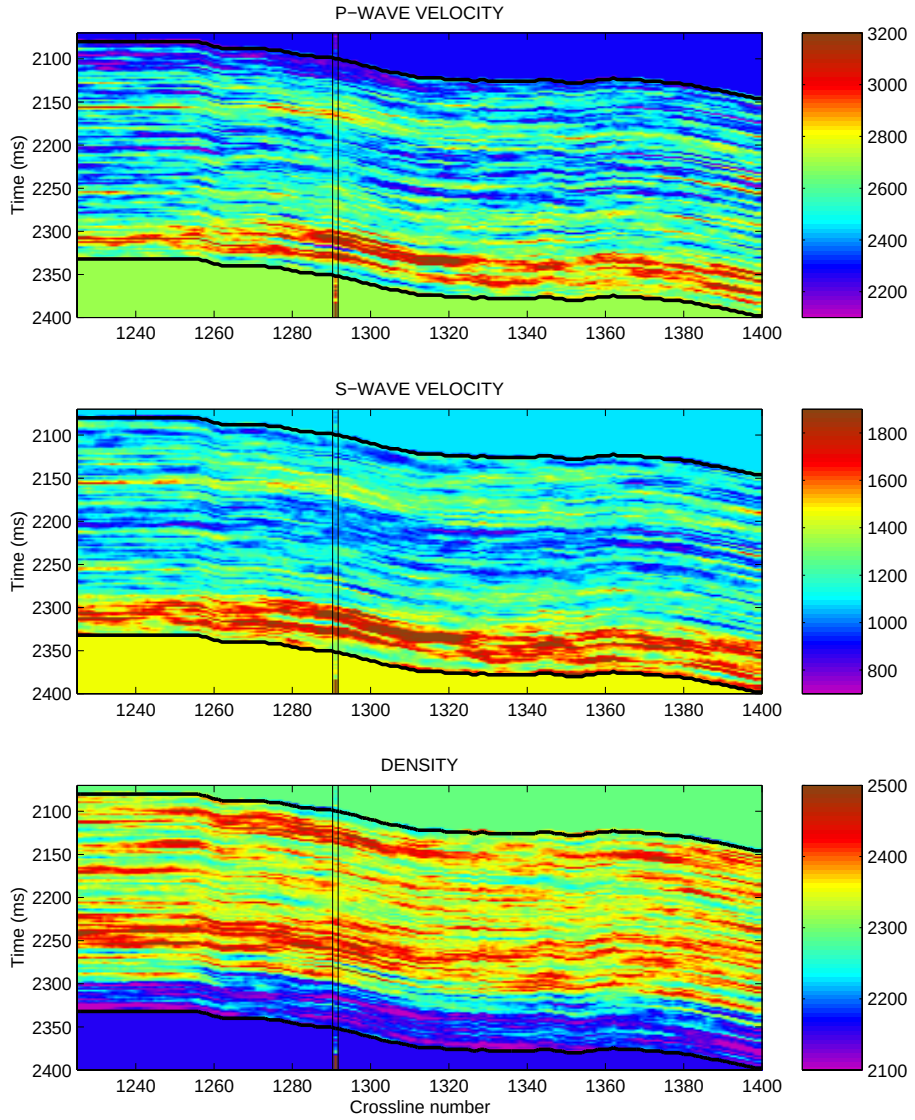


Figure 8: The simulated solution conditioned to the well logs.