

Non-linear least-squares inversion with data-driven Bayesian regularization

Tor Erik Rabben* and Bjørn Ursin

Department of Petroleum Engineering and Applied Geophysics, Norwegian University of Science and Technology, 7491 Trondheim, Norway

* *E-mail: torerik@ntnu.no*

SUMMARY

The non-linear inverse problem is formulated in a Bayesian framework. The multivariate normal distribution is assumed in both the noise and prior distributions. However, only the structures of the covariance matrices have to be specified, estimation of the variance levels is included in the inversion procedure. The maximum a posteriori approximation is derived, and the final result is a weighted least-squares inversion algorithm with the ratio between the variance levels as an adaptive, data-driven regularization factor, hence the name Bayesian regularization. The algorithm is tested on inversion of seismic reflection amplitudes and compared with the L-curve approach for choosing the regularization parameter. The Bayesian regularization results in a better regularization value in only a fraction of the time.

Key words: Inverse theory – Probability distributions – Acoustic properties

1 INTRODUCTION

Inverse problems, in geophysics and other disciplines, are often ill-posed and non-unique. Ill-posed means that small changes in the measurements can cause large changes in the solution. In a situation where the measured data is contaminated with noise this would be a

serious problem. The non-uniqueness problem can arise if the forward model is too simple or if the coverage of the measurements is insufficient. In any of these situations regularization is necessary. Regularizing the inverse problem means finding a physically meaningful stable solution (Tenorio, 2001).

Historically, the most famous inversion method is least squares (also known as regression analysis), see e.g. Tarantola (1987) or Lines & Treitel (1984), and the goal is to find the set of parameters that minimizes the square of the misfit. The straight-forward solution to solve it is to find the gradient and set it equal to zero. This results in the normal equations. In statistical nomenclature this method is equal to assuming a Gaussian distribution of the error and maximizing the likelihood (ML estimate). In case of a linear forward model the normal equations will have a simple, analytic form. However, if the forward model is non-linear the least-squares problem is solved by Taylor expanding the gradient and creating an iterative solution algorithm which (hopefully) converges to a global minimum.

A very common regularization technique for least-squares problems is Tikhonov regularization (Tikhonov & Arsenin, 1977; Tenorio, 2001) (also known as ridge regression) which seeks to minimize both the residual and a property of the solution alone. Typical properties of the solution can be its derivative to ensure smoothness or the distance from an a priori expected solution. Statistically, this can be seen as allowing a bias in order to reduce the variance in the solution (Golub et al., 1979). The equivalence to Tikhonov regularization in the statistical nomenclature is to maximize the posterior distribution (MAP estimate) when assuming Gaussian error and prior distributions. The challenge in the Tikhonov regularization is the trade-off between minimizing the residuals or the parameter norms – a trade-off between trusting measurements or a priori information. Several methods exist and the trade-off parameters is commonly denoted λ^2 , a notation which will be adapted here.

An intuitive solution is to parametrically display the residual and parameter norms as a function of different λ^2 -values. Typically this will give a figure with the shape of an L, and therefore the name L-curve (Lawson & Hanson, 1974; Hansen, 1992). The “perfect” λ^2 is in the corner point where the curvature is largest since this corresponds to a good

trade-off between minimization of the two norms. The singular value decomposition (SVD) (Golub & Van Loan, 1996) is a way to form a generalized inverse operator to solve the linear inverse problem. This procedure can easily be extended to include regularization by either truncating some singular values (Hansen, 1987) or applying filter factors (Lines & Treitel, 1984; Aster et al., 2004). The latter can be shown to be mathematically equivalent to solving the normal equations (Lines & Treitel, 1984). In both situations the basic concept is to remove amplification of eigenvectors corresponding to singular values below the noise level. Other approaches to discriminate between measurements and prior information are the generalized cross-validation (Golub et al., 1979), the chi-square test (Snedecor & Cochran, 1989; Aster et al., 2004) or the maximum entropy condition (Aster et al., 2004), coming from different statistical consideration.

A common feature for all the methods mentioned here, to find the trade-off parameter λ^2 , is that they are computer intensive in case of a large problem sizes. They all require the inversion performed for a range of possible values before finding the “best” value. A better solution would be to have the measurements itself determine λ^2 in an automatic, robust and effective procedure.

In this paper we present a method which is adaptive, data-driven, based on a sound theoretical background, and more efficient than its competitors. We start by defining the Bayesian model where we follow closely the work of Rabben et al. (2008). The noise and the prior are multivariate Gaussian distributions with given structure of the covariance matrices. The scaling factors of these covariance matrices are included as stochastic parameters in the inversion procedure (Buland & Omre, 2003). Based on this model we iteratively compute the MAP estimate of the model parameters. The trade-off parameter λ^2 is the ratio between the two scale parameters of the covariance matrices. It is updated in each iteration, and the new value depends on the data misfit and the model misfit from the prior. The final result is a least-squares algorithm with adaptive and data-driven regularization.

Hansen and O’Leary (1993) did a comparison of several of the methods mentioned above, and concluded that L-curve was more robust than the main competitor, generalized cross-

validation. We will therefore use this method as a comparison for the new inversion algorithm. Our numerical example will be inversion of seismic reflection amplitudes, both only PP inversion and joint PP and PS inversion, following the work of Rabben et al. (2008).

2 BAYESIAN MODEL

In order to formulate our inversion procedure we start by introducing the non-linear forward problem

$$\mathbf{d} = \mathbf{f}(\mathbf{m}) + \mathbf{e}. \quad (1)$$

Here, \mathbf{d} is a vector containing a set of measurements, \mathbf{m} is a vector of model parameters, \mathbf{f} is a non-linear function, and \mathbf{e} is a general noise term. The inversion problem is, given \mathbf{d} and \mathbf{f} , to find \mathbf{m} . The corresponding Bayesian formulation of the inverse problem is, via Bayes' rule, given by

$$\pi(\mathbf{m}|\mathbf{d}) \propto \pi(\mathbf{d}|\mathbf{m}) \pi(\mathbf{m}), \quad (2)$$

where $\pi(\cdot)$ represents any probability distribution. In Bayesian inversion we are not only searching for one optimal \mathbf{m} , but the full statistical distribution of \mathbf{m} given the measured data \mathbf{d} . This is the left hand side of (2) and is known as the posterior distribution. It is proportional to the product of $\pi(\mathbf{d}|\mathbf{m})$ (the likelihood model, analog to the forward model) and $\pi(\mathbf{m})$ (the prior model representation). To find the posterior model we need to assign probability distributions to the noise \mathbf{e} and to the model parameters \mathbf{m} . Common and convenient choices are the multivariate normal distribution (defined in Appendix A)

$$\pi(\mathbf{e}) = \mathcal{N}(\mathbf{e}; 0, \Sigma_e) \quad (3)$$

$$\pi(\mathbf{m}) = \mathcal{N}(\mathbf{m}; \boldsymbol{\mu}_m, \Sigma_m). \quad (4)$$

Under the assumption of a known, deterministic forward model \mathbf{f} we easily find the likelihood to be $\pi(\mathbf{d}|\mathbf{m}) = \mathcal{N}(\mathbf{d}; \mathbf{f}(\mathbf{m}), \Sigma_e)$ using (1) and (3). In this formulation the choice of covariance matrices is very important since it greatly will influence the optimal solution. Our approach

is to define the covariance matrices as (Buland & Omre, 2003)

$$\Sigma_e = \sigma_e^2 \mathbf{S}_e \tag{5}$$

$$\Sigma_m = \sigma_m^2 \mathbf{S}_m, \tag{6}$$

where \mathbf{S}_e and \mathbf{S}_m are known correlation matrices, and then include the estimation of the scalar random variables σ_e^2 and σ_m^2 as a part of the inversion procedure. As a consequence, we need prior distributions on σ_e^2 and σ_m^2 , and we choose the inverse gamma distributions

$$\pi(\sigma_e^2) = \mathcal{IG}(\sigma_e^2; \alpha_e, \beta_e) \tag{7}$$

$$\pi(\sigma_m^2) = \mathcal{IG}(\sigma_m^2; \alpha_m, \beta_m), \tag{8}$$

where $\alpha_e, \beta_e, \alpha_m$ and β_m (known as hyperparameters (Robert, 2007)) are scalar constants, see Appendix A for further definitions. The inverse gamma distribution is flexible and defined for positive values and can thereby be adapted to varying prior knowledge. Moreover, it makes the mathematical treatment of the resulting posterior easier. It is a convenient (although maybe not optimal) choice as the normal distribution is for the medium parameters and the measurement noise. With these two new model parameters, our Bayesian inverse problem now becomes to estimate the joint posterior distribution

$$\pi(\mathbf{m}, \sigma_e^2, \sigma_m^2 | \mathbf{d}) \propto \pi(\mathbf{d} | \mathbf{m}, \sigma_e^2, \sigma_m^2) \pi(\mathbf{m}, \sigma_e^2, \sigma_m^2). \tag{9}$$

This equation will constitute the modelling basis for our inversion algorithm.

3 MAXIMUM A POSTERIORI SOLUTION

There are several ways to assess posterior information. Rabben et al. (2008) showed how to sample the posterior (9) and hence assessing the full posterior distribution through a MCMC Metropolis-Hastings sampling approach. However, this is complicated and computationally expensive since the forward model is non-linear. We will therefore search for only the most likely solution,

$$\arg \max_{\mathbf{m}, \sigma_e^2, \sigma_m^2} \pi(\mathbf{m}, \sigma_e^2, \sigma_m^2 | \mathbf{d}), \tag{10}$$

also known as the maximum a posteriori (MAP) solution or posterior mode. Further, instead of trying to assess the joint posterior $\pi(\mathbf{m}, \sigma_e^2, \sigma_m^2 | \mathbf{d})$ we will update each parameter sequentially (known as Gibbs steps) in an iterative algorithm. We therefore need the three posterior expressions (one parameter conditioned on the two others)

$$\pi(\mathbf{m} | \mathbf{d}, \sigma_e^2, \sigma_m^2) \propto \pi(\mathbf{d} | \mathbf{m}, \sigma_e^2) \pi(\mathbf{m} | \sigma_m^2) \quad (11)$$

$$\pi(\sigma_e^2 | \mathbf{d}, \mathbf{m}) \propto \pi(\mathbf{d} | \mathbf{m}, \sigma_e^2) \pi(\sigma_e^2) \quad (12)$$

$$\pi(\sigma_m^2 | \mathbf{m}) \propto \pi(\mathbf{m} | \sigma_m^2) \pi(\sigma_m^2) \quad (13)$$

or

$$\pi(\mathbf{m} | \mathbf{d}, \sigma_e^2, \sigma_m^2) \propto \mathcal{N}(\mathbf{d}; \mathbf{f}(\mathbf{m}), \sigma_e^2 \mathbf{S}_e) \mathcal{N}(\mathbf{m}; \boldsymbol{\mu}_m, \sigma_m^2 \mathbf{S}_m) \quad (14)$$

$$\pi(\sigma_e^2 | \mathbf{d}, \mathbf{m}) \propto \mathcal{N}(\mathbf{d}; \mathbf{f}(\mathbf{m}), \sigma_e^2 \mathbf{S}_e) \mathcal{IG}(\sigma_e^2; \alpha_e, \beta_e) \quad (15)$$

$$\pi(\sigma_m^2 | \mathbf{m}) \propto \mathcal{N}(\mathbf{m}; \boldsymbol{\mu}_m, \sigma_m^2 \mathbf{S}_m) \mathcal{IG}(\sigma_m^2; \alpha_m, \beta_m). \quad (16)$$

For the posterior (14) we write

$$\begin{aligned} \pi(\mathbf{m} | \mathbf{d}, \sigma_e^2, \sigma_m^2) \propto & \exp \left\{ -\frac{1}{2\sigma_e^2} [\mathbf{d} - \mathbf{f}(\mathbf{m})]^T \mathbf{S}_e^{-1} [\mathbf{d} - \mathbf{f}(\mathbf{m})] \right\} \\ & \times \exp \left\{ -\frac{1}{2\sigma_m^2} [\mathbf{m} - \boldsymbol{\mu}_m]^T \mathbf{S}_m^{-1} [\mathbf{m} - \boldsymbol{\mu}_m] \right\}. \end{aligned} \quad (17)$$

Maximizing the posterior is equal to minimizing the expression

$$\psi = \|\mathbf{d} - \mathbf{f}(\mathbf{m})\|_{\mathbf{S}_e^{-1}}^2 + \frac{\sigma_e^2}{\sigma_m^2} \|\mathbf{m} - \boldsymbol{\mu}_m\|_{\mathbf{S}_m^{-1}}^2, \quad (18)$$

which is a non-linear weighted least-squares problem. The minimum is reached when the gradient of ψ is zero, and by expanding it in a Taylor series we find the iterative solution algorithm

$$\begin{aligned} \mathbf{m}_{k+1} &= \mathbf{m}_k - \mathbf{H}_k^{-1} \mathbf{g}_k \\ &= \mathbf{m}_k - (\mathbf{J}^T \mathbf{S}_e^{-1} \mathbf{J} + \lambda^2 \mathbf{S}_m^{-1})^{-1} (\lambda^2 \mathbf{S}_m^{-1} \Delta \mathbf{m}_\mu - \mathbf{J}^T \mathbf{S}_e^{-1} \Delta \mathbf{d}), \end{aligned} \quad (19)$$

where $\mathbf{J} = \partial \mathbf{f} / \partial \mathbf{m}^T |_{\mathbf{m}_k}$, $\Delta \mathbf{d} = \mathbf{d} - \mathbf{f}(\mathbf{m}_k)$, $\Delta \mathbf{m}_\mu = \mathbf{m}_k - \boldsymbol{\mu}_m$ and $\lambda^2 = \sigma_e^2 / \sigma_m^2$. By omitting the first term in the gradient and assuming $\mathbf{S}_e = \mathbf{S}_m = \mathbf{I}$ it reduces to the famous Levenberg-Marquardt algorithm (Levenberg, 1944; Marquardt, 1963).

The posterior (15) and (16) can, by using the definition of the normal and inverse gamma

distribution (Appendix A), be written

$$\pi(\sigma_e^2 | \mathbf{d}, \mathbf{m}) \propto \mathcal{IG} \left(\sigma_e^2; \alpha_e + \frac{1}{2}n_e, \beta_e + \frac{1}{2} \|\Delta \mathbf{d}\|_{\mathbf{S}_e^{-1}}^2 \right) \quad (20)$$

$$\pi(\sigma_m^2 | \mathbf{m}) \propto \mathcal{IG} \left(\sigma_m^2; \alpha_m + \frac{1}{2}n_m, \beta_m + \frac{1}{2} \|\Delta \mathbf{m}_\mu\|_{\mathbf{S}_m^{-1}}^2 \right), \quad (21)$$

where n_e and n_m are the number of elements in \mathbf{d} and \mathbf{m} . In other words, the posterior distributions of σ_e^2 and σ_m^2 are still inverse gamma distributed, only with modified parameters.

The MAP solution of an inverse gamma distribution $\mathcal{IG}(\sigma^2; \alpha', \beta')$ is $\sigma^2 = \beta' / (1 + \alpha')$ (see (A5)), and with this result we find the MAP of (15) and (16) to be

$$\sigma_e^2 = \frac{\beta_e + \frac{1}{2} \|\Delta \mathbf{d}\|_{\mathbf{S}_e^{-1}}^2}{1 + \alpha_e + \frac{1}{2}n_e} \quad (22)$$

$$\sigma_m^2 = \frac{\beta_m + \frac{1}{2} \|\Delta \mathbf{m}_\mu\|_{\mathbf{S}_m^{-1}}^2}{1 + \alpha_m + \frac{1}{2}n_m}. \quad (23)$$

To speed up the algorithm we will perform only one update of \mathbf{m} using (19) before updating the MAP of σ_e^2 and σ_m^2 . The final expression for λ^2 reads

$$\lambda_{k+1}^2 = \frac{\sigma_{e,k+1}^2}{\sigma_{m,k+1}^2} = \frac{\beta_e + \frac{1}{2} \|\Delta \mathbf{d}\|_{\mathbf{S}_e^{-1}}^2}{\beta_m + \frac{1}{2} \|\Delta \mathbf{m}_\mu\|_{\mathbf{S}_m^{-1}}^2} \cdot \frac{1 + \alpha_m + \frac{1}{2}n_m}{1 + \alpha_e + \frac{1}{2}n_e}. \quad (24)$$

Eqs (19) and (24) constitute our inversion algorithm. In addition to the parameters needed in the weighted least-squares algorithm, we also have to assign the hyperparameters α 's and β 's in the two inverse gamma distributions. Since σ_e^2 and σ_m^2 appears relative to each other, it will be convenient to scale the correlation matrices \mathbf{S}_e and \mathbf{S}_m in (5) and (6) such that the determinant of both equals one.

When assigning values to the hyperparameters we can start by looking at eqs (7) and (8) which are distributions of the a priori knowledge about σ_e^2 and σ_m^2 . The safe choice would be a ‘‘reasonable’’ mean and a large variance such that the data itself can decide. However, it is not obvious how to choose this just by looking at eqs (A3) and (A4). We will therefor use (24) to guide us. Firstly, we will assume the dimensionality to be large such that $n_m \gg 1 + \alpha_m$ and $n_d \gg 1 + \alpha_e$ and set $\alpha_m = \alpha_e = 0$, resulting in

$$\lambda_{k+1}^2 = \frac{n_m}{n_e} \frac{\beta_e + \frac{1}{2} \|\Delta \mathbf{d}\|_{\mathbf{S}_e^{-1}}^2}{\beta_m + \frac{1}{2} \|\Delta \mathbf{m}_\mu\|_{\mathbf{S}_m^{-1}}^2} \quad (25)$$

From here, if we assume both β 's to be zero it becomes

$$\lambda_{k+1}^2 = \frac{n_m}{n_e} \frac{\|\Delta \mathbf{d}\|_{\mathbf{S}_e^{-1}}^2}{\|\Delta \mathbf{m}_\mu\|_{\mathbf{S}_m^{-1}}^2}, \quad (26)$$

which is a purely data-driven damping. However, one should be careful here since both of the terms $\|\Delta \mathbf{d}\|_{\mathbf{S}_e^{-1}}^2$ and $\|\Delta \mathbf{m}_\mu\|_{\mathbf{S}_m^{-1}}^2$ can be zero. If either one of the two norms approaches zero, it will lead to λ^2 approaching the undesirable value of zero or infinite. In this setting, the β 's act as stabilizing terms by defining a priori lower bounds on the two norms.

Lastly, the total opposite choice for the β 's is $\beta_e \gg \|\Delta \mathbf{d}\|_{\mathbf{S}_e^{-1}}^2$ and $\beta_m \gg \|\Delta \mathbf{m}_\mu\|_{\mathbf{S}_m^{-1}}^2$ which leads to

$$\lambda_{k+1}^2 \approx \frac{n_m}{n_e} \frac{\beta_e}{\beta_m}, \quad (27)$$

a solution driven totally by the hyperparameters. Needless to say, this choice has to be avoided in order for our method to work.

4 NUMERICAL EXAMPLE: INVERSION OF SEISMIC REFLECTION AMPLITUDES

To demonstrate our algorithm we apply it to the numerical example used in Rabben et al. (2008). It is a synthetic problem where the true model contains large contrasts, and are chosen to enhance non-linear effects. Our method for comparison is the L-curve which is a logarithmic plot ($\|\mathbf{d} - \mathbf{f}(\mathbf{m})\|_{\mathbf{S}_e^{-1}}^2, \|\mathbf{m} - \boldsymbol{\mu}_m\|_{\mathbf{S}_m^{-1}}^2$) for a range of regularization values λ^2 . The point on the graph which has the largest curvature, the corner point, is considered the ‘‘correct’’ regularization value.

The forward model is isotropic and the medium parameters \mathbf{m} are defined over a 100×100 lattice with the 3 elastic moduli in each point, resulting in $n_m = 3 \times 10^4$. For the measurements we have two cases: only PP reflections and both PP and PS reflections. In both cases the exact Zoeppritz equations are used to generate the reflection amplitudes before correlated noise is added, this noise is generated using the correct correlation matrix in (3). For PP inversion we have 4 angles, and for joint PP and PS a total 7 angles, giving $n_e = 4 \times 10^4$ and $n_e = 7 \times 10^4$ respectively.

When inverting we use only the quadratic approximations to the Zoeppritz equation (Stovas & Ursin, 2003; Rabben et al., 2008) as our forward model \mathbf{f} . The a priori expected mean, $\boldsymbol{\mu}_m$, is half of the true model. For the correlation matrices \mathbf{S}_e and \mathbf{S}_m we have included spatial correlations, but no correlations within medium parameters or reflection angles.

In the first case, inversion of only PP reflection amplitudes, the obvious first choice for β_e and β_m is zero for both, eq. (26). However, this solution converges to the prior because of our initial λ_0^2 . This can easily be fixed by assigning a relative small value to β_m . We have chosen $\beta_m = 5$, and in Fig. A1 our Bayesian regularization is plotted together with the L-curve. The black, dotted line in the figure is the value of β_m . We see that the classical L-shape is not present, this is because of the spatial correlation which prevents the solution from growing very large. However, our solution converges to a very plausible value, with a final value $\lambda^2 = 2.4 \times 10^{-3}$. In Fig. A2 we see a zoom of the left end of the L-curve.

Although $(\beta_e = 0, \beta_m = 5)$ is our final choice for the inversion, we will also see how two other choices will influence the solution. In a situation where we have less confidence in the measurements we will assign a non-zero value to β_e also. In Fig. A3 we have used the values $(\beta_e = 1, \beta_m = 5)$, and we see that the algorithm converges to a solution within the rectangle defined by β_e and β_m . Based on this we can interpret β_e and β_m as a priori noise-level estimates. However, in this case the algorithm is still strongly data-driven. If we compare with the situation in Fig. A4 $(\beta_e = 10, \beta_m = 100)$, we see that there the algorithm converges in a very few iterations, and the final λ^2 -value will be very close to the estimate from eq. (27). From purely geometrical considerations we can conclude that if the point (β_e, β_m) is above or close to the L-curve, our algorithm will not be data-driven but controlled mainly by prior information.

We will now continue with the solution from Fig. A1. Since we know the truth we can access the true bias, this is displayed in Fig. A5, plotted together with true bias in the L-curve calculations. Here we see that the optimal regularization value is $\lambda^2 \approx 10^{-2}$, somewhat higher than our algorithm but far from the L-curve solution.

The convergence of our algorithm is also interesting. Fig. A6 shows the update $\|\mathbf{m}_k -$

\mathbf{m}_{k-1} and the damping factor λ^2 in each iteration. The medium parameters \mathbf{m} does not converge until the damping factor has. We also see that our algorithm converges after 17 iterations. As a comparison, the L-curve in Fig. A1 which consists of 49 different values of λ^2 , adds up to a total of 215 iterations. In Table A1 we have summarized this together with the total number of Conjugate Gradients (Saad, 2003) iterations in order to solve the matrix inversion in (19). We clearly see that our method is superior also when it comes to computational time, not only to find the regularization parameter. To conclude the PP example we compare the bias in our MAP estimate with the bias in the mean (Rabben et al., 2008) in Fig. A7. For the contrast in P-wave impedance the results are the same, for S-wave impedance the mean is slightly better, while for contrasts in density our result is best. In other words, the two methods yield very similar results.

Our second example, joint inversion of PP and PS reflections, yields very similar results to the previous example. This time we plot both our method and the L-curve together in Figs A8 and A9, and again we see that the L-curve breaks down while our Bayesian regularization converges to a very plausible value. In Fig. A10 we use the known truth to calculate the true bias in the two approaches. Also this time the damping factor is slightly too low, but far better than the L-curve. However, if we compare with Fig. A5 we see that it performs better than for PP inversion. The bias in the inversion is also lower in the case of a too low damping factor. This is due to the forward model, joint inversion is less ill-posed than PP inversion.

When it comes to convergence we see that the only major difference in Fig. A11 compared to Fig. A6 is the number of iteration. In our two examples joint inversion converges in 13 iterations while PP inversion requires 17, again due to a not so ill-posed problem. Table A2 confirms this by looking at both number of least-squares and conjugate gradients iterations, and comparing with Table A1. More important, it also shows how computationally superior our algorithm is to the L-curve approach. To conclude our numerical examples we compare the bias in our MAP estimate with the bias in the mean (Rabben et al., 2008) in Fig.

A12. Again we see that the two methods yield very similar results with contrasts in density impedance slightly better and the opposite for the S-wave impedance.

5 DISCUSSION

We have used the Bayesian model from Rabben et al. (2008), but have to distinctly different approaches on how to assess posterior information. While they used a sampling methodology, we have used an optimization strategy. The advantage with sampling of the posterior is the ability to quantify uncertainties together with expected value and most likely solution, compared to our method which only finds the most likely solution. In this setting, their method is definitely preferred. However, it does not come for free. The sampling method is very computer intensive. In the example used, our method converges in minutes while the sampling algorithm may require more than one day to generate the large number of samples needed.

Our method has its strength for non-linear problems. Since we have to alternate between updating the medium parameters \mathbf{m} and the dampening factor λ^2 until convergence, the method requires more than one iteration - even for linear problems. Therefore, for linear problems which converges in one iteration, the workload of computing the L-curve would be about the same as our Bayesian regularization approach when the number of λ^2 values in the L-curve equals the number of iterations needed in our method. However, for non-linear problems each point on the L-curve will be as costly as our method.

An other important point to make is the use of covariance matrices \mathbf{S}_e and \mathbf{S}_m . With a wrong assumption here, the misfit function (18) will be sensitive to changes in only a subset of the measurements and model parameters. This will clearly be a violation of our assumptions, and our method (specially in the case of eq. (26)) may not perform as expected. As a minimum, one should include correlation matrices with relative variances on the diagonal.

6 CONCLUSION

We have formulated the non-linear inverse problem in a Bayesian setting where we, in addition to the medium parameters, have included estimation of the prior and noise level variance. Searching for the most likely solution, the maximum a posteriori solution, results in a weighted non-linear least-squares algorithm with an adaptive, data-driven regularization factor. The origin from a Bayesian formulation is reflected in the term Bayesian regularization.

To test the algorithm we have compared it with the L-curve approach by applying both approaches to the problem non-linear inversion of reflection amplitudes. From the L-curve we see that our method converges to a very plausible regularization factor. We have also demonstrated how different strategies for choosing the hyperparameters will influence the inversion result. The parameters β_e and β_m are a priori lower bounds of the data and prior misfit. In a situation where the point $(\beta_e = 0, \beta_m = 5)$ lies far below the L-curve, the algorithm will be purely data-driven, while when the point is close to or above the curve, the solution will be more and more constrained by the hyperparameters. For non-linear problems the new method is computationally more efficient than the L-curve approach.

7 ACKNOWLEDGMENTS

We wish to thank BP, Schlumberger, StatoilHydro and The Research Council of Norway for their support through the Uncertainty in Reservoir Evaluation (URE) project. Bjørn Ursin thanks StatoilHydro via the VISTA project for financial support. We thank Colin G. Farquharson for constructive comments in the review process.

REFERENCES

- Aster, R., Borchers, B., & Thurber, C., 2004. *Parameter Estimation and Inverse Problems*, Elsevier Academic Press, Burlington.
- Buland, A. & Omre, H., 2003. Joint AVO inversion, wavelet estimation and noise-level estimation using a spatially coupled hierarchical Bayesian model, *Geophysical Prospecting*, **51**, 531–550.

- Golub, G. H. & Van Loan, C. F., 1996. *Matrix Computations*, The Johns Hopkins University Press, Baltimore.
- Golub, G. H., Heath, M., & Wahba, G., 1979. Generalized cross-validation as a method for choosing a good ridge parameter, *Technometrics*, **21**, 215–223.
- Hansen, P. C., 1987. The truncated SVD as a method for regularization, *BIT*, **27**, 534–553.
- Hansen, P. C., 1992. Analysis of discrete ill-posed problems by means of the L-curve, *SIAM Review*, **34**, 561–580.
- Hansen, P. C. & O’Leary, D. P., 1993. The use of the L-curve in the regularization of discrete ill-posed problems, *SIAM Journal on Scientific Computing*, **14**, 1487–1503.
- Lawson, C. L. & Hanson, R. J., 1974. *Solving Least Squares Problems*, Prentice-Hall, Englewood Cliffs.
- Levenberg, K., 1944. A method for the solution of certain non-linear problems in least squares, *Quarterly of Applied Mathematics*, **2**, 164–168.
- Lines, L. R. & Treitel, S., 1984. A review of least-squares inversion and its application to geophysical problems, *Geophysical Prospecting*, **32**, 159–186.
- Marquardt, D. W., 1963. An algorithm for least-squares estimation of nonlinear parameters, *SIAM Journal on Applied Mathematics*, **11**, 431–441.
- Rabben, T. E., Tjelmeland, H., & Ursin, B., 2008. Non-linear Bayesian joint inversion of seismic reflection coefficients, *Geophysical Journal International*, Accepted, available at OnlineEarly.
- Robert, C., 2007. *The Bayesian Choice*, Springer, New York.
- Saad, Y., 2003. *Iterative Methods for Sparse Linear Systems*, SIAM, Philadelphia.
- Snedecor, G. W. & Cochran, W. G., 1989. *Statistical Methods*, Iowa State University Press, Ames.
- Stovas, A. & Ursin, B., 2003. Reflection and transmission responses of layered transversely isotropic viscoelastic media, *Geophysical Prospecting*, **51**, 447–477.
- Tarantola, A., 1987. *Inverse Problem Theory*, Elsevier, Amsterdam.
- Tenorio, L., 2001. Statistical regularization of inverse problems, *SIAM Review*, **43**, 347–366.
- Tikhonov, A. N. & Arsenin, V. Y., 1977. *Solutions of Ill-Posed Problems*, John Wiley & Sons, New York.

APPENDIX A: STATISTICAL DISTRIBUTIONS

A multivariate Gaussian variable \mathbf{x} with expectation vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ has the probability function

$$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{n/2} |\boldsymbol{\Sigma}|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\}, \quad (\text{A1})$$

where n is the dimension of \mathbf{x} .

The inverse gamma probability function is

$$\mathcal{IG}(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\frac{1}{x}\right)^{\alpha+1} e^{-\beta/x} \quad (\text{A2})$$

where $x \geq 0, \alpha > 0$ and $\beta > 0$. Its mean, variance and mode are

$$\text{E } \mathcal{IG}(x; \alpha, \beta) = \frac{\beta}{\alpha - 1} \quad (\text{A3})$$

$$\text{Var } \mathcal{IG}(x; \alpha, \beta) = \frac{\beta^2}{(\alpha - 1)^2(\alpha - 2)} \quad (\text{A4})$$

$$\max_x \mathcal{IG}(x; \alpha, \beta) = \frac{\beta}{\alpha + 1}. \quad (\text{A5})$$

Given the prior distribution $\sigma^2 \sim \mathcal{IG}(\alpha, \beta)$ and measurements $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 \mathbf{S})$, the posterior distribution of σ^2 is

$$\begin{aligned} \pi(\sigma^2 | \mathbf{x}) &\propto \pi(\mathbf{x} | \sigma^2) \pi(\sigma^2) \\ &= \mathcal{N}(\boldsymbol{\mu}, \sigma^2 \mathbf{S}) \mathcal{IG}(\alpha, \beta) \\ &\propto \frac{1}{(\sigma^2)^{n/2} |\mathbf{S}|^{1/2}} \exp \left\{ -\frac{1}{2} \sigma^{-2} (\mathbf{x} - \boldsymbol{\mu})^T \mathbf{S}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\} \\ &\quad \times \left(\frac{1}{\sigma^2}\right)^{\alpha+1} \exp \left\{ -\frac{\beta}{\sigma^2} \right\} \\ &\propto \left(\frac{1}{\sigma^2}\right)^{\alpha+1+n/2} \exp \left\{ -\sigma^{-2} \left(\beta + s^2 \frac{n}{2} \right) \right\} \\ &\propto \mathcal{IG} \left(\sigma^2 \middle| \alpha + \frac{n}{2}, \beta + s^2 \frac{n}{2} \right) \end{aligned} \quad (\text{A6})$$

where

$$s^2 = \frac{1}{n} (\mathbf{x} - \boldsymbol{\mu})^T \mathbf{S}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \quad (\text{A7})$$

and n is the dimension of \mathbf{x} . Clearly, the posterior is also inverse gamma but with modified parameters.

This paper has been produced using the Blackwell Publishing GJI L^AT_EX₂e class file.

LIST OF FIGURES

- A1 Bayesian regularization for PP inversion and comparison with the L-curve. The red x-marks show how our algorithm converges for $(\beta_e = 0, \beta_m = 5)$, and the thick x-marks is the final iteration, with $\lambda^2 = 2.4 \times 10^{-3}$. The L-curve is made using λ^2 -values from $10^{-4.5}$ to $10^{1.5}$ with constant logarithmic increments. The smallest values of λ^2 is to the left on the curve. The black, dotted line in the figure is the value of β_m , and the black square indicates the zoomed area in Fig. A2.
- A2 Zoom of the black square in Fig. A1.
- A3 Bayesian regularization for PP inversion and comparison with the L-curve. The red x-marks show how our algorithm converges for $(\beta_e = 1, \beta_m = 5)$, and the thick x-marks are the final iteration. The black circle and dotted lines are the values of β_e and β_m .
- A4 Bayesian regularization for PP inversion and comparison with the L-curve. The red x-marks show how our algorithm converges for $(\beta_e = 10, \beta_m = 100)$, and the thick x-marks are the final iteration. The black circle and dotted lines are the values of β_e and β_m .
- A5 True bias in the Bayesian regularization and in the L-curve calculations, both for PP inversion.
- A6 Convergence of the Bayesian regularization algorithm for inversion of PP reflection amplitudes. The black, dotted line is the convergence criterion for the update.
- A7 Absolute value of bias in the medium parameters \mathbf{m} for PP inversion. The left column is our MAP estimate while the right column (Mean) is reproduced from Rabben et al. (2008). Some high values in the contrast in S-wave impedance are clipped in order to visually enhance the differences.
- A8 Bayesian regularization for joint PP and PS inversion and comparison with the L-curve. For the L-curve we have again used λ^2 values from $10^{-4.5}$ to $10^{1.5}$. The red x-marks show how our algorithm converges for $(\beta_e = 0, \beta_m = 5)$, and the thick x-mark to the left is the final iteration, with $\lambda^2 = 7.1 \times 10^{-3}$. The black, dotted line in the figure is the value of β_m , and the black square indicates the zoomed area in Fig. A9.
- A9 Zoom of the black square in Fig. A8.
- A10 True bias in the Bayesian regularization and in the L-curve calculations, both for joint PP and PS inversion.
- A11 Convergence of the Bayesian regularization algorithm for inversion of joint PP and PS reflection amplitudes. The black, dotted line is the convergence criterion for the update.
- A12 Absolute value of bias in the medium parameters \mathbf{m} for joint PP and PS inversion. The left column is our MAP estimate while the right column (Mean) is reproduced from Rabben et al. (2008). A few high values in the contrast in S-wave impedance are clipped in order to visually enhance the differences.

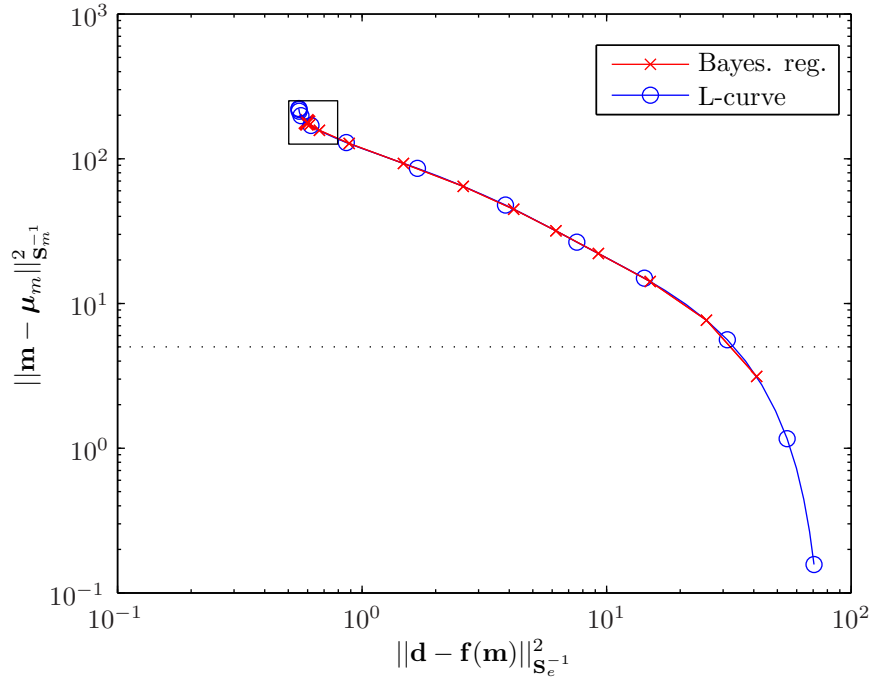


Figure A1. Bayesian regularization for PP inversion and comparison with the L-curve. The red x-marks show how our algorithm converges for $(\beta_e = 0, \beta_m = 5)$, and the thick x-mark is the final iteration, with $\lambda^2 = 2.4 \times 10^{-3}$. The L-curve is made using λ^2 -values from $10^{-4.5}$ to $10^{1.5}$ with constant logarithmic increments. The smallest values of λ^2 is to the left on the curve. The black, dotted line in the figure is the value of β_m , and the black square indicates the zoomed area in Fig. A2.

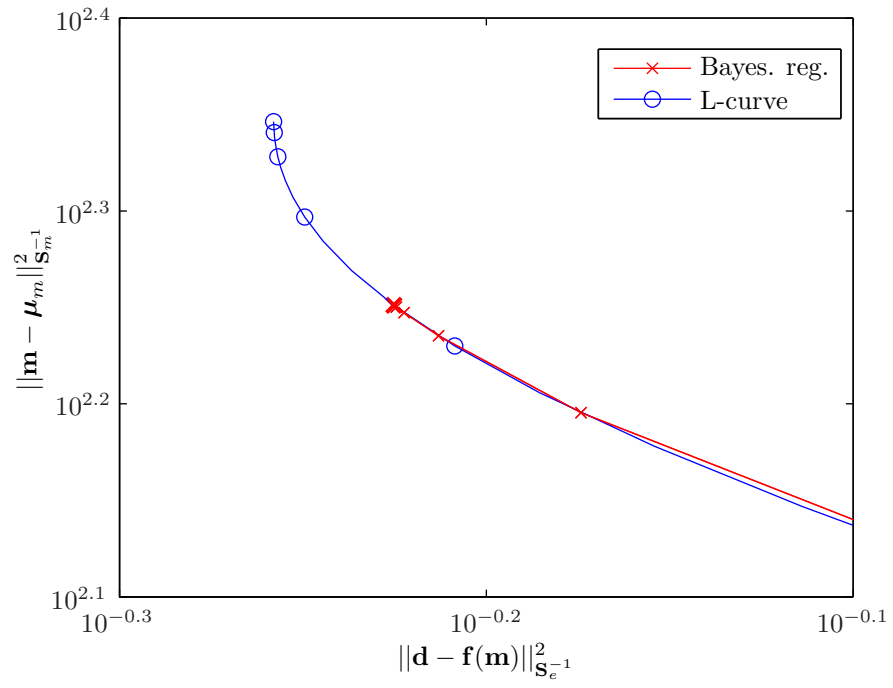


Figure A2. Zoom of the black square in Fig. A1.

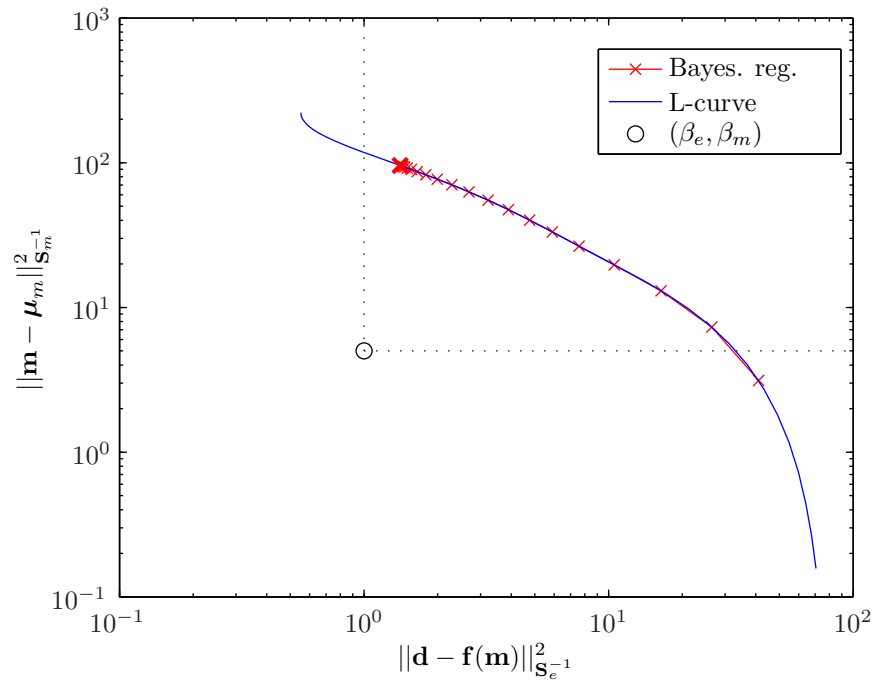


Figure A3. Bayesian regularization for PP inversion and comparison with the L-curve. The red x-marks show how our algorithm converges for $(\beta_e = 1, \beta_m = 5)$, and the thick x-marks are the final iteration. The black circle and dotted lines are the values of β_e and β_m .

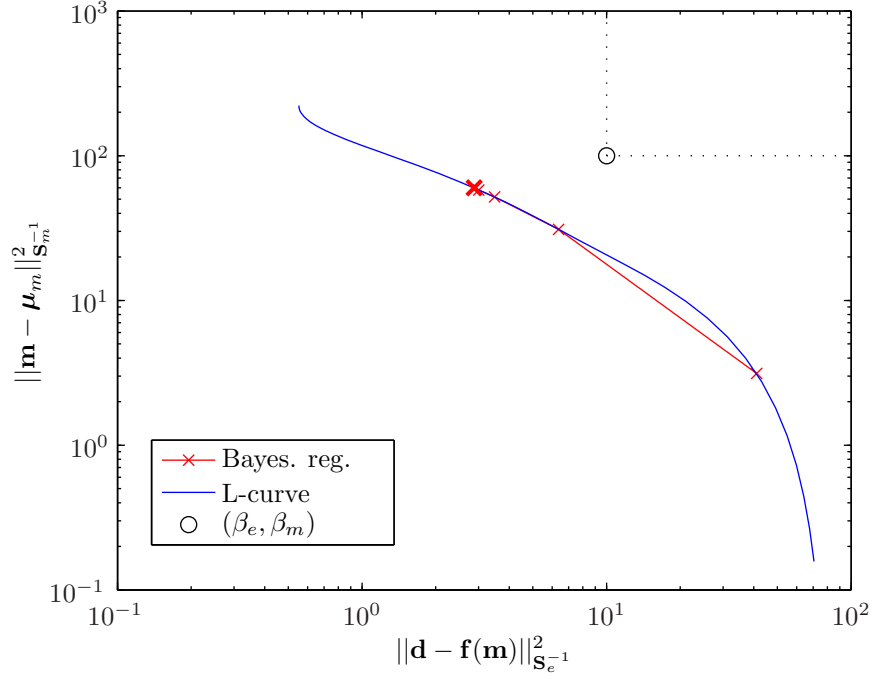


Figure A4. Bayesian regularization for PP inversion and comparison with the L-curve. The red x-marks show how our algorithm converges for $(\beta_e = 10, \beta_m = 100)$, and the thick x-marks are the final iteration. The black circle and dotted lines are the values of β_e and β_m .

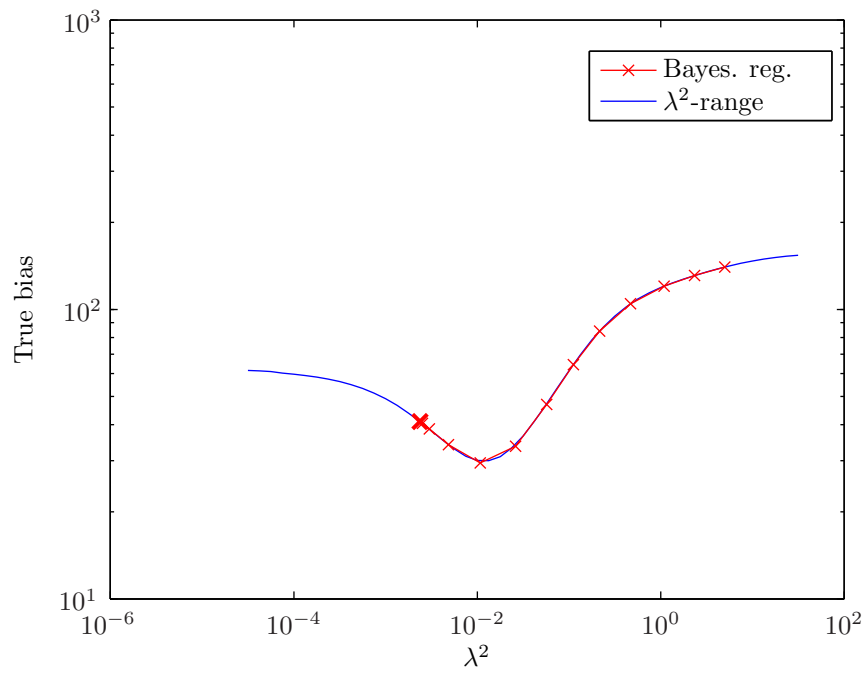


Figure A5. True bias in the Bayesian regularization and in the L-curve calculations, both for PP inversion.

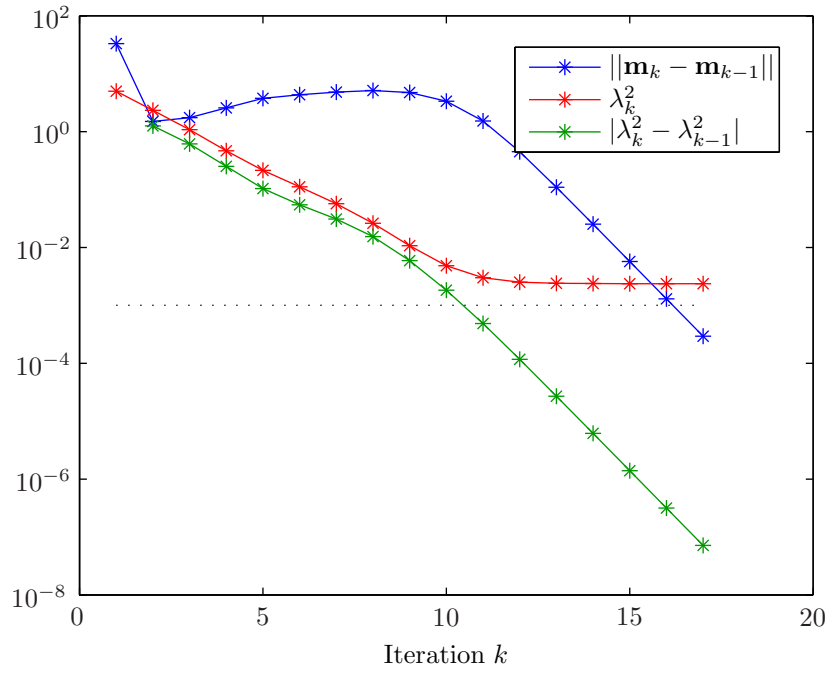


Figure A6. Convergence of the Bayesian regularization algorithm for inversion of PP reflection amplitudes. The black, dotted line is the convergence criterion for the update.

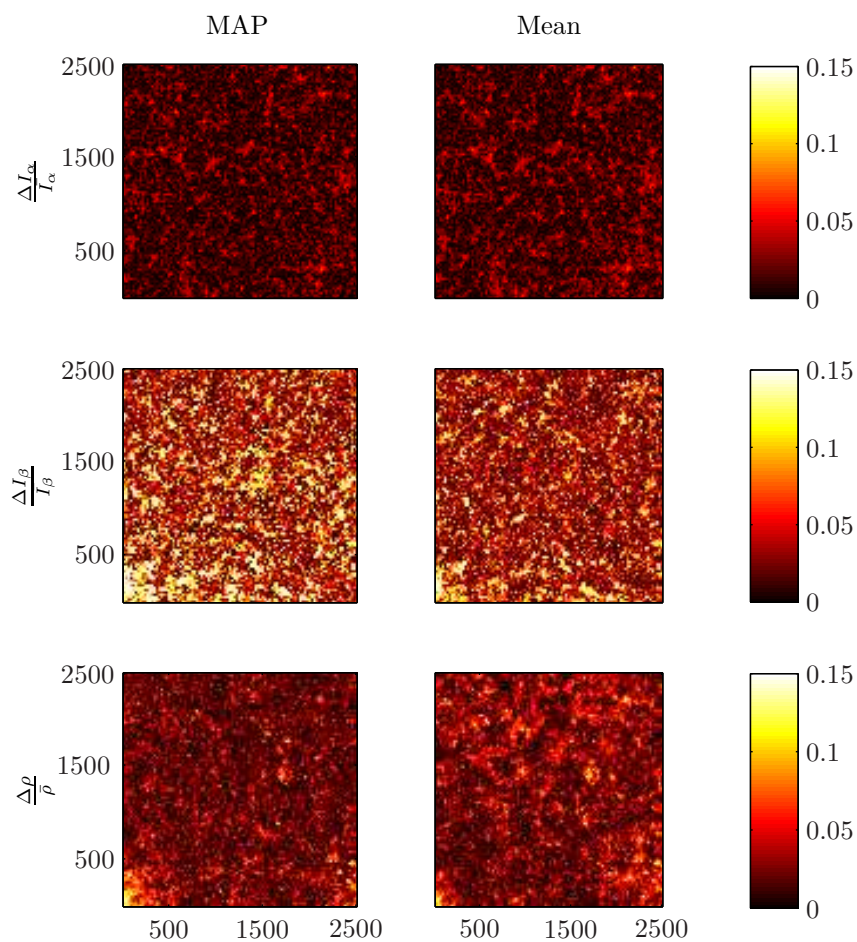


Figure A7. Absolute value of bias in the medium parameters \mathbf{m} for PP inversion. The left column is our MAP estimate while the right column (Mean) is reproduced from Rabben et al. (2008). Some high values in the contrast in S-wave impedance are clipped in order to visually enhance the differences.

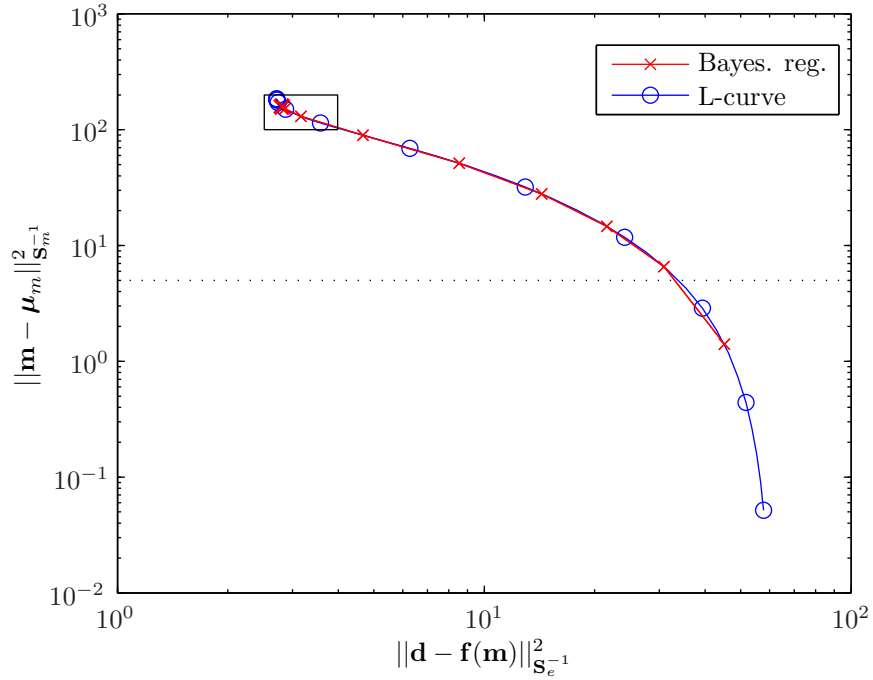


Figure A8. Bayesian regularization for joint PP and PS inversion and comparison with the L-curve. For the L-curve we have again used λ^2 values from $10^{-4.5}$ to $10^{1.5}$. The red x-marks show how our algorithm converges for $(\beta_e = 0, \beta_m = 5)$, and the thick x-mark to the left is the final iteration, with $\lambda^2 = 7.1 \times 10^{-3}$. The black, dotted line in the figure is the value of β_m , and the black square indicates the zoomed area in Fig. A9.

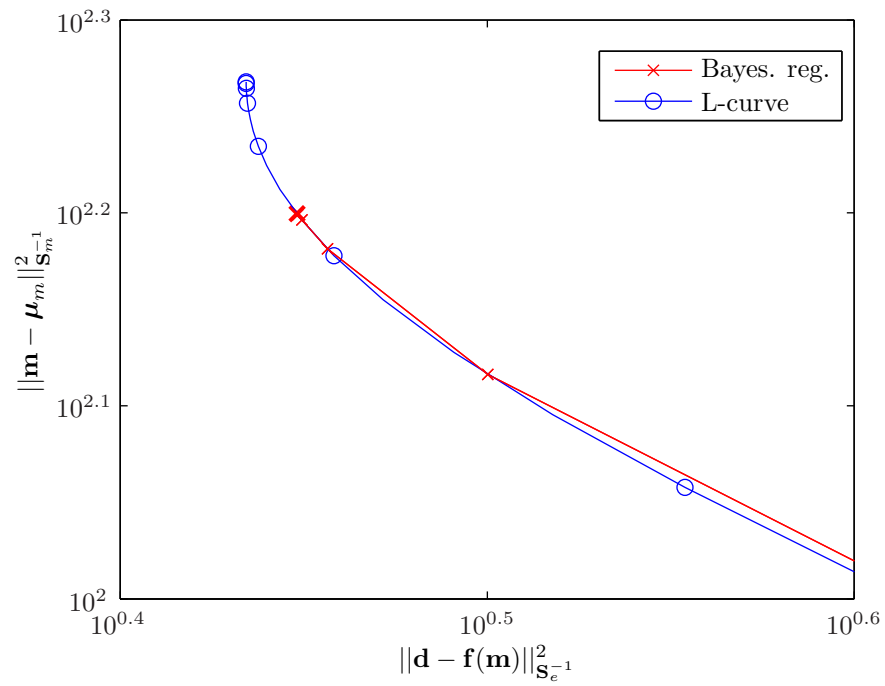


Figure A9. Zoom of the black square in Fig. A8.

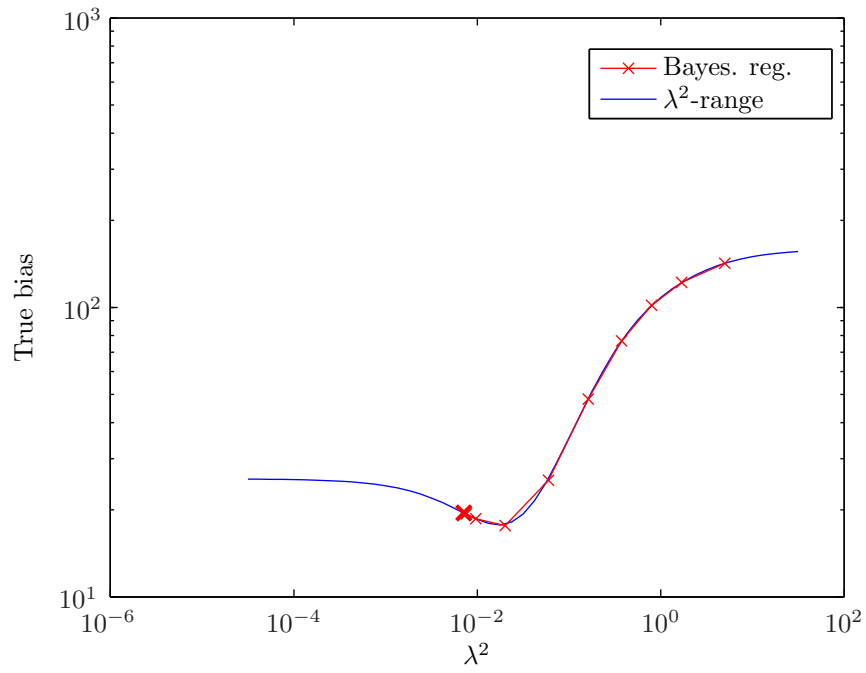


Figure A10. True bias in the Bayesian regularization and in the L-curve calculations, both for joint PP and PS inversion.

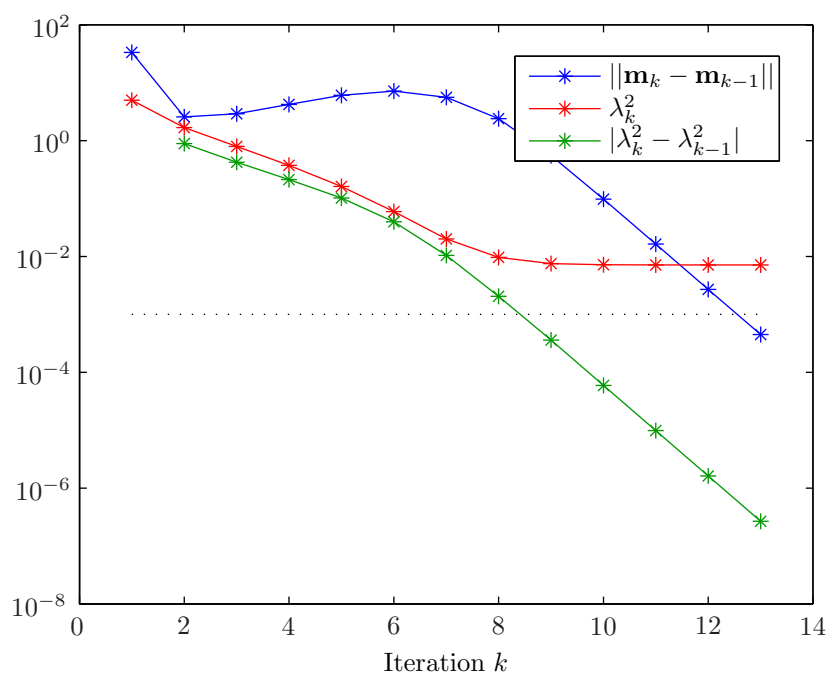


Figure A11. Convergence of the Bayesian regularization algorithm for inversion of joint PP and PS reflection amplitudes. The black, dotted line is the convergence criterion for the update.

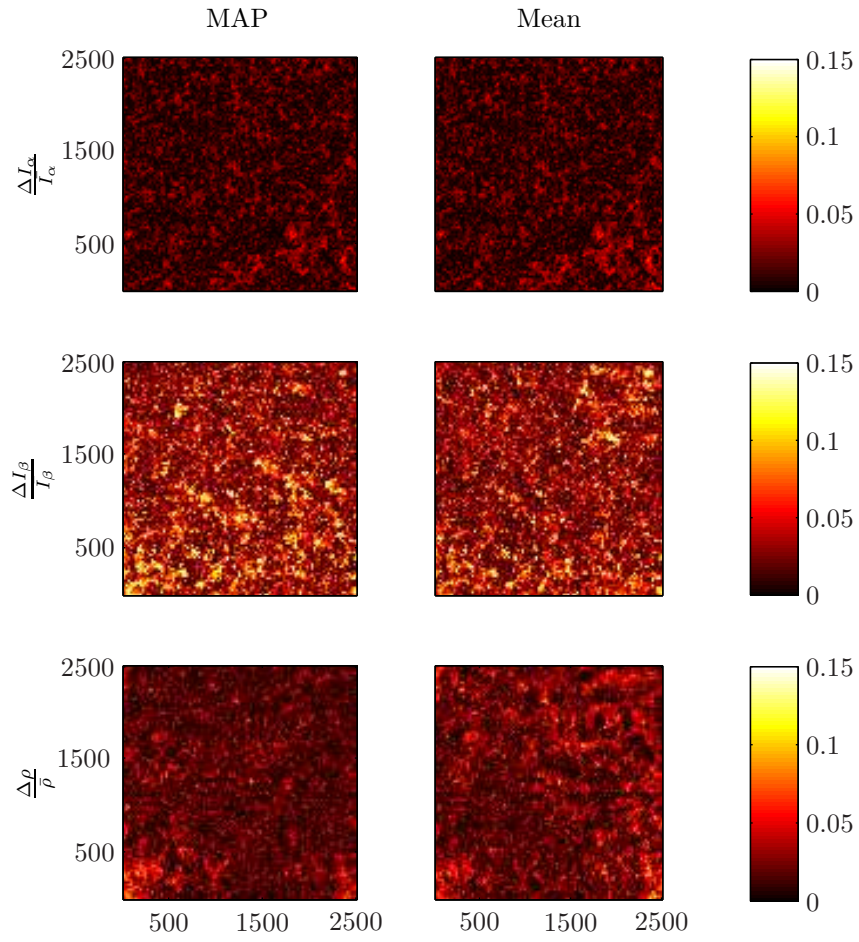


Figure A12. Absolute value of bias in the medium parameters \mathbf{m} for joint PP and PS inversion. The left column is our MAP estimate while the right column (Mean) is reproduced from Rabben et al. (2008). A few high values in the contrast in S-wave impedance are clipped in order to visually enhance the differences.

LIST OF TABLES

A1 Comparison of the workload for the Bayesian regularization and the L-curve approaches in the PP example. LS is the number of least-squares iterations, while CG is the total number of conjugate gradient iterations.

A2 Comparison of the workload for the Bayesian regularization and the L-curve approaches in the joint PP and PS example. LS is the number of least-squares iterations, while CG is the total number of conjugate gradient iterations.

Table A1. Comparison of the workload for the Bayesian regularization and the L-curve approaches in the PP example. LS is the number of least-squares iterations, while CG is the total number of conjugate gradient iterations.

	LS	CG
Bayes. reg.	17	2930
L-curve	295	55048

Table A2. Comparison of the workload for the Bayesian regularization and the L-curve approaches in the joint PP and PS example. LS is the number of least-squares iterations, while CG is the total number of conjugate gradient iterations.

	LS	CG
Bayes. reg.	13	1205
L-curve	215	20546