JOINT PARAMETER AND STATE ESTIMATION FOR WAVE-BASED IMAGING AND INVERSION

Tristan van Leeuwen

Mathematical Institute, Utrecht University, the Netherlands

ABSTRACT

In many applications, such as exploration geophysics, seismology and ultrasound imaging, waves are harnessed to image the interior of an object. We can pose the image formation process as a non-linear data-fitting problem: fit the coefficients of a wave-equation such that its solution fits the observations approximately. This allows one to effectively deal with errors in the observations.

However, a simple wave-equation most likely does not represent the physics properly. For example, the equation may only capture one wave-type or the source-term may include (unknown) random effects. In such cases, it is not desirable to solve the wave-equation exactly. Instead, we can formulate a joint estimation problem: find the field and the coefficients such that they obey both physics and observations approximately. In this formulation we put the physics and the observations on equal footing, allowing both errors in the model as well as the observations.

In this paper, I discuss the implications of such a joint approach and discuss the possibility of estimating the covariance matrices corresponding to errors in the observations and the physics.

Index Terms**—** Wave-equation, state estimation, model error, covariance estimation

1. INTRODUCTION

Many inverse problems in science and engineering are cast as *parameter estimation* problems, where the goal is to find parameters in a pre-defined model such that the predictions fit the observations. Features in the observations that are not well-predicted by the (simplified) model are typically removed prior to attempting a fit. These problems often involve relatively few parameters, a non-linear model, and redundant observations.

Another class of inverse problems is *state estimation*, where the aim is to estimate the (initial) state of a partially observed system. These problems occur, for example, in weather prediction where one wants to estimate the temperature at time t_0 everywhere based on a sparse set of observations at times $t_0, t_1, \ldots$. The models used to interpolate the data are typically linear, but the number of parameters to be estimated is large. In these applications, it is less common to remove unwanted features from the data. Instead, a lot of effort goes into estimating the so-called *model-error*, that quantifies to what extend we can expect to be able to fit the observations [1].

Inverse problems that involve both aspects can typically be cast as a partial-differential-equation (PDE) constrained optimization problem, where the goal is to find both the state (the solution of the PDE) and the parameters (coefficients of the PDE) such that the sampled state fits the observations. Model-errors are typically ignored here, allowing one to effectively eliminate the state by solving the PDE and pose the problem as a pure parameter estimation problem [2, 3]. Allowing for both errors in the observations as well as the model is challenging for large-scale applications since it entails optimizing over both the parameters and the state simultaneously.

In this paper I review a joint parameter and state estimation algorithm for PDE-constrained optimization [4] and extend it to include estimation of the covariance corresponding to model and data errors. The approach is specifically geared towards a *multi-experiment* setting, where multiple experiments yield independent observations of the same system, possibly for different inputs. The underlying assumptions of the proposed algorithm are that both model and data errors follow a Normal distribution with zero mean. A compelling application is seismic imaging, where model errors can be thought to arise from unknown random sources. A typical setup is illustrated in figure 1.

The outline of the paper is as follows. In section 2, I introduce the formulation of the joint estimation problem and discuss how the covariance matrices can be estimated as part of the optimization procedure. Then, I discuss a basic algorithm for joint estimation and discuss various extensions in section 3. Numerical examples on a seismic reflection problem are presented in section 4 and section 5 concludes the paper.

This research was in part financially supported by the Netherlands Organisation of Scientific Research (NWO) as part of research programme 613.009.032.

978-1-5090-4117-6/17/$31.00 ©2017 IEEE

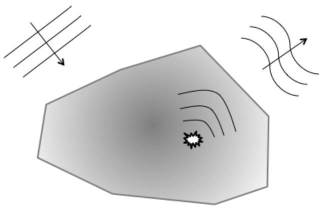

Fig. 1. Typical imaging setup, where a medium is insonified (top-left) and its response (top-right) is recorded. Aside from measurement noise, there is a noise-source in the form of an un-modelled source (center) inside the medium.

2. THEORY

The relation between the parameters, state and observations are captured by the *process* and *measurement* models,

$$\begin{aligned} A(m)u &= q + \eta, && \text{(process model)} \\ Pu &= d + \epsilon, && \text{(measurement model)} \end{aligned}$$

where $A(m)$ is the system matrix, $m \in \mathbb{R}^M$ are the parameters of interest, $u \in \mathbb{C}^N$ is the state, $q \in \mathbb{C}^N$ is a source term, $P \in \mathbb{C}^{K\times N}$ is the sampling operator, $d \in \mathbb{C}^K$ are the observations and $\eta \sim \mathcal{N}(0, E)$ and $\epsilon \sim \mathcal{N}(0, H)$ are Gaussian noise terms in the proces and measurement respectively with covariance matrices E and H. In wave-based imaging applications $A(m)$ could for example be a discretization of the scalar Helmholtz operator $\omega^2 m + \nabla^2$. Imaging applications are characterized by a relatively high number of wavelengths in the domain of interest (say 10^2), dictating the use of $N = \mathcal{O}(10^6)$ gridpoints for two-dimensional imaging (i.e., 10 gridpoints per wavelength). The observations are typically done by a one-dimensional array, which with the same sampling criterion would require $K = \mathcal{O}(10^3)$. We further assume that multiple observations $\{d_l\}_{l=1}^L$ of the same system are obtained, possibly for different inputs $\{q_l\}_{l=1}^L$. For an optimal (full) sampling we would need $L = \mathcal{O}(10^3)$.

2.1. Joint parameter and state estimation

The conventional approach to dealing with parameter estimation problems involving PDEs on this scale is to eliminate the process model explicitly and solving a reduced problem in m alone

$$\min_m \sum_{l=1}^L \|PA(m)^{-1}q_l - d_l\|_H^2,$$

where $\|r\|_W = r^T W^{-1} r$ is a weighted norm [2]. This formulation corresponds to a maximum likelihood estimation (MLE) of m under the assumption that $\eta = 0$. In principle one could formulate a MLE problem to account for noise in the process model as well using a Gaussian mixture model. The effective covariance of η is given by $A(m)^{-1}EA(m)^{-T}$ in this case. The dependence of the covariance on m makes this a less attractive option from an optimization point-of-view.

An alternative route is to cast the joint estimation problem as

$$\min_{m,U} \sum_{l=1}^L \|Pu_l - d_l\|_H^2 + \|A(m)u_l - q_l\|_E^2, \qquad (1)$$

where $U = [u_1, u_2, \ldots, u_L]$ contains the states for all the experiments. Applying a non-linear optimization method to this problem directly is not attractive because of the dimensionality of the problem. However, an efficient algorithm may be devised by eliminating the states u_l. Instead of solving the states directly from the PDE, they are obtained by solving the following normal equations

$$\begin{aligned} \left(A(m)^T E^{-1} A(m) + P^T H^{-1} P\right) u_l = \\ A(m)^T E^{-1} q_l + P^T H^{-1} d_l, \end{aligned}$$

for each l independently. This is equivalent to a Kalman smoother [5]. Plugging the solutions $u_l(m)$ back into (1) yields a reduced problem in m alone. More details on this approach can be found in [6, 4].

2.2. Estimating covariance matrices

To estimate the covariance matrices, we formulate an *extended least-squares problem* [7]

$$\begin{aligned} \min_{m,U,H,E} & L\log(|H|) + L\log(|E|) \\ & + \sum_{l=1}^L \|Pu_l - d_l\|_H^2 + \|A(m)u_l - q_l\|_E^2, \qquad (2) \end{aligned}$$

where $|\cdot|$ denotes the determinant. At first glance this makes the problem even worse, since we now need to optimize over large (dense) matrices. For a fixed m and U, however, we find the following closed-form solutions

$$\widehat{H} = \frac{1}{L}\left(PU - D\right)\left(PU - D\right)^T, \qquad (3)$$

$$\widehat{E} = \frac{1}{L}\left(A(m)U - Q\right)\left(A(m)U - Q\right)^T, \qquad (4)$$

which are effectively sample-variance estimates of the residuals [8]. This approach can be easily extended to accomodate the estimation of structured covariance matrices. For example, if we restrict to diagonal covariance matrices we can estimate the diagonals by computing the squares of the residuals, i.e., $L^{-1}\sum_{l=1}^L \left(Pu_l - d_l\right)^2$. If we want to move away from diagonal approximations of the covariance matrices, we need to impose some additional restrictions. In fact, as stated here, the estimate for E is likely to be rank deficient as in general $L < N$. In practice we typically even have only a small number of experiments ($L < K$) making the estimate of H rank deficient as well. A promising approach is impose sparsity constraints on the inverse of the covariance matrix [9, 10].

2.3. Variable projection

We now formulate the joint estimation problem as an optimization over the parameters m and U alone

$$\min_{m,U} f(m,U), \tag{5}$$

where $f(m,U)$ can be evaluated by substituting the covariance estimates (3)-(4) in (2). The gradients of this reduced objective with respect to m and u_k are now given by

$$\nabla_m f(m,U) = \sum_{l=1}^{L} G(m,u_l)^T \widehat{E}^{-1} \left(A(m)u_l - q_l\right), \tag{6}$$

$$\begin{aligned}\nabla_{u_l} f(m,U) &= P^T \widehat{H}^{-1} \left(Pu_l - d_l\right) \\ &\quad + A(m)^T \widehat{E}^{-1} \left(A(m)u_l - q_l\right),\end{aligned} \tag{7}$$

where $G(m,u) = \frac{\partial A(m)u}{\partial m}$ is the Jacobian matrix of $A(m)u$. This matrix is typically sparse and can be easily computed for any given A. It is remarkable that the sensitivity of the covariance matrices w.r.t (m,U) does not enter into the gradient expression. This follows from the fact that we are differentiating an optimal value function [11]. The Hessian of f is given by the Schur complement of the Hessian of the objective in (2) [12]. We thus find the following expressions

$$\begin{aligned}&\nabla^2_m f(m,U) = \\ &\qquad \sum_{l=1}^{L} G(m,u_l)^T \widehat{E}^{-1} G(m,u_l) - \text{correction term},\end{aligned} \tag{8}$$

$$\begin{aligned}&\nabla^2_{u_l} f(m,U) = \\ &\qquad P^T \widehat{H}^{-1} P + A(m)^T \widehat{E}^{-1} A(m) - \text{correction term},\end{aligned} \tag{9}$$

where the correction terms capture the sensitivity of E and H w.r.t. m and U.

It can be shown that a (local) minimum of f, together with the estimated covariance matrices (4)-(3), are a local minimum of the extended least-squares problem (2) [8, 4].

3. ALGORITHM

Using the expressions for the gradient and Hessian presented above, we can employ a Newton-like method to update (m,U) simultaneously. Every function evaluation of f then entails estimating the covariance matrices for the current (m,U).

The challenge is to devise an algorithm that avoids storing all the states and works with a structured approximation of the covariance matrices. To avoid storing all the states, we note that the extended least-squares problem (2) is quadratic in U, and this permits a closed-form solution. Having eliminated E and H from the problem we loose this property due to the dependence of the covariance matrices on U. However, if we ignore the correction term in (9) then the Newton update for u_k is given by

$$\begin{aligned}u_k := &\left(P^T \widehat{H}^{-1} P + A(m)^T \widehat{E}^{-1} A(m)\right)^{-1} \\ &\qquad \left(P^T \widehat{H}^{-1} d + A(m)^T \widehat{E}^{-1} q\right).\end{aligned}$$

Thus, we can safely discard previous iterates of U as the updated U only depends on previous states through the covariance matrices. We can update m using our favourite Hessian approximation (e.g., L-BFGS) using the gradient expression derived earlier. Convergence of the overall algorithm is ensured by the simple fact that we are effectively using a gradient-descent method in (m,U) with a positive-definite Hessian approximation. A basic version of this algorithm is listed in Algorithm 1. Note that we can accumulate all quantities on-the-fly without computing all states simultaneously.

Algorithm 1 basic algorithm

```
Input: Initial parameter, $m$, and initial covariance matrices, $E$, $H$.
Output: Parameter and state estimates $(m,U)$ and estimated covariance matrices $E$, $H$
  while not converged do
    Compute states:
    $U = \left(P^T H^{-1} P + A(m)^T E^{-1} A(m)\right)^{-1} \left(P^T H^{-1} D + A(m)^T E^{-1} Q\right)$
    Update parameters:
    $m := m - \alpha \nabla_m f(m,U)$
    Compute sample covariance matrices:
    $H := (PU - D)(PU - D)^T$
    $E := (A(m)U - Q)(A(m)U - Q)^T$
  end while
```

This algorithm can be applied directly when we restrict the covariance matrices to be diagonal. For estimating the full covariance matrix, it is not feasible to use the expression presented here and we may need to add some regularization. On top of that, it would be more efficient to update the inverse of the covariance matrices in order to avoid recomputing the inverse when solving for U.

4. RESULTS

We test the proposed algorithm on a seismic reflection problem, where A is a standard finite-difference discretization of the scalar Helmholtz equation with Sommerfeld boundary conditions for frequencies $2, 3, 4, 5, 6$ Hz (i.e., A is a block diagonal matrix). The sources, q_l, are point-sources located near the top boundary of the domain. The sampling operator P samples the wavefield at the receiver locations, near the top boundary of the domain. A point source is added in the center of the domain to generate noise in the process model.

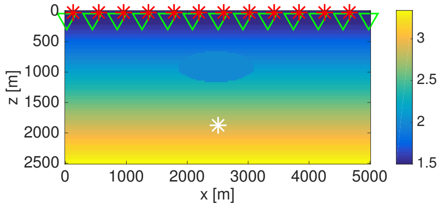

Fig. 2. Soundspeed (km/s) as a function of z, x. The goal is to recover the anomaly around $(1000, 2500)$ from the response of the sources (red *) as recorded by the receiver array (green ∇). The additional source of noise is located at $(2000, 2500)$ (white *)

The dimensions of the problem are $N = (101 \times 201) \times 5$ (gridpoints × number of frequencies), $K = 49$, $L = 49$. The true parameters (soundspeed) are shown in figure 2.

We compare the results of conventional and joint estimation on data generated with and without noise on the process model. For the joint inversion we use Algorithm 1, with a Borzilai-Borwein steplength and keep the covariance matrices fixed. For conventional inversion we use a slightly modified version of Algorithm 1, where the state is obtained as $U = A(m)^{-1}Q$. The results from an inversion with constant covariance matrices ($E = I$, $H = I$) are shown in figure 3. The results on noisy data clearly show the imprint of the noise source. Having obtained an initial estimate of m and U from the noisy data, we estimate the covariance matrix E and show the results in figure 4. In particular, the joint method produces an estimate of the covariance matrix that can be used to further improve the reconstruction.

5. CONCLUSIONS AND DISCUSSION

We showed how model-errors can be incorporated in a joint parameter and state-estimation framework. A numerical example on a seismic reflection problem illustrates the feasibility of the method; the estimated covariance contains a clear imprint of the noise in the process model which can be exploited to further improve the results. So far, we have illustrated the method as a two-step approach, where the covariance is estimated a-posteriori for jointly reconstructed paramaters and states. For fully automatic on-the-fly estimation, more robust algorithms that include appropriate regularization are needed.

The joint parameter and state estimation method with diagonal covariance matrices can in principle be scaled to large-scale problems. Extending the approach to other structured approximations of the covariance matrices (e.g., sparse or low rank), requires more sophisticated estimation of the covariance matrix. This would increase the computational cost dramatically, since we re-estimate the covariance at each iteration. An obvious solution would be to estimate the covariance only once every few iterations.

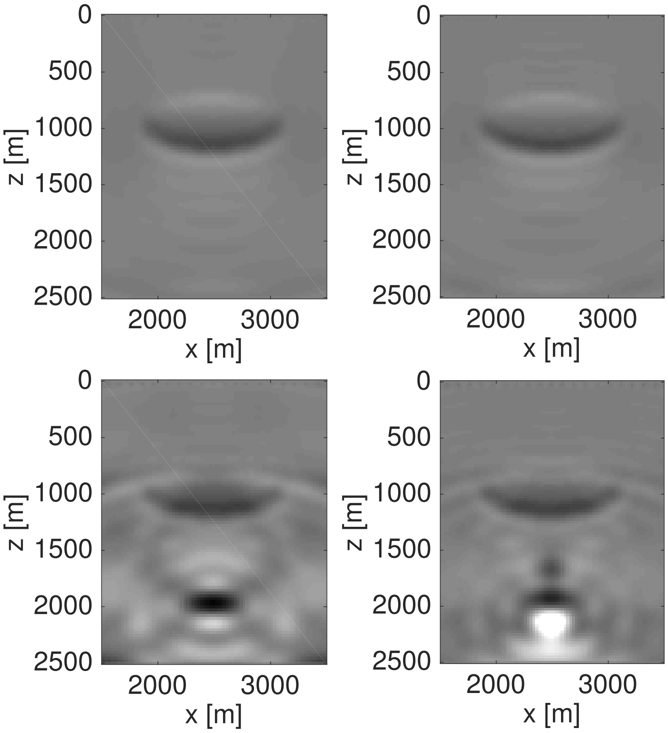

Fig. 3. Top: reconstruction of the anomaly from noiseless data using the conventional (left) and joint (right) method. Bottom: reconstruction using the conventional (left) and joint (right) method. The imprint of the noise source is clearly present in both results.

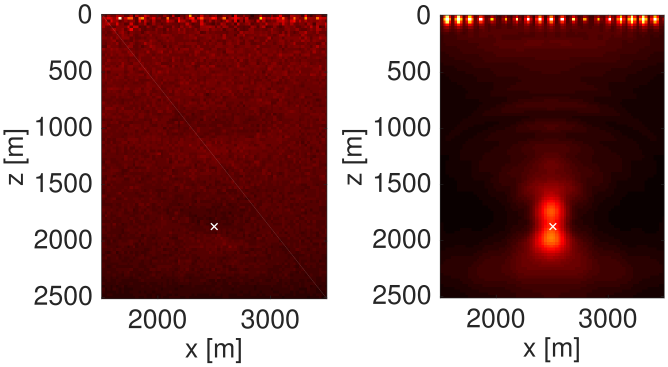

Fig. 4. Diagonal of the estimate of the covariance matrix E corresponding to the results shown in figure 3 (bottom). The estimate resulting from the joint estimation (right) clearly indicates the presence of the noise source, whereas the one resulting from the conventional estimate contains no information at all.

6. REFERENCES

[1] Sebastian Reich and Colin Cotter, *Probabilistic forecasting and Bayesian data assimilation*, Cambridge University Press, 2015.

[2] Eldad Haber, Uri M Ascher, and Doug Oldenburg, "On optimization techniques for solving nonlinear inverse problems," *Inverse Problems*, vol. 16, no. 5, pp. 1263–1280, oct 2000.

[3] E Haber and U M Ascher, "Preconditioned all-at-once methods for large, sparse parameter estimation problems," *Inverse Problems*, vol. 17, no. 6, pp. 1847–1864, dec 2001.

[4] T van Leeuwen and Felix J Herrmann, "A penalty method for PDE-constrained optimization in inverse problems," *Inverse Problems*, vol. 32, no. 1, pp. 015007, jan 2016.

[5] Aleksandr Y. Aravkin, James V. Burke, and Gianluigi Pillonetto, "Optimization viewpoint on kalman smoothing with applications to robust and sparse estimation," in *Compressed Sensing & Sparse Filtering*, Berlin, Heidelberg, 2014, pp. 237–280, Springer Berlin Heidelberg.

[6] Tristan van Leeuwen and Felix J Herrmann, "Mitigating local minima in full-waveform inversion by expanding the search space," Tech. Rep. 1, jul 2013.

[7] Bradley M. Bell, James V. Burke, and Alan Schumitzky, "A relative weighting method for estimating parameters and variances in multiple data sets," *Computational Statistics & Data Analysis*, vol. 22, no. 2, pp. 119–135, jul 1996.

[8] Aleksandr Y. Aravkin and Tristan van Leeuwen, "Estimating nuisance parameters in inverse problems," *Inverse Problems*, vol. 28, no. 11, pp. 115016, nov 2012.

[9] CJ Hsieh and MA Sustik, "Sparse inverse covariance matrix estimation using quadratic approximation," *Advances in Neural Information Processing Systems*, vol. 24, pp. 2330–2338, 2011.

[10] Cho-Jui Hsieh, Matyas A Sustik, Inderjit S Dhillon, Pradeep K Ravikumar, and Russell Poldrack, "BIG & QUIC: Sparse inverse covariance estimation for a million variables," in *Advances in Neural Information Processing Systems*, vol. 26, pp. 3165–3173. 2013.

[11] Bradley M Bell and James V Burke, "Algorithmic differentiation of implicit functions and optimal values," in *Lecture Notes in Computational Science and Engineering*, 2008, vol. 64 LNCSE, pp. 67–77.

[12] Axel Ruhe and Per Ake Wedin, "Algorithms for Separable Nonlinear Least Squares Problems," *SIAM review*, vol. 22, no. 3, pp. 318–337, 1980.