

# **ROBUST ENSEMBLE KALMAN FILTER FOR HIGH DIMENSIONAL DATA**

INGE MYRSETH

Department of Mathematical Sciences,  
Norwegian University of Science and Technology, Norway.

## **ABSTRACT**

*This paper introduces the Hierarchical Ensemble Kalman Filter as an extension of the Ensemble Kalman Filter. The extension is constructed to be robust with respect to estimation uncertainty and rank issues which can arise in problems involving Time lapse seismic data where the number of observations is large. The methodology is derived in a fully hierarchical Bayesian manner. An example shows promising results where the hierarchical approach provides more stable solutions than the EnKF in a problem where the EnKF has been known to have trouble.*

## **INTRODUCTION**

Reservoir engineers evaluating reservoir properties are aiming to maximize the recovery. One of the strategies to achieve this target is known as History Matching (HM). This is done by conditioning a model of the reservoir on historical production data until it reproduces the past behavior of the reservoir. When this is done, the model can be used to simulate future reservoir behavior. The forecast should include uncertainty, and due to the nonlinear nature of the reservoir models, this is done by repeated reservoir simulations. Traditionally this is done by tuning individual parameters to minimize the difference between observed production history and simulated production. This method does not account for uncertainty and also has the problem that it conditions on all available data simultaneously. This means that an entirely new model has to be matched when new data are available.

The Ensemble Kalman Filter (EnKF) first introduced in Evensen (1994) is a method that copes with these issues. It conditions sequentially on the data and can readily be updated as new data are available. The method uses an initial ensemble of reservoir models (states) which are simulated forward in time and sequentially conditions on data as they are available. The state covariance matrix at any given time is estimated based on the simulated reservoir realizations, and the ensemble members are conditioned on data at that time via an expression involving the estimated state covariance matrix. While the reservoir may be modeled as a large

system with millions of grid nodes traditional HM usually involves a small number of reservoir observations (water cut, bottom hole pressure etc.).

History matching time lapse seismic data is a problem where the number of observations is large and higher than the number of ensemble members. This may introduce rank complications in the covariance estimation and the EnKF might break down because of this. We propose a robust Bayesian hierarchical extension to deal with this problem. We impose prior distributions on the mean and covariance of the ensemble at any given time, and use distributions from the Gauss-conjugate class of distributions. This entails that we can analytically calculate full rank covariance estimates and account for the uncertainty in the covariance estimate. Another feature is that introducing prior information on the correlation structure is more robust towards edge effects introduced by localization.

## MODEL FORMULATION

Let  $[x_0, x_1, \dots, x_{t+1}]$  be a multivariate time series of unknown states. Further, let  $[d_0, d_1, \dots, d_t]$  be an associated time series of observations. Let the dimension of each  $x_i$  and  $d_j$  be known as  $n_x$  and  $n_d$  respectively. The objective is to make inference on the states based on the observations, and, be able to determine the forecast  $x_{t+1}$  given  $[d_0, d_1, \dots, d_t]$ .

Let the prior model for  $[x_0, x_1, \dots, x_{t+1}]$  be:

$$f(x_0, x_1, \dots, x_{t+1}) = f(x_0) \prod_{i=0}^t f(x_{i+1} | x_0, x_1, \dots, x_i) = f(x_0) \prod_{i=0}^t f(x_{i+1} | x_i)$$

where  $f(x_0)$  is a known pdf for the initial state and  $f(x_{i+1} | x_i) = n(w_i(x_i), \Sigma_x)$  is a Gaussian distribution with expectation given by  $w_i(\cdot)$ , a set of given forward functions, and  $\Sigma_x$  a known covariance matrix. Note that the prior model is Markovian, ie each state conditioned on the past is only dependent on the previous state.

Let the likelihood model for  $[d_0, d_1, \dots, d_t]$  conditioned on  $[x_0, x_1, \dots, x_{t+1}]$  be:

$$f(d_0, d_1, \dots, d_t | x_0, x_1, \dots, x_{t+1}) = \prod_{i=0}^t f(d_i | x_0, x_1, \dots, x_{t+1}) = \prod_{i=0}^t f(d_i | x_i)$$

where  $f(d_i | x_i) = n(Hx_i, \Sigma_d)$  and  $H$  is a given observation design matrix and  $\Sigma_d$  is a known observation covariance matrix. The combination of the prior and likelihood define a hidden Markov process as shown in the graph in Figure 1.

The posterior stochastic model can be expressed as:

$$\begin{aligned} f(x_0, x_1, \dots, x_{t+1} | d_0, d_1, \dots, d_t) &= \text{const} \times f(d_0, d_1, \dots, d_t | x_0, x_1, \dots, x_{t+1}) \times f(x_0, x_1, \dots, x_{t+1}) \\ &= \text{const} \times f(d_0 | x_0) f(x_0) \left[ \prod_{i=0}^{t-1} f(d_{i+1} | x_{i+1}) f(x_{i+1} | x_i) \right] f(x_{t+1} | x_t) \end{aligned}$$

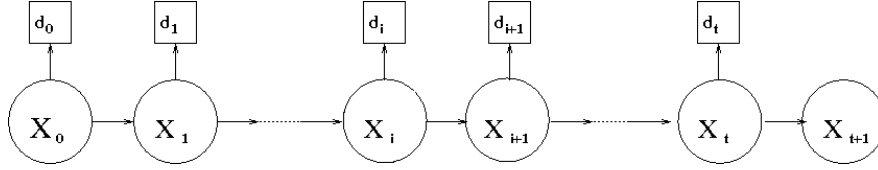


Figure 1: Hidden Markov process

where  $\text{const}$  is a normalizing constant which can be hard to assess. The forecast problem can be expressed as:

$$f(x_{t+1}|d_0, d_1, \dots, d_t) = \int \dots \int f(x_0, x_1, \dots, x_{t+1}|d_0, d_1, \dots, d_t) dx_0 dx_1 \dots dx_t.$$

For a general non linear forward function this problem might be very hard, however an analytical solution can be obtained through the Kalman Filter if the forward function is linear. Solution strategies for the non linear model include recursive approaches and the Ensemble Kalman Filter. A more detailed description of the Ensemble Kalman Filter follows. See Evensen (2007) for a thorough introduction.

### TRADITIONAL ENKF-ALGORITHM

The Ensemble Kalman Filter uses an ensemble of realizations of states to explore the hidden Markov process in Figure 1 in a manner where the ensemble members capture the dynamics of the problem through the forward function and are updated with respect to observations. At time  $t = 0$  an initial ensemble is generated iid from  $f(x_0)$ , the pdf of the initial state,

$$e_0^u : \{x_{0,j}^u; j = 1, \dots, m\}$$

Note that the superscript  $u$  refers to unconditioned whereas the superscript  $c$  will refer to conditioned. In between observations the ensemble is advanced via the forward function until observations are available

$$e_{i+1}^u : \{x_{i+1,j}^u = w_i(x_{i,j}^c) + u_{x,j}; j = 1, \dots, m\}$$

where  $u_{x,j}; j = 1, \dots, m$  iid  $n(0, \Sigma_x)$ . At this time the ensemble is updated as

$$e_i^c : \{x_{i,j}^c = x_{i,j}^u + S_i H^T (H S_i H^T + \Sigma_d)^{-1} (d_i + u_{d,j} - H x_{i,j}^u); j = 1, \dots, m\} \quad (1)$$

where

$$S_i = \frac{1}{m-1} \sum_{j=1}^m (x_{i,j}^u - a_i)(x_{i,j}^u - a_i)^T \quad \text{and} \quad a_i = \sum_{j=1}^m \frac{x_{i,j}^u}{m}$$

are the ensemble covariance and mean and  $u_{d,j}; j = 1, \dots, m$  iid  $n(0, \Sigma_d)$ . The update is based on the following relationship

$$f(x_i^c) = f(x_i^u | d_i) = \text{const} \times f(d_i | x_i^u) f(x_i^u).$$

In the case where the forward function is linear  $f(x_i^d)$  will be Gaussian and thereby the pdf for the conditioned ensemble will also be Gaussian. However for a non linear forward function the Ensemble Kalman Filter is justified by approximating  $f(x_i^d)$  with the Gaussian distribution  $n(\mu_i, \Sigma_i)$ , where the unknown parameters are estimated by the ensemble mean,  $a_i$  and the ensemble covariance,  $S_i$ . However the ensemble mean is not explicitly part of the update. Some of the features that define the success of the EnKF are as follows. Solutions will be affected by model approximation and the non linearity of the forward function. If  $f(x_i^d)$  is not well represented by a Gaussian approximation the EnKF might perform poorly. However for a unimodal  $f(x_i^d)$  the Gaussian approximation is fairly robust. For the multimodal case Zafari and Reynolds (2005) found that the EnKF has problems. In some cases where the traditional covariance estimate is used for heavy tailed distributions it might be unreliable even though it is unbiased. Additionally, if the forward function is highly non linear the solutions might be unstable.

In the case where the number of ensemble members is smaller than the number of observations or the dimension of  $x_i$  the estimated covariance matrix  $S_i$  will be a low rank estimate of  $\Sigma_i$ . Related to the update, equation 1, the term  $S_i H^T$  will be low rank and the information in the observations might be distributed somewhat randomly in the space of  $x_i$ . This might introduce artifacts in the solution. Skjervheim et al. (2005) present a numerical procedure to deal with this problem under the defined EnKF model.

In some applications rather than updating all state variables with regard to all observations, a localized update scheme where only a subset of state variables in the vicinity of each observation is updated with regard to that observation is necessary due to computational limitations. However, artifacts may be introduced in the state variables due to localization.

A hierarchical extension of the EnKF is presented in the next section. It is constructed to be more robust towards estimation uncertainty, rank deficiency and localization problems and differs from other approaches in that the model is redefined.

## HIERARCHICAL ENKF-ALGORITHM

The main approximation in the EnKF appears when  $f(x_i^d)$  is approximated by the Gaussian distribution  $n(\mu_i, \Sigma_i)$ . In the traditional approach  $\mu_i$  and  $\Sigma_i$  are estimated by the ensemble mean and covariance matrix. The Hierarchical Ensemble Kalman Filter (HEnKF) extends the traditional EnKF in the sense that  $\mu_i$  and  $\Sigma_i$  are considered random variables and prior stochastic models have to be assigned to them. Figure 2 shows the extended hidden Markov process. Assign the following prior distributions from the Gaussian conjugate family of distributions. Let  $\mu_i | \Sigma_i$  be Gaussian with expectation  $\xi_i$  and variance scaling parameter  $\eta_i > 0$ , ie  $n(\xi_i, \eta_i \Sigma_i)$ . Let  $\Sigma_i$  be inverse Wishart distributed with  $(n_x \times n_x)$ -matrix parameter  $\Psi_i$  and degrees of freedom  $\nu_i > n_x + 1$ , ie  $iw(\Psi_i, \nu_i)$ . These priors can be adapted to the

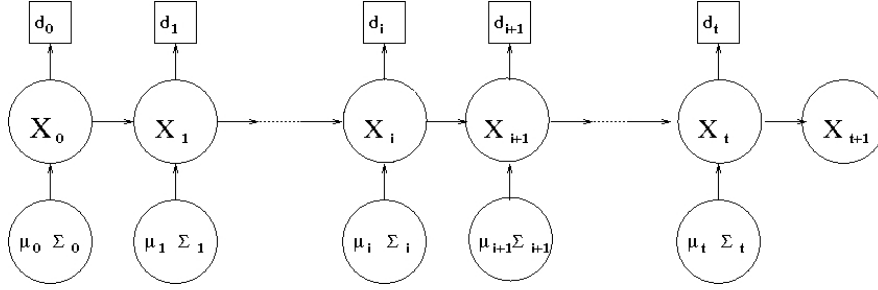


Figure 2: Hidden hierarchical Markov process

ensemble in the following manner.

$$\hat{f}(x_i^u | e_i^u) = \int \int \hat{f}(x_i^u | \mu_i, \Sigma_i, e_i^u) f(\mu_i | \Sigma_i, e_i^u) f(\Sigma_i | e_i^u) d\mu_i d\Sigma_i$$

where

$$\begin{aligned} \hat{f}(x_i^u | \mu_i, \Sigma_i, e_i^u) &= n(\mu_i, \Sigma_i), \\ f(\mu_i | \Sigma_i, e_i^u) &= n\left(\frac{1}{1+m\eta_i} \xi_i + \frac{m\eta_i}{1+m\eta_i} a_i, \frac{\eta_i}{1+m\eta_i} \Sigma_i\right), \\ f(\Sigma_i | e_i^u) &= iw(\Psi + (m-1)S_i + \left(\frac{1}{m} + \eta_i\right)^{-1} (a_i - \xi_i)(a_i - \xi_i)^T, \nu_i + m), \end{aligned}$$

and  $a_i$  and  $S_i$  are the ensemble mean and covariance. Note that the adapted expectation of  $\mu_i$  is a weighted average of prior expectation and the ensemble mean and that when  $m \rightarrow \infty$   $\mu_i$  coincides with the ensemble mean without variance. The adapted expectation of  $\Sigma_i$  is a weighted average of the prior expectation and the ensemble covariance and when  $m \rightarrow \infty$   $\Sigma_i$  coincides with the ensemble covariance without variance. The HEnKF-algorithm explores the model through an ensemble in a manner similar to the EnKF-algorithm. At time  $t = 0$  an initial ensemble

$$e_0^u : \{x_{0,j}^u; j = 1, \dots, m\}$$

is generated iid from  $f(x_0)$ , the pdf of the initial state. In between observations the ensemble is advanced via the forward function until observations are available

$$e_{i+1}^u : \{x_{i+1,j}^u = w_i(x_{i,j}^c) + u_{x,j}; j = 1, \dots, m\}$$

where  $u_{x,j}; j = 1, \dots, m$  iid  $n(0, \Sigma_x)$ . At this time the ensemble is updated as

$$e_i^c : \{x_{i,j}^c = x_{i,j}^u + \Sigma_{i,j} H^T (H \Sigma_{i,j} H^T + \Sigma_d)^{-1} (d_i + u_{d,j} - H x_{i,j}^u); j = 1, \dots, m\}$$

where  $u_{d,j}; j = 1, \dots, m$  iid  $n(0, \Sigma_d)$ , and

$$\Sigma_{i,j}; j = 1, \dots, m \text{ iid } iw(\Psi + (m-1)S_i + \left(\frac{1}{m} + \eta_i\right)^{-1} (a_i - \xi_i)(a_i - \xi_i)^T, \nu_i + m).$$

The update is an approximation since  $\Sigma_i$  is generated from  $f(\Sigma_i | \eta_i)$  and not from  $f(\Sigma_i | \eta_i, d_i)$ . However it is a reasonable approximation since the information in  $d_i$

will be assimilated into the ensemble at the next time step. Note that  $\Sigma_{i,j}$  is ensured to be positive definite and full rank and thereby resolving issues regarding rank deficiency and estimation uncertainty. However this is resolved by imposing prior information on the ensemble. Note further that when the ensemble size  $m \rightarrow \infty$  the HEnKF and the EnKF coincide.

The prior parameters at the initial time should be related to the known pdf at time  $t = 0$ . That is,  $E\{\mu_0|\Sigma_0\} = \xi_0 = E\{x_0\}$  and  $E\{\Sigma_0\} = \frac{\Psi_0}{v_0 - n_x - 1} = \text{Var}\{x_0\}$ . The parameters  $\eta_0$  and  $v_0$  should be chosen in accordance with uncertainty associated with the prior assessment. For the next states,  $i = 1, \dots, t$  the following recursive procedure for assessment of the priors based on the recursive nature of the problem are reasonable:

$$\begin{aligned} E\{\mu_{i+1}|\Sigma_{i+1}\} &= \xi_{i+1} = w_i(\xi_i), \\ \eta_{i+1} &= \eta_i, \quad v_{i+1} = v_i. \\ (v_{i+1} - n_x - 1)E\{\Sigma_{i+1}\} &= \Psi_{i+1} = \Psi_i + \Psi_i H^T (H \Psi_i H^T + \Sigma_d)^{-1} H \Psi_i. \end{aligned}$$

### EXAMPLE

In applications the forward model is usually non linear. However for clear results, the example features a linear forward function. Consider the following model. The variable of interest  $\mathbf{x}$  is a time series of dimension  $n_x = 100$ . Observations are available at time steps  $t = 0, \dots, 10$ , and forecast to time  $t = 11$  is the objective. The forward function is linear;  $w_i(x_i) = Ax_i$ , where  $A$  is a  $(100 \times 100)$  matrix. Note that there is no error associated with the forward function. The matrix  $A$  acts as a 10 node sliding average for the first 50 nodes while it is just the identity for the last 50 nodes. The initial pdf at time  $t = 0$  is a zero-centered Gaussian with covariance matrix defined by the covariance function  $c(i - j) = 20 \exp\{3 \frac{|i-j|}{20}\}$ , where  $(i, j)$  are nodes in  $x_i$ . The observations are given as  $d_i|x_i = x_i + u_d$ . That is, every node is observed directly at all times with variance  $\sigma_d^2 = 20$ . A reference realization  $\mathbf{x}^R$  and reference observations  $\mathbf{d}^R$  have been generated and Figure 3 shows  $x_{10}^R$  and  $d_{10}^R$ . The initial prior parameters are chosen in accordance with the Gaussian distribution used when generating the reference case. The variance scaling parameter  $\eta_0$  is chosen large such that the influence of the mean terms in the hierarchical covariance estimate is negligible. The degrees of freedom parameter is  $v_0 = n_x + 30 = 130$ . To reflect a worst case scenario for the EnKF, the number of ensemble members in this exercise is  $m = 10$ .

Figure 4 shows the mean of one run of the EnKF- and HEnKF-algorithms respectively with 95%-marginal confidence bounds in the two upper displays. The gray line is the reference at time  $t = 11$ . The two middle plots show the mean of 10 runs of the EnKF- and HEnKF-algorithms and the analytical solution (gray). The two lower plots show weights and covariance for node 50 when updating at time  $t = 10$  for EnKF (black dotted), HEnKF (darker gray) and the analytical solution (gray). From the two upper most plots it is clear that the EnKF-solution misses the target and underestimates the uncertainty of the solution. The HEnKF-solution on the other hand does a better job as it hits the target and provides more realistic

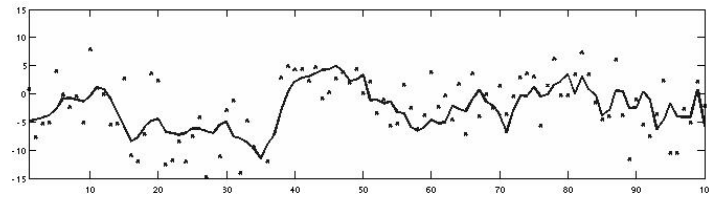


Figure 3: Reference realization for the linear case. The reference truth at time  $t = 10$ ,  $x_{10}^R$  (line) and reference observations  $d_{10}^R$ .

forecast intervals. The two middle displays expose the large variability in between runs for the EnKF, while the HEnKF appears more stable. The two lower plots establish that the unwanted effects in the EnKF solution are due to a lacking covariance structure and almost uniform weighting. The HEnKF shows an overall good covariance estimate, somewhat over estimating due to not accounting for the smoothing effect in the forward function. The weights appear trustworthy.

This example, which is of the worst case scenario for the EnKF exposes vulnerability to estimation uncertainty and rank deficiency. The HEnKF provides stable results in this case due to the more robust covariance estimates made possible by incorporating prior information in a hierarchical manner.

## CONCLUSIONS

The Hierarchical Ensemble Kalman Filter has been introduced as a robust extension to the Ensemble Kalman Filter. The extension is constructed to be robust with respect to estimation uncertainty and rank issues which can arise in problems involving Time Lapse seismic data where the number of observations is large. Examples show promising results where the hierarchical approach provides more stable solutions than the EnKF in problems where the EnKF has been known to have trouble.

## ACKNOWLEDGMENTS

This work is funded by the partners in the URE-initiative.

## REFERENCES

- Evensen, G (1994). *Sequential data assimilation with nonlinear quasi-geostrophic model using monte carlo methods to forecast error statistics*. In Journal of Geophysical Research, vol. 99, pp. 143–162.
- Evensen, G (2007). *Data Assimilation; The Ensemble Kalman Filter*. Springer.
- Skjervheim, J, Evensen, G, Aanonsen, S, Ruud, B and Johansen, T (2005). *Incorporating 4D seismic data in reservoir simulation models using ensemble kalman filter*. In SPE, vol. SPE 95789.
- Zafari, M and Reynolds, A (2005). *Assessing the uncertainty in reservoir description and performance predictions with the ensemble kalman filter*. In SPE, vol. SPE 95750.

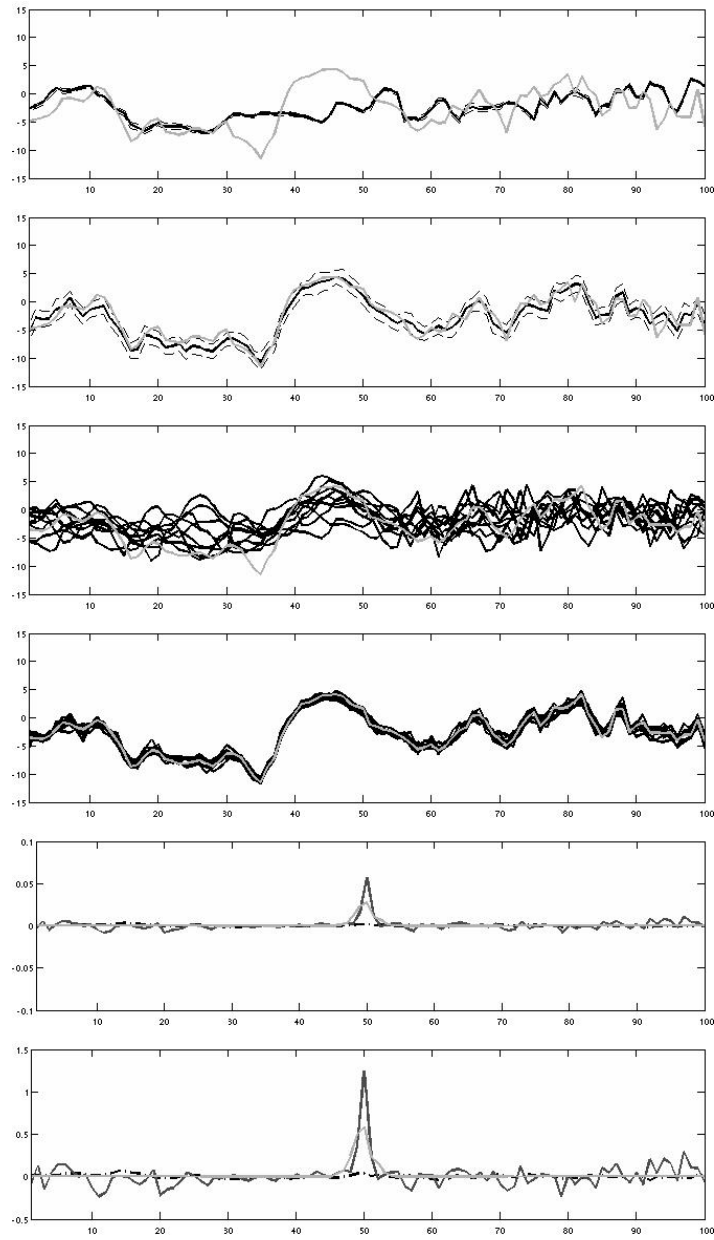


Figure 4: The two upper most plots show the mean of one run of the EnKF- and HEnKF-algorithms respectively with 95%-marginal confidence bounds. The gray line is the reference at time  $t = 11$ . The two middle plots show the mean of 10 runs of the EnKF- and HEnKF-algorithms and the analytical solution (gray). The two lower plots show weights and covariance for node 50 when updating at time  $t = 10$  for EnKF (black dotted), HEnKF (darker gray) and the analytical solution (lighter gray).