Ensemble Kalman filter regularization using leave-one-out data cross-validation

Lautaro Rayo and Ibrahim Hoteit

Citation: AIP Conference Proceedings **1479**, 1247 (2012); doi: 10.1063/1.4756379
View online: http://dx.doi.org/10.1063/1.4756379
View Table of Contents: http://scitation.aip.org/content/aip/proceeding/aipcp/1479?ver=pdfcov
Published by the AIP Publishing

Ensemble Kalman Filter Regularization Using Leave-One-Out Data Cross-Validation

Lautaro Rayo and Ibrahim Hoteit

King Abdullah University of Science and Technology (KAUST), MCSE, Thuwal 23955-6900. Saudi Arabia

Abstract. In this work, the classical leave-one-out cross-validation method for selecting a regularization parameter for the Tikhonov problem is implemented within the EnKF framework. Following the original concept, the regularization parameter is selected such that it minimizes the predictive error. Some ideas about the implementation, suitability and conceptual interest of the method are discussed. Finally, what will be called the *data cross-validation regularized EnKF* (dCVr-EnKF) is implemented in a 2D 2-phase synthetic oil reservoir experiment and the results analyzed.

Keywords: Ensemble Kalman Filter, Cross-Validation, Regularization
PACS: 02.30.Zz, 02.70.Tt, 02.70.Uu

INTRODUCTION

Designed for data assimilation in non-linear systems, the Ensemble Kalman Filter (EnKF) [1] is a Monte Carlo implementation of the classical Kalman Filter [2]. That is to say, the state statistics are estimated, propagated in time and updated through an ensemble representation. The EnKF is a sequential algorithm in which the ensemble of state realizations is integrated forward in time following the model dynamics up to a certain time when data becomes available. At that moment, the ensemble is conditioned to, and only to, the newly arrived data. Once the ensemble of states has been updated, a new integration step begins. Most likely because of its relative success and simple implementation, the EnKF has been widely used for data assimilation in multiphase-flow models in porous media, see for example [3] and [4] for a review. The EnKF reduces the complexity of the inverse problem from the size of the state vector to the number of ensemble realizations. In exchange, the method is intrinsically affected by sampling errors in high dimensional systems, which in turn affect the filter performance [5]. Given the sequential nature of the EnKF, sampling errors in the forecast statistics caused by the use of a finite number of ensemble realizations may lead to a systematic underestimation of the uncertainty associated with the filter update through the assimilation process [6]. This may cause the filter to poorly acquaint for newly arrived data and eventually, to diverge [7].

Beyond sampling errors, the EnKF performance is also likely to be severely affected by a wrong prescription of the errors associated with the noisy data being assimilated. In order to address this issue, an extra degree of freedom can be added to the proposed inverse problem in the form of a regularization parameter. The goal is to tune a probably mis-specified observation error variance against a mis-represented forecast error covariance. Several approaches to adaptively estimate a regularization parameter have been proposed in the literature, see for example [8] and [9]. In this work, a data-space cross-validation method is proposed.

ENKF REGULARIZATION USING DATA CROSS-VALIDATION

Before each assimilation step, the EnKF with the Tikhonov regularization is solved, where the solution is damped by a regularizaton parameter ρ and biased towards the forecast ensemble mean state $\overline{\psi^f}$:

$$\min_{\psi} J(\psi;\rho) = \frac{1}{\rho^2}\|\psi - \overline{\psi^f}\|^2_{\hat{B}^{-g}} + \|H\psi - d\|^2_{R^{-1}}, \tag{1}$$

where $\hat{B}$ is the low-rank forecast background error covariance matrix and $\hat{B}^{-g}$ its general inverse. H is a linear operator extracting the expected observations out of the augmented state vector ψ. d is the vector of observations and R the observational error variance, assumed diagonal with coefficients r_i. The solution $\overline{\psi^a}$ of (1) is similar to the well-known

Numerical Analysis and Applied Mathematics ICNAAM 2012
AIP Conf. Proc. 1479, 1247-1250 (2012); doi: 10.1063/1.4756379
© 2012 American Institute of Physics 978-0-7354-1091-6/$30.00

EnKF update equation, where now the regularization parameter ρ is included:

$$\overline{\psi^a} = \overline{\psi^f} + \rho^2 \hat{B} H^T (R + \rho^2 H \hat{B} H^T)^{-1} (d - H\overline{\psi^f}) = \overline{\psi^f} + K_{(\rho)} d_*, \tag{2}$$

where $d_* = d - H\overline{\psi^f}$. The idea of the ordinary cross-validation approach [10] is to leave one, say *k-th*, observation out of the minimization problem in (1), so that the assimilated state $\overline{\psi^a}$ is now computed solving:

$$\min_{\psi} J(\psi;\rho) = \frac{1}{\rho^2} \| \psi - \overline{\psi^f} \|^2_{\hat{B}-g} + \sum_{i \neq k}^{m} ((H\psi)_i - d_i)^2 r_i^{-1}. \tag{3}$$

The solution of (3) will instead be called ψ_ρ^k and the notation $\Delta\psi_\rho^k = \psi_\rho^k - \overline{\psi^f}$ is introduced. The left-out observation can then be used for validation, and ideally the assimilated would accurately predict the missing data value. In the *leave-one-out* approach, ρ is chosen such that the predictive error V_0, measuring the total predictive error for all $\psi_\rho^k : k = 1, \ldots, m$, is minimized:

$$V_0(\rho) = \sum_{k=1}^{m} ([H\Delta\psi_\rho^k]_k - [d_*]_k)^2 r_k^{-1} \to \min_{\rho}. \tag{4}$$

For a given value of ρ, computing $V_o(\rho)$ would require solving m equations like (3). Generalized cross validation is a way to simplify this computation [10]. Let:

$$\tilde{d}_i = \begin{cases} (H\psi_\rho^k)_k & i = k \\ d_i & i \neq k. \end{cases} \tag{5}$$

Noticing that $(H\psi_\rho^k)_k = \tilde{d}_k$, then ψ_ρ^k , the solution of (3), also solves equation (1) when replacing d by the modified data $\tilde{d}$. Therefore, using the solution of such equation, shown in (2), one can write $\Delta\psi_\rho^k = K(\rho)\tilde{d}_*$. This expression will be used to derive a much simpler analytical formula for the predictive error $V_0(\rho)$. Indeed, by definition of $\tilde{d}$:

$$HK_{(\rho)}(\tilde{d}_* - d_*) = [HK_{(\rho)}]_{:,k}([\tilde{d}_*]_k - [d_*]_k), \tag{6}$$

where $[\;]_{:,k}$ and $[\;]_k$ denote respectively the *k-th* column of a matrix, or element of a vector (no summation). Taking the *k-th* component of the vectorial equality above:

$$[HK_{(\rho)}(\tilde{d}_* - d_*)]_k = [HK_{(\rho)}\tilde{d}_*]_k - [HK_{(\rho)}d_*]_k = [HK_{(\rho)}]_{k,k}([\tilde{d}_*]_k - [d_*]_k). \tag{7}$$

On the other hand, it is easy to see that $[HK_{(\rho)}\tilde{d}_*]_k = [\tilde{d}_*]_k = [H\Delta\psi_\rho^k]_k$ and that $[HK_{(\rho)}d_*]_k = [H\Delta\psi_\rho]_k$. Replacing these two expressions in (7) leads:

$$[HK_{(\rho)}]_{k,k} = \frac{[H\Delta\psi_\rho^k]_k - [H\Delta\psi_\rho]_k}{[H\Delta\psi_\rho^k]_k - [d_*]_k}. \tag{8}$$

Substracting both sides of equation (8) to 1, rearranging terms so that $[H\Delta\psi_\rho^k]_k - [d_*]_k$ is cleared on one side, and replacing it in the equation for the predictive error in (4), it is possible to write:

$$V_{0(\rho)} = \sum_{k=1}^{m} \left(\frac{[H\Delta\psi_\rho]_k - [d_*]_k}{1 - [HK_{(\rho)}]_{k,k}} \right)^2 r_k^{-1} = \sum_{k=1}^{m} \left(\frac{[(I - HK_{(\rho)})d_*]_k}{1 - [HK_{(\rho)}]_{k,k}} \right)^2 r_k^{-1}, \tag{9}$$

where now the estimation of $V_{0(\rho)}$ does not require solving the minimization problem in (4) m times, and can instead be easily computed given any ensemble forecast. Notice that the evaluation of (9) only requires the computation of the Kalman gain in the data space: $HK_{(\rho)} = \rho^2 H\hat{B}H^T(R + \rho^2 H\hat{B}H^T)^{-1}$; being therefore fairly inexpensive for reservoir problems where the number of observations assimilated is generally small. This could not be true for problems assimilating seismic data, but is definitely appropriate for most well log data-fed problems. Assimilation of larger data sets could be tackled using computationally cheaper cross-validation methods, see for example V-Fold CV [11].

The projection of the Kalman gain into the data space, $HK_{(\rho)}$, can be interpreted as the data resolution matrix from the theory of inverse problems [10]. As ρ increases, the measurement error covariance R is given less weight relative to the observational background error covariance $H\hat{B}H^T$, and the data will be more closely fit. Indeed, as $\rho \to \infty$, $\|HK_{(\rho)} - I\| \to 0$ and the noisy data is fit exactly. From equation (9), the predictive error will decrease as the regularization parameter increases as long as the correspondant fit compensates the loss in data predictability. Namely, the method prescribes *overfitting* if any marginal increase in the fit cannot be explained by the rest of the data. On the other hand, as ρ decreases, $H\hat{B}H^T$ is given less weight and eventually the incoming data will be ignored. The minimization of the predictive error will prescribe model *overconfidence* if any further increase in data predictability does not compensate the overall loosening of the data fit.

In certain applications, as the one we will present bellow, the regularization parameter found by minimizing the predictive error may provide an unphysical update at certain assimilation step. In order to avoid this, we set thresholds for a maximal ρ_t and minimal $1/\rho_t$ regularization parameter such that $0 < 1/\rho_t < 1 < \rho_t$, and study the effect of tuning this bound.

IMPLEMENTATION ON A 2D OIL-WATER SYNTHETIC RESERVOIR MODEL

The dCVr-EnKF was applied to a water flooding oil reservoir assimilation problem. The synthetic two-dimensional model is 720 x 720 x 30 ft. (equations averaged in $\hat{z}$ direction) and was discretized using 24 x 24 x 1 gridblocks. At the initial time, all pores are fully saturated with oil. A set of 4 producers located on the corners of the reservoir are opened inmediatelly after, while a central water injector maintains the reservoir pressure during the simulation. The producers and injector are rate-controlled. The synthetic model was run for 1560 days. Both the true log-permeability field and the ensemble of priors were sampled using the sequential gaussian simulation scheme [12] over an isotropic variogram estimated from the same arbitrary data at well locations. The true permeability field was sampled using a gaussian variogram with sill 0.6 and range 200 ft., while the ensemble of priors with a sill of 0.4 and a range of 100 ft. The porosity field in both cases was set constant and equal to 0.15. The observational error was assumed independent and normally distributed with mean 0 and standard deviation 3 stb/day both for oil and water production rates and 3 psi for bottom hole pressures. The augmented state vector contains 2320 variables; 576 values for pressure, horizontal and vertical transmissibility and water saturation, 4 well connectivity factors and expected simulated observations on the production wells: 4 water production rates (WPR), 4 oil production rates (OPR) and 4 bottom hole pressures (BHP).

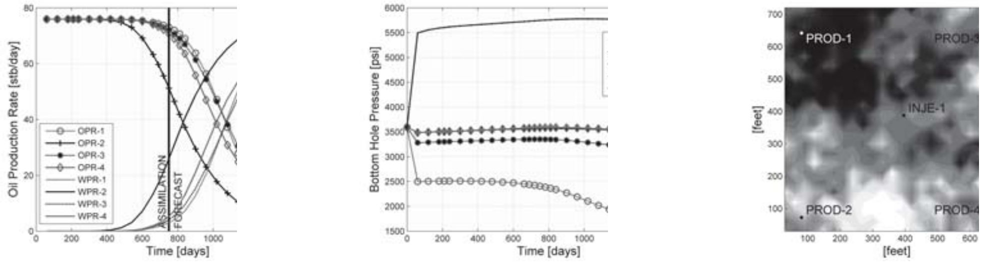

FIGURE 1. **Synthetic model. [left]** oil and water production rates, vertical line indicates the time where the shown data stops being assimilated. **[center]** bottom hole pressure. **[right]** log-permeability field for the true synthetic case and well locations.

As depicted in figure 1[left], data from the true synthetic model was assimilated at 15 reporting times: the first one at time 60 days and the final one at 750 days. The ensemble of assimilated model realizations was then integrated from time zero on what we call a rerun. Figure 2 shows the field oil production rate (FOPR), defined as the sum of the rates from all the 4 producers, forecasted by 50 members without data assimilation [left] and the rerun of such members after assimilation [center] using the standard EnKF_{50} (with constant regularization parameter $\rho = 1$).

RESULTS & DISCUSSION

The dCVr-EnKF method proposed improved the quality of the assimilated model when compared to the standard EnKF, providing better forecasts and more accurate estimates of the original permeability field. The performance of the dCVr-EnKF was found sensitive to the threshhold parameter ρ_t. Once ρ_t is well tuned, the dCVr-EnKF outperformed the standard EnKF with any constant regularization parameter. The root mean square error (RMS) of the FOPR (Field Oil Production Rate, combined production from all wells) relative to the true synthetic model (FOPR-RMS) was integrated from time 750 up to time 1560 as an estimator of the assimilated model forecast quality. Figure 3 [left]

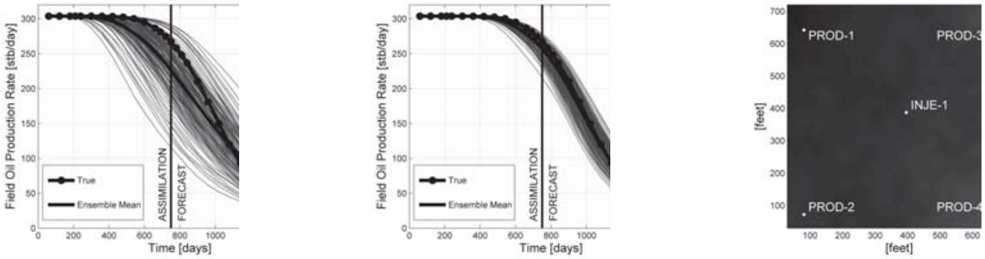

FIGURE 2. **EnKF$_{96}$ run:** standard ($\rho = 1$) 96 members EnKF. [left] FOPR of the initial non-assimilated set of perturbed realizations. [center] FOPR of the rerun of the assimilated realizations. [right] initial log-permeability ensemble mean.

compares the integrated RMS of the FOPR as it results from the standard EnKF with and without regularization and from the dCVr-EnKF using 40,56,72 and 96 ensemble members. The results are shown as a fraction of the 96 members standard EnKF$_{96}$. Horizontal lines show the performance of the standard EnKF for reference. The best threshold parameter for the dCVr-EnKF$_{96}$ was $\rho_t{*} = 1.50$, resulting in an improvement of 20% respect to the standard EnKF$_{96}$. The best forecast achieved using dCVr-EnKF$_{72}$ compares to the standard EnKF$_{96}$ and the same is observed for dCVr-EnKF$_{54}$ in comparison to the standard EnKF$_{72}$. Tuning the threshold parameter of the dCVr-EnKF$_{40}$ outperforms the standard EnKF$_{54}$ and shows comparable results to the standard EnKF$_{72}$.

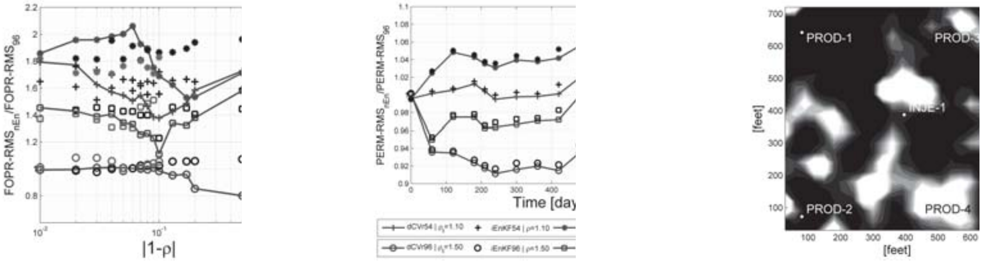

FIGURE 3. **dCVr-EnKF results:** [left] dCVr-EnKF method forecast quality as a function of ρ_t and compared to the EnKF with constant regularization parameter: iEnKF$_{nEn}$ for $\rho > 1$ and dEnKF$_{nEn}$ for $\rho < 1$. [center] see text for details [right] dCVr-EnKF$_{96}$ assimilated ensemble mean log-permeability field.

The best threshold parameter for the dCVr-EnKF was chosen from Figure 3 [left] for each number of ensemble members tested. The assimilation history of the log-perm field corresponding to these tuned dCVr-EnKF runs are shown in Figure 3 [center] and compared to the EnKF with constant regularization parameter equal to the tuned threshhold parameter. The estimation error was evaluated using the RMS of the log-perm field relative to the true synthetic model. The results are shown as a fraction of the standard EnKF$_{96}$. The dCVr-EnKF gives a better estimate of the log-perm field but does not prevent the over-shooting observed in the runs with 40, 54 and 72 ensemble members.

Acknowledgments: The authors acknowledge Schlumberger for the donation of multiple ECLIPSE licenses.

REFERENCES

1. G. Evensen, *Journal of Geophysical Research* **99(C5)**, 10143–10162 (1994).
2. R. Kalman, *Transactions of the ASME - Journal of Basic Engineering* **82(D)**, 35–45 (1960).
3. S. I. Aanonsen, G. Naevdal, D. S. Oliver, A. C. Reynolds, and B. Valles, *SPE Journal* **14(3)**, 393–412 (2009).
4. D. S. Oliver, and Y. Chen, *Computational Geosciences* **15**, 185–221 (2011).
5. G. Evensen, *Data Assimilation, The Ensemble Kalman Filter*, Springer, 2009, 2nd edn.
6. P. J. van Leeuwen, *American Meteorological Society* **127**, 1374–1377 (1999).
7. T. M. Hamill, J. S. Whitaker, and C. Snyder, *Monthly Weather Review* **129**, 2776–2790 (2001).
8. H. Li, E. Kalnay, and T. Miyoshi, *Q.J.R. Meteorol. Soc.* **135**, 523–533 (2009).
9. X. Wang, and C. H. Bishop, *Journal of the Atmospheric Sciences* **60**, 1140–1158 (2003).
10. R. C. Aster, B. Borchers, and C. H. Thurber, *Parameter Estimation and Inverse Problems*, Elsevier Academic Press, 2005.
11. S. Geisser, *Journal of the American Statistical Association* **60**, 320–328 (1975).
12. C. V. Deutsch, and A. G. Journel, *GSLIB Geostatistical Software Library and User's Guide*, Oxford University Press, 1992.