Robust method for EnKF in the presence of observation outliers/Multivariate localization methods for EnKF

Mikyoung Jun

Department of Statistics
Texas A&M University

May 19, 2015
Setup

- \( y_t \in \mathbb{R}^p \): observations at time \( t \)
- \( x_t \in \mathbb{R}^n \): unobservable state at time \( t \)
- Nonlinear system eq. \( x_t = M(x_{t-1}) + e_t \)
- Observation eq. \( y_t = H_t x_t + \epsilon_t \)
- System error \( e_t \sim N_n(0, Q_t) \)
- Observation error \( \epsilon_t \sim N_p(0, R_t) \)
- System and observation errors are uncorrelated.
- \( M, H_t, Q_t, R_t \) assumed to be known.
Definitions

- **$M$-member background ensemble**
  \[ \mathbf{x}^b = \{ \mathbf{x}^{b(k)} : k = 1, \ldots, M \} \in \mathbb{R}^{n \times M} \]

- background mean \( \bar{\mathbf{x}}^b = \frac{1}{M} \sum_{k=1}^{M} \mathbf{x}^{b(k)} \)

- ensemble-based estimate of the background error covariance
  \[ \mathbf{P}^b = \frac{1}{M-1} \sum_{k=1}^{M} \mathbf{x}^{b(k)} [\mathbf{x}^{b(k)}]^T, \text{ where } \mathbf{x}^{b(k)} = \mathbf{x}^{b(k)} - \bar{\mathbf{x}}^b \]

- analysis mean \( \bar{\mathbf{x}}^a = \bar{\mathbf{x}}^b + \mathbf{K}(\mathbf{y} - \mathbf{H}\bar{\mathbf{x}}^b) \)

- analysis covariance \( \mathbf{P}^a = (\mathbf{I} - \mathbf{K}\mathbf{H})\mathbf{P}^b \)

- Kalman gain \( \mathbf{K} = \mathbf{P}^b\mathbf{H}^T (\mathbf{H}\mathbf{P}^b\mathbf{H} + \mathbf{R})^{-1} \)
Outline

Part 1: Multivariate Localization for EnKF

Part 2: Robust EnKF in the presence of observation outliers
Motivation

Joint work with Soojin Roh, Istvan Szunyogh, and Marc Genton

- Localization: Schur (elementwise) product of $P^b$ and a localization matrix from a compactly supported correlation function $\rho(\cdot)$
- In statistics, localization function is called “taper” or compactly supported covariance function; needs to be positive definite
- For multivariate state variables, current practice is to apply the same localization function to each “block” of $P^b$
- Does it not matter? ($K = (P^bH^T(HP^bH + R)^{-1}$)
- Kang et al. (2011, JGR) zeroes out covariances between physically unrelated variables (not about positive-definiteness)
Motivation

- Problem of rank deficiency:
  - Localization matrix \( \begin{pmatrix} L & L \\ L & L \end{pmatrix} \)
  - Problem is more serious when \( P_{ij}^b \)'s are “significantly” non-zero
- We need \( \rho(\cdot) = \{\rho_{ij}(\cdot)\}_{i,j=1,...,N} \): matrix-valued correlation (positive definite) function, \( N \) : number of state variables – multivariate version of Gaspari-Cohn functions?
- In statistics literature, not many known such “valid” \( \rho \) (parametric) functions are available yet
One simple idea

- Use $\rho_{ij}(\cdot) = \beta_{ij} \cdot \rho(\cdot)$ with $|\beta_{ij}| < 1$, $|\beta_{ji}| < 1$, and $\beta_{ii} = \beta_{jj} = 1$.

- $\begin{pmatrix} 1 & \beta \\ \beta & 1 \end{pmatrix}$ is positive-definite and of full rank for any $\beta$ with $|\beta| < 1$.

- For $\rho$, use any localization functions in Gaspari and Cohn (1999).

  - Example:

    $\rho(d;c) = \begin{cases} 
    -\frac{1}{4}(|d|/c)^5 + \frac{1}{2}(d/c)^4 + \frac{5}{8}(|d|/c)^2 - \frac{5}{3}(d/c)^2 + 1, & 0 \leq |d| \leq c; \\
    \frac{1}{12}(|d|/c)^5 - \frac{1}{2}(d/c)^4 + \frac{5}{8}(|d|/c)^3 + \frac{5}{3}(d/c)^2 - 5(|d|/c) + 4 - \frac{2}{3}c/|d|, & c \leq |d| \leq 2c; \\
    0, & 2c \leq |d| 
    \end{cases}$

and $\beta_{11} = \beta_{22} = 1$, $0 \leq \beta_{ij} \leq 1$.

- This multivariate localization function is **separable** in the sense that

  multivariate component (in the above example, $\begin{pmatrix} 1 & \beta \\ \beta & 1 \end{pmatrix}$)

  and

  localization function (in the above example, $\rho$)

  are factored
Another idea

- Use one of a few multivariate compactly supported correlation functions available in statistics literature.
  - e.g. Bivariate Askey function (Porcu et al. 2012)
    \[ \rho_{ij}(d; \nu, c) = \beta_{ij} \left(1 - \frac{d}{c}\right)^{\nu + \mu_{ij}}, \]
  - \( c > 0, \mu_{12} = \mu_{21} \geq \frac{1}{2} (\mu_{11} + \mu_{22}), \nu \geq \left[ \frac{1}{2} s \right] + 2, \) and \( s \) is space dimension.
  - \( |\beta_{ij}| \leq \frac{\Gamma(1+\mu_{12})}{\Gamma(1+\nu+\mu_{12})} \sqrt{\frac{\Gamma(1+\nu+\mu_{11})\Gamma(1+\nu+\mu_{22})}{\Gamma(1+\mu_{11})\Gamma(1+\mu_{22})}}, \beta_{ii} = \beta_{jj} = 1 \)
  - \( |\beta_{ij}| \leq 1 \) if \( \mu_{11} = \mu_{22}. \)
Experiment with bivariate Lorenz Model

- $X_k$ and $Y_{j,k}$ are equally spaced on a latitude circle ($j = 1, \ldots, J$ and $k = 1, \ldots, K$).
- With boundary conditions $X_{k \pm K} = X_K$, $Y_{j,k \pm K} = Y_{j,k}$, $Y_{j-J,k} = Y_{j,k-1}$, and $Y_{j+J,k} = Y_{j,k+1}$,

$$\frac{dX_k}{dt} = -X_{k-1}(X_{k-2} - X_{k+1}) - X_k - \frac{ha}{b} \sum_{j=1}^{J} Y_{j,k} + F,$$

$$\frac{dY_{j,k}}{dt} = -abY_{j+1,k}(Y_{j+2,k} - Y_{j-1,k}) - aY_{j,k} + \frac{ha}{b}X_k$$
Part 1: Multivariate Localization for EnKF

Experiment details

- True model states generated by a long time step integration of the model
- Initialize ensembles by adding Gaussian noise to the true state
- We discard first 3000 time steps
- Simulated observations are generated by adding mean zero noise (variance 0.02 for $X$ and 0.005 for $Y$) to the truth
- 20 ensemble members are used (we tested 40 ensembles as well)
- Covariance inflation of 1.015
- RMSE calculated using last 1000 time steps and we repeat 50 times to produce boxplots
Bivariate Lorenz Model

- 36 variables of $X$, 360 variables of $Y$, $a = 10$, $b = 10$, $h = 2$

locations

longitudinal profiles
Part 1: Multivariate Localization for EnKF

Experiment set up

- **Two scenarios** for observation
  1. Observe 20% of $X$ and 90% of $Y$ at locations where $X$ is not observed.
  2. Fully observe $X$ and $Y$

- **Four localization schemes**
  
  - **S1** No localization
  - **S2** No localization and let $P_{12}^b = P_{21}^b = 0$.
  - **S3** localize $P_{11}^b$ and $P_{22}^b$, but let $P_{12}^b = P_{21}^b = 0$.
  - **S4** localize $P_{11}^b, P_{22}^b, P_{12}^b, P_{21}^b$
Localization (S4)

1. Gaspari-Cohn function: \( \rho_{ij}(d; c) = \beta_{ij} \rho(d; c) \), \( i, j = 1, 2 \), where

\[
\rho(d; c) = \begin{cases} 
-\frac{1}{4} (|d|/c)^5 + \frac{1}{2} (d/c)^4 + \frac{5}{8} (|d|/c)^3 - \frac{5}{3} (d/c)^2 + 1, & 0 \leq |d| \leq c; \\
\frac{1}{12} (|d|/c)^5 - \frac{1}{2} (d/c)^4 + \frac{5}{8} (|d|/c)^3 + \frac{5}{3} (d/c)^2 - 5(|d|/c) + 4 - \frac{2}{3} c/|d|, & c \leq |d| \leq 2c; \\
0, & 2c \leq |d| 
\end{cases}
\]

and \( \beta_{11} = \beta_{22} = 1, 0 \leq \beta_{ij} \leq 1. \) (support=2c)

2. Bivariate Askey function

\[
\rho_{ij}(d; c) = \beta_{ij} \left( 1 - \frac{|d|}{c} \right)^{\nu + \mu_{ij}}, \quad i, j = 1, 2
\]

with \( \mu_{11} = 0, \mu_{22} = 2, \mu_{ij} = 1, \nu = 3, \) and \( \beta_{11} = \beta_{22} = 1, \)

\( 0 \leq \beta_{ij} \leq 0.7. \) (support=c)
Results for $X$ in scenario 1

Part 1: Multivariate Localization for EnKF

**Support 50**

- Gaspari–Cohn
- Askey

**Support 70**

**Support 100**

**Support 160**

RMSE

M. Jun (TAMU)
Results for $X$ in scenario 1

- S1 vs S2: ignoring cross-covariance is better than not doing localization
- S3 is worse than S1
- S4 with $\beta = 0.01$ performs the best regardless of localization radius
- Askey seems to be better than the Gaspari-Cohn function
Part 1: Multivariate Localization for EnKF

Results for $Y$ in scenario 1

- **support 50**: {support values, RMSE values}
- **support 70**: {support values, RMSE values}
- **support 100**: {support values, RMSE values}
- **support 160**: {support values, RMSE values}

Key:
- Gaspari–Cohn
- Askey

M. Jun (TAMU)  Multivariate Localization/Robust Methods for EnKF  May 19  17 / 32
Results for $Y$ in scenario 1

- Askey clearly performs better
- Smaller localization radius is advantageous
- S3 is better than S1 or S2 and S4 performs only slightly better (only for lowest value of support)
Results for X in scenario 2

Part 1: Multivariate Localization for EnKF

M. Jun (TAMU)  Multivariate Localization/Robust Methods for EnKF  May 19
Results for X in scenario 2

- Less sensitive to the localization radius compared to Scenario 1
- State estimates are more accurate
- S3 and S4 are comparable and better than S1 and S2
Results for $Y$ in scenario 2

Part 1: Multivariate Localization for EnKF

![Graphs showing RMSE for different supports and localization methods](image)

- Support 50
- Support 70
- Support 100
- Support 160

Legend:
- Gaspari–Cohn
- Askey

- $5e^{-3}$
- $1e^{-2}$
- $0.1$
- $0.4$
- $0.7$
- $1$
Results for $\mathbf{Y}$ in scenario 2

- Best result by Askey with a short localization radius
- S3 and S4 perform similarly
Some issues

- More flexible multivariate tapers? (different localization length for each state variable)
- Estimation of “tuning parameters”
- Experiment with more realistic system
Outline

Part 1: Multivariate Localization for EnKF

Part 2: Robust EnKF in the presence of observation outliers
Motivation

Joint work with **Soojin Roh, Istvan Szunyogh, Marc Genton**, and **Ibrahim Hoteit** (MWR 2013)

- Ensemble Kalman filter is known to be not robust to observational outliers (Ruckdeschel 2010, Luo and Hoteit 2011)
Robust Ensemble Kalman Filter (REnKF)

- What’s been done in practice: usually discard suspicious observations.
- Huberization (Ruckdeschel 2010)

\[ \hat{x}^a = \bar{x}^b + K G_c(y - H\bar{x}^b), \]

where for any \( c \in \mathbb{R}^p_+ \) and \( u \in \mathbb{R}^p \), the Huber function \( G_c(u) \) is

\[
\{ G_c(u) \}_i = \begin{cases} 
  u_i, & \text{if } |u_i| < c_i, \\
  c_i, & \text{if } u_i \geq c_i, \\
  -c_i, & \text{if } u_i \leq -c_i.
\end{cases}
\]
Choosing $c$

1. **Efficiency criterion**

\[
\frac{E|\mathbf{x} - \bar{x}^a|^2}{E|\mathbf{x} - \hat{x}^a|^2} = \delta,
\]

for a given efficiency $\delta \in (0, 1)$.

Denominator = $E| (\mathbf{x} - \bar{x}^b) - (\mathbf{K}_i G_{ci}) [\mathbf{H} (\mathbf{x} - \bar{x}^b) + \epsilon]_i |^2$, where $\mathbf{x} - \bar{x}^b \sim N_n(0, \mathbf{P}^b)$.

2. **Radius criterion**

\[
E(\|(\mathbf{y} - \mathbf{H}\bar{x}^b)_i| - c_i)_+ : c_i = r : (1 - r),
\]

for a given radius $r \in (0, 1)$. Here, $r$ is a proportion of the amount of clipping in the innovation.

- **Computing a common $c$**
  - Use $\lim_{t \to \infty} \mathbf{P}_t^b$ with sufficiently large ensemble size.
Effects of Outliers

1. Additive outliers (AO)

\[ y_t = Hx_t + \xi_t + \epsilon_t, \]

where \( \xi_t \in \mathbb{R}^p \) is the outlying value.

2. Innovations outliers (IO)

\[ \epsilon_t \sim (1 - \alpha)N_p(0, R_t) + \alpha N_p(0, k_t \cdot R_t), \]

where \( 0 < \alpha < 1, k_t = \text{diag}(k_{t1}, \ldots, k_{tp}) \), and some \( k_{ti} > 1 \).
Lorenz model

- 40-variable Lorenz 96 model with $F = 8$
- Observation eq. $y_t = x_t + \epsilon_t$ with $\epsilon_t \sim N_{40}(0, 0.05 \cdot I_{40})$
- Additive outliers $\xi_t = 10$ at variable $x_{11}$ at $t = 71 - 73$
- Traditional EnKF vs Huberizing vs Discarding

\[ \delta = 0.999 \quad r = 0.001 \]
Part 2: Robust EnKF in the presence of observation outliers

Lorenz model

- Additive outliers $\xi_{71} = 10$ at variables $x_{11}$
- Traditional EnKF vs Huberizing vs Discarding

![Box plots showing bias for various $\delta$ and $r$.]
Part 2: Robust EnKF in the presence of observation outliers

Lorenz model

- Innovations outliers with $k_{71} = 100$ and $\alpha = 0.2$ at variables $x_{11}$
- Traditional EnKF vs Huberizing vs Discarding

![Box plots showing bias for various $\delta$ and $r$ values.](image-url)
Summary

- REnKF reduces the estimation bias at the expense of increasing the error variance.
- Huberizing filter performs better than discarding suspect observations.
- Multivariate extension.