Particle Filters and EnsKF

NCAR, June 2010

Peter Bickel

Statistics Dept. UC Berkeley

(Joint work with Jeff Anderson, Thomas Bengtsson, Jing Lei, Bo Li,

Chris Snyder in different parts)
Outline

1 Schematics of climate forecasting
2 State space models and correspondence
3 Issues with application in forecasting
4 Kalman filter and EnsKF
5 Particle filters
6 Collapse of particle filters
7 The importance of localization
8 NLEAF and NLEAFq as compromises between particle filters and EnsKF
Main References


Climate Forecasting Schematic

- $X_t$: “state of climate”
- Modelled as dynamical system
  \[ D[X_t] = h(X_{t-1}) \]  
  $D$ = differential operator
- $X_t$ – very high dimensional. Dimension of order $10^3$ or higher.
Climate Forecasting Schematic

- $Y_t$: “observed variables”
- Simplest stochastic model
  \[ Y_t = HX_t + e \]
- Very high-dimensional
- $H$ linear
- $e_t$ Gaussian $\perp X_t$, $e_t \sim N(0, R)$. 
Actual:

1) Approximate (1) by computer model

2) Generate ensemble of $n$ starting points $\{x_{t-1}^{(j)u}\}_{j=1}^n$

3) Use computer model to generate forecast ensemble $\{x_t^{(j)}\}_{j=1}^n$

4) “Assimilate” $y_t$ to produce update ensemble $\{x_t^{(j)u}\}_{j=1}^n$
State Space Models

- A state space model consists of two sequences of random variables.
  - A (hidden) Markov chain \( (X_t : t \geq 0) \)
    \[ X_t \mid X_{t-1} \sim q(X_{t-1}, \cdot), \]
  - and a sequence of observations \( (Y_t : t \geq 1) \),
    \[ Y_t \mid X_t \sim g(\cdot ; X_t). \]
- Parameters: \( (q, g, p_0) \). \( p_0 \): the initial distribution of \( X_0 \).
- Also known as the hidden Markov model if \( X_t \) is discrete.

\[
\begin{align*}
\ldots & \rightarrow X_{t-1} \rightarrow X_t \rightarrow X_{t+1} \rightarrow \ldots \\
\downarrow & \quad \downarrow \quad \downarrow \\
\ldots & \rightarrow Y_{t-1} \rightarrow Y_t \rightarrow Y_{t+1} \rightarrow \ldots
\end{align*}
\]
Example of State Space Models

- In data assimilation
  - $X_t$: vector of the true weather condition.
  - $Y_t$: the observed weather data.
  - $q(\cdot, \cdot)$ determined by the dynamics, apparently delta function.

- Other applications: speech recognition, target tracking, DNA and protein sequences, finance, etc.
The filtering recursion

\[ \cdots \longrightarrow X_{t-1} \overset{q}{\longrightarrow} X_t \longrightarrow X_{t+1} \longrightarrow \cdots \]

\[ \downarrow \quad g \quad \downarrow \quad \quad \downarrow \quad \downarrow \]

\[ \cdots \quad Y_{t-1} \quad Y_t \quad Y_{t+1} \quad \cdots \]

- Start from \( p_{0|0} = p_0 \).
- Forecast: for any \( t \geq 1 \),
  \[ p_{t|t-1}(x_t) = \int p_{t-1|t-1}(x_{t-1})q(x_{t-1}, x_t)dx_{t-1}. \]
- Update: for any \( t \geq 1 \),
  \[ p_{t|t}(x_t|y_t) \propto p_{t|t-1}(x_t)g(y_t; x_t). \]
Correspondence of Schematic to State Space Models

1 \( \{ x_0^{(j)} \} \) sample from \( p_{0|0} \)

2 \( \{ x_t^{(j)} \} \) sample from \( p_{t|t-1} \)

3 (want) \( \{ x_t^{(j)u} \} \) from \( p_{t|t}(x_t|y_t) \)
Take $H = \text{Identity}$

\[
p_{t|t}(x_t|y_t) = \frac{p_{t|t-1}(x_t)g(y_t; x_t)}{\int p_{t|t-1}(x)g(y_t; x)dx}
\]

Problems:

1. $q$ is a delta function – not a real problem
2. $p_{t|t-1}(x_t)$ no analytic representation
3. $x$ is very high dimensional.
Correspondence of Schematic to State Space Models

Note: A sample from $p_{t|t-1}$ can be turned into one from $p_{t|t}(x_t|y_t)$ by rejective sampling.

a) Sample $x^{(1)}, \ldots, x^{(n)} \ N \gg n$ from $p_{t|t-1}$

b) Toss coin with probability of heads $= \frac{g(y_t; x^{(j)})}{\max_x g(y_t; x)}$

to determine whether $x^{(j)}$ is acceptable

c) $x^{(1)}_t = \text{first accepted } x^{(j)}$

But $N$ needs to be impossibly large.
The Gaussian Case: Kalman Filter

• Suppose: Gaussian forecast; linear observation

\[ X_{t \mid Y_{1}^{t-1}} \sim N(\mu_{t \mid t-1}, \Sigma_{t \mid t-1}); \]
\[ Y_{t} = HX_{t} + \epsilon_{t}, \epsilon_{t} \sim N(0, R). \]

• Update:

\[ X_{t \mid Y_{1}^{t}} \sim N(\mu_{t}, \Sigma_{t}). \]
\[ \mu_{t} = \mu_{t \mid t-1} + K_{t}(y_{t} - H\mu_{t \mid t-1}) \]
\[ \Sigma_{t} = (I - KH)\Sigma_{t \mid t-1} \]
where
\[ K_{t} = \Sigma_{t \mid t-1}H^{T}(H\Sigma_{t \mid t-1}H^{T} + R)^{-1}. \]
The Lorenz 63 system

- State vector $Z_\tau \in \mathbb{R}^3$.
  - $\frac{dZ_{\tau,1}}{d\tau} = \sigma(Z_{\tau,2} - Z_{\tau,1})$,
  - $\frac{dZ_{\tau,2}}{d\tau} = Z_{\tau,1}(\rho - Z_{\tau,3}) - Z_{\tau,2}$,
  - $\frac{dZ_{\tau,3}}{d\tau} = Z_{\tau,1}Z_{\tau,2} - \beta Z_{\tau,3}$.
- $\sigma = 10$, $\beta = 8/3$, $\rho = 28$. 
The ensemble Kalman filter (EnKF, Evensen 94)

- Assuming linear observation with Gaussian noise:
  \[ Y_t = HX_t + \epsilon_t, \quad \epsilon_t \sim N(0, R). \]
- Idea: pretend \( X_t \) to be Gaussian; use only linear relationship.

The EnKF update (with perturbed observation)

1. Start with a sample \( \{x_t^{(j)}\}_{j=1}^n \) from \( \hat{p}_{t|t-1} \).
2. Estimate the linear regression coefficient of \( X_{t|t-1} \) on \( Y_t \):
   \[ \hat{K}_t = \hat{\Sigma}_{t|t-1} H^T (H\hat{\Sigma}_{t|t-1} H^T + R)^{-1}, \] with sample cov \( \hat{\Sigma}_{t|t-1} \).
3. Background observations: \( y_t^{(j)} = Hx_t^{(j)} + \epsilon_t^{(j)}, \epsilon_t^{(j)} \sim N(0, R). \)
4. Update: \( x_t^{(j)u} = x_t^{(j)} + \hat{K}_t(y_t - y_t^{(j)}). \)

Remark: Another class of EnKF based on the square root filter is also widely used (Anderson 01, Bishop et al 01). These are different for population than EnKF above (for comparison see Lawson & Hansen 04; Lei & Bickel 10).
The EnKF

a) Isn’t the same as Kalman filter unless \( p_{t|t-1}(x_t) \) is Gaussian.

b) Doesn’t estimate \( p_{t|t}(x_t|y_t) \).

If \( n = \infty \), distribution of EnKF ensemble \( \tilde{p}_{t|t} \) is that of

\[
(I - K)X_t + K(Y_t + \epsilon)|Y_t = y_t,
\]

where \( X_t \sim p_{t|t-1}(\cdot) \), \( Y_t = X_t + e_t \), \( e \sim N(0, R) \) independent of \( (X_t, Y_t) \), which disagrees with that of \( X_t|Y_t = y_t \) except for Gaussian \( X_t \).
Sequential Monte Carlo filters

A fully nonparametric method (Gordon et al. 93; Liu & Chen 98).

- Main idea: sample \( \{ x_t^{(j)} \}_{j=1}^n \) independently from

\[
\hat{p}_{t|t}(x_t) \propto \hat{p}_{t|t-1}(x_t)g(y_t; x_t)
\]

\[
= \frac{1}{n} \sum_{j=1}^{n} q(x_{t-1}^{(j)}, x_t)g(y_t; x_t).
\]

- Many clever sampling techniques make SMC filters useful in practice (e.g., Künsch 05).

In climate forecasting: \( q \) not available in parametric form.
The Particle Filter: A Principled Approximation

(Most naive version)

• Given ensemble \( \mathcal{E}_t = \{x_t^{(j)}\}_{j=1}^n \)

• Importance sample proportional to \( g(y_t; x_t^{(j)}) \) from \( \mathcal{E}_t \), \( n \) times with replacement.

• Get:

\[
\mathcal{E}_t^{(j)u} = \{x_t^{(j)u}\}_{j=1}^n \text{ with } \{x_t^{(j)u}\} \text{ iid}
\]

\[
w_{\ell t} \equiv P[x_t^{(j)u} = x_t^\ell] = \frac{\varphi(y_t - x_t^\ell; R)}{\sum_{\ell=1}^n \varphi(y_t - x_t^\ell; R)}, \text{ where } \varphi(z; R) \text{ is density of } N(0, R).
\]
Theoretical Support for Method

(See, e.g., Kunsch (2005, Ann. Stat.))

If $t$ is fixed, $n \to \infty$, and

$$P^*(A) = n^{-1} \sum_{j=1}^{n} 1\{x_t^{(j)u} \in A\}$$

then $P^* \Rightarrow P \sim p_t|x_t(y_t)$.

In fact, even if $t \to \infty$ at a polynomial rate as $n \to \infty$. 
Collapse of Particle Filters

\[ w_{\ell_0 t} \equiv \max_{\ell} w_{\ell t} \to 1, \]

so that \( \sum_{\ell \neq \ell_0} w_{\ell t} \to 0. \)
The Difficulty

- High dimension (Snyder, Bengtsson et al, MWR, 2007)
- More subtly, high effective dimension.
Prototype

(i) \( X_{d \times 1} \sim \mathcal{N}(0, J_d) \)

(ii) \( Y = X + \varepsilon, \varepsilon \perp X, \varepsilon \sim \mathcal{N}(0, J_d) \)

\[
X \leftrightarrow \varepsilon_{1l}^{d}
\]

\[
Y|X \leftrightarrow \prod_{j=1}^{d} \phi(y_j - x_j) = q_{t}(y|x)
\]
\section*{Simulation}

\( \mathcal{N}_X \equiv \text{dimension of } X \)
\( \mathcal{N}_e \equiv \text{ensemble size} \)
\( \mathcal{N}_Y \equiv \text{dimension of } Y \)
\( X \sim \mathcal{N}(0, I) \)
\( Y = X + \epsilon \)
Forecast ensemble \( X^{(j)} \) iid \( \mathcal{N}(0, I) \)
\( B = 10^3 \) simulations.
Prototype

Histogram of max $w_i$ for $N_x = 10, 30, 100$ and $N_e = 10^3$ from the particle-filter simulations described in text

$[N_e = 10^3, x^t \sim N(0, I), N_y = N_x, H = I$ and $\epsilon \sim N(0, I)]$

$N_x \equiv d, N_e \equiv B, \frac{B}{d} = 100, 33, 10, \frac{\log_e B}{d} = 0, 69, 0.23, 0.069$
Claims

A. If \( \frac{\log B}{d} \rightarrow \infty \) and \( g \) is bounded,

\[
\left| \frac{1}{B} \sum_{b=1}^{B} g \left( \mathbf{x}^{(b)} \right) - \mathbb{E} g (\mathbf{X} | \mathbf{Y}) \right| \rightarrow_p 0
\]

where \( \mathbb{E} \) is under correct \( P(\mathbf{X} | \mathbf{Y}) \).
B. In prototype situation, if \( \frac{\log B}{d} \to 0 \),

\[
\mathbb{E}(w_{(B)}) = 1 - \frac{2}{\sqrt{5}} \sqrt{\frac{\log B}{d}} (1 + o(1)).
\]

In this case, for \( g \) bounded, continuous,

\[
\frac{1}{B} \sum_{b=1}^{B} g \left( X^{(b)} \right) \Rightarrow g \left( X^{(k)} \right)
\]

where \( k = \arg\min_{j} |X^{(j)}| \), \( X^{(j)} \in \mathcal{E}_{1}, X^{(j)} \sim \tilde{p}_{1} \)

• If $X_{p \times 1}$ is Gaussian but has support of dimension $d < p$, then collapse is determined by $d$ not $p$.

• Roughly we expect collapse to occur quickly unless $n \gg e^d$. 
Dimension Reduction Is Necessary

- Coordinates of $X$ represent spatial positions.
- Distant coordinates are approximately independent.
- Neighboring $X$ coordinates independent of far $Y$ coordinates given neighboring $Y$ coordinates.
The Lorenz 96 System

- State vector $Z_\tau \in \mathbb{R}^{40}$.
- For $k = 1, \ldots, 40$, 
  \[
  \frac{dZ_{\tau,k}}{d\tau} = (Z_{\tau,k+1} - Z_{\tau,k-2})Z_{\tau,k-1} - Z_{\tau,k} + 8.
  \]
Localization

- Lorenz 96 small blocks of adjacent coordinates
- Needed for EnKF as well
- Amount of localization: local block size from 7 to 9.
- Particle filter requires pasting together of pieces – not succeeded
A Reformulation of EnKF

• A simple fact: If \((X, Y)\) is Gaussian, then for any \(y, y'\),

\[
(X | Y = y) - E(X | Y = y) \stackrel{d}{=} (X | Y = y') - E(X | Y = y').
\]

• As a result, if \((x', y') \sim (X, Y)\), then for all \(y\)

\[
x' - E(X | Y = y') + E(X | Y = y) \sim (X | Y = y).
\]

• In EnKF, \(y^j_t = Hx^j_{t|t-1} + \epsilon^j_t\), so \((x^j_{t|t-1}, y^j_t) \sim (X_{t|t-1}, Y_t)\).

\[
x^j_t = x^j_{t|t-1} + \hat{K}_t(y_t - y^j_t)
= x^j_{t|t-1} - \hat{E}(X_t | y^j_t, p_{t|t-1}) + \hat{E}(X_t | y_t, p_{t|t-1})
\]

is approximately a random sample from \((X_t | y_t, p_{t|t-1})\), and
the Kalman adjustment \(\hat{K}_t y = \hat{E}(X_t | y, p_{t|t-1})\).
Reduce the “First Order Bias”

- The true conditional expectation is

\[
E(X_t|y_t, p_{t|t-1}) = \int xp_{t|t-1}(x)g(y_t; x)dx \quad \int p_{t|t-1}(x)g(y_t; x)dx.
\]

- A non-parametric estimator (“Importance Sampling”, Hammersley & Handscomb 65):

\[
\hat{E}(X_t|y_t, p_{t|t-1}) = \frac{\sum_{j=1}^{n} x^j_t|t-1} {\sum_{j=1}^{n} g(y_t; x^j_t|t-1)}
\]

\[= E (\text{particle filter update ensemble})
\]

(1)
The non-linear ensemble adjustment filter (NLEAF)

The NLEAF update step

Given the prediction sample \( \{x_{t|t-1}^j\}_{j=1}^n \sim \hat{p}_{t|t-1} \),

1. Compute \( y_t^j = Hx_{t|t-1}^j + \epsilon_t^j \) for \( j = 1, \ldots, n \).

2. Estimate \( \hat{E}(X_t|y_j, p_{t|t-1}) \) using importance sampling, for \( y = y_t, y_t^j, j = 1, \ldots, n \).

3. Update: \( x_t^j = x_{t|t-1}^j - \hat{E}(X_t|y_t^j, \hat{p}_{t|t-1}) + \hat{E}(X_t|y_t, \hat{p}_{t|t-1}) \).
Reduce Dimensionality

Localization is still needed but relevant dimension is that of $Y$ (localized) $+ 1$ only, not $(X, Y)$ localized.
Simulation set up

- Discretize: \( x_t = z_{\Delta t}, \ t = 0, 1, 2, \ldots \).
- Observation \( y_t = Hx_t + \epsilon_t, \ \epsilon_t \sim N(0, \sigma^2 I) \).
- \( y_t \) generated from hidden true trajectory \( x_t^* \).
- Approximation error:
  \[ \| \frac{1}{n} \sum_j x_t^j - x_t^* \|_2. \]
- Three different levels of difficulty:
  - Hard case (Bengtsson et al 03): \( \Delta = 0.4; H \) incomplete; \( \sigma^2 = 0.5 \).
  - Easy case (Ott et al 04): \( \Delta = 0.05; H = I; \sigma^2 = 1 \).
  - Intermediate case: \( \Delta = 0.05; H \) incomplete; \( \sigma^2 = 2 \).
- Ensemble size \( n \) must be a concern.
Numerical results for Lorenz 96 model: hard case

- $\Delta = 0.4$, $y_{t,k} = x_{t,2k-1} + \epsilon_{t,k}$, $\epsilon_{t,k} \sim N(0, 0.5)$, $k = 1, \ldots, 20$.

<table>
<thead>
<tr>
<th></th>
<th>NLEAF</th>
<th>EnKF</th>
<th>XEnsF</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>0.65</td>
<td>0.97</td>
<td>0.92</td>
</tr>
<tr>
<td>med</td>
<td>0.63</td>
<td>0.88</td>
<td>0.85</td>
</tr>
<tr>
<td>std</td>
<td>0.20</td>
<td>0.32</td>
<td>0.31</td>
</tr>
</tbody>
</table>

Table: Summary of App. Err. with $T=2000$, $n=400$.

Results of EnKF and XEnsF: Bengtsson et al (03).
Numerical results for Lorenz 96 model: intermediate case

- $\Delta = 0.05$, $y_{t,k} = x_{t,2k-1} + \epsilon_{t,k}$, $\epsilon_{t,k} \sim N(0,2)$, $k = 1, \ldots, 20$. 

RMSE vs ensemble size (LETKF: Ott et al 04)
Conclusion

- Particle filters have promise primarily when
  
  a) There has been successful localization or other drastic dimension reduction.
  
  b) There are well specified parametric models physics based or good approximations.

- There is room for compromises between linear filters and particle filters.