Statistics for Model/Data Fusion: An Introduction

Christopher K. Wikle

Department of Statistics

University of Missouri-Columbia
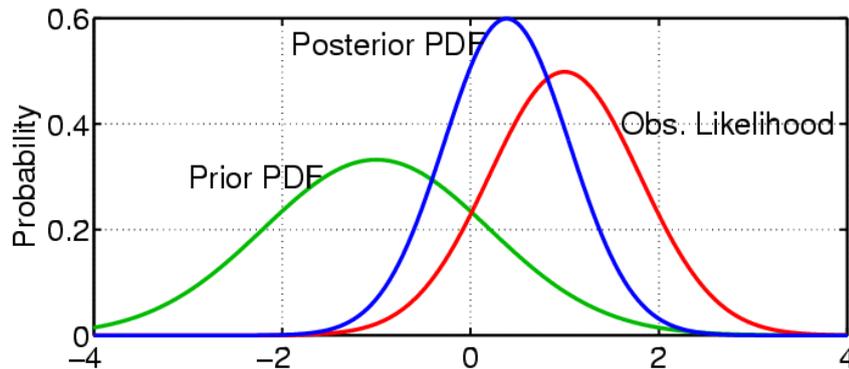
Talk Outline

- Simple Model/Data fusion (Data Assimilation) from a Bayesian Perspective

    – univariate motivation

    – multivariate (kriging, optimal interpolation)

- Sequential Approaches

    – Kalman filter

- Monte Carlo Methods

    – Sequential Importance Sampler (Particle Filter)

    – Ensemble Kalman Filter

NOTE: Many thanks to Jeff Anderson (NCAR) for the use of some of his figures (the nice ones!) from his talk "Ensemble Filtering for Geophysical Data Assimilation" at:

`http://www.samsi.info/200405/data/activity/workshop/index.html`

What is Data Assimilation?

- Combining Information

  - "interpolating fields for subsequent use as initial data in a model integration" (Bennett, 2002)

  - "statistical combination of observations and short-range forecasts" (Kalnay, 2003)

  - "using all the available information, to define as accurate as possible the state" (Talagrand, 1997)

- *Statistical Perspective:* Fusing data (observations) with prior knowledge (e.g., physical laws; model output) to get an estimate of the (distribution of) the true state of the physical system

  - Need statistical model for observations (data model; direct or indirect)

  - Need (prior) statistical model for the system (process model)

    * deterministic model (with additive errors)
    * long term history of observations (e.g., climatology)

Tropical Wind Example

NSCAT Winds: Nov 7 00Z 1998

NCEP Analysis Winds: Nov 7 00Z 1998
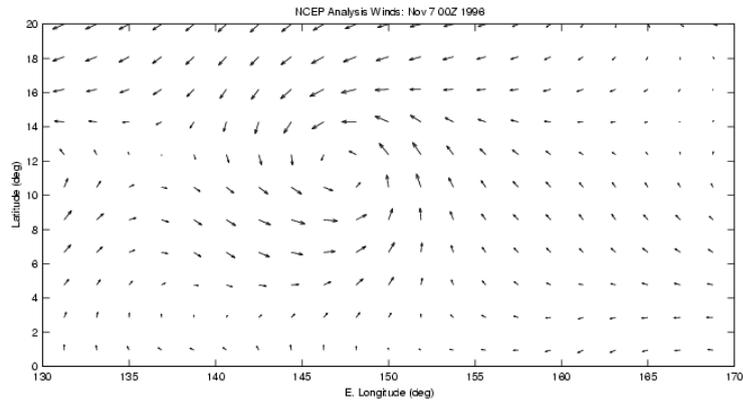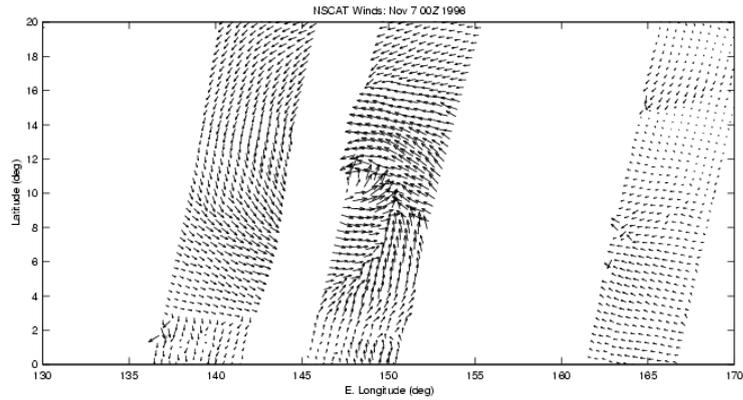
Bayesian Modeling

Notation:

- X - unobservable quantities of interest (e.g., true value of wind component)
- Y - observed data (e.g., wind observations at various locations)

The full probability model can always be factored into components:

$$
\begin{aligned}
p(X, Y) &= p(Y|X)p(X) \\
&= p(X|Y)p(Y)
\end{aligned}
$$

and thus

$$
p(X|Y) = \frac{p(Y|X)p(X)}{p(Y)} \quad \text{(Bayes' Rule)}
$$

provided $p(Y) \neq 0$

Primary Components of Bayesian Model

$$p(X|Y) = \frac{p(Y|X)p(X)}{p(Y)} \quad \text{(Bayes' Rule)}$$

- Prior distribution: $p(X)$

  - may be informative or non-informative, subjective or objective
  - we like to think of it as a "process or parameter model"
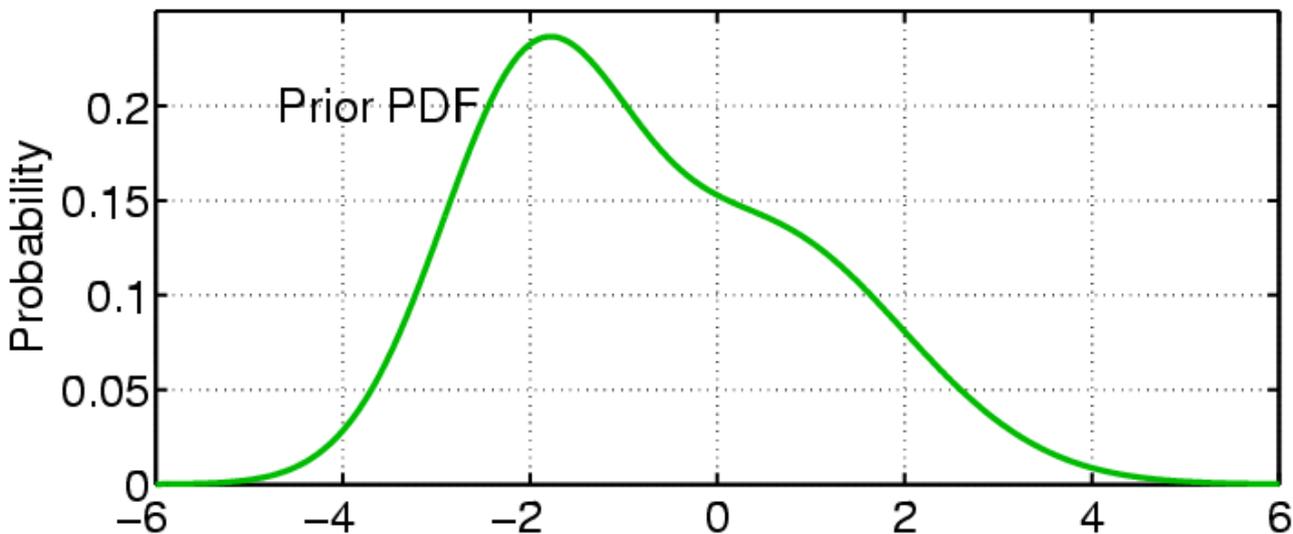
$$p(X|Y) = \frac{p(Y|X)p(X)}{p(Y)} \quad \text{(Bayes' Rule)}$$

- Data distribution: $p(Y|X)$

  – observation model, sampling distribution, measurement model
  – if viewed as a function of X, it is known as a likelihood function, $L(X|Y)$.
  – **KEY:** think of the data *conditioned* upon the process X (relatively simple)
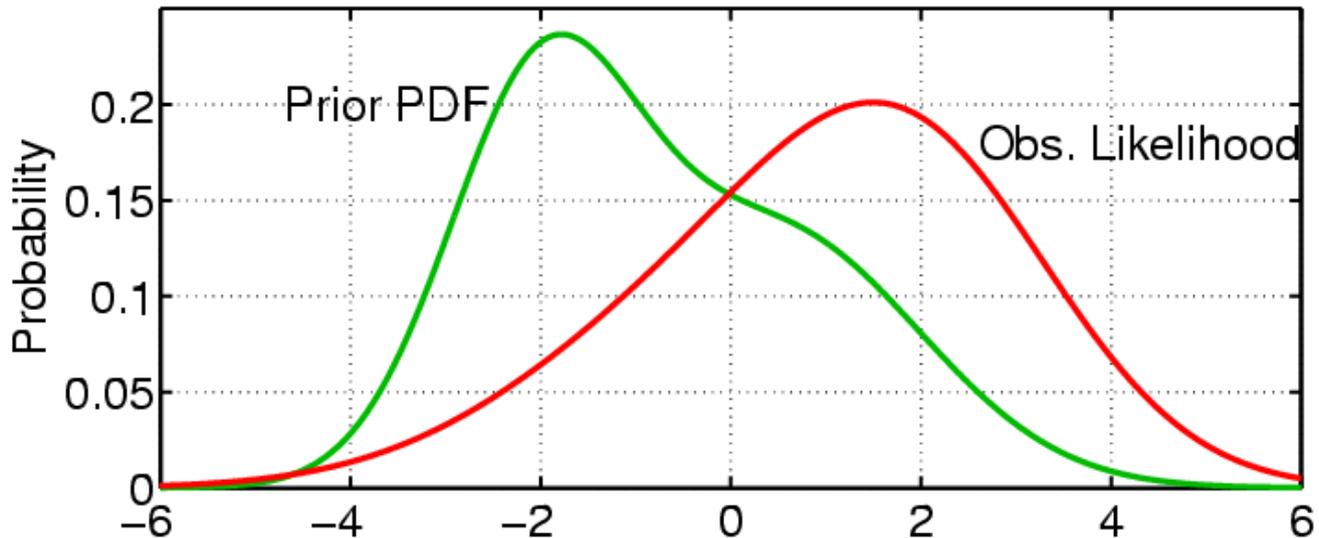
$$p(X|Y) = \frac{p(Y|X)p(X)}{p(Y)} \quad \text{(Bayes' Rule)}$$

- Marginal distribution: $p(Y) = \int p(Y|X)p(X)dX$ (assuming continuous X; analogous form for X discrete)

  - Also known as the *prior predictive distribution*
  - does not depend on X
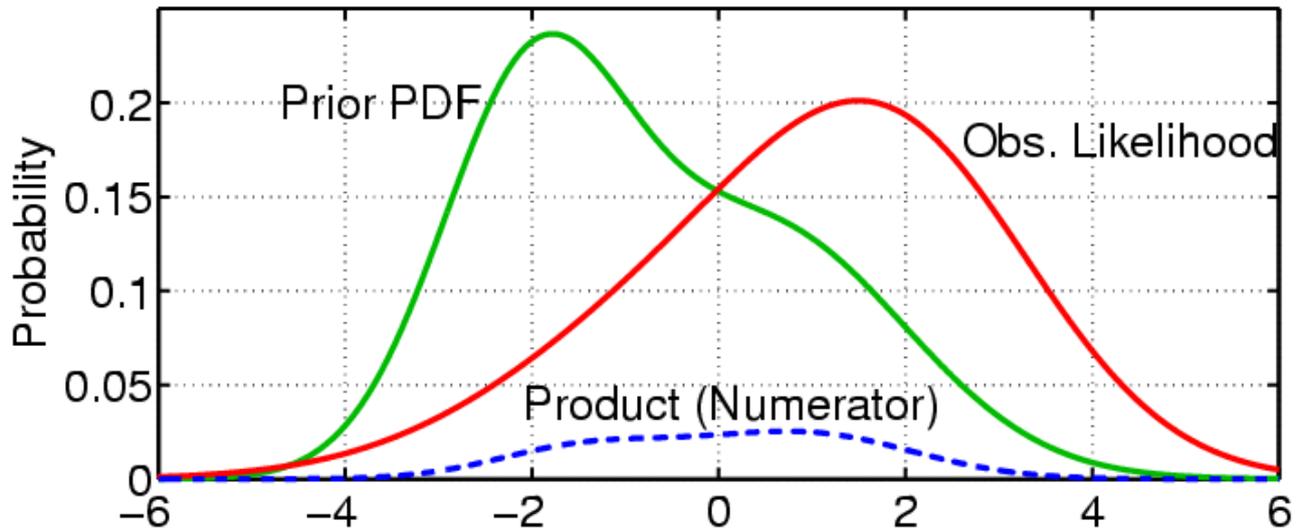
$$p(X|Y) = \frac{p(Y|X)p(X)}{p(Y)} \quad \text{(Bayes' Rule)}$$

- Marginal distribution: $p(Y) = \int p(Y|X)p(X)dX$ (assuming continuous X; analogous form for X discrete)

  – Also known as the *prior predictive distribution*

  – does not depend on X
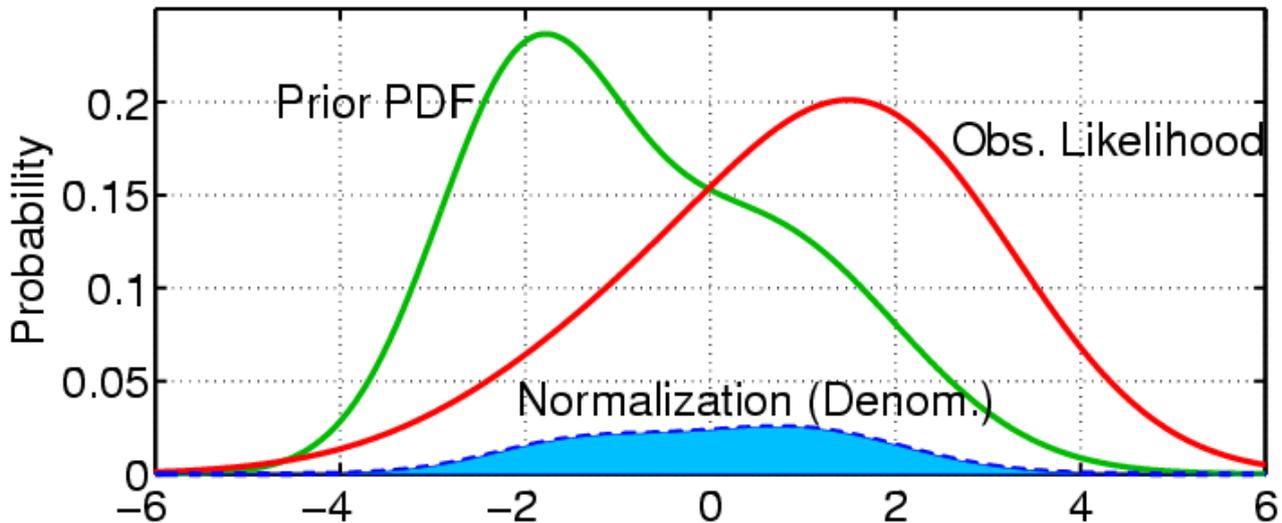
$$p(X|Y) = \frac{p(Y|X)p(X)}{p(Y)} \quad \text{(Bayes' Rule)}$$

- Posterior distribution: $p(X|Y)$

  – Our primary interest for inference
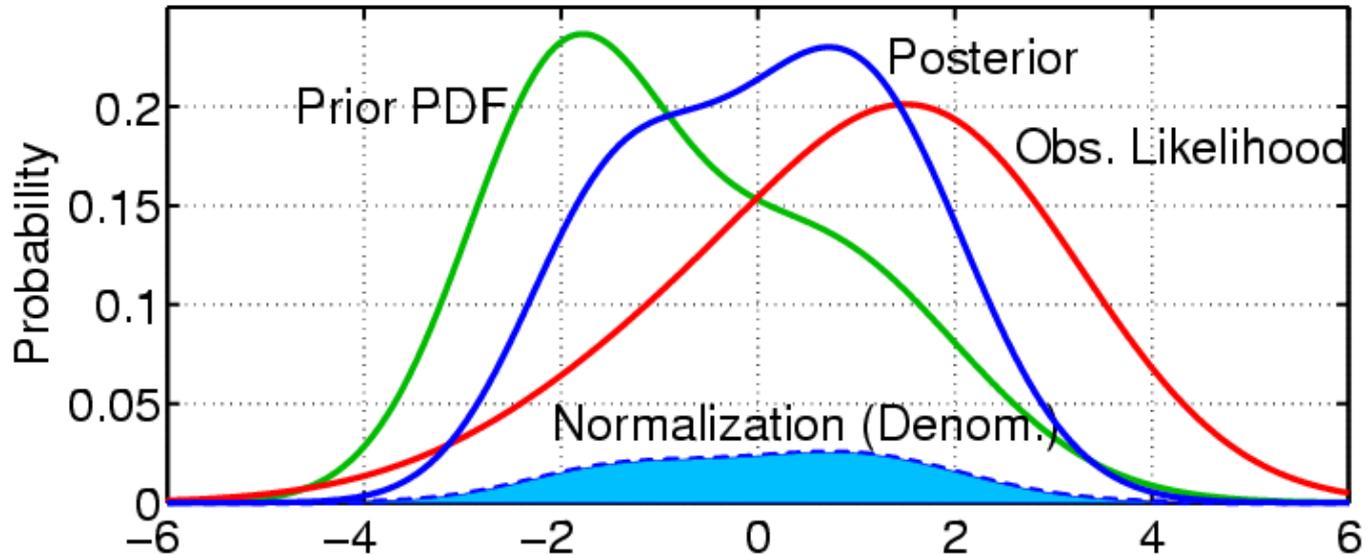
Summary: Bayesian Inference

To summarize, Bayesian statistics is based on a simple rule: **Bayes' Rule**

$$p(X|Y) = \frac{p(Y|X)p(X)}{\int p(Y|X)p(X)dX} \propto p(Y|X)p(X)$$

posterior distribution $\propto$ "likelihood" $\times$ prior distribution

$$p(process|data) \propto p(data|process)p(process)$$

- All inference for X based on the posterior, $p(X|Y)$

- Combines prior information and data

- (normalizing constant) $p(Y) = \int p(Y|X)p(X)dX$

- It does not alter the result if we multiply the likelihood by any constant (or any function of Y alone.) In this sense we can view the likelihood as any constant multiple of $p(Y|X)$.

- Note that as a function of Y the likelihood is a pdf but as a function of X it is not ($\int p(Y|X)dX$ may not be finite).

Simple Example: Normal Data and Prior Distributions

Say we are interested in the univariate state variable: x (e.g., u-component of wind at some location)

We have prior distribution: $x \sim N(\mu, \tau^2)$ (e.g., from a forecast model)

Conditioned on the true value of the process, x, we have n independent but noisy observations $\mathbf{y} = (y_1, \ldots, y_n)'$ and thus the data model: $y_i | x \sim N(x, \sigma^2)$. Then,

$$p(\mathbf{y}|x) = \prod_{i=1}^{n} (1/\sqrt{2\pi\sigma^2}) \exp\{-1/2(y_i - x)^2/\sigma^2\}$$

$$\propto \exp\{-1/2 \sum_{i=1}^{n} (y_i - x)^2/\sigma^2\}$$

From Bayes' rule: $p(x|\mathbf{y}) \propto p(\mathbf{y}|x)p(x)$
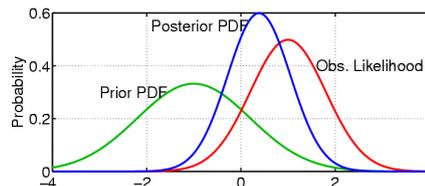
Univariate Normal-Normal Example (cont.)

Using the data and prior models,

$$p(x|\mathbf{y}) \propto \exp\{-1/2[\sum_{i=1}^{n}(y_i - x)^2/\sigma^2 + (x - \mu)^2/\tau^2]\}$$
$$\propto \exp\{-1/2[x^2(n/\sigma^2 + 1/\tau^2) - 2(\sum y_i/\sigma^2 + \mu/\tau^2)x]\}$$

NOTE: this is just the product of two Gaussian distributions. It can be shown (completing the square) that this product is also a Gaussian, so the posterior is:

$$x|\mathbf{y} \sim N((n/\sigma^2 + 1/\tau^2)^{-1}(\sum y_i/\sigma^2 + \mu/\tau^2), (n/\sigma^2 + 1/\tau^2)^{-1})$$

Examining the Posterior Mean/Variance

We showed that the posterior distribution for Normal data and Normal prior is:

$$x|\mathbf{y} \sim N\left((n/\sigma^2 + 1/\tau^2)^{-1}(\sum_{i=1}^{n} y_i/\sigma^2 + \mu/\tau^2), (n/\sigma^2 + 1/\tau^2)^{-1}\right),$$

The posterior variance is:

$$\begin{aligned}
\text{var}(x|\mathbf{y}) &= (n/\sigma^2 + 1/\tau^2)^{-1} \\
&= \left(1 - \frac{n\tau^2}{\sigma^2 + n\tau^2}\right)\tau^2
\end{aligned}$$

We call the inverse of the variance the *precision*. Thus, for normal data and normal prior, each with known precision, the posterior precision equals the prior precision plus the data (mean) precision.

Posterior Mean (cont.)

Note that we can write the posterior mean:

$$E(x|\mathbf{y}) = \frac{\sigma^2 \tau^2}{\sigma^2 + n\tau^2}(n\bar{y}/\sigma^2 + \mu/\tau^2)$$

$$= w_y \bar{y} + w_\mu \mu$$

where $\bar{y} = \Sigma_i y_i/n$, $w_y = n\tau^2/(n\tau^2 + \sigma^2)$, $w_\mu = \sigma^2/(n\tau^2 + \sigma^2)$, $w_y + w_\mu = 1$.

- Posterior mean is a weighted average of the prior mean (μ) and data mean($\bar{y}$).

- If our prior beliefs are uncertain ($\tau^2 \to \infty$) then the likelihood swamps the prior: $p(x|\mathbf{y}) \to N(\bar{y}, \sigma^2/n)$ (frequentist sampling dist)

- If we have lots of data ($n \to \infty$) then $p(x|\mathbf{y}) \to N(\bar{y}, 0)$

- If we have little data ($n \to 0$) then $p(x|\mathbf{y}) \to N(\mu, \tau^2)$

- If $\sigma^2 = \tau^2$, prior mean has same weight as 1 extra observation of value μ.

- If $\tau^2 \to 0$, the prior is infinitely more precise than the data

- if $\sigma^2 \to 0$, the data are perfectly precise

The posterior mean can also be written:

$$E(x|\mathbf{y}) = \mu + \left(\frac{n\tau^2}{\sigma^2 + n\tau^2}\right)(\bar{y} - \mu)$$

$$= \mu + K(\bar{y} - \mu),$$

where we say that the prior mean (μ) is adjusted toward the sample mean ($\bar{y}$) and $K = (n\tau^2)/(\sigma^2 + n\tau^2)$ is the "**gain**".

Analogously, the posterior variance can be rewritten:

$$\text{var}(x|\mathbf{y}) = (1 - K)\tau^2$$

where the posterior variance is updated from the prior variance according to the gain, K.

These last two equations are critical for understanding data assimilation.

Numerical Example 1

Say our prior distribution is $\mathbf{x} \sim N(20, 3)$, and our data model is $y_i|x \sim N(x, 1)$. In this case the data are relatively precise compared to the prior. We have two observations $\mathbf{y} = (19, 23)'$.

posterior mean = 20 + (6/7)(21-20) = 20.86
posterior variance = (1 - 6/7)3 = 0.43
The posterior distribution is: $x|\mathbf{y} \sim N(20.86, 0.43)$
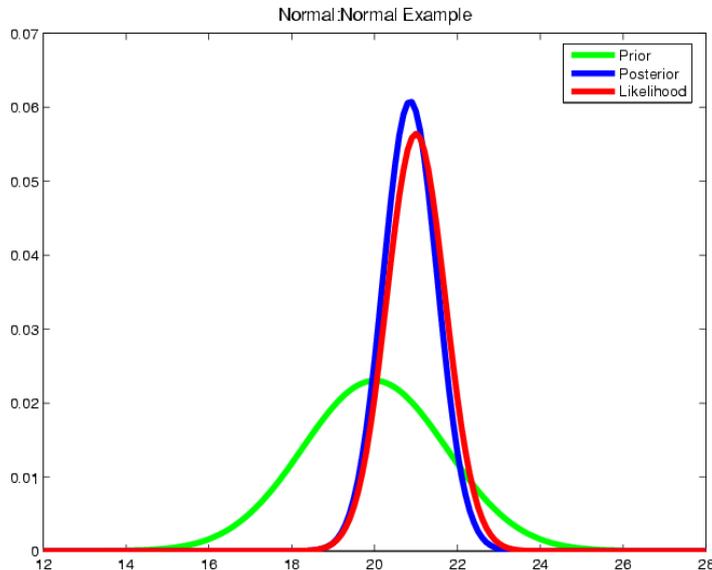
Numerical Example 2

Say our prior distribution is $\mathbf{x} \sim N(20, 3)$, and our data model is $y_i|x \sim N(x, 10)$. In this case the data are relatively imprecise compared to the prior. We have two observations $\mathbf{y} = (19, 23)'$.

posterior mean = 20 + (6/16)(21-20) = 20.375
posterior variance = (1 - 6/16)3 = 1.875
The posterior distribution is: $x|\mathbf{y} \sim N(20.375, 1.875)$
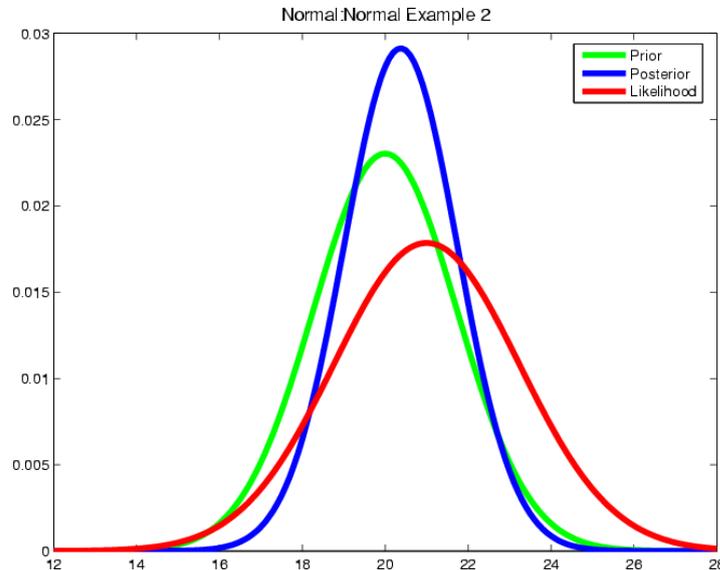
Mixture Priors

Suppose we have likelihood $p(y_i|x)$ and $p_1(x)$ and $p_2(x)$ are both conjugate densities that give rise to the posteriors $p_1(x|y_i)$ and $p_2(x|y_i)$, respectively.

Let w_1 and w_2 be any non-negative real numbers such that $w_1 + w_2 = 1$, and write the *mixture prior*:

$$p(x) = w_1 p_1(x) + w_2 p_2(x).$$
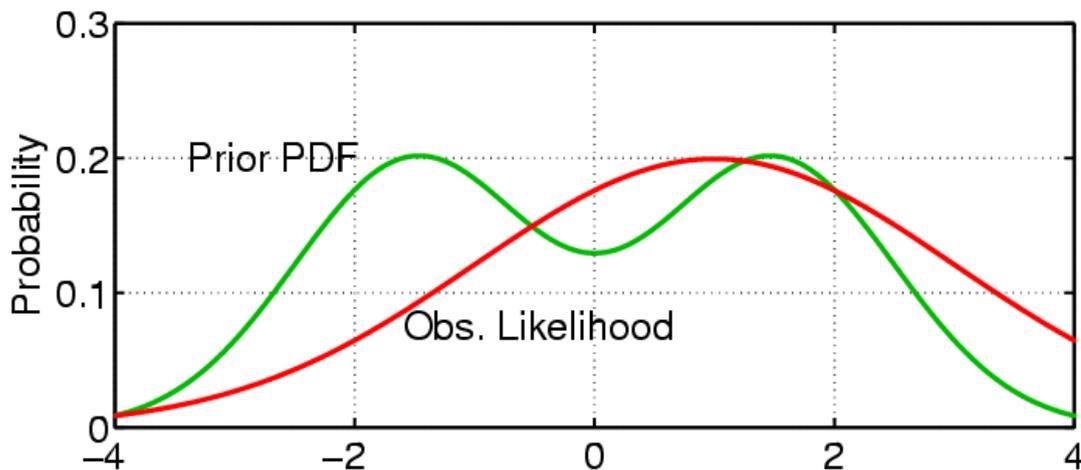
Then, it is easy to show that the posterior corresponding to $p(x)$ is:

$$p(x|\mathbf{y}) = w_1^* p_1(x|\mathbf{y}) + w_2^* p_2(x|\mathbf{y}),$$

where

$$w_i^* \propto w_i \int p(\mathbf{y}|x)p_i(x)dx, \quad i = 1, 2$$

with the constant of proportionality such that $w_1^* + w_2^* = 1$. (Note, generally, we can take any convex combination of two or more conjugate priors corresponding to a convex combination of respective posteriors.)
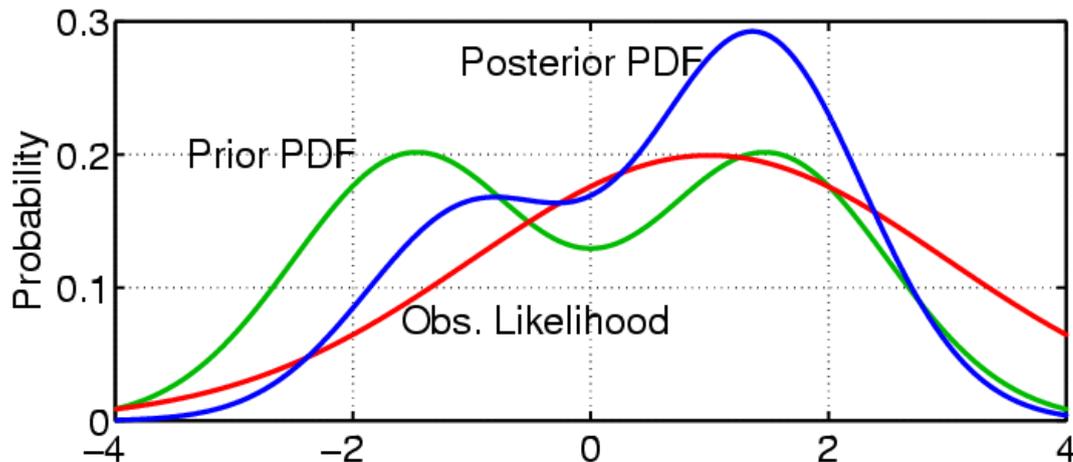
Multivariate Normal-Normal Case

Assume we are interested in the $n \times 1$ vector process $\mathbf{x}$ (e.g., u-winds at several locations), that has prior distribution:

$$\mathbf{x} \sim N(\boldsymbol{\mu}, \mathbf{P}),$$

where for now we assume that the mean $\boldsymbol{\mu}$ and variance/covariance matrix $\mathbf{P}$ are known.

In addition, we observe the $p \times 1$ data vector $\mathbf{y}$ and assume the following data model:

$$\mathbf{y}|\mathbf{x} \sim N(\mathbf{Hx}, \mathbf{R}),$$

where the $p \times n$ observation matrix $\mathbf{H}$, that maps the process (e.g., true wind component) to the observations (e.g., wind observations), and the observation error covariance matrix $\mathbf{R}$ are assumed to be known.

Multivariate Normal-Normal (cont.)

We are interested in the posterior distribution of $\mathbf{x}|\mathbf{y}$ which is given by:

$$p(\mathbf{x}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{x})p(\mathbf{x})$$

As with the univariate case, the product of two normals in this context is also normal, and thus the posterior distribution is:

$$\mathbf{x}|\mathbf{y} \sim N((\mathbf{H}'\mathbf{R}^{-1}\mathbf{H} + \mathbf{P}^{-1})^{-1}(\mathbf{H}'\mathbf{R}^{-1}\mathbf{y} + \mathbf{P}^{-1}\boldsymbol{\mu}), (\mathbf{H}'\mathbf{R}^{-1}\mathbf{H} + \mathbf{P}^{-1})^{-1})$$

Multivariate Normal-Normal Posterior Mean/Variance

As with the univariate case, we can rewrite the posterior mean:

$$E(\mathbf{x}|\mathbf{y}) = (\mathbf{H'R}^{-1}\mathbf{H} + \mathbf{P}^{-1})^{-1}\mathbf{H'R}^{-1}\mathbf{y} + (\mathbf{H'R}^{-1}\mathbf{H} + \mathbf{P}^{-1})^{-1}\mathbf{P}^{-1}\boldsymbol{\mu}$$

$$= \boldsymbol{\mu} + \mathbf{K}(\mathbf{y} - \mathbf{H}\boldsymbol{\mu})$$

where $\mathbf{K} = \mathbf{PH'}(\mathbf{R} + \mathbf{HPH'})^{-1}$ is the "**gain**".

Similarly, the posterior variance/covariance matrix can be written:

$$\text{var}(\mathbf{x}|\mathbf{y}) = (\mathbf{H'R}^{-1}\mathbf{H} + \mathbf{P}^{-1})^{-1}$$
$$= (\mathbf{I} - \mathbf{KH})\mathbf{P}$$

These formulas are the basis of DA!

Relationship to Kriging/Optimal Interpolation

We consider the relationship to Kriging (geostatistics)/Optimal Interpolation (meteorology,oceanography) by a simple example.

Assume $\mathbf{x} = [x(s_1), x(s_2), x(s_3)]'$ at spatial locations $s_i, i = 1, 2, 3$. Also, assume we have observations at s_2 and s_3 but not s_1: $\mathbf{y} = [y(s_2), y(s_3)]'$ and thus $\mathbf{H}$ is defined as:

$$\mathbf{H} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Assume the prior covariance matrix that describes the (forecast) error covariance matrix is given by:

$$\mathbf{P} = \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix}$$

Note that even though we only have observations for locations 2 and 3, it is critical that we have the covariance information between all state locations of interest (e.g., 1,2 and 3).

Kriging/OI example (cont.)

In this case, the "gain" is given by:

$$\mathbf{K} = \mathbf{PH}'(\mathbf{R} + \mathbf{HPH}')^{-1} = \begin{pmatrix} c_{12} & c_{13} \\ c_{22} & c_{23} \\ c_{32} & c_{33} \end{pmatrix} \left(\mathbf{R} + \begin{pmatrix} c_{22} & c_{23} \\ c_{32} & c_{33} \end{pmatrix} \right)^{-1}$$

For simplicity, assume $\mathbf{R} = \sigma^2 \mathbf{I}$ (i.e., independent measurement error; "nugget effect" in kriging).

Then, for example, the marginal posterior mean for $x(s_1)$ is given by:

$$E(x(s_1)|\mathbf{y}) = \mu(s_1) + w_{12}(y(s_2) - \mu(s_2)) + w_{13}(y(s_3) - \mu(s_3))$$

where the interpolation weights, $\mathbf{w}_1 = [w_{12}, w_{13}]'$, are given by:

$$\mathbf{w}_1' = \begin{pmatrix} c_{12} & c_{13} \end{pmatrix} \begin{pmatrix} c_{22} + \sigma^2 & c_{23} \\ c_{32} & c_{33} + \sigma^2 \end{pmatrix}^{-1}$$

Thus, the prior mean is adjusted by a weighted combination of the anomalies (difference between observation and prior mean) at each data location.

The marginal mean-squared prediction error (posterior variance) is given by:

$$\text{var}(x(s_1)|\mathbf{y}) = c_{11} - \begin{pmatrix} c_{12} & c_{13} \end{pmatrix} \begin{pmatrix} c_{22} + \sigma^2 & c_{23} \\ c_{32} & c_{33} + \sigma^2 \end{pmatrix}^{-1} \begin{pmatrix} c_{12} \\ c_{13} \end{pmatrix}$$

NOTE: Such spatial prediction (interpolation) is the optimal (best linear unbiased) prediction (assuming the parameters, $\mathbf{R}$, $\mathbf{Q}$ are known.)

- In spatial statistics this is known as *simple kriging* (Matheron 1963)

- In atmospheric/oceanographic science this is known as *optimal interpolation* (Gandin 1963).

It is relatively simple to accommodate more complicated (unknown) prior means (*ordinary kriging* if mean is constant but unknown; *universal kriging* if mean is a linear function of covariates). These methods are easily expressed in the framework of *linear mixed models* in statistics or as variational (optimization) problems.

Numerical Example

Assume we have two observations $y_2 = 16$, $y_3 = 23$ and we are interested predicting the true process x_i at these locations and a third location x_1. Our prior mean is $\mu_i = 18$, $i = 1, 2, 3$ and our prior covariance matrix is:

$$\mathbf{P} = \begin{pmatrix} 1 & .61 & .22 \\ .61 & 1 & .37 \\ .22 & .37 & 1 \end{pmatrix}$$

Our measurement error covariance matrix is $\mathbf{R} = 0.5\mathbf{I}$

In this case, our gain (interpolation weights) is (are):

$$\mathbf{K} = \begin{pmatrix} 0.3914 & 0.0528 \\ 0.6453 & 0.0870 \\ 0.0870 & 0.6453 \end{pmatrix}$$

Numerical Example (cont.)

Our posterior mean is:

$$E(\mathbf{x}|\mathbf{y}) = \begin{pmatrix} 17.4810 \\ 17.1442 \\ 21.0527 \end{pmatrix}$$

with posterior covariance:

$$\text{var}(\mathbf{x}|\mathbf{y}) = \begin{pmatrix} 0.7508 & 0.1957 & 0.0264 \\ 0.1957 & 0.3227 & 0.0435 \\ 0.0264 & 0.0435 & 0.3227 \end{pmatrix}$$

- the optimal prediction at location 1 gives more weight to observation 2 than 3 (since it is more highly correlated (i.e., closer))

- the prediction variance at location 1 is greater than location 2 and 3 since there is no data for location 1

"Real" Example

Consider an NCEP u-wind forecast over the tropical Pacific for 00 UTC, 7 Nov 1996: (This is our prior mean.)

NCEP U-Wind
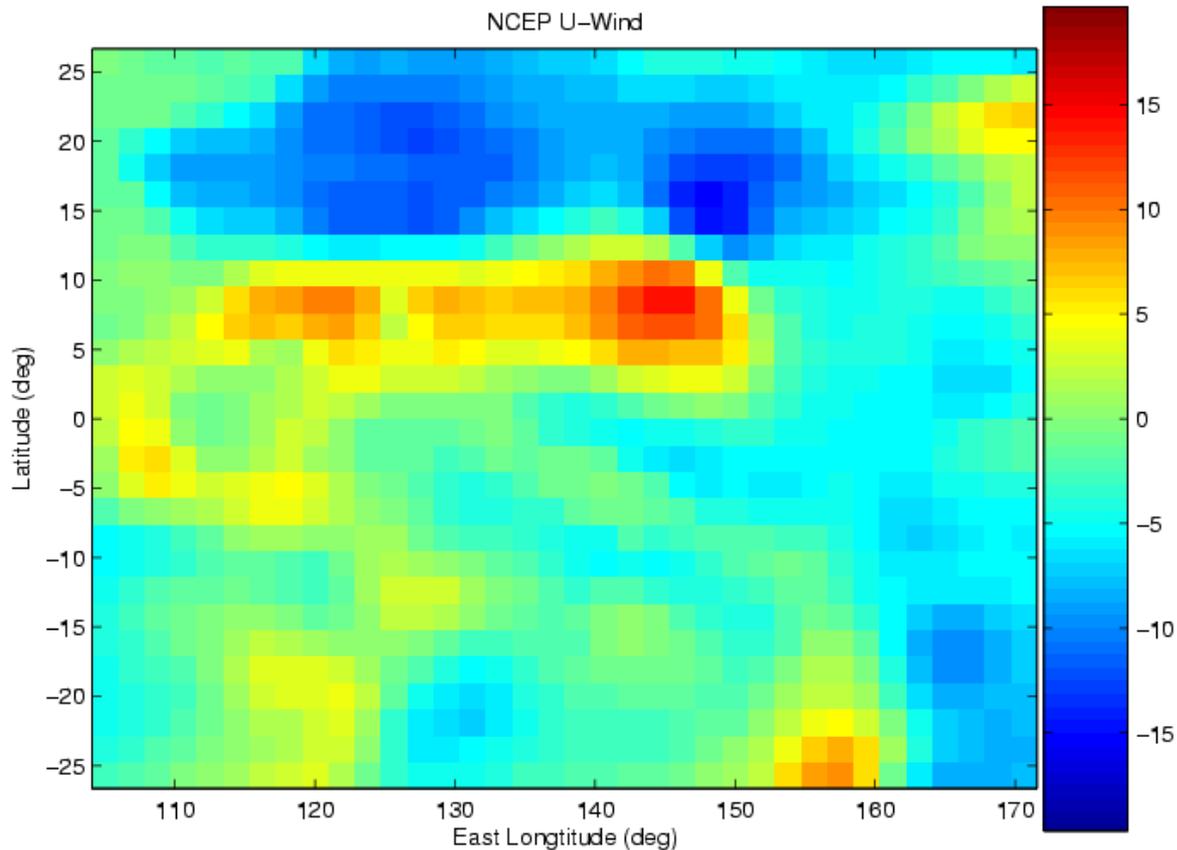
Example (cont.)

Consider data from NSCAT scatterometer at the same time (5473 observations):
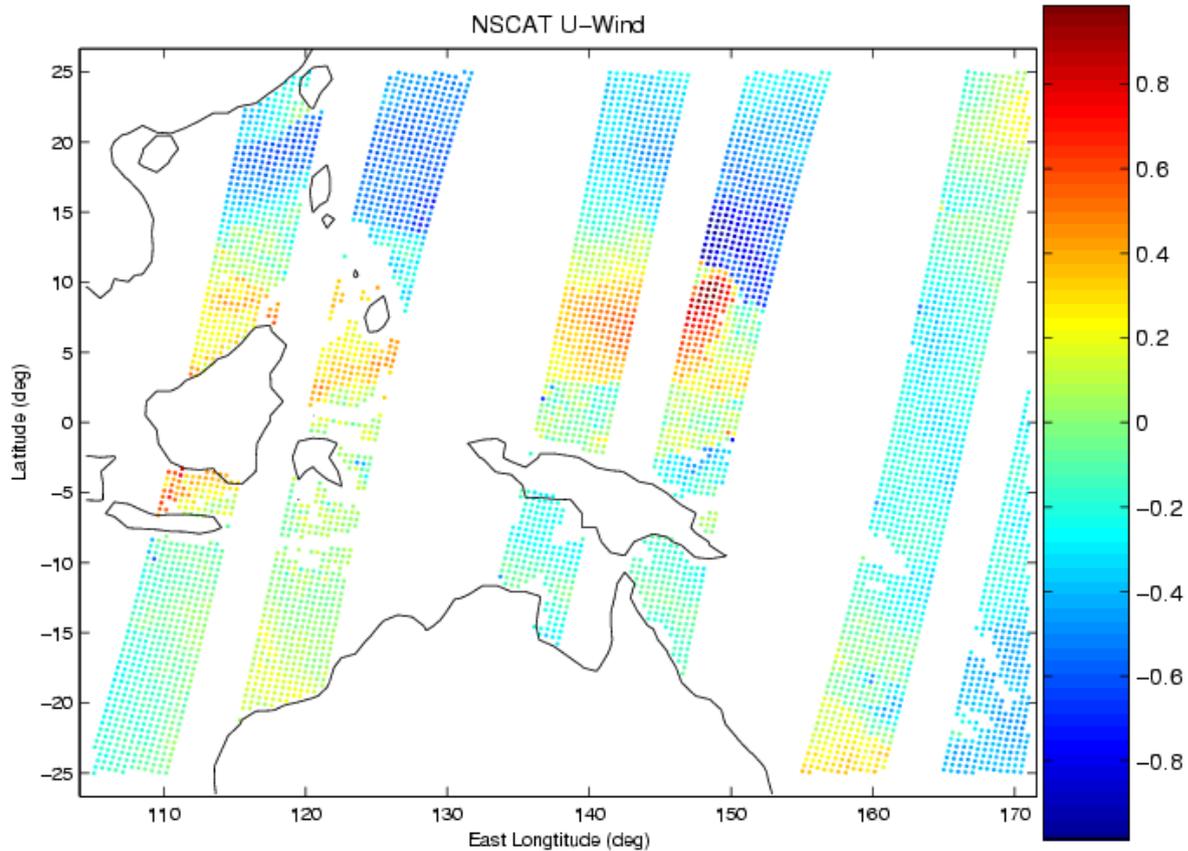
Example (cont.)

Posterior Mean:

Posterior Mean U-Wind
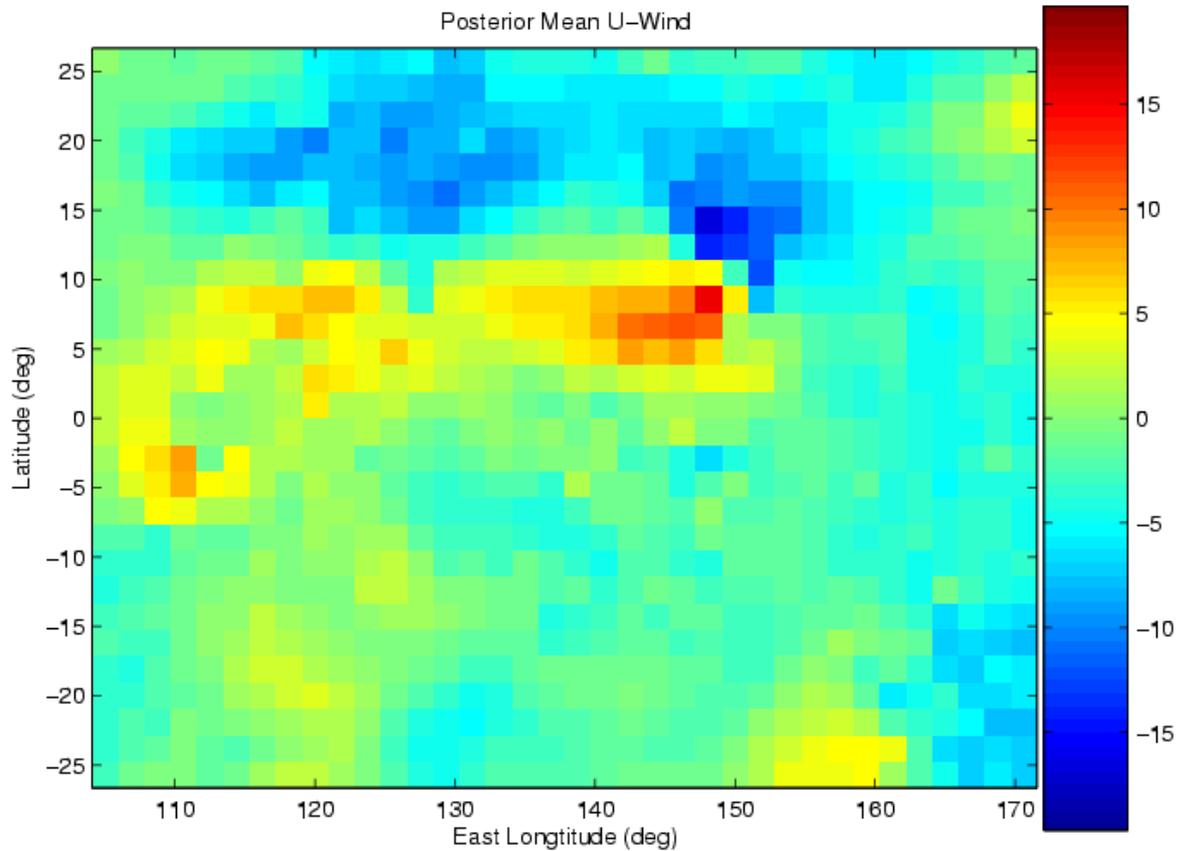
Example (cont.)

Posterior Variance:
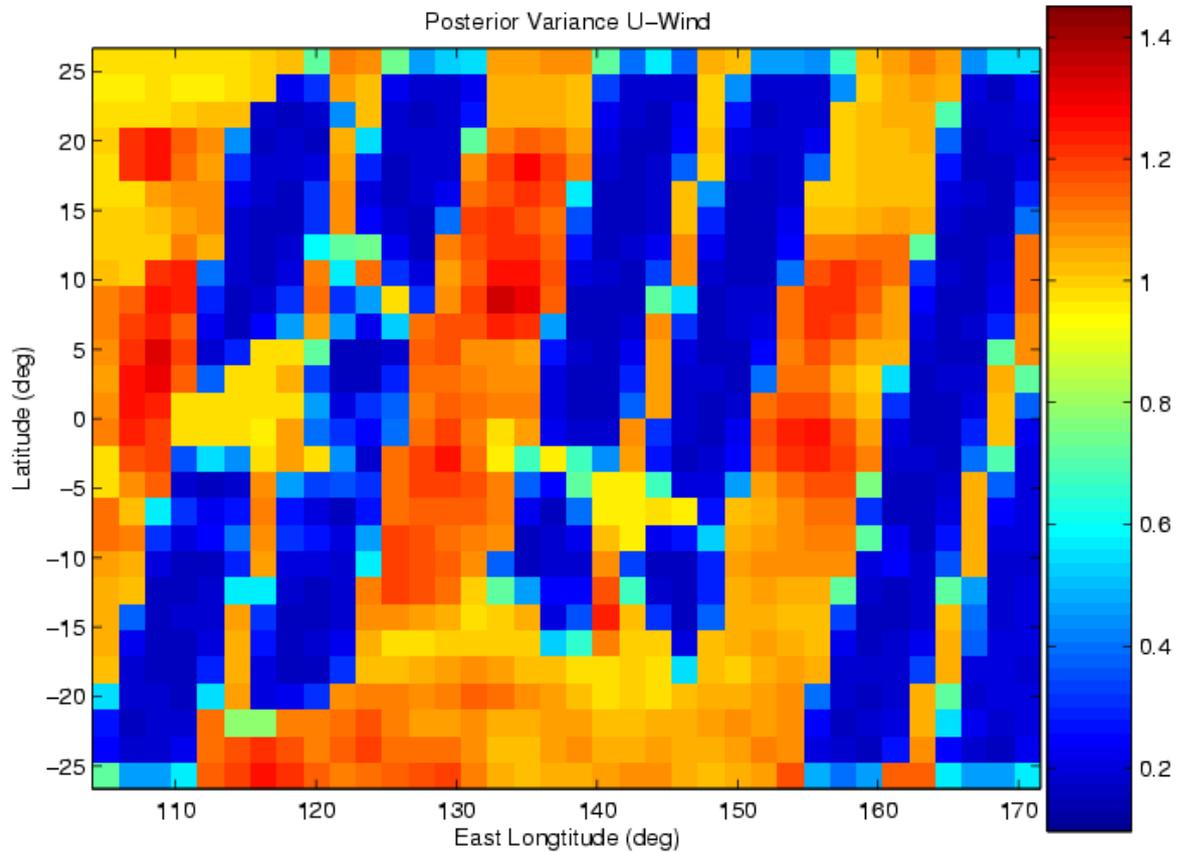
Connections to Variational Approaches

It is well-known (e.g., Lorenc 1986; Talagrand 1997) that the optimal interpolation problem can equivalently be posed as a variational problem.

In particular, the posterior mode (and mean in this case) of the multivariate normal/normal model is also found by minimizing the objective function:

$$J(\mathbf{x}) = (\mathbf{y} - \mathbf{Hx})'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{Hx}) + (\mathbf{x} - \boldsymbol{\mu})'\mathbf{P}^{-1}(\mathbf{x} - \boldsymbol{\mu})$$

Although formally equivalent to the Bayes formulation, for high-dimensional processes it is often more computationally efficient to approach the problem from this variational perspective.

Sequential Approaches

In many respects, it would be more efficient if as new data becomes available, one could simply update the previous (optimal) estimate of the state process without having to start from scratch.

The *Kalman filter* is an ideal framework for such sequential updating. The Kalman filter can be derived from many different perspectives. Here, we take a Bayesian (probabilistic) perspective.

Select References: Kalman 1960, Jazwinski 1970, Ghil 1989, Ghil and Malanotte-Rizzoli 1991, West and Harrison 1997, and *many* others!

Filtering, Smoothing, and Forecasting

Consider a vector process:

$$\mathbf{x}_t \equiv [x(s_1; t), \dots, x(s_n; t)]'$$

at n states (e.g., spatial locations) $s_j, \; j = 1, \dots, n$ and time t.
Consider p_t-dimensional observation vectors at time t:

$$\mathbf{y}_t \equiv [y(r_1; t), \dots, y(r_{p_t}; t)]'$$

We are interested in the distribution:

$$p(\mathbf{x}_t | \mathbf{y}_1, \dots, \mathbf{y}_T)$$

If

$$
\begin{aligned}
t = T &\;\rightarrow\; \textbf{filtering}(analysis) \\
1 < t < T &\;\rightarrow\; \textbf{smoothing}(4Dvar) \\
t > T &\;\rightarrow\; \textbf{forecasting}(prediction)
\end{aligned}
$$

Kalman Filter: Linear Systems

Consider the measurement (data) model:

$$\mathbf{y}_t = \mathbf{H}_t \mathbf{x}_t + \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \sim \ indep \ N(\mathbf{0}, \mathbf{R}_t) \tag{1}$$

where $\mathbf{H}_t$ is the observation operator that maps the process to the observations, and $\mathbf{R}_t$ is the (potentially) time-varying observation (measurement) error covariance matrix.

Also, consider the evolution (or process) model:

$$\mathbf{x}_{t+1} = \mathbf{M}_t \mathbf{x}_t + \boldsymbol{\eta}_t, \quad \boldsymbol{\eta}_t \sim \ indep \ N(\mathbf{0}, \mathbf{Q}_t) \tag{2}$$

where $\mathbf{M}_t$ is the (linear) model operator or propagator that maps the evolution of the process in time, and $\mathbf{Q}_t$ is a noise covariance matrix perhaps representing stochastic forcing or features not resolved by the model.

Typically, it is assumed that the measurement and model noise processes are independent. We have also assumed noise processes have zero mean, although this need not be the case in general.

Kalman Filter: Notation

Let,

$$\mathbf{Y}_t \equiv \{\mathbf{y}_1, \ldots, \mathbf{y}_t\}, \qquad \mathbf{X}_t \equiv \{\mathbf{x}_0, \ldots, \mathbf{x}_t\}$$

Define the conditional expectations for "analysis"(filter) and "forecast":

$$\mathbf{x}_{t|t} \equiv E[\mathbf{x}_t|\mathbf{Y}_t]$$

$$\mathbf{x}_{t|t-1} \equiv E[\mathbf{x}_t|\mathbf{Y}_{t-1}]$$

Similarly, define the conditional error covariance matrices for analysis and forecast, respectively:

$$\mathbf{P}_{t|t} = E[(\mathbf{x}_t - \mathbf{x}_{t|t})(\mathbf{x}_t - \mathbf{x}_{t|t})'|\mathbf{Y}_t]$$

$$\mathbf{P}_{t|t-1} = E[(\mathbf{x}_t - \mathbf{x}_{t|t-1})(\mathbf{x}_t - \mathbf{x}_{t|t-1})'|\mathbf{Y}_{t-1}]$$

Then, for normal error models (as considered here), we define the filter (analysis) and forecast distributions, respectively:

$$\mathbf{x}_t|\mathbf{Y}_t \sim N(\mathbf{x}_{t|t}, \mathbf{P}_{t|t}) \tag{3}$$

$$\mathbf{x}_t|\mathbf{Y}_{t-1} \sim N(\mathbf{x}_{t|t-1}, \mathbf{P}_{t|t-1}) \tag{4}$$

Kalman Filter Derivation

The forecast distribution (4) can be obtained via Bayes' rule when one considers it as just the *posterior predictive distribution*:

$$
\begin{aligned}
p(\mathbf{x}_t|\mathbf{Y}_{t-1}) &= \int p(\mathbf{x}_t, \mathbf{x}_{t-1}|\mathbf{Y}_{t-1})d\mathbf{x}_{t-1} \\
&= \int p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathbf{Y}_{t-1})d\mathbf{x}_{t-1}
\end{aligned}
$$

where the first distribution on the RHS is just the process evolution model (2) and the second distribution on the RHS is the posterior (analysis) distribution for the previous time (3).

Using conditional expectation and conditional variance arguments,

$$
\mathbf{x}_{t|t-1} = E(\mathbf{x}_t|\mathbf{Y}_{t-1}) = E(E(\mathbf{x}_t|\mathbf{x}_{t-1})|\mathbf{Y}_{t-1}) = E(\mathbf{M}_t\mathbf{x}_{t-1}|\mathbf{Y}_{t-1}) = \mathbf{M}_t\mathbf{x}_{t-1|t-1}
$$
$$
(5)
$$

$$
\begin{aligned}
\mathbf{P}_{t|t-1} &= \mathrm{var}(\mathbf{x}_t|\mathbf{Y}_{t-1}) = E(\mathrm{var}(\mathbf{x}_t|\mathbf{x}_{t-1})|\mathbf{Y}_{t-1}) + \mathrm{var}(E(\mathbf{x}_t|\mathbf{x}_{t-1})|\mathbf{Y}_{t-1}) \\
&= E(\mathbf{Q}_t|\mathbf{Y}_{t-1}) + \mathrm{var}(\mathbf{M}_t\mathbf{x}_{t-1}|\mathbf{Y}_{t-1}) \\
&= \mathbf{Q}_t + \mathbf{M}_t\mathbf{P}_{t-1|t-1}\mathbf{M}_t'
\end{aligned}
$$
$$
(6)
$$

Similarly, we can obtain the analysis distribution by Bayes' rule:

$$
\begin{aligned}
p(\mathbf{x}_t|\mathbf{Y}_t) &= p(\mathbf{x}_t|\mathbf{y}_t, \mathbf{Y}_{t-1}) \\
&\propto p(\mathbf{y}_t|\mathbf{x}_t, \mathbf{Y}_{t-1})p(\mathbf{x}_t|\mathbf{Y}_{t-1}) \\
&= p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{Y}_{t-1}) \tag{7}
\end{aligned}
$$

where the first dist on the RHS of (7) is just the data model (1) and the 2nd dist on the RHS is the forecast distribution (4) with mean (5) and variance (6).

As before for the normal:normal case, we get

$$
\mathbf{x}_t|\mathbf{Y}_t \sim N\left((\mathbf{H}_t'\mathbf{R}_t^{-1}\mathbf{H}_t + \mathbf{P}_{t|t-1}^{-1})^{-1}(\mathbf{H}_t'\mathbf{R}_t^{-1}\mathbf{y}_t + \mathbf{P}_{t|t-1}^{-1}\mathbf{x}_{t|t-1}), (\mathbf{H}_t'\mathbf{R}_t^{-1}\mathbf{H}_t + \mathbf{P}_{t|t-1}^{-1})^{-1}\right) \tag{8}
$$

Using the same matrix derivation as for the non-sequential case, we can write equivalently the mean and variance of (8):

$$
\mathbf{x}_{t|t} = \mathbf{x}_{t|t-1} + \mathbf{K}_t(\mathbf{y}_t - \mathbf{H}_t\mathbf{x}_{t|t-1}) \tag{9}
$$

$$
\mathbf{P}_{t|t} = (\mathbf{I} - \mathbf{K}_t\mathbf{H}_t)\mathbf{P}_{t|t-1} \tag{10}
$$

where the Kalman gain is given by

$$
\mathbf{K}_t = \mathbf{P}_{t|t-1}\mathbf{H}_t'(\mathbf{H}_t'\mathbf{P}_{t|t-1}\mathbf{H}_t + \mathbf{R}_t)^{-1}. \tag{11}
$$

Kalman Filter Algorithm

Given the parameter matrices $\mathbf{H}_t$, $\mathbf{M}_t$, $\mathbf{Q}_t$, $\mathbf{R}_t$ for $t = 1, \ldots, T$ and initial conditions (or background state) $\hat{\mathbf{x}}_{0|0} \equiv \mathbf{x}^b$, $\hat{\mathbf{P}}_{0|0} \equiv \mathbf{P}^b$, one can use the following Kalman filter algorithm to obtain sequential estimates of the state and associated covariance matrices:

for $t = 1$ to T

1. get forecasts $\hat{\mathbf{x}}_{t|t-1}$ and $\hat{\mathbf{P}}_{t|t-1}$ from (5) and (6), respectively

2. get gain $\hat{\mathbf{K}}_t$, and analysis $\hat{\mathbf{x}}_{t|t}$, and $\hat{\mathbf{P}}_{t|t}$ from (11), (9) and (10), respectively

end

NOTE: In Statistics, one does not typically know the parameter matrices (especially, $\mathbf{M}_t$, $\mathbf{Q}_t$ and $\mathbf{R}_t$). In cases where these are not time-varying, and the dimensionality n is relatively low, one can use the E-M algorithm (Shumway and Stoffer, 1982) or numerical maximum likelihood methods (Gupta and Mehra, 1974) to obtain estimates. **See Jon Stroud's talks at this workshop.** Such approaches are not practical for high-dimensional problems.

Simple Kalman Filter Example

Assume we have the univariate measurement model:

$$y_t = x_t + \epsilon_t, \quad \epsilon_t \sim N(0, .1)$$

(i.e., $R = .1$, $H = 1$)

Also assume we have the forecast (or process) model:

$$x_{t+1} = 0.7x_t + \eta_t, \quad \eta_t \sim N(0, .5)$$

(i.e., $Q = .5$, $M = .7$)

Given an initial condition $x_0 \sim N(0, 1)$ we simulated x_t and y_t for $t = 1, \ldots, 100$. In addition, we let the data at times 40 to 43 and 80 to 83 be missing.

Our goal: get back the filtered "state" x_t for all times, $t = 1, \ldots, 100$ given the data y_t.

Simulated "truth" and "data"

Simulation from AR(1) Process
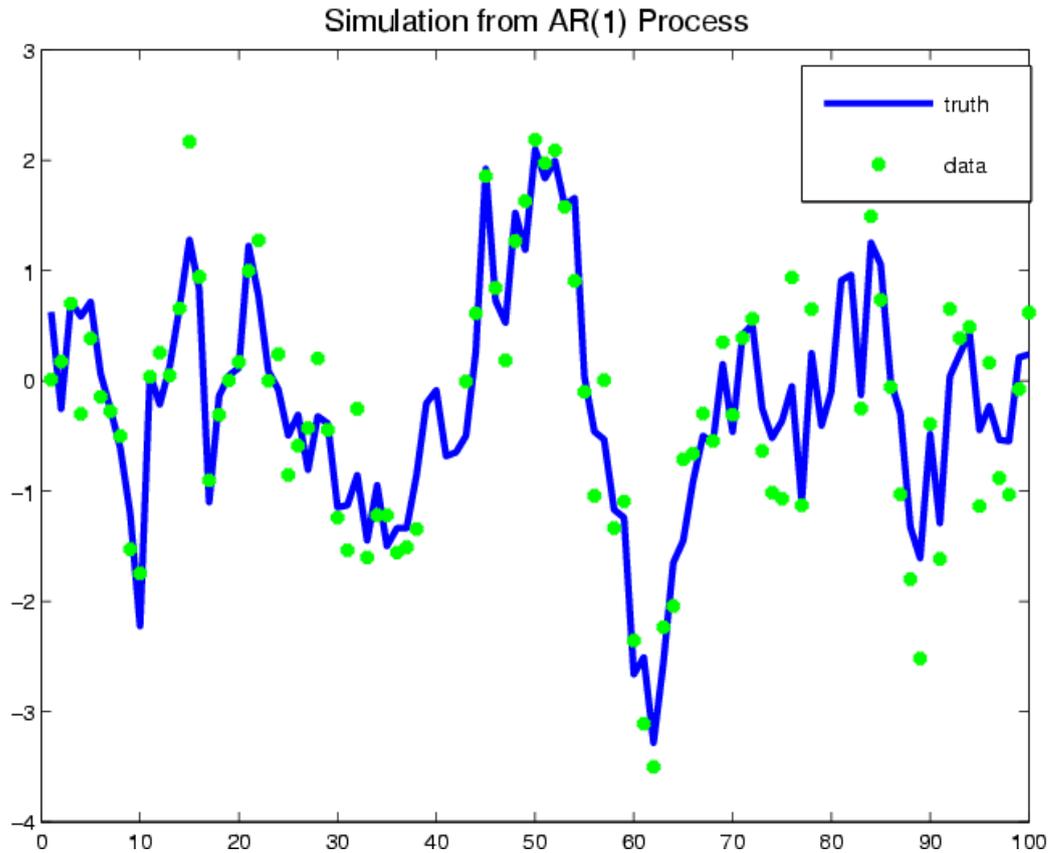
Simple KF Example: Kalman Filter Algorithm

The Kalman Filter is simple to program for this univariate example ($Q = .5$, $R = .1$, $M = .7$, $H = 1$):

1. set $\hat{x}_{0|0} = 0$, $\hat{P}_{0|0} = Q = .5$

2. For $t = 1, \ldots, 100$

   (a) $\hat{x}_{t|t-1} = M\hat{x}_{t-1|t-1}$

   (b) $\hat{P}_{t|t-1} = Q + M\hat{P}_{t-1|t-1}M$

   (c) $\hat{K}_t = \hat{P}_{t|t-1}H[H\hat{P}_{t|t-1}H + R]^{-1}$

   (d) $\hat{x}_{t|t} = \hat{x}_{t|t-1} + \hat{K}_i[y_t - H\hat{x}_{t|t-1}]$

   (e) $\hat{P}_{t|t} = [1 - \hat{K}_t H]\hat{P}_{t|t-1}$

NOTE: if there is a missing observations y_t then the steps (c)-(e) are skipped and $\hat{x}_{t|t} = \hat{x}_{t|t-1}$, $\hat{P}_{t|t} = \hat{P}_{t|t-1}$.

Simple KF Example: State Recovery

Simulation from AR(1) Process
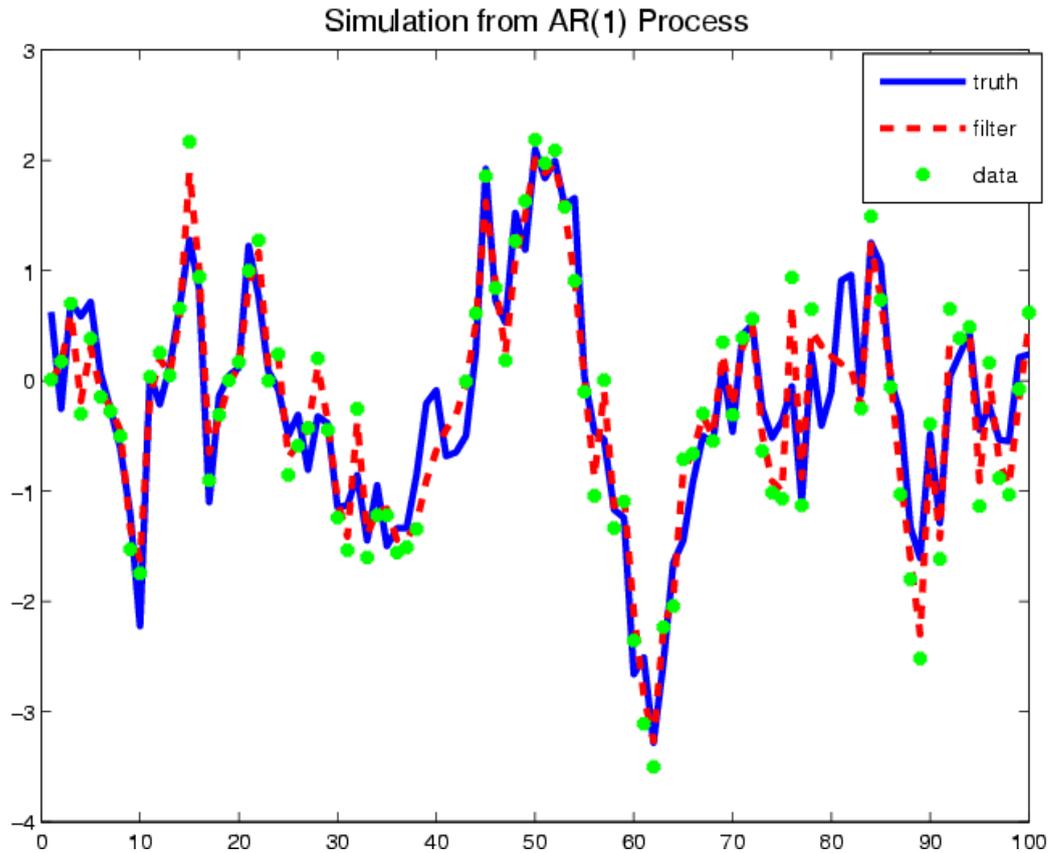
Simple KF Example: Prediction Variance

KF Prediction Variance
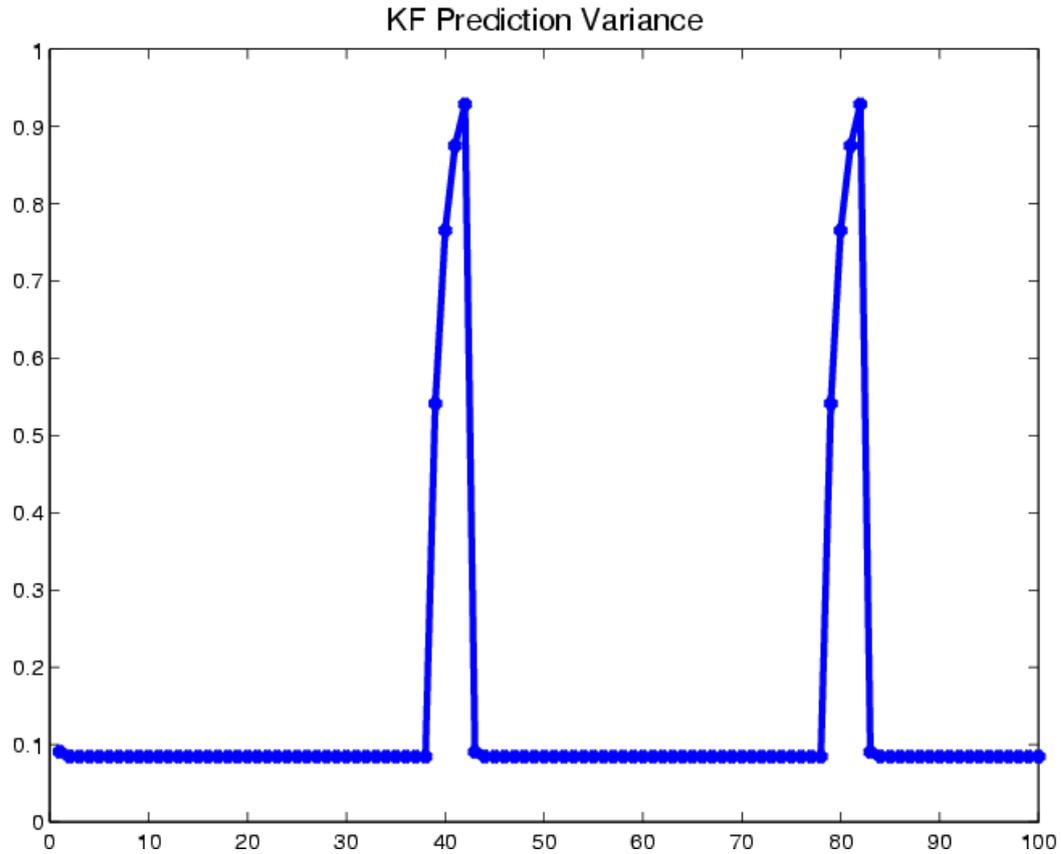
Nonlinear and Non-Gaussian Models

One traditional approach to handling nonlinear observation and/or evolution models is by local (tangent linear) linearization of the model and evolution operators. In the sequential case, this is known as *extended Kalman filtering*.

Additionally, some types of non-Gaussian error structures can be accommodated by allowing error structures to be convex mixtures of Gaussian distributions.

In principle, one can accommodate nonlinear and non-Gaussian processes directly in the Bayes context. Recall, the forecast and analysis distributions can be obtained from

$$p(\mathbf{x}_t|\mathbf{Y}_{t-1}) = \int p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathbf{Y}_{t-1})d\mathbf{x}_{t-1}$$

$$p(\mathbf{x}_t|\mathbf{Y}_t) \propto p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{Y}_{t-1})$$

These are valid regardless of the distribution form or nonlinearity in the conditional mean. However, typically there is no closed-form for these distributions and in high-dimensions, the dimensionality of the integrals prohibit direct numerical solution.

Alternatively, we may utilize Monte Carlo methods to evaluate these integrals.

Monte Carlo Sampling and Bayesian Statistics

Recall, a main use of Monte Carlo is estimation of integrals (or expectations for probability models).

As before, let the data be represented by:

$$Y_t \equiv \{y_1, \ldots, y_t\}$$

and let the state-process be represented by:

$$X_t \equiv \{x_0, \ldots, x_t\}$$

Let f be a function of the state variable of interest X_t and assume a Bayesian context in which we have data Y_t and are interested in:

$$E(f(X_t)|Y_t) = \int f(X_t)p(X_t|Y_t)dX_t = \frac{\int f(X_t)p(Y_t|X_t)p(X_t)dX_t}{\int p(Y_t|X_t)p(X_t)dX_t}$$

(assuming the integrals exist)

A Monte Carlo (MC) estimate can be obtained:

1. generate N pseudo-random realizations, $X_t^{(i)}$ from $p(X_t|Y_t)$, $i = 1, \ldots, N$

2. evaluate f for each realization and compute the arithmetic average of the results, $\hat{E}(f(X_t)|Y_t) = (1/N) \Sigma_{i=1}^{N} f(X_t^{(i)})$

Under independent sampling this average converges (a.s.) to $E(f(X_t)|Y_t)$ as $N \to \infty$. [This holds if realizations are stationary (or ergodic) though not necessarily independent. Also, the rate of convergence is independent of the dimensionality of the integrand].

Note, we can also approximate the distribution $p(X_t|Y_t)$ by:

$$\hat{p}(X_t|Y_t) \equiv p^N(X_t|Y_t) = \frac{1}{N} \sum_{i=1}^{N} \delta_{X_t^{(i)}}$$

where $\delta_{X_t^{(i)}}$ is a Dirac-delta mass at $X_t^{(i)}$.

Recall the familiar sequential update distributions:

$$p(x_t|Y_{t-1}) = \int p(x_t|x_{t-1})p(x_{t-1}|Y_{t-1})dx_{t-1}$$

$$p(x_t|Y_t) \propto p(y_t|x_t)p(x_t|Y_{t-1})$$

Assuming we can obtain samples $x_{t-1|t-1}^{(i)}$, $i = 1, \ldots, N$ from $p(x_{t-1}|Y_{t-1})$, then a sequential MC algorithm would consist of the following steps:

- (i)

$$p^N(x_{t-1}|Y_{t-1}) = \frac{1}{N} \sum_{i=1}^{N} \delta_{x_{t-1|t-1}^{(i)}}$$

- (ii)

$$p^N(x_t|Y_{t-1}) = \frac{1}{N} \sum_{i=1}^{N} p(x_t|x_{t-1|t-1}^{(i)})$$

- (iii)

$$p^N(x_t|Y_t) \propto p(y_t|x_t) \sum_{i=1}^{N} p(x_t|x_{t-1|t-1}^{(i)})/N$$

Importance Sampling

When direct simulation of $p(X_t|Y_t)$ is difficult/impossible, we can use **Importance Sampling**. The idea is:

- Consider another distribution with the same support as $p(X_t|Y_t)$, say $q(X_t|Y_t)$, that is comparatively easy to sample from.

- Generate N samples $X_t^{(i)}$ from $q(X_t|Y_t)$ and evaluate $f(X_t)$ for each

- To use these to estimate $E(f(X_t)|Y)$ we must weight each sample (ensemble) member to adjust for the fact that the samples are *not* from the target posterior:

$$\hat{E}(f(X_t)|Y_t) = \sum_{i=1}^{N} w_i f(X_t^{(i)}),$$

and, we can approximate the posterior distribution by:

$$p^N(X_t|Y_t) = \sum_{i=1}^{N} w_i \delta_{X_t^{(i)}}$$

Importance Sampling Weights

How does one get the IS weights?

$$E(f(X_t)|Y_t) = \int f(X_t) \frac{p(X_t|Y_t)}{q(X_t|Y_t)} q(X_t|Y_t) dX_t$$

$$= \int f(X_t) w_t(X_t) q(X_t|Y_t) dX_t$$

where

$$w_t(X_t) = \frac{p(X_t|Y_t)}{q(X_t|Y_t)}$$

are the unnormalized importance sampling weights and

$$\tilde{w}_t(X_t^{(i)}) \equiv \frac{w_t(X^{(i)})}{\sum_{j=1}^{N} w_t(X^{(j)})}$$

are the normalized IS weights (note, by using these normalized weights, we *do not need to know the normalizer for $p(X_t|Y_t)$ or $q(X_t|Y_t)$*)!

Importance Distribution

There are many reasonable IS proposal distributions $q(\ )$ that one could use. An intuitive one in our case is to choose the "prior" or forward model distribution:

$$q(\mathbf{X}_t|\mathbf{Y}_t) = p(\mathbf{X}_t) = p(\mathbf{x}_0) \prod_{k=1}^{t} p(\mathbf{x}_k|\mathbf{x}_{k-1})$$

In this case, the unnormalized weights are just given by the likelihood:

$$w_t(X_t) = \frac{p(X_t|Y_t)}{p(X_t)} = \frac{p(Y_t|X_t)p(X_t)}{p(X_t)} = p(Y_t|X_t)$$

These weights can be computed iteratively:

$$w_t(X_t) = p(Y_t|X_t) = \prod_{k=1}^{t} p(Y_k|X_k) = p(y_t|x_t)w_{t-1}(X_{t-1})$$

This then allows one to implement the importance sampler sequentially.

Sequential Importance Sampler

- (i) Start with the samples (i.e., **particles**) and weights from the posterior at time $t - 1$:

$$p^N(x_{t-1}|Y_{t-1}) = \sum_{i=1}^{N} \delta_{x_{t-1|t-1}^{(i)}} w_{t-1}^{(i)}$$

- (ii) For each particle, simulate from the forecast evolution (transition) density:

$$p^N(x_t|Y_{t-1}) = \sum_{i=1}^{N} p(x_t|x_{t-1|t-1}^{(i)}) w_{t-1}^{(i)}$$

- (iii) Reweight each particle according to the likelihood (the particles that are "closer" to the data get higher weights), and resample to obtain updated points and weights (and thus a representation of the posterior) at time t:

$$p^N(x_t|Y_t) \propto p(y_t|x_t) \sum_{i=1}^{N} p(x_t|x_{t-1|t-1}^{(i)}) w_{t-1}^{(i)}$$

Unfortunately, as t increases the importance weights degenerate (i.e., only one particle has non-zero importance weight) and the posterior is not adequately represented.

One way to address this problem is to eliminate the particles having low importance weights and to multiply particles that have high weights (e.g., Gordon et al. 1993).

There are many approaches to dealing with this degeneracy problem and it is still an active area of research. See the review in Doucet et al. 2001 .

Bootstrap Filter Algorithm (Doucet, de Freitas, Gordon, 2001)

1. Initialization, $t = 0$

   - for $i = 1, \ldots N$ sample $x_{0|0}^{(i)} \sim p(x_{0|0})$ and set $t = 1$

2. Importance sampling step

   - for $i = 1, \ldots, N$ sample $\tilde{x}_t^{(i)} \sim p(x_t | x_{t-1}^{(i)})$ and set $\tilde{X}_t^{(i)} = \{X_{t-1}^{(i)}, \tilde{x}_t^{(i)}\}$
   - for $i = 1, \ldots, N$ evaluate the importance weights $\tilde{w}_t^{(i)} = p(y_t | \tilde{x}_t^{(i)})$ (note: in this algorithm these weights are not proportional to the weights at the previous time $(t-1)$ due to the resampling in Step 3, which induces equal weights on the resample)
   - normalize IS weights

3. Selection step

   - resample with replacement N particles $\{X_t^{(i)} : i = 1, \ldots, N\}$ from the set $\{\tilde{X}_t^{(i)} : i = 1, \ldots, N\}$ according to importance weights.
   - set $t = t + 1$ and go to step 2

Ensemble Kalman Filter

Sequential importance sampling (or **particle filtering**) can be used in the non-linear/non-Gaussian filtering problem in principle (see Doucet et al. 2001 for a comprehensive overview). However, these approaches have serious problems (e.g., degeneracy of the weights) in very high dimensional problems (like one would see for an atmospheric DA problem). This methodology is still the subject of intense research.

A closely related idea is to use Monte Carlo in the context of the Kalman filter. One loses some of the generality of the importance sampling approach but gains much in terms of computational efficiency.

These "ensemble" based assimilation algorithms are known as **ensemble Kalman filters** (e.g., Evensen 1994; Evensen and van Leeuven 1996, and many, many others!). The basic approach uses Monte Carlo samples to approximate the forecast distribution (but, critically, uses the nonlinear forward model). In particular, one estimates the the prior (forecast) means and variance/covariance matrices with the Monte Carlo samples (ensemble). These are then used in the linear KF update formulas to obtain the analysis distribution.

"Clasical" EnKF Approach: Approximation of SIS

- (i) Assume one has N independent samples from the posterior at time $t-1$ (i.e., weights, $w_{t-1}^{(i)} = 1/N$):

$$p^N(x_{t-1}|Y_{t-1}) = \sum_{i=1}^{N} \delta_{x_{t-1|t-1}^{(i)}} \times 1/N$$

- (ii) Since it is assumed that we have independent samples from the posterior at time $t-1$, we can use straight MC to obtain samples from the forecast distribution (again, like SIS but with weights $= 1/N$):

$$p^N(x_t|Y_{t-1}) = \sum_{i=1}^{N} p(x_t|x_{t-1|t-1}^{(i)}) \times 1/N$$

NOTE: we assume this forecast distribution can be characterized by its first two moments (or, equivalently, that it is normal with mean $x_{t|t-1}^{(i)}$ and (estimated) variance/covariance matrix $\hat{P}_{t|t-1}$).

- (iii) The update (posterior at time t) is then given by:

$$p^N(x_t|Y_t) \propto p(y_t|x_t) \sum_{i=1}^{N} N(x_t|x_{t|t-1}^{(i)}, \hat{P}_{t|t-1}) \times 1/N$$

Note, if we assume the measurement distribution is normal, then this mixture can be computed exactly using the KF update equations.

EnKF Algorithm

1. Choose initial estimates of the filter mean and variance: $\hat{\mathbf{x}}_{0|0}$ and $\hat{\mathbf{P}}_{0|0}$

2. for $t = 1, \ldots, T$

    (a) Generate N independent samples, $\mathbf{x}_{t-1|t-1}^{(i)}$ from:

    $$\mathbf{x}_{t-1|t-1}^{(i)} \sim N(\hat{\mathbf{x}}_{t-1|t-1}, \hat{\mathbf{P}}_{t-1|t-1}), \quad i = 1, \ldots, N$$

    (b) Forecast each of the samples from (a) forward using the evolution model: e.g.,

    $$\mathbf{x}_{t|t-1}^{(i)} = m(\mathbf{x}_{t-1|t-1}^{(i)}) + \boldsymbol{\eta}_t^{(i)}, \quad \boldsymbol{\eta}_t^{(i)} \sim N(\mathbf{0}, \mathbf{Q})$$

    NOTE: in problems where there is assumed no model error, then one just evolves the sample forward using the $m(\ )$ model evolution, but no additive noise.

    (c) Use the forecast samples to calculate a sample forecast covariance matrix:

    $$\hat{\mathbf{P}}_{t|t-1} = \frac{1}{N-1} \sum_{i=1}^{N} (\mathbf{x}_{t|t-1}^{(i)} - \hat{\mathbf{x}}_{t|t-1})(\mathbf{x}_{t|t-1}^{(i)} - \hat{\mathbf{x}}_{t|t-1})'$$

    where $\hat{\mathbf{x}}_{t|t-1} = (1/N) \Sigma_{i=1}^{N} \mathbf{x}_{t|t-1}^{(i)}$

(d) Draw N samples $(\mathbf{y}_t^i)$ from the measurement distribution:

$$\mathbf{y}_t^{(i)} \sim N(\mathbf{y}_t, \mathbf{R}), \quad i = 1, \ldots, N$$

NOTE: this was not part of the original EnKF algorithm but was later shown to be necessary to obtain posterior distributions with adequate spread.

(e) Use the Kalman Filter update equations to update each forecast sample given the sampled observations:

$$\mathbf{x}_{t|t}^{(i)} = \mathbf{x}_{t|t-1}^{(i)} + \mathbf{K}_t(\mathbf{y}_t^{(i)} - \mathbf{H}_t \mathbf{x}_{t|t-1}^{(i)})$$

where

$$\mathbf{K}_t = \hat{\mathbf{P}}_{t|t-1} \mathbf{H}_t'(\mathbf{H}_t \hat{\mathbf{P}}_{t|t-1} \mathbf{H}_t' + \mathbf{R})^{-1}$$

In addition,

$$\hat{\mathbf{P}}_{t|t} = \frac{1}{N-1} \sum_{i=1}^{N} (\mathbf{x}_{t|t}^{(i)} - \hat{\mathbf{x}}_{t|t})(\mathbf{x}_{t|t}^{(i)} - \hat{\mathbf{x}}_{t|t})'$$

where $\hat{\mathbf{x}}_{t|t} = (1/N) \sum_{i=1}^{N} \mathbf{x}_{t|t}^{(i)}$.

EnKF Algorithm: Modifications

- Typically, N is relatively small because it is too expensive to run the forward model. $N = 100$ is often sufficient.

- *Problem*: Covariance estimates are not stable and not of full rank when the dimension of $\mathbf{x}_t$ is larger than N. The standard "fix" is to consider the **shur product** (or **Hadamard product**). This is an "element-by-element" multiplication of the ensemble estimated covariance matrix by a correlation matrix $\mathbf{S}$ that has "compact support". That is, use $\mathbf{P}_{t|t-1} \circ \mathbf{S}$ where $\circ$ is the Shur product and $\mathbf{S}$ is a correlation matrix.

  (NOTE: the product of a covariance matrix and a correlation matrix is a covariance matrix.)

- *Computational Efficiency*: It is computationally more efficient to calculate the elements of $\mathbf{K}_t$ directly, rather than $\mathbf{P}_{t|t-1}$:

$$\hat{\mathbf{P}}_{t|t-1}\mathbf{H}_t' = \frac{1}{N-1} \sum_{i=1}^{N} (\mathbf{x}_{t|t-1}^{(i)} - \mathbf{x}_{t|t-1})(\mathbf{H}_t\mathbf{x}_{t|t-1}^{(i)} - \overline{\mathbf{H}\mathbf{x}_{t|t-1}})'$$

$$\mathbf{H}_t\hat{\mathbf{P}}_{t|t-1}\mathbf{H}_t' = \frac{1}{N-1} \sum_{i=1}^{N} (\mathbf{H}_t\mathbf{x}_{t|t-1}^{(i)} - \overline{\mathbf{H}\mathbf{x}_{t|t-1}})(\mathbf{H}_t\mathbf{x}_{t|t-1}^{(i)} - \overline{\mathbf{H}\mathbf{x}_{t|t-1}})'$$

Simple AR(1) Example

From before, recall the simple AR(1) simulation example:
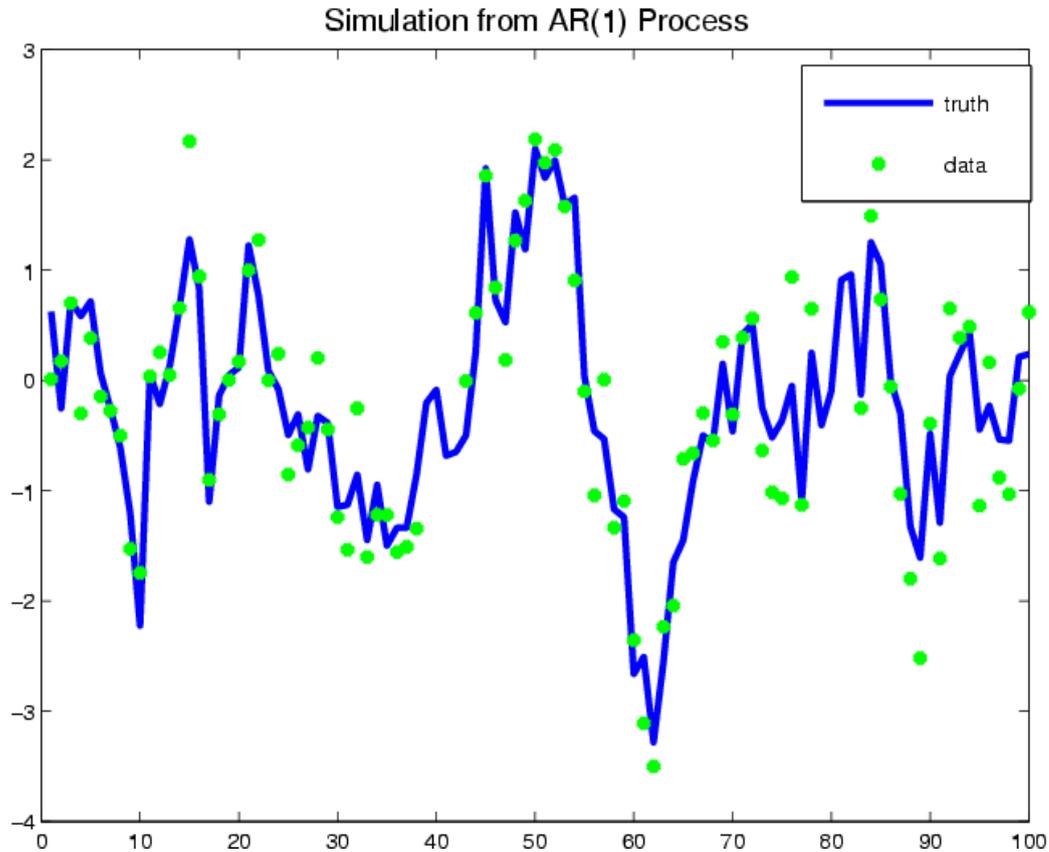
EnKF Results

Consider the EnKF filter results:
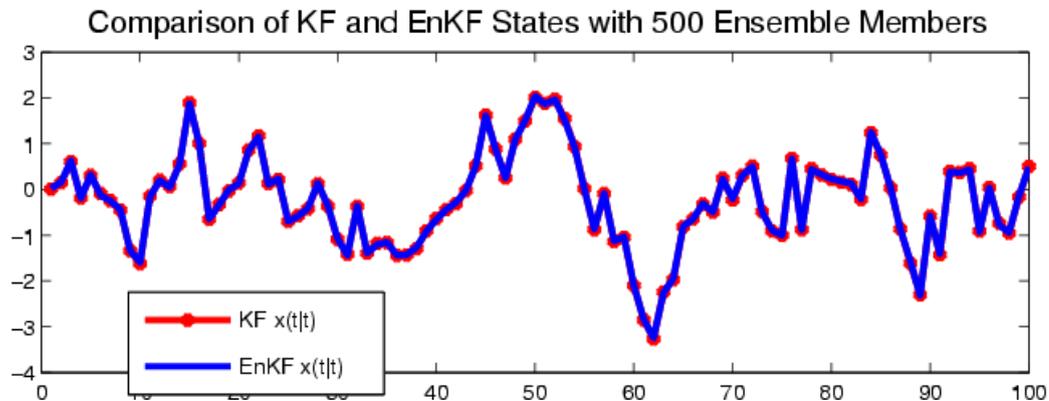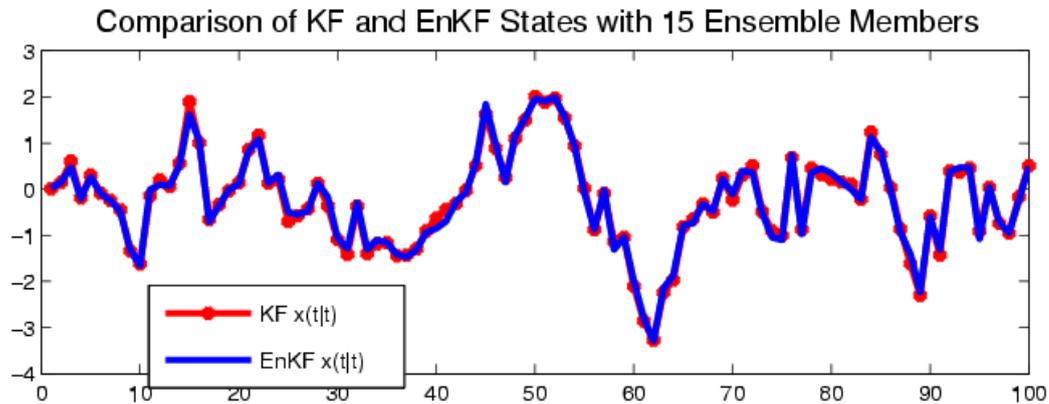
EnKF Results

Consider the EnKF filter results:
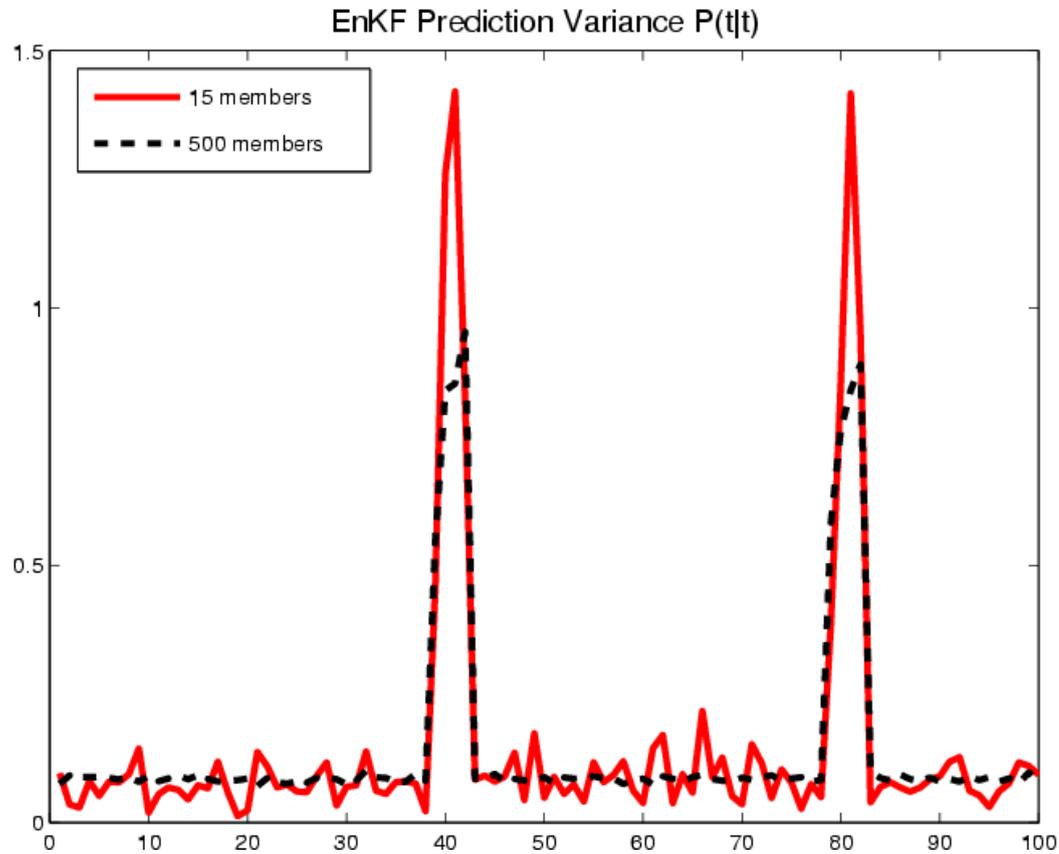
Future Directions

- Monte Carlo based filters are beginning to be implemented by several operational centers.

- There are still many issues (statistical and operational) that are to be worked out concerning these methods.

  - type of samples
  - parameter uncertainty
  - smoothing problem
  - efficiency
  - algorithms

- Given the computational efficiency of these approaches, and their performance, they are sure to become a dominant approach for model/data fusion.