Data Assimilation Basics: Part II

Analysis of Data Assimilation Systems

Ed Zaron

Department of Civil and Environmental Engineering
Portland State University
Portland, Oregon, USA

GODAE Summer School
Perth, Australia
January 2010

Contents

Overview

Last time, we used a minimum variance criterion to derive optimal interpolation and the Kalman filter.

This time, we will use a Bayesian approach with multivariate Gaussian probability distributions assumed.

Our approach will permit the derivation of so-called variational data assimilation methods, and show how these are related to maximum likelihood and minimum variance criteria.

Optimal Interpolation

Recall,

$$\mathbf{y}^o = \mathbf{H}\mathbf{x}^t + \epsilon \quad \text{and} \quad \mathbf{R} = \langle \epsilon\epsilon^T \rangle \in \mathrm{R}^{M \times M}.$$

Make stronger assumption that ϵ is multivariate Gaussian:

$$P_\epsilon(\epsilon) = \frac{1}{\sqrt{(2\pi)^M det(\mathbf{R})}} \exp\left(-\frac{1}{2}\epsilon^T \mathbf{R}^{-1}\epsilon \right)$$

But $P_\epsilon(\epsilon) = P_\mathbf{y}(\mathbf{y}^o|\mathbf{x}^t)$,

$$P_\mathbf{y}(\mathbf{y}^o|\mathbf{x}^t) = \frac{1}{\sqrt{(2\pi)^M det(\mathbf{R})}} \exp\left(-\frac{1}{2}(\mathbf{y}^o - \mathbf{H}\mathbf{x}^t)^T \mathbf{R}^{-1}(\mathbf{y}^o - \mathbf{H}\mathbf{x}^t) \right)$$

Optimal Interpolation

Previously, we assumed $\mathbf{x}^f = \langle \mathbf{x}^t \rangle$ was given.

Now suppose $\mathbf{x}^t$ is Gaussian,

$$P_{\mathbf{x}}(\mathbf{x}^t) = \frac{1}{\sqrt{(2\pi)^N det(\mathbf{B})}} \exp\left(-\frac{1}{2}(\mathbf{x}^t - \mathbf{x}^f)^T \mathbf{B}^{-1} (\mathbf{x}^t - \mathbf{x}^f)\right)$$

with $\mathbf{B} = (\mathbf{x}^t - \mathbf{x}^f)(\mathbf{x}^t - \mathbf{x}^f)^T \in \mathrm{R}^{N \times N}$.

Optimal Interpolation

Assuming ϵ and $\mathbf{x}^t$ are independent, Bayes' Theorem says,

$$P_{\mathbf{x}}(\mathbf{x}^t|\mathbf{y}^o) = \frac{P_{\mathbf{y}}(\mathbf{y}^o|\mathbf{x}^t)P_{\mathbf{x}}(\mathbf{x}^t)}{P_{\mathbf{y}}(\mathbf{y}^o)}$$

The denominator, $P_{\mathbf{y}}(\mathbf{y}^o) = \int P_{\mathbf{y}}(\mathbf{y}^o|\mathbf{x}^t)P_{\mathbf{x}}(\mathbf{x}^t)d\mathbf{x}^t$, is independent of $\mathbf{x}^t$.

$$P_{\mathbf{x}}(\mathbf{x}^t|\mathbf{y}^o) \propto \exp\left(-\frac{1}{2}(\mathbf{y}^o - \mathbf{H}\mathbf{x}^t)^T\mathbf{R}^{-1}(\mathbf{y}^o - \mathbf{H}\mathbf{x}^t)\right)$$
$$\times \exp\left(-\frac{1}{2}(\mathbf{x}^t - \mathbf{x}^f)^T\mathbf{B}^{-1}(\mathbf{x}^t - \mathbf{x}^f)\right)$$

Optimal Interpolation

Let

$$\mathcal{J}(\mathbf{x}^t;\mathbf{y}^o) = (\mathbf{y}^o - \mathbf{H}\mathbf{x}^t)^T\mathbf{R}^{-1}(\mathbf{y}^o - \mathbf{H}\mathbf{x}^t) + (\mathbf{x}^t - \mathbf{x}^f)^T\mathbf{B}^{-1}(\mathbf{x}^t - \mathbf{x}^f)$$

Then

$$P_{\mathbf{x}}(\mathbf{x}^t|\mathbf{y}^o) \propto \exp\left(-\frac{1}{2}\mathcal{J}(\mathbf{x}^t;\mathbf{y}^o)\right)$$

The maximum likelihood estimate for $\mathbf{x}^t$ given $\mathbf{y}^o$ is the minimizer of $\mathcal{J}$, and it coincides with the mean and median estimators *when Gaussian statistics and a linear measurement operator,* $\mathbf{H}$, *are assumed.*

Optimal Interpolation

So-called variational methods seek to minimize an objective function such as $\mathcal{J}$ rather than satisfy a minimum-variance criterion.

The let $\mathbf{x}^a$ be the value of $\mathbf{x}^t$ such that $\mathcal{J}(\mathbf{x}^t;\mathbf{y}^o)$ is minimum:

$$\begin{aligned}\frac{1}{2}\frac{\partial\mathcal{J}(\mathbf{x}^t=\mathbf{x}^a;\mathbf{y}^o)}{\partial\mathbf{x}^t} =& 0\\ =& -\mathbf{H}^T\mathbf{R}^{-1}(\mathbf{y}^o-\mathbf{H}\mathbf{x}^a)+\mathbf{B}^{-1}(\mathbf{x}^a-\mathbf{x}^f)\end{aligned}$$

Optimal Interpolation

Solving for $\mathbf{x}^a$:

$$(\mathbf{B}^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})\mathbf{x}^a = \mathbf{B}^{-1}\mathbf{x}^f + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{y}^o$$
$$(\mathbf{I} + \mathbf{B}\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})\mathbf{x}^a = \mathbf{x}^f + \mathbf{B}\mathbf{H}^T\mathbf{R}^{-1}\mathbf{y}^o$$

The latter equation has been called the "primal form of the variational data assimilation" problem.

It is equivalent to $\mathbf{B}\dfrac{\partial \mathcal{J}}{\partial \mathbf{x}^t} = 0$.

Optimal Interpolation

The so-called "dual form of variational data assimilation" is obtained by noticing the above is equivalent to,

$$\mathbf{x}^a = \mathbf{x}^f + \mathbf{B}\mathbf{H}^T\mathbf{w}$$
$$(\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R})\mathbf{w} = \mathbf{y}^o - \mathbf{H}\mathbf{x}^f,$$

where $\mathbf{w} \in \mathrm{R}^{M\times 1}$ is said to be in the dual or observation space.

Optimal Interpolation

Exercise 1: Show that $\mathbf{x}^a$ derived here is the same as $\mathbf{x}^a = \mathbf{x}^f + \mathbf{K}(\mathbf{y}^o - \mathbf{H}\mathbf{x}^f)$, derived in Part I.

Exercise 2: Derive the "dual formulation" by applying the Sherman-Morrison-Woodbury formula to find the inverse of $(\mathbf{B}^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})$.

Exercise 3: The Hessian matrix of second derivatives of $\mathcal{J}$ with respect to $\mathbf{x}^t$ is

$$\begin{aligned}\mathcal{H} =& \frac{\partial^2 \mathcal{J}}{\partial \mathbf{x}^t \partial \mathbf{x}^t} \\ =& \mathbf{B}^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\end{aligned}$$

Use the Sherman-Morrison-Woodbury formula to show $\mathbf{B}^a = \mathcal{H}^{-1}$.

Remarks

- The "dual formulation" is deeper than it appears. It can be applied to infinite-dimensional state vectors, e.g., $\mathbf{x}^t$ a function rather than a vector, see Bennett (1992,2002).
- Smoothing splines, interpolation by radial basis functions, etc. can be developed from the same viewpoint as above, see Wahba (1990).
- The literature on non-parametric estimation has developed methods for handling inequality constraints, rank-deficient $\mathbf{B}$, and model-building. See Wahba's web site; likewise, see Candes, Tao, Donoho, and others in the "compressive sampling" community.
- The above derivation is in the context of univariate spatial optimal interpolation. Multivariate interpolation in space-time follows the same approach.

Variational Methods: OI revisited

Variational Methods: time-dependent assimilation

Solution Strategies

Summary: Analysis of Data Assimilation Systems

Appendix: Pictoral Examples

Approaches to Time-Dependent Assimilation

- Kalman filter (Brasseur lectures)
- Optimal interpolation with augmented state space
- 4D-Var “weak” and “strong” versions (Moore lectures)
- Generalized inversion

Requires a general notation for representing time-dependent data assimilation in order to analyze solution algorithms.

Variational Formulation of Time-Dependent Assimilation

dynamics

Consider

$$\begin{aligned} \mathbf{x}^t(t_i) =& \mathcal{M}_i[\mathbf{x}^t(t_{i-1})] + \eta_i, \quad \text{for} \quad i \in \{1, \ldots, T\} \\ \mathbf{x}^t(t_0) =& \mathbf{x}_0 + \eta_0 \quad \text{initial conditions.} \end{aligned}$$

Use $\mathcal{M}_i$ rather than $\mathbf{M}_i$ to denote possibly nonlinear dynamics.

Covariance of system noise is $\mathbf{Q}_j = \langle \eta_j \eta_j^T \rangle$.

Covariance of initial condition error is $\mathbf{Q}_0 = \langle \eta_0 \eta_0^T \rangle$.

Note: η_j is uncorrelated in time.

Variational Formulation of Time-Dependent Assimilation
data

Data are given as

$$\mathbf{y}_j^o = \mathbf{H}_j \mathbf{x}_j^t + \epsilon_j \quad \text{for} \quad j \in \{1, \ldots, T\}.$$

$\mathbf{H}_j \in \mathrm{R}^{M_j \times N}$ are observation operators, possibly different at each time.

Covariance of observational error is $\mathbf{R}_j = \langle \epsilon_j \epsilon_j^T \rangle$.

Variational Formulation of Time-Dependent Assimilation

objective function

The objective function (cost function, negative log-likelihood function) is:

$$\begin{aligned}
\mathcal{J}(\{\mathbf{x}_j^t\}) = & \sum_{j=1}^{T} \left(\mathbf{x}^t(t_j) - \mathcal{M}_j[\mathbf{x}^t(t_{j-1})]\right)^T \mathbf{Q}_j^{-1} \eta_j^T & \text{(dynamics)} \\
& + \left(\mathbf{x}^t(t_0) - \mathbf{x}_0]\right)^T \mathbf{Q}_0^{-1} \left(\mathbf{x}^t(t_0) - \mathbf{x}_0]\right)^T & \text{(i.c.'s)} \\
& + \sum_{j=1}^{T} \left(\mathbf{y}_j^o - \mathbf{H}_j \mathbf{x}^t(t_j)\right)^T \mathbf{R}_j^{-1} \left(\mathbf{y}_j^o - \mathbf{H}_j \mathbf{x}^t(t_j)\right) & \text{(data)}
\end{aligned}$$

Let $\{\mathbf{x}_j^a\}_{j=1}^T$ denote a minimizer of $\mathcal{J}$.

Introduce an adjoint variable, $\lambda_i = \mathbf{Q}_i^{-1}\eta_i$.

The minimizer of $\mathcal{J}$ solves the following system of equations, which express $\partial\mathcal{J}(\mathbf{x})/\partial\mathbf{x} = 0$:

$$\lambda_i = \mathbf{M}_{i+1}^T\lambda_{i+1} + \sum_{j=1}^{T}\mathbf{H}_j^T\mathbf{R}_j^{-1}\left(\mathbf{y}_j^o - \mathbf{H}_j\mathbf{x}^a(t_j)\right) \quad (1)$$

$$\lambda_T = 0 \quad (2)$$

$$\mathbf{x}^a(t_i) = \mathcal{M}_i[\mathbf{x}^a(t_{i-1})] + \mathbf{Q}_i\lambda_i \quad (3)$$

$$\mathbf{x}^a(t_0) = \mathbf{x}_0 + \mathbf{Q}_0\lambda_0 \quad (4)$$

$\mathbf{M}_j = \dfrac{\partial\mathcal{M}_j[\mathbf{x}_j]}{\partial\mathbf{x}_j}$ is the so-called "adjoint model."

Variational Formulation of Time-Dependent Assimilation

Remarks:

- For linear $\mathcal{M}_j[\mathbf{x}] = \mathbf{M}_j\mathbf{x}$, the unique minimizer of $\mathcal{J}$ and the Kalman filter are the same at the final time, $j = T$.
- For linear $\mathcal{M}_j$, the minimizer of $\mathcal{J}$ may be derived as a variance-minimizing estimator, known as the Kalman smoother or Rauch-Tung-Striebel (RTS) smoother.
- For nonlinear $\mathcal{M}_j$, the minimizer of $\mathcal{J}$ may not be unique.
- The above system is the "primal form" of variational data assimilation. The "dual form" decouples the λ and $\mathbf{x}^a$ equations.
- There is a precise analogy between the primal and dual forms of time-dependent variational data assimilation presented above, and the primal and dual forms of optimal interpolation.

Summary

At this point we could spend an hour each on boundary conditions, derivation of extremal conditions, discrete vs. continuous formulations, etc.

Subsequent lectures will touch on some of these issues in the context of specific applications.

We have developed just enough to illustrate the issues involved in various "data assimilation methods" which are just solvers for approximate maximum likelihood or variance-minimizing estimators.

Common Themes and Issues

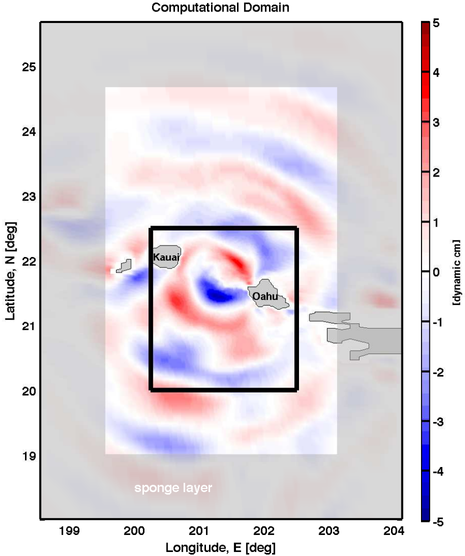
- Dimensionality
  - Large dimension of state trajectory, N
  - Large dimension of state covariance, $N \times N$
  - Large dimension of observation space, M
- Nonlinearity of dynamics or measurement operators
- Non-normality of error statistics

Dimensionality

size of state vectors

Frequency-domain model:

$$N \approx N_x \times N_y \times N_z \times N_{var}$$
$$\approx 250 \times 400 \times 60 \times 3$$
$$\approx 18 \times 10^6$$

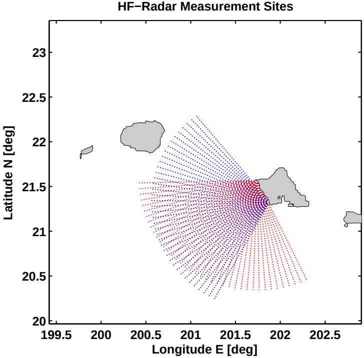


Zaron et al (2009) Baroclinic tidal generation in Kauai Channel ... . *Dyn. Atm. Oc.*, 48, 93–120

Dimensionality

size of observation vectors

$$M \approx 1800$$



Dimensionality

size of state and obs vectors

[Image from BlueLink].

$$N \approx N_x \times N_y \times N_z \times N_T \times N_{var} \times \alpha_{land}$$

$$M \approx N_{sites}$$

Dimensionality

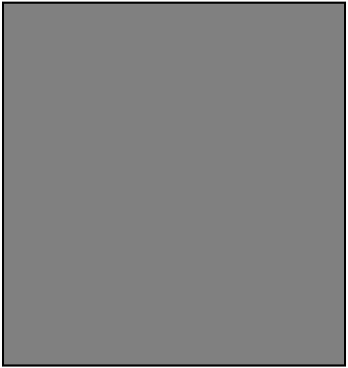
State vector:
$\mathbf{x}^a, \mathbf{x}^t, \mathbf{x}^f \in \mathrm{R}^{N \times 1}$

Observation vector:
$\mathbf{y}^o \in \mathrm{R}^{M \times 1}$

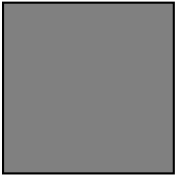
Note: not to scale!

Dimensionality

Forecast error covariance:
$\mathbf{B}, \mathbf{B}^{-1} \in \mathrm{R}^{N \times N}$



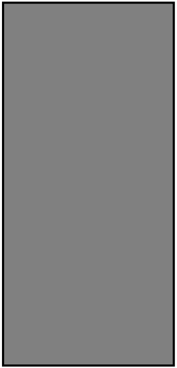
Observation error covariance:
$\mathbf{R}, \mathbf{R}^{-1} \in \mathrm{R}^{M \times 1}$



Dimensionality

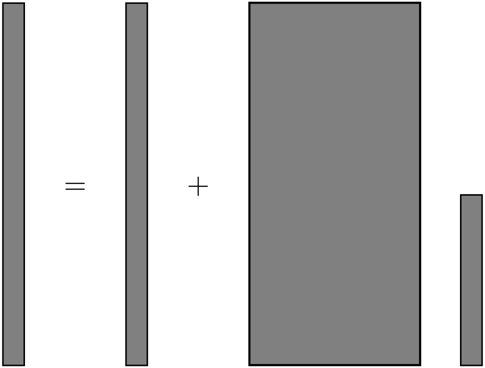
Measurement operator:
$\mathbf{H} \in \mathrm{R}^{M \times N}$

Kalman gain:
$\mathbf{K} \in \mathrm{R}^{N \times M}$



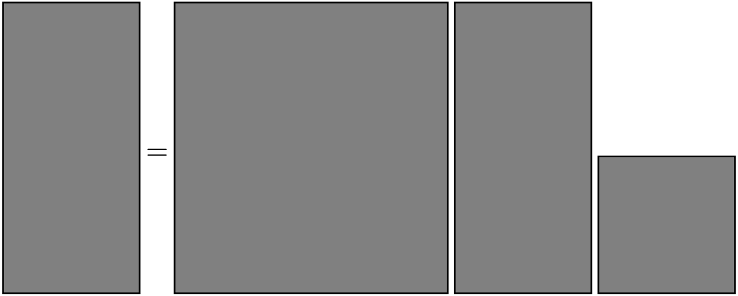
Optimal Interpolation

$$\mathbf{x}^a = \mathbf{x}^f + \mathbf{K}(\mathbf{y}^o - \mathbf{H}\mathbf{x}^f)$$



Optimal Interpolation

$$\mathbf{K} = \mathbf{B}\mathbf{H}^T\mathbf{P}^{-1}$$

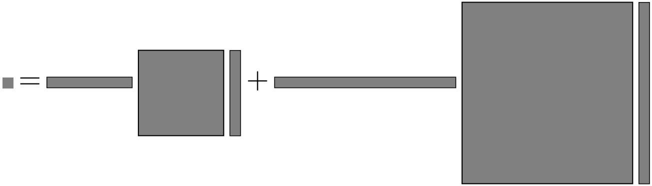


Typical approach (Daley, 1991):

- Define region of influence; approximate $\mathbf{P}^{-1}$
- Treat $\mathbf{B}$ as an operator, e.g., $B_{ij} = \sigma^2 \exp\left(-\frac{D(i,j)^2}{2L^2}\right)$

Variational Methods

$$\mathcal{J}(\mathbf{x}^t;\mathbf{y}^o) = (\mathbf{y}^o - \mathbf{H}\mathbf{x}^t)^T\mathbf{R}^{-1}(\mathbf{y}^o - \mathbf{H}\mathbf{x}^t) + (\mathbf{x}^t - \mathbf{x}^f)^T\mathbf{B}^{-1}(\mathbf{x}^t - \mathbf{x}^f)$$

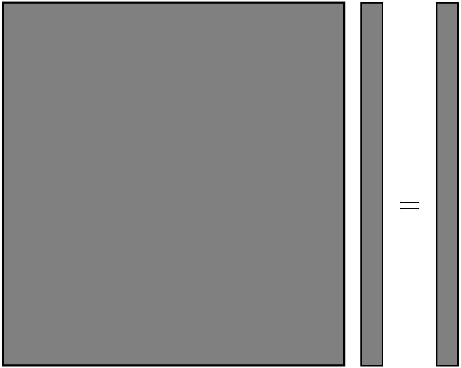


Naive approach would require $\mathbf{B}^{-1}$!

Variational Methods

"Primal formulation":

$$(\mathbf{I} + \mathbf{B}\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})\mathbf{x}^a = \mathbf{x}^f + \mathbf{B}\mathbf{H}^T\mathbf{R}^{-1}\mathbf{y}^o$$



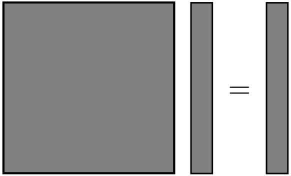
- Treat left-hand side matrix as an operator, never explicitly form large matrices.
- $\mathbf{R}^{-1}$ easily computed if data errors are uncorrelated.

Variational Methods

"Dual formulation":

$$\mathbf{x}^a = \mathbf{x}^f + \mathbf{B}\mathbf{H}^T\mathbf{w}$$

$$(\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R})\mathbf{w} = \mathbf{y}^o - \mathbf{H}\mathbf{x}^f,$$



- Smaller linear system to solve.
- "Representer algorithm" computes $\mathbf{H}\mathbf{B}\mathbf{H}^T$ directly.
- Compute $(\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R})^{-1}$ with simpler model and use as preconditioner.
- AKA: "indirect representer algorithm," "accelerated representers," "physical space assimilation."

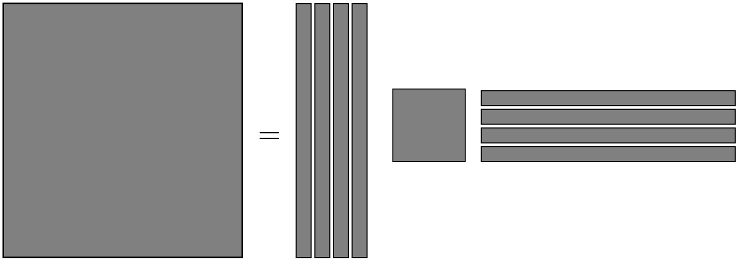
Model Reduction

Directly reduce N by projecting model dynamics onto reduced-dimensional subspace:

- Lorenz model is reduced description of atmosphere, e.g., Gauthier (1992), Miller (1994), Evensen and Fario (1994), ...
- Model reduction using empirical orthogonal functions (EOFs) from ensemble simulations, e.g., Vermueulen and Heemink (2006).
- Model reduction using EOFs in data space, $\mathbf{y}^o - \mathbf{H}\mathbf{x}^a$, e.g., Daescu and Navon (2007).
- Nonlinear model surrogate from neural net, e.g., Frolov et al (2009)

Reduced rank approximations

Store and update orthogonal decomposition $\mathbf{B} = \mathbf{U}\Lambda\mathbf{U}^T$



- Singular evolutive extended Kalman filter, Pham et al (1998)
- Error Subspace Statistical Estimation (ESSE), Lermusiaux and Robinson (1999)
- Adjoint-free variational assimilation, Logutov and Lermusiaux (2008)

Incremental Formulation

Courtier et al (1994), in the context of 4D-Var, suggested

- Use full physics or resolution to compute $\mathbf{x}^f$.
- Use reduced physics or resolution in gradient calculations or analysis increments.

Evaluate $\mathbf{B} \approx \widehat{\mathbf{B}}$ and $\mathbf{H} \approx \widehat{\mathbf{H}}$ with a simple, fast model.

$$\mathbf{x}^a = \mathbf{x}^f + \mathbf{S}^{-1}\widehat{\mathbf{B}}\widehat{\mathbf{H}}^T\mathbf{w} \quad (5)$$

$$(\widehat{\mathbf{H}}\widehat{\mathbf{B}}\widehat{\mathbf{H}}^T + \mathbf{R})\mathbf{w} = \mathbf{y}^o - \mathbf{H}\mathbf{x}^f. \quad (6)$$

$\mathbf{S}^{-1}$ is rank-deficient projection operator from reduced space to full space.

Reduced Basis

When M is too large, Egbert et al (1994) suggested using a reduced set of basis vectors for the analysis increment.

Let $\overline{\mathbf{H}} \in \mathrm{R}^{P\times N}$ denote subset of rows of $\mathbf{H} \in \mathrm{R}^{M\times N}$, corresponding to the $P < M$ members of the basis set.

$$\begin{aligned} \mathbf{x}^a =& \mathbf{x}^f + \mathbf{B}\overline{\mathbf{H}}^T\mathbf{w} \\ (\overline{\mathbf{H}}\mathbf{B}\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\mathbf{B}\overline{\mathbf{H}}^T + \overline{\mathbf{H}}\mathbf{B}\overline{\mathbf{H}}^T)\mathbf{w} =& \overline{\mathbf{H}}\mathbf{B}\mathbf{H}^T\mathbf{R}^{-1}(\mathbf{y}^o - \mathbf{H}\mathbf{x}^f). \end{aligned}$$

Solve using direct inversion of $P \times P$ matrix.

Ensemble Methods

Most direct treatment of nonlinearity by forming

$$\mathbf{B}^f \approx \frac{1}{N_E} \sum_{k=1}^{N_E} (\mathbf{x}^e_{(k)} - \mathbf{x}^f)(\mathbf{x}^e_{(k)} - \mathbf{x}^f)^T$$

from an ensemble of nonlinear model runs,

$$\begin{aligned} \mathbf{x}^e_{(k)}(t_i) =& \mathcal{M}_i[\mathbf{x}^e_{(k)}(t_{i-1})] + (\eta_i)_{(k)}, \quad \text{for} \quad i \in \{1, \ldots, T\} \\ \mathbf{x}^e_{(k)}(t_0) =& \mathbf{x}_0 + (\eta_0)_{(k)}. \end{aligned}$$

See Evensen (1994) for the Ensemble Kalman Filter (EnKF).

Ensemble Methods

Anderson et al (2009) presents a recent overview with pointers to the (large) primary literature:

- Effective rank of $\mathbf{B}$ is N_E.
- Off diagonal elements of $\mathbf{B}$ are inaccurate; must "taper" the spatial correlations.
- Loss of statistical independence of $\mathbf{B}^a$ at subsequent times $\Longrightarrow$ underestimate analysis error $\Longrightarrow$ filter "lock-on"
- So-called particle filters propagate the entire probability distribution function of $\mathbf{x}$.

Nonlinearity

Treatment of nonlinearity (of ocean models and measurement systems) is important, but not easily discussed in generality.

Approaches:

- Ignore it; limit focus to linear problems (e.g., deep ocean tidal modeling).
- Tangent-linearize dynamical operators around the current analysis $\Longrightarrow$ Extended Kalman Filter (EKF); primal formulation of variational methods.
- Use other linear approximations $\Longrightarrow$ incremental formulations, Bennett's bounded iterate strategy in dual formulation.
- Retain nonlinearity $\Longrightarrow$ Monte Carlo methods, particle filters.

Other Topics in Analysis of Data Assimilation Systems

- Covariance modeling: $\mathbf{B}$ and $\mathbf{Q}$
- Calibration and tuning
- $\mathbf{R}$: instrumental error vs. error of representativeness
- Observational impact
- Hypothesis testing

Analysis of Data Assimilation Systems

The differences between data assimilation "methods" often amount to differences in solution strategy.

Virtually all systems are built on approximations that

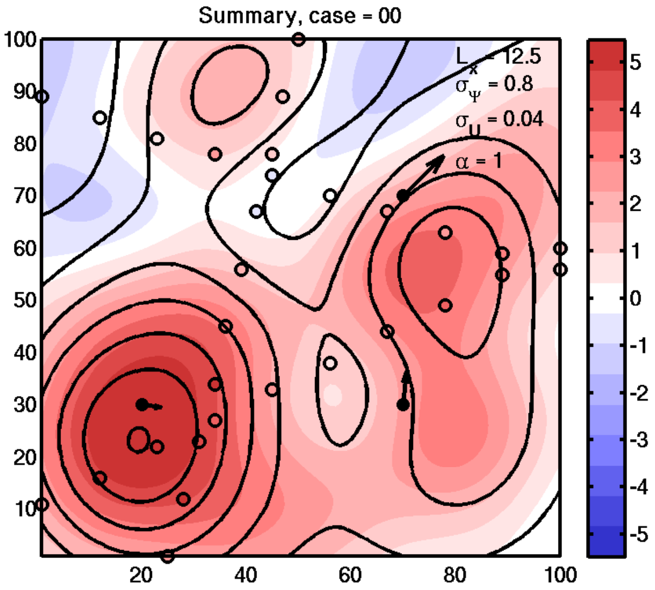
- Make dynamical or statistical linearizations;
- Reduce the effective degrees of freedom from N to M or fewer; and
- Simplify the representation of $N \times N$ covariances by parameterization or rank-reduction.

Analysis of Data Assimilation Systems

Analysis should consider the following factors:

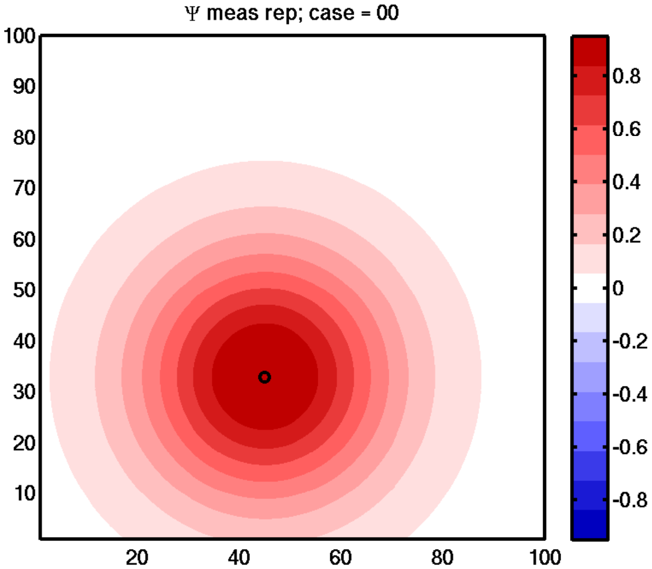
- The structure of the estimation problem – optimal interpolation, sequential estimation, fixed-interval smoothing;
- The practicability of solution algorithms;
- The hypothesized sources of system noise; and
- The ability to perform calibration and validation.

Example from Part I



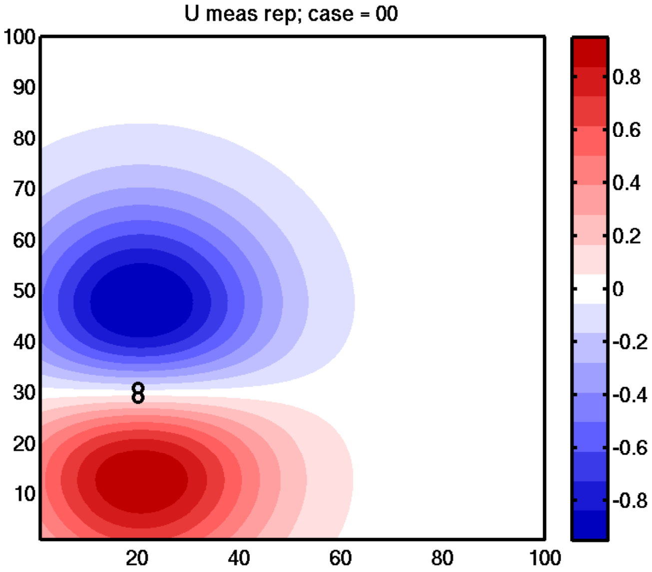
A Representer Function: $\Psi(x, y)$ measurement

A column of $\mathbf{BH}^T$ in $\mathbf{x}^a = \mathbf{x}^f + \mathbf{BH}^T\mathbf{w}$

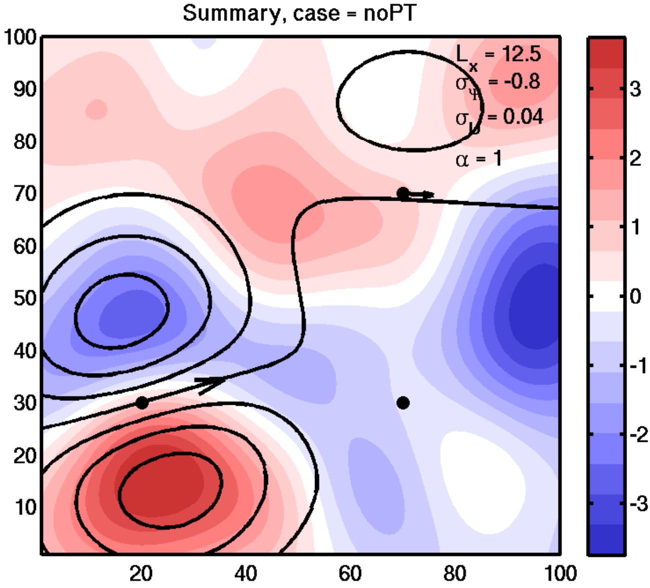


A Representer Function: $u(x,y) = -\partial\Psi/\partial y$ measurement

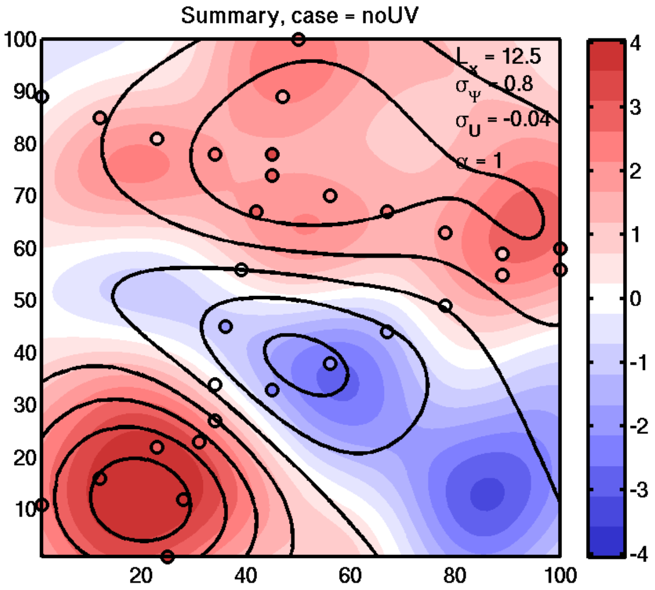
A column of $\mathbf{BH}^T$ in $\mathbf{x}^a = \mathbf{x}^f + \mathbf{BH}^T\mathbf{w}$



(u, v) Measurements

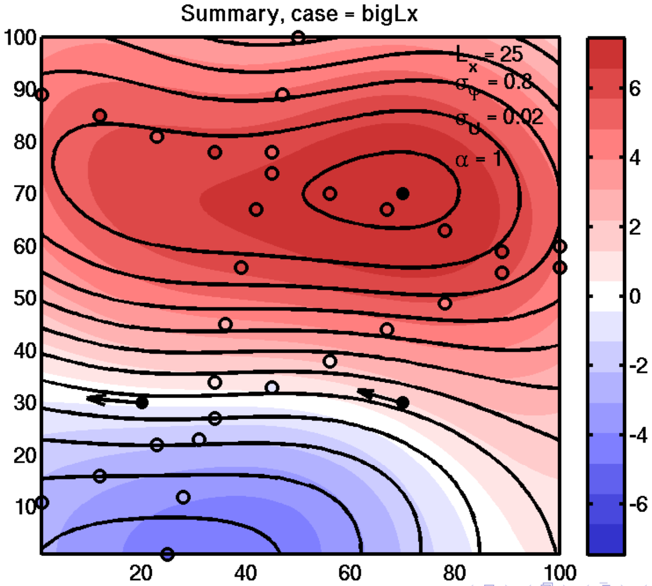


Ψ Measurements



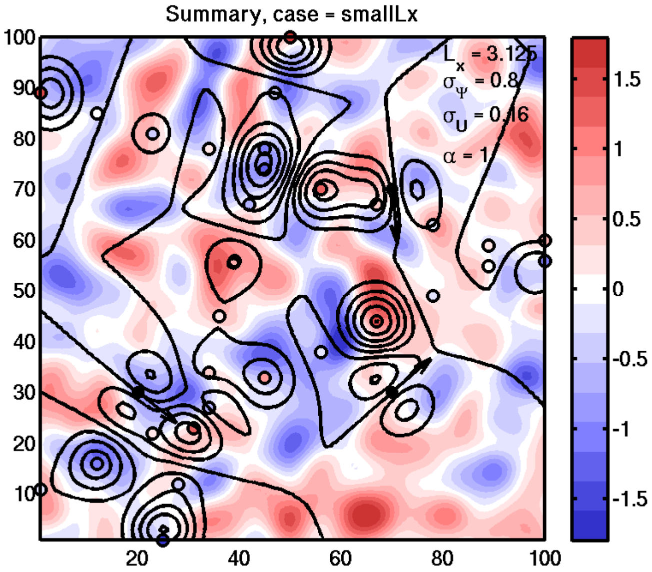
Large L_x

few d.o.f., well-estimated



Small L_x

many d.o.f., poorly-estimated



GODAE/BlueLink Course Page

Updated lecture notes, overhead slides, and matlab software:

http://web.cecs.pdx.edu/~zaron/GODAE/pub/SummerSchool.html