Multi-scale Hydrological Data Assimilation in Layered Media

Juan M. Restrepo
Department of Mathematics
Physics Department
University of Arizona
Collaborators:

- Daniel Tartakovsky, UCSD
- Michael Holst, UCSD
TIME SERIES

Estimation Problems:

Given a random time series \( \{z(t): t < t_0\} \)
\( z(t) \in \mathbb{R}^N \)

- **Prediction:**
  
  Estimate \( \{z(t): t > t_0\} \)

- **Filtering (Nudiction):**
  
  Estimate \( \{z(t_0)\} \)

- **Smoothing (Retrodiction):**
  
  Estimate \( \{z(t): t \cdot t_0\} \)
Turning a model into a state estimation problem

Example:

\[ \partial_t u(z,t) = \nu \partial_{zz} u(z,t) + f(t) \]
\[ u(z,0) = u_0(z) \]
\[ u(0,t) = g(t) \quad u(1,t) = h(t) \]

Discretizing:

\[ x(t) \begin{bmatrix} u_1(t) \\ u_2(t) \\ \vdots \\ u_N(t) \end{bmatrix} \]

is the state variable, obeying

\[ x(t+\delta t) = A x(t) + B q(t) \]
\[ x(t) = A x(t-\delta t) + B q(t-\delta t) \]

\[ \ldots \]

Leads to LINEAR PROBLEM:

\[ L(x(0),\ldots,x(t-\delta t),x(t),x(t+\delta t),\ldots,x(t_f),\ldots, \]
\[ Bq(t), Bq(t+\delta t),\ldots,t) = 0 \]

\[ x(t) \in \mathbb{R}^N \quad B q \in \mathbb{R}^N \]
Statement of the Problem

MODEL (Langevin Problem):

\[ dx(t) = f(x(t), t)dt + (2D)^{1/2}(x, l)W(t), \quad l > l_0, \]
\[ x(t_0) = x_0. \]
\[ x, f, dW \in \mathbb{R}^N, \]

DATA:

\[ y(t_m) = h(x_m) + [2R(x_m, t)]^{1/2}\epsilon_m \]
where \( m = 1, 2, \ldots, M \)
\[ h, \epsilon : \mathbb{R}^N \rightarrow \mathbb{R}^{N_y} \]
GOAL: estimate moments

(at least) find mean conditioned on data:
\[ x_S(t) = E[ x(t)| y_1, ..., y_M] \]
and
Covariance matrix (uncertainty)
\[ C_S(t) = E[(x(t)-x_S(t))(x(t) -x_S(t))^\top|y_1, ..., y_M] \]

The conditional mean \( x_S(t) \) minimizes
\[ \text{tr} \ C_S(t) = E[| (x(t)-x_S(t))|^2 |y_1, ..., y_M] \]
It is termed the smoother estimate.
Optimal Estimate of Discretized Linear Model with Gaussian Noise

Let $z_i = u(x_i)$ where $x_i \in \Omega$

$Bz + nm = F$
$Dz + nd = Y$

OR

$Mz + N = T$

$\min_z J = <N^T N>$

Least Squares, SVD (Kalman)
A Nonlinear Example

Stochastic Dynamics (Langevin Problem):
\[ dx(t) = f(x(t)) \, dt + \kappa \, dW(t) \]
with
\[ V(x) = -2x^2 + x^4 \]
\[ f(x) = -V'(x) = 4x(1-x^2) \]
\[ \kappa = 0.5 \]

Measurements:
at times \( m \Delta t, \ m=1,\ldots, M \) one observes
\[ y_m := X(t_m) + \rho \, N_m \]
to have measured values \( Y_m, \ m=1,2,\ldots,M \)
Kolmogorov Equation

\[ \frac{\partial P}{\partial t} = -\partial_x [f(x) P] + \kappa^2 \partial_{xx} P/2 \]

\[ P(x,t)_{t=1} = P_s(x) \]
Observations

\[ Y_m \approx y(t_m) \]
BAYESIAN STATEMENT

- \( P(X|D) \) / Prior £ Likelihood
- Use data for the likelihood
- Use model for the prior

\[
P(X|D) \sim \exp(-A_{\text{data}}) \exp(-A_{\text{model}})
\]
Extended Kalman Filter

EKF Results (Hiller et al '94) Langevin Problem
R (Variance of observation errors)

(a) $R=0.01$
$\Delta t = 1$

(b) $R=0.04$
$\Delta t = 1$

(c) $R=0.04$
$\Delta t = 0.25$
Alternative Approaches

- KSP: optimal, but impractical
- ADJOINT/4D-VAR: optimal on linear/Gaussian
  - (Restrepo, Leaf, Griewank, SIAM J. Sci Comp 1995)
- Mean Field Variational Method
  - (Eyink, Restrepo, Alexander, Physica D, 2003)
- enKF (ensemble Kalman Filter)
- Particle Method
- Path Integral Method
Path Integral Method

- Related to simulated annealing
- It could be developed as a black box
- Simple to implement
- Can handle nonlinear/non-Gaussian problems
- Calculates sample moments

**PROBLEM**: Relies on MC!!!
\[ dx(t) = f(x(t), t)dt + [2D(x, t)]^{1/2}dW(t), \quad t > t_0, \]
\[ x(t_0) = x_0. \]

Discretized using explicit Euler-Maruyama scheme

\[ x_{k+1} = x_k + f(x_k, t_k)\delta t + (2D)^{1/2}(x_k, t_k)(W(t_k + \delta t) - W(t_k)), \]
\[ k = 0, 1, 2, \ldots \]
\[ x_{k=0} = x_0. \]
Let $\eta(t_k) = W(t_k + \delta t) - W(t_k)$, at times $t_k, \quad k=0,1,2,...,$

Suppose $\eta(t_k)$ is Gaussian

\[
\text{Prob } \eta(t) \sim \exp\left(-\frac{1}{2} \sum_k |\eta(t_k)|^2\right).
\]

Hence $\exp(-A_{\text{dyn}})$, for $t = t_0, t_1, ..., t_T$

\[
A_{\text{dyn}} = \frac{1}{4} \sum_{k=0}^{T-1} \left[ \frac{(x_{k+1} - x_k)}{\delta t} - f(x_k, t_k) \right] D^{-1}(x_k, t_k) \left[ \frac{(x_{k+1} - x_k)}{\delta t} - f(x_k, t_k) \right]
\]
\[ A_{\text{dyn}} \coloneqq \sum_{k=0}^{T-1} \left[ \left( (x_{k+1} - x_k)/\delta t - f(x_k, t_k) \right) \right] \quad \text{D}^{-1}(x_k, t_k) \left[ \left( (x_{k+1} - x_k)/\delta t - f(x_k, t_k) \right) \right]/4 \]

To include influence of observations use Bayes' rule. This modifies Action:

\[ A_{\text{obs}} = \sum_{m=1}^{M} \left[ h(x(t_m) - y(t_m)) \right] \quad \text{R}^{-1}[h(x(t_m)) - y(t_m)] \]

The Total Action:

\[ A = A_{\text{dyn}} + A_{\text{obs}} \]

The Action is like the log-likelihood.
PIMC Filter Results
Estimation Applied To Steady State Hydrology

- Estimate hydraulic head in domain
- Estimate material properties in domain
- Estimate “best” boundary values
- Estimate all of the above
Simplest Boundary Value Problem

**MODEL:**
\[-\nabla \cdot [K(x)\nabla u] - f(x) = n(x) \quad x \in \Omega\]
\[
u|_{\partial \Omega_i} \quad \text{continuous}
\]
\[
\hat{n} \cdot K(x)\nabla u|_{\partial \Omega_i} \quad \text{continuous}
\]
\[
u|_{\partial \Omega} = U + \theta(x)
\]

**DATA:**
\[
Y(x) = T(u, K) + \rho
\]
\[n(x), \theta(x), \rho(x)\] are known statistical quantities
OUR APPROACH

- USE DATA-DRIVEN CLASSIFICATION: estimates partitioning into homogeneous layers.
  Support Vector Machines
- DISCRETIZE Variational formulation for the model plus constraint (via Lagrange multiplier): constrained minimum satisfies E-L. Coupling of each subproblem is automatically satisfied.
  Weak form (using Dirichlet energy)
- SOLVE nonlinear system in each subdomain: Newton
Data-Driven Classification

Estimate the boundaries between heterogeneous geologic facies

- Data
  \[ K_i = K(x_i), \text{ e.g., conductivity} \]
  \[ h_{jk} = h(x_j, t_k), \text{ e.g., head} \]

- Data are sparse

- Measurements are well differentiated

Measurements of system parameters \((K)\) \(\implies\) forward FD problem
Measurements of system states \((h)\) \(\implies\) inverse FD problem
- Assign indicators to data,
  \[ I(x_i) = 1(0) \quad \text{if} \quad x_i \in M_1(M_2) \]
- \( \mathcal{I}(x, \alpha) \equiv \) an estimate of \( I(x) \)
- \( \min R = \int \| I - \mathcal{I} \| dP(I, \{x\}_{i=1}^N) \)

- **Geostatistics (Kriging)**
  1. the \( L^2 \) norm
  2. the indicator function \( I(x) \) is a random field, and
  3. the choice of sampling locations \( \{x_i\}_{i=1}^N \) as deterministic.  \( \implies \)
  4. Variance: \( \sigma_I^2 = \int (I - \mathcal{I})^2 dP(I) \)

- **SVMs**
  1. the \( L^1 \) norm
  2. the indicator function \( I(x) \) as deterministic, and
  3. the choice of sampling locations \( \{x_i\}_{i=1}^N \) as random.  \( \implies \)
  4. Expected risk: \( \min R_{\exp} = \int |I - \mathcal{I}| dP(\{x\}_{i=1}^N) \)
Support Vector Machines

- Alternative to Kriging
- Very good alternative when sample densities are too low for Kriging
- Highly automated
- Can be incorporated in the solver problem
Heterogeneous Sub-Surface

In each subdomain $i = 1, 2, \ldots, M$

\[
K(x, \omega) = \exp \left[ \sum_{j=1}^{\infty} \kappa_j(\omega) \phi_j(x) \right]
\]

\[
u(x, \omega) = \sum_{j=1}^{\infty} \mu_j(\omega) \phi_j(x)
\]

\[
-\nabla \cdot (K \nabla u) - \bar{f} = n(x, \omega)
\]

\[
E(n) = 0
\]

\[
E(n(x)n(y)) = g(|x - y|)
\]
(Weak) Variational Formulation

- Let $P:=[u,K]$

- Use standard machinery to solve nonlinear problem but use weighted norms (locally in each subdomain).

- Use Newton solver but decide whether to do global estimate of partial estimates (increasing or decreasing the uncertainty in each subdomain).

- Use Galerkin discretization of Newton Systems.
Weak Formulation (no noise)

\[
\phi(P) = \frac{1}{2} \|T(P) - y\|^2 + S(P - P_0)
\]

Dirichlet-like Energy

\[
S(P) = \sum_{i=1}^{M} \left[ \int_{\Omega_i} \left( \frac{1}{2} |\nabla P|^2 + kP^2 \right) dx \right]
\]

\[
G(P) = -\nabla \cdot (K \nabla P) - f = 0
\]

\[
\Phi(P, \Lambda) = \phi(P) - \langle \Lambda, G(P) \rangle
\]

LEADS TO: Find \([ P, \Lambda ]^T\) such that

\[
\langle \Phi'(P), v \rangle - \langle \Lambda, G'(P)v \rangle = 0, \quad \forall \, v \in X
\]

\[
\langle v, G(P) \rangle = 0, \quad \forall \, v \in Y^*
\]

X, Y* Banach spaces
Newton Solution

Find corrections \([\pi, \lambda]^\top\) to \([P, \Lambda]^\top\)

\[
\begin{bmatrix}
\phi'''(P) - [G'''(P)^T \Lambda] & -G''(P)^T \\
G'(P) & 0
\end{bmatrix}
\begin{bmatrix}
\pi \\
\lambda
\end{bmatrix} =
\begin{bmatrix}
-\phi'(P) + G''(P)^T \Lambda \\
-G(P)
\end{bmatrix}
\]

To find Hessians and Jacobians, use ADIFOR/C
Final Comments

- Model error formulation vs. closure?
- Already existing nonlinear solvers.
- Weak formulation automatically takes care of boundary conditions at the layer interfaces.
- Can give a-priori estimates of error.
- Unlike Inverse Method (Tikhonov, e.g.) problem is greatly more numerically stable.
- Use PIMC (see Restrepo, 2007) to benchmark results.
- Constrain number of SVM subdomains to the Newton solve.
Further Information:

http://www.physics.arizona.edu/~restrepo