Data Assimilation and Forecasting
Data Assimilation

- Trendy, good to put in research proposals
  - Less controversial than Bayes, despite being the same thing
- No agreed upon definition in biological sciences
  - You can always say you're doing it
- Well funded
  - National Ecological Observatory Network (NEON)
    - Network of 62 sites, budget of $433M over first 6 yrs
- Biology in a “data-rich” era
DA in Weather Forecasting

• More precise definition of Data Assimilation in atmospheric sciences
  – Subset of what biologists refer to as DA
• Estimation of the current state of the system based on the current data and the previous model prediction
• Under-constrained state-variable estimation
• Fundamental component of analysis/forecast cycle
Characteristics of Data Assimilation Problems

- Often multiple types/sources of data
- May be interested in estimating parameters, state variables, or both
- May be static or dynamic (time series)
  - May be “offline” or “online” / sequential
  - Updating estimates with new information
- Typically working with more complex process models (e.g. simulations)
- Model may be deterministic (parameters known, no process error) or stochastic
Fig. 1.1.1 A view of data assimilation.
Fig. 1.2.1 A classification of models.
Fig. 1.2.2 Classes of models used in data assimilation
Large parts of DA are things you already know how to do!

- “Offline” estimation
  - Everything in the class so far
- Prediction with uncertainty (CI / PI)
- Estimation of state variables (latent variables)
- Nonlinear models
- Partitioning uncertainty / Random effects
- Spatial and temporal models
- Multiple data
Fig. 1.4.1 A classification of the estimation problem.
What do we not yet know?

- The “online”/sequential/recursive problem
  - New data arrive
  - Need to update model
  - Want to avoid recomputing whole analysis
    - Especially because these models are usually big/slow
- Fundamentally a Bayesian problem
  - Previous posterior prediction becomes new prior
- observed position $z$
- predicted position $x^f$
- estimated position $\hat{x}$
- uncertainty in prediction
- uncertainty in observation
Dynamic Models: Uncertainty

- Parameter uncertainty
- Initial Conditions (state)
- Drivers / Boundary Conditions
  - Scenario uncertainty
- Process error
- Measurement uncertainty / Data model
- Random effects
Fig. 2.8.1 Predicted and observed path of Hurricane Donna, where the predictions were based on slight differences in the initial conditions.
Fig. 2.9.1 Probability density functions (pdf) of the stochastic/dynamic solution to \( \frac{dx}{dt} = x \). \( P_0(x) \) and \( P_T(x) \) are the pdf's at \( t = 0 \) and \( t = T \), respectively.
Data Assimilation Models

• Consider the simplest possible DA problem
  – Want to estimate the state of a single state variable $X$
  – We have $n$ independent but noisy observations $Y = \{Y_1, Y_2, \ldots, Y_n\}$ that we assume are Normally distributed around the true state $X$
    \[ Y \sim N(X, \sigma^2) \]
  – We have the posterior predictive distribution (the one you use to calculate the PI) from our process model
    \[ X \sim N(\mu, \tau^2) \]
  – Want to estimate $X | Y$
• Treat the model prediction as our prior
\[ X|Y \sim N \left( Y|X, \sigma^2 \right) N \left( X|\mu, \tau^2 \right) \]

• Recall the general solution
\[ X|Y \sim N \left( \left( n/\sigma^2 + 1/\tau^2 \right)^{-1} \left( \sum \frac{y_i}{\sigma^2} + \frac{\mu}{\tau^2} \right), \left( n/\sigma^2 + 1/\tau^2 \right)^{-1} \right) \]

• We can rearrange the mean and variance to be
\[ E \left[ X|Y \right] = \mu + K (\bar{Y} - \mu) \]
\[ Var \left[ X|Y \right] = (1 - K) \tau^2 \]

• Where K is referred to as the “gain”
\[ K = \tau^2 / (\sigma^2 / n + \tau^2) \]
Fig. 1. Posterior distribution with normal prior and normal likelihood; relatively precise data.
DA “Analysis” Step

- Generalize linear univariate problem to
  - \((n \times 1)\) vector of X's
  - \((p \times 1)\) vector of Y's
  - \((p \times n)\) observation matrix H
  - \((p \times p)\) observation error matrix R
  - \((n \times n)\) forecast covariance matrix P

\[
X|Y \sim N(Y|HX, R) N(X|\mu, P)
\]
\[ X|Y \sim N(Y|HX, R)N(X|\mu, P) \]

- Solves to be

\[ X|Y \sim N\left(\left(H^T R^{-1} H + P^{-1}\right)^{-1} \left(H^T R^{-1} Y + P^{-1}\right), \left(H^T R^{-1} H + P^{-1}\right)^{-1}\right) \]

- Mean and variance simplify to

\[ E[X|Y] = \mu + K(Y - H\mu) \]

\[ \text{Var}[X|Y] = \left(I - KH\right)P \]

\[ K = PH^T \left(R + HPH^T\right)^{-1} \]
Example

- Assume X = \{x1, x2, x3\}, Y = \{y2, y3\}, and the measurement error is iid \( R = \sigma^2 I \)

\[
H = \begin{bmatrix}
0 & 1 & 0 \\
0 & 0 & 1 \\
\end{bmatrix}
\]

- The posterior mean for the unobserved \( x_1 \) is

\[
E[ x_1 ] = \mu_1 + w_{12} ( y_2 - \mu_2 ) + w_{13} ( y_3 - \mu_3 )
\]

\[
w_1^T = (p_{12} \quad p_{13}) \left( \begin{array}{cc}
p_{22} + \sigma^2 & p_{23} \\
p_{32} & p_{33} + \sigma^2 \end{array} \right)^{-1}
\]

If X's are locations and P is a spatial covariance matrix, model is equivalent to Kriging.
Sequential Data Assimilation

• Alternate between Analysis step and Forecast step
• For Analysis step assume a time series of $Y$'s that only become available one $\hat{Y}_t$ at a time
• For Forecast step, assume a linear process model with Normal process error

\[
X_t = M_t X_{t-1} + \nu_t
\]

\[
\nu_t \sim N(0, Q_t)
\]
Forecast Step

- The posterior distribution of $X_t$ given $X_{t-1}$ and the Y's up to $t-1$ is multivariate normal with

\[
X_{t|t-1} = E[ X_t | Y_{1:t-1} ] = M_t X_{t-1|t-1}
\]

\[
P_{t|t-1} = \text{Var} [ X_t | Y_{1:t-1} ] = Q_t + M_t P_{t-1|t-1} M_t^T
\]
Analysis Step

• The posterior distribution of Xt given Xt-1 and the current Y is multivariate normal with mean and variance

\[ X_{t|t} = X_{t|t-1} + K_t \left( Y_t - H_t X_{t|t-1} \right) \]

\[ P_{t|t} = \left( I - K_t H_t \right) P_{t|t-1} \]

\[ K_t = P_{t|t-1} H_t^T \left( H_t P_{t|t-1} H_t^T + R_t \right)^{-1} \]

• This is equivalent to the multivariate model defined before but with the Forecast step as the prior
Kalman Filter

- The linear Normal data assimilation model just introduced is known as the Kalman Filter

- Important to note that the posterior distribution of X at each step only depended upon the PREVIOUS state, the current Forecast, and the current Data

- Updating does not require access to whole data set, computationally efficient

- Assumes all parameters (H, R, M, Q) are known
Fig. 4. First-order autoregressive process simulation Kalman filter state estimates, truth and data.
Fig. 5. First-order autoregressive process simulation Kalman filter variance estimates.
Extensions to the Kalman Filter

• Assumption of Normal error, linear models fairly restrictive
  – How do we relax this assumption?

• Extended Kalman Filter
  – Based on local linearization of the model

• Monte Carlo Methods
  – Ensemble Kalman Filter
  – State Space Model
Ensemble Kalman Filter (EnKF)

- Uses Monte Carlo samples to approximate forecast distribution
- Draw m samples from the posterior at t-1
- Forecast: run process model + process error for sample to generate posterior predictive dist'n

\[ p^m(X_t|Y_{1:t-1}) = \frac{1}{m} \sum p(X_t|X_{t-1}^{i}|t-1) \]

- Analysis:

\[ p^m(X_t|Y_{1:t}) \propto \frac{1}{m} p(Y_t|X_t) \sum p(X_t|X_{t-1}^{i}|t-1) \]
Fig. 30.1.1 A view of ensemble filtering.
EnKF: Nonlinear process model, Linear Gaussian observation error

- Special case if $Y = N(HX, R)$
- Forecast:
  $$X^i_{t|t-1} = N(f(X^i_{t-1|t-1}), Q)$$

- Approx Forecast dist'n with Normal with sample mean & sample covariance (only true if linear)
- Use standard Analysis step from Kalman Filter
  - Normal prior + Normal Likelihood
Fig. 6. First-order autoregressive process simulation Kalman filter state estimates and EnKF filter state estimates based on 10 ensemble members.
"True" state of ocean for the model, given its resolution and physics

As close to "true" state as observation density and observation error allow

Model forecast

Small correction to short-term forecast
Ensemble Analyses

• More generally, ensembles can be used to propagate many forms of uncertainty besides just initial state
  – Parameters
  – Covariates / Drivers / Scenarios
  – Model choice

• For example, IPCC climate forecasts are based on ensembles of different models for a discrete set of scenarios
Ensemble Statistics

• Ensemble means across models tend to perform better than any even the best individual model

• Can calculate variance, quantiles, etc same as a CI or PI (depending on whether you add the data model uncertainty back in)
State Space Model Revisited

- Previously have considered the State Space model in an “offline” mode
  - Likelihood of $X_t$ depends on $X_{t-1}$ AND $X_{t+1}$
- Easy to implement in an “online” mode where $X_t$ depends only on $X_{t-1}$
- Flexible framework for either state estimation or both state and parameter estimation
- At each Analysis step, treat parameter posteriors from previous step as priors for this step
Fig. 9. MCMC posterior mean and 2.5% and 97.5%-tiles for AR(1) simulation described in Section 3.1.1.