DATA ASSIMILATION

LET'S CONSIDER TO HAVE A SYSTEM GIVEN BY:

$$\frac{d\underline{y}}{dt} = f(t, \underline{y})$$

WITH THE INITIAL STATE OF THE SYSTEM GIVEN BY:

$$\underline{y}(0) = \underline{y}_0$$

PROVIDED THAT $f(t, \underline{y})$ IS WELL BEHAVED, I.E. CONTINUOUS AND DIFFERENTIABLE, THEN A SOLUTION TO THE PROBLEM CAN BE UNIQUELY DETERMINED. WE HAVE AN N-DIMENSIONAL VECTOR WITH N INITIAL CONDITIONS SO THE PROBLEM IS WELL POSED.

UP TO NOW WE HAVE ASSUMED TO BE ABLE TO PERFECTLY PRESCRIBE BOTH, BUT ACTUALLY THIS IS IMPOSSIBLE.

GEORGE BOX: "ALL MODELS ARE WRONG BUT SOMEONE IS USEFUL"

WE HAVE TO ACCOUNT FOR UNMODELLED DYNAMICS AND ERROR IN THE PRESCRIBED INITIAL CONDITIONS

$$\frac{d}{dt}\underline{y} = f(t, \underline{y}) + \underline{q}_1(t) \qquad (\cdot)$$
$$\underline{y}(0) = \underline{y}_0 + \underline{q}_2(t)$$

FOR A STRONGLY NONLINEAR SYSTEM THE PRESENCE OF q_1 AND q_2 MAKE IMPOSSIBLE TO USE $(\cdot)$ FOR AN ACCURATE FORECASTING OF THE STATE OF THE SYSTEM. WE NEED TO FIND AN EFFECTIVE STRATEGY FOR DEALING WITH THE ERROR EFFECTS.

THE IDEA OF DATA ASSIMILATION IS TO ASSIMILATE EXPERIMENTAL MEASUREMENTS DIRECTLY INTO THE MODEL IN ORDER TO INFORM THE DYNAMICS. CONSIDERING A SET OF M MEASUREMENTS:

$$g(t, \underline{y}) + \underline{q}_3 = 0 \qquad (\cdot\cdot)$$

WHERE ALSO HERE WE HAVE MEASUREMENT ERROR ASSOCIATED FOR EXAMPLE TO THE NOISE.

THE ADDITION OF $(\cdot\cdot)$ MAKES THE SYSTEM OVERDETERMINED AND SO NO SOLUTION EXISTS IN GENERAL

WE WANT TO FIND A WAY TO MINIMIZE THE VARIANCE AND TO FORMULATE THE PROBLEM AS QUADRATIC IN ORDER TO OBTAIN A CONVEX OPTIMIZATION PROBLEM.

$$J(\underline{y}) = \int_0^T \int_0^T \underline{q}_1^T(t_1) \underline{\underline{W}}_1 \underline{q}_1(t_2)\, dt_1 dt_2 + \underline{q}_2^T \underline{\underline{W}}_2 \underline{q}_2 + \underline{q}_3^T \underline{\underline{W}}_3 \underline{q}_3$$

x - prediction from the model

y - prediction from measurement.

THE IDEA IS TO COMBINE THESE TWO MEASUREMENT TO OBTAIN A BETTER PREDICTION FOR THE TRUE x.

FOR SIMPLICITY WE WILL CONSIDER GAUSSIAN DISTRIBUTED RANDOM VARIABLES. THE PROBABILITY OF FINDING x CONDITIONED ON HAVING MEASURED y IS GIVEN BY THE BAYES RULE.

$$p(x|y) = \frac{p(y|x)\, p(x)}{p(y)}$$

WE CONSIDER THE FOLLOWING PROBABILITY DENSITY DISTRIBUTIONS:

$$p(y|x) = C_1 \exp\left[-\frac{1}{2}\left(\frac{y-x}{\sigma_y}\right)^2\right]$$

$$p(x) = C_2 \exp\left[-\frac{1}{2}\left(\frac{x-x_0}{\sigma_0}\right)^2\right]$$

WHERE σ_Y IS THE ERROR VARIANCE FOR THE OBSERVATION, x_0 IS THE PREDICTED MODEL MEAN AND σ_0 IS THE ASSOCIATED ERROR VARIANCE.

$$p(x|y) = C_3 \exp\left[-\frac{1}{2}\left(\frac{y-x}{\sigma_Y}\right)^2\right] \exp\left[-\frac{1}{2}\left(\frac{x-x_0}{\sigma_0}\right)^2\right]$$

OUR GOAL IS TO MODIFY x FROM ITS DEFAULT VALUE OF x_0 IN LIGHT OF OUR OBSERVATION y.

$$J(x) = -\log[p(x|y)] + \log(C_3) = \frac{1}{2}\left(\frac{y-x}{\sigma_Y}\right)^2 + \frac{1}{2}\left(\frac{x-x_0}{\sigma_0}\right)^2$$

THE MINIMUM IS OBTAINED:

$$\frac{d}{dx} J(\bar{x}) = 0$$

$$\bar{x} = \left(\frac{\sigma_y^2}{\sigma_y^2 + \sigma_0^2}\right) x_0 + \left(\frac{\sigma_0^2}{\sigma_y^2 + \sigma_0^2}\right) y$$

PERFECT SENSOR: $\sigma_Y = 0 \rightarrow \bar{x} = y$

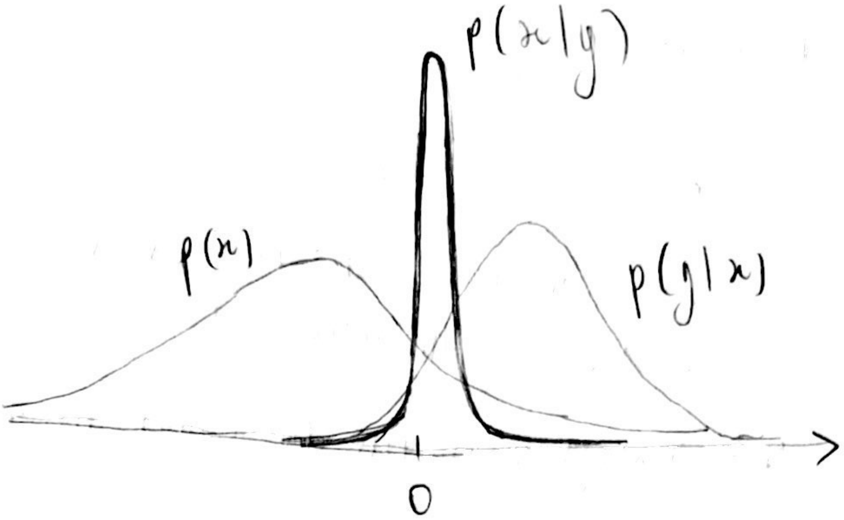
PERFECT MODEL: $\sigma_0 = 0 \rightarrow \bar{x} = x_0$

THE ERROR VARIANCE FOR $\bar{x}$ CAN BE COMPUTED.

$$\bar{\sigma}^2 = \frac{\sigma_0^2}{1 + (\sigma_0^2/\sigma_Y^2)} = \frac{\sigma_Y^2}{1 + (\sigma_Y^2/\sigma_0^2)} < \sigma_0^2, \sigma_Y^2$$

THE VARIANCE OF THE ASSIMILATED MODEL IS LESS THAN THE VARIANCE OF THE MODEL ALONE ~~OR~~ OF THE MEASUREMENT ALONE

BELOW A FIGURE TO BETTER UNDERSTAND THE EFFECT.



WE CAN ALSO EXPRESS THE PREVIOUS EQUATION AS:

$$\bar{x} = x_0 + K\underbrace{(y - x_0)}_{\text{INNOVATION}}$$

WHERE

$$K = \frac{\sigma_0^2}{\sigma_0^2 + \sigma_y^2} \leq 1 \qquad \text{KALMAN FILTER}$$

THE PREDICTED VALUE OF x IS A LINEAR COMBINATION OF ITS MODEL PREDICTION x_0 AND THE INNOVATION.

IN GENERAL, THE MEASUREMENTS ARE NOT ALIGNED WITH THE GRID USED TO EVALUATE OUR MODEL. THE QUESTION IS HOW TO OVERLAY THE IRREGULARLY SPACED OBSERVATIONS ONTO THE REGULARLY SPACED STATE VARIABLES. WE CAN MAP THE STATE VECTOR TO THE OBSERVATIONS AS:

$$\underline{y}(t) = \underline{\underline{H}}\,\underline{x}(t) + \underline{q}_3$$

A ONE DIMENSIONAL KALMAN FILTER

WE WILL CONSIDER HOW TO DERIVE THE PROJECTION OPERATOR $\underline{\underline{H}}$.

$$x_{k+1} = f(x_k) + q_{k+1}$$

WHERE $f(x_k)$ IS THE FLOW MAP. THE MODEL APPROXIMATION OF THE GIVEN SYSTEM IS:

$$x_{o_{k+1}} = f(x_{o_k})$$

WHERE x_{o_k} IS THE BEST ESTIMATE OF THE STATE AT TIME t_k AND $x_{o_{k+1}}$ IS THE FORECAST OF THE DYNAMICS.

THE ERROR BETWEEN THE TRUTH AND THE FORECAST AT TIME t_{k+1} IS:

$$x_{k+1} - x_{o_{k+1}} = f(x_k) - f(x_{o_k}) + q_{k+1}$$

BY TAYLOR EXPANDING $f(x_k)$ AROUND x_{o_k}:

$$x_{k+1} - x_{o_{k+1}} = (x_k - x_{o_k}) f'(x_{o_k}) + \frac{1}{2}(x_k - x_{o_k})^2 f''(x_{o_k}) + \frac{1}{6}(x_k - x_{o_k})^3 f'''(x_{o_k}) + \dots + q_{k+1}$$

THE ERROR VARIANCE IS COMPUTED AS:

$$\mathbb{E}\left[(x_{k+1} - x_{o_{k+1}})^2\right] = \mathbb{E}\left[(x_k - x_{o_k})^2\right](f'(x_{o_k}))^2 + \text{h.o.t.} + \mathbb{E}\left[q_{k+1}^2\right]$$

HIGHER ORDER TERMS

IN FIRST APPROXIMATION WE CAN NEGLECT THE h.o.t.. DEFINING:

$$P_{k+1} = \mathbb{E}\left[(x_{k+1} - x_{o_{k+1}})^2\right]$$

$$P_k = \mathbb{E}\left[(x_k - x_{o_k})^2\right]$$

WE OBTAIN:

$$P_{k+1} = P_k (f'(x_{o_k}))^2 + \mathbb{E}\left[q_{k+1}^2\right]$$

ACCOUNT FOR THE DYNAMICS

ACCOUNT FOR ERRORS IN ESTIMATING THE INITIAL STATE

TO MAKE A DATA ASSIMILATED PREDICTION, $\bar{x}_{k+1}$.

$$\bar{x}_{k+1} = x_{o_{k+1}} + K_{k+1}(y_{k+1} - x_{o_{k+1}})$$

WHERE:

$$K_{k+1} = \frac{P_{k+1}}{P_{k+1} + R}$$

OBSERVATION ERROR VARIANCE

EXTENDED KALMAN FILTER (EKF)

1) THE VECTOR EKF

NOW WE HAVE

$$\underline{x}_{k+1} = f(\underline{x}_k) + \underline{q}_{k+1}$$

$$\underline{x}_{o_{k+1}} = f(\underline{x}_{o_k})$$

WHERE $\underline{x} \in \mathbb{R}^n$, $\underline{y} \in \mathbb{R}^m$ AND USUALLY $m << n$.

BY TAYLOR EXPANDING NOW WE WILL OBTAIN THE COVARIANCE EVOLUTION.

$$P_{k+1} = \underbrace{J(f)}_{\uparrow \text{ JACOBIAN}} P_k J(f)^T + Q$$

$$\underline{x}_{k+1} = x_{o_{k+1}} + \underline{\underline{K}}_{k+1}(y_{k+1} - \underline{\underline{H}}\,\underline{x}_{o_{k+1}})$$

WHERE THE KALMAN GAIN IS NOW.

$$\underline{\underline{K}}_{k+1} = \frac{\underline{\underline{P}}_{k+1}\underline{\underline{H}}^T}{\underline{\underline{H}}\,\underline{\underline{P}}_{k+1}\underline{\underline{H}}^T + \underline{\underline{R}}}$$

SO, $\underline{\underline{H}}$ SERVES TO OVERLAYING THE DATA MEASUREMENT LOCATIONS WITH THE GRID USED FOR COMPUTATIONALLY EVOLVING THE MODEL

DYNAMICS FORWARD IN TIME.

THE DRAWBACK OF THIS EKF IS THE COMPUTATIONAL ISSUE ASSOCIATED TO THE COMPUTATION OF THE INNOVATION WHEN THE STATE VECTOR IS VERY HIGH. ONE POSSIBILITY TO OVERCOME THIS PROBLEM IS TO USE AN ENSAMBLE OF KALMAN FILTERS (EnKF) IN WHICH THE DOMAIN IS DIVIDED INTO SMALLER SUBDOMAIN

ONE INTERESTING APPLICATION OF DATA ASSIMILATION IS TO THE LORENZ EQUATIONS, A SIMPLIFIED MODEL OF CONVECTIVE DRIVEN ATMOSPHERIC MOTION. THIS IS AN EXAMPLE OF CAHOTIC BEHAVIOR SINCE A SMALL CHANGE IN INITIAL CONDITIONS PRODUCE A DRIFT OF THE TRAJECTORY.

$$x' = \sigma(y - x)$$

$$y' = rx - y - xz$$

$$z' = xy - bz$$