STA 4103/5107 Computational Methods in Statistics II

Department of Statistics
Florida State University

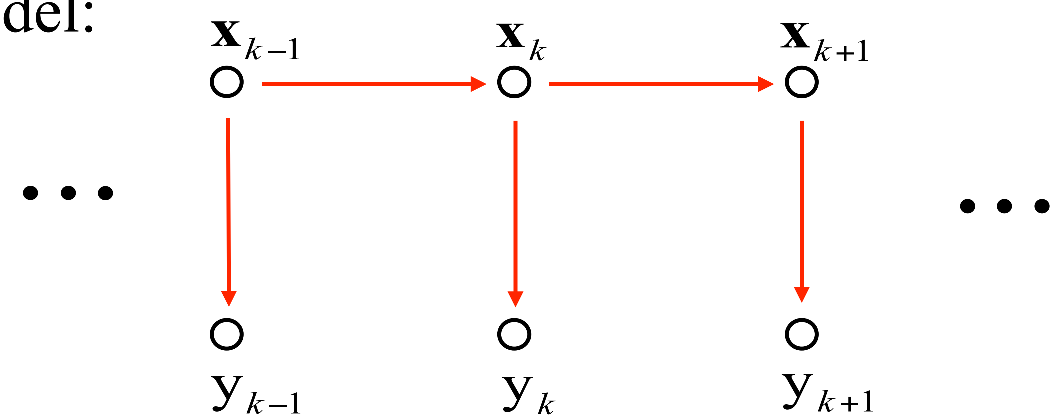
Class 11
February 14, 2017

Review: Nonlinear Filtering Problem

- *State vector* $x_k \in \mathbf{R}^d$, observation vector $y_k \in \mathbf{R}^c$:
- **State Equation (prior):** $\boldsymbol{x_k = \mathrm{F}(x_{k-1}) + w_k}$

  Observation equation (likelihood): $\boldsymbol{y_k = \mathrm{G}(x_k) + q_k}$
- Independence Assumptions:

  $f(x_k \mid x_1, \ldots, x_{k-1}) = f(x_k \mid x_{k-1}),\ f(y_k \mid y_1, \ldots, y_{k-1}, x_k) = f(y_k \mid x_k).$
- Graphical model:



Review: Prediction and Update Equations

- **Goal: Estimate**

$$f(x_k \mid y_1, y_2, \ldots, y_k)$$

- **Prediction equation**

$$f(x_k \mid y_1, y_2, \ldots, y_{k-1}) = \int_{x_{k-1}} f(x_k, x_{k-1} \mid y_1, y_2, \ldots, y_{k-1})\, dx_{k-1}$$

$$= \int_{x_{k-1}} f(x_k \mid x_{k-1}) f(x_{k-1} \mid y_1, y_2, \ldots, y_{k-1})\, dx_{k-1}$$

- **Update equation**

$$f(x_k \mid y_1, y_2, \ldots, y_k) = \frac{f(x_k, y_1, y_2, \ldots, y_k)}{f(y_1, y_2, \ldots, y_k)}$$

$$= \frac{f(y_k \mid x_k) f(x_k \mid y_1, y_2, \ldots, y_{k-1})}{f(y_k \mid y_1, y_2, \ldots, y_{k-1})}$$

Review: Notation in Kalman Filter

- $x_k \in R^d$: internal state at *k*th frame (hidden random variable, e.g. position of the object in the image).

  $X_k = [x_1, x_2, \ldots, x_k]^T$: history up to time step *k*

- $y_k \in R^c$: measurement at *k*th frame (observable random variable, e.g. the given image).

  $Y_k = [y_1, y_2, \ldots, y_k]^T$: history up to time step *k*

- Goal:

  Estimating the posterior probability $p(x_k \mid Y_k)$

Kalman Filter: Likelihood Model

- *Generative model* for the observation:

$$y_k = H_k \ x_k + q_k$$

$$H_k \in R^{c \times d}, \quad q_k \sim N(0, Q_k), \quad Q_k \in R^{c \times c}$$

- The likelihood model is equivalent to that

$$y_k \mid x_k \sim N(H_k \ x_k, Q_k)$$

- The conditional probability has explicit form:

$$p(y_k | x_k) = \frac{1}{((2\pi)^c \det(Q_k))^{1/2}} \exp(-\frac{1}{2}(y_k - H_k x_k)^T Q_k^{-1} (y_k - H_k x_k))$$

Kalman Filter: Prior Model

- *Temporal prior* of the state:

$$x_k = A_k \ x_{k-1} + w_k$$

$$A_k \in R^{d \times d}, \quad w_k \sim N(0, W_k), \quad W_k \in R^{d \times d}$$

- The prior model is equivalent to that

$$x_k \mid x_{k-1} \sim N(A_k \ x_{k-1}, W_k)$$

- The conditional probability has explicit form:

$$p(x_k | x_{k-1}) = \frac{1}{((2\pi)^d \det(W_k))^{1/2}} \exp(-\frac{1}{2}(x_k - A_k x_{k-1})^T W_k^{-1}(x_k - A_k x_{k-1}))$$

Kalman Filter Model

Definition:

System Equation:

$$x_k = A_k x_{k-1} + w_k, \qquad w_k \in N(0, W_k)$$

$$k = 2, 3, \cdots$$

Measurement Equation:

$$y_k = H_k x_k + q_k, \qquad q_k \in N(0, Q_k)$$

$$k = 1, 2, \cdots$$

Assumption:

All random variables have Gaussian distributions and they are linearly related.

Learning Kalman Model

- In practice, the parameters in the model need to be estimated from training data. (In training data, we know both hidden states and measurements.)

- Common simplification: $\mathrm{A}_k, \mathrm{H}_k, \mathrm{W}_k, \mathrm{Q}_k$ are constant over time (independent of k).

- The A, H, W, Q can be estimated by maximizing the joint probability $p(\mathrm{X}_M, Y_M)$. In fact,

$$
\begin{aligned}
p(\mathrm{X}_M, \mathrm{Y}_M) &= p(\mathrm{X}_M) p(\mathrm{Y}_M | \mathrm{X}_M) \\
&= [p(\mathrm{x}_1) \prod_{k=2}^{M} p(\mathrm{x}_k | \mathrm{x}_{k-1})][\prod_{k=1}^{M} p(\mathrm{y}_k | \mathrm{x}_k)]
\end{aligned}
$$

Splitting the Joint Distribution

$$\arg\max_{A,W,H,Q} p(X_M, Y_M)$$

$$= \{\arg\max_{A,W} p(X_M), \arg\max_{H,Q} p(Y_M | X_M)\}$$

$$= \{\arg\min_{A,W} f(A,W), \arg\min_{H,Q} g(H,Q)\}$$

where

$$f(A,W) = -\alpha \log p(X_M) = \sum_{k=2}^{M} [\log(\det W) + (x_k - Ax_{k-1})^T W^{-1} (x_k - Ax_{k-1})],$$

$$g(H,Q) = -\beta \log p(Y_M | X_M) = \sum_{k=1}^{M} [\log(\det Q) + (y_k - Hx_k)^T Q^{-1} (y_k - Hx_k)].$$

How to optimize functions with matrix variables ?

Matrix Calculus

- Definition: assume $X = (x_{ij})_{mn}$, then

$$d / dX = \begin{pmatrix} d / dx_{11} & \cdots & d / dx_{1n} \\ \vdots & \ddots & \vdots \\ d / dx_{m1} & \cdots & d / dx_{mn} \end{pmatrix}$$

- Quadratic Products:

$$d/d\mathbf{X}\ ((\mathbf{Xa+b})^T\mathbf{C}(\mathbf{Xa+b})) = (\mathbf{C+C}^T)(\mathbf{Xa+b})\mathbf{a}^T$$

- Determinant: $d/d\mathbf{X}\ (\log(\det(\mathbf{X}))) = \mathbf{X}^{-T}$
- Inverse: $d/d\mathbf{X}\ (\mathbf{a}^T\mathbf{X}^{-1}\mathbf{b}) = -\mathbf{X}^{-T}\mathbf{ab}^T\mathbf{X}^{-T}$

(upper case: matrix, lower case: column vector)

http://www.ee.ic.ac.uk/hp/staff/dmb/matrix/calculus.html

Detailed Steps

Show one example on prior probability:

$$f(\mathrm{A},\mathrm{W}) = \sum_{k=2}^{M} [\log(\det \mathrm{W}) + (\mathrm{x}_k - \mathrm{A}\mathrm{x}_{k-1})^T \mathrm{W}^{-1} (\mathrm{x}_k - \mathrm{A}\mathrm{x}_{k-1})]$$

i) $$\frac{\partial}{\partial \mathrm{A}} f(\mathrm{A},\mathrm{W}) = \sum_{k=2}^{M} (\mathrm{W}^{-1} + \mathrm{W}^{-1})(\mathrm{x}_k - \mathrm{A}\mathrm{x}_{k-1})(-\mathrm{x}_{k-1})^T = 0$$

$$\Rightarrow \sum_{k=2}^{M} \mathrm{x}_k \mathrm{x}_{k-1}^T = \mathrm{A} \sum_{k=2}^{M} \mathrm{x}_{k-1} \mathrm{x}_{k-1}^T \quad \Rightarrow \quad \mathrm{A} = \left(\sum_{k=2}^{M} \mathrm{x}_k \mathrm{x}_{k-1}^T \right) \left(\sum_{k=2}^{M} \mathrm{x}_{k-1} \mathrm{x}_{k-1}^T \right)^{-1}$$

ii) $$\frac{\partial}{\partial \mathrm{W}} f(\mathrm{A},\mathrm{W}) = \sum_{k=2}^{M} (\mathrm{W}^{-1} - \mathrm{W}^{-1} (\mathrm{x}_k - \mathrm{A}\mathrm{x}_{k-1})(\mathrm{x}_k - \mathrm{A}\mathrm{x}_{k-1})^T \mathrm{W}^{-1}) = 0$$

$$\Rightarrow \quad \mathrm{W} = \sum_{k=2}^{M} (\mathrm{x}_k - \mathrm{A}\mathrm{x}_{k-1})(\mathrm{x}_k - \mathrm{A}\mathrm{x}_{k-1})^T / (M-1)$$

Closed-form Solutions

$$A = \left(\sum_{k=2}^{M} x_k x_{k-1}^T\right)\left(\sum_{k=2}^{M} x_{k-1} x_{k-1}^T\right)^{-1},$$

$$W = \frac{1}{M-1}\left(\sum_{k=2}^{M} x_k x_k^T - A\sum_{k=2}^{M} x_{k-1} x_k^T\right),$$

$$H = \left(\sum_{k=1}^{M} y_k x_k^T\right)\left(\sum_{k=1}^{M} x_k x_k^T\right)^{-1},$$

$$Q = \frac{1}{M}\left(\sum_{k=1}^{M} y_k y_k^T - H\sum_{k=1}^{M} x_k y_k^T\right).$$

Recursive Estimation

$$\boxed{p(\mathrm{x}_k | \mathrm{Y}_k) = \kappa\, p(\mathrm{y}_k | \mathrm{x}_k) \int p(\mathrm{x}_k | \mathrm{x}_{k-1}) p(\mathrm{x}_{k-1} | \mathrm{Y}_{k-1}) d\mathrm{x}_{k-1}}$$

Time update:

posterior at previous step:

$$p(\mathrm{x}_{k-1} | \mathrm{Y}_{k-1})$$

temporal prior:

$$p(\mathrm{x}_k | \mathrm{x}_{k-1})$$

prior distribution:

$$p(\mathrm{x}_k | \mathrm{Y}_{k-1}) = \int p(\mathrm{x}_k | \mathrm{x}_{k-1}) p(\mathrm{x}_{k-1} | \mathrm{Y}_{k-1}) d\mathrm{x}_{k-1}$$

Measurement update:

prior distribution:

$$p(\mathrm{x}_k | \mathrm{Y}_{k-1})$$

likelihood:

$$p(\mathrm{y}_k | \mathrm{x}_k)$$

posterior distribution:

$$p(\mathrm{x}_k | \mathrm{Y}_k) = \kappa\, p(\mathrm{y}_k | \mathrm{x}_k) p(\mathrm{x}_k | \mathrm{Y}_{k-1})$$

Basic Properties of Normal Distributions

- $\mathrm{x},\ \mathrm{y} \in \mathrm{R}^d$ independent random vectors,

$$\mathrm{x} \sim N(0, \mathrm{A}), \quad \mathrm{y} \sim N(0, \mathrm{B}),$$

then, i) for any matrix $\mathrm{C} \in \mathrm{R}^{d \times d}$, $\mathrm{Cx} \sim N(0, \mathrm{CAC}^T)$;

ii) $\mathrm{x} + \mathrm{y} \sim N(0, \mathrm{A} + \mathrm{B})$.

These two properties can be derived by the following rules:

i)

$$Cov(C\mathrm{x}) = C(Cov(\mathrm{x}))C^T = CAC^T.$$

ii)

$$Cov(\mathrm{x} + \mathrm{y}) = Cov(\mathrm{x}) + Cov(\mathrm{y}) = \mathrm{A} + \mathrm{B}.$$

Kalman Filtering, Step I: Time Update

Assume: $\mathrm{x}_{k-1} \mid \mathrm{Y}_{k-1} \sim N(\hat{\mathrm{x}}_{k-1}, \mathrm{P}_{k-1})$

$$\Leftrightarrow \quad \mathrm{x}_{k\text{-}1} = \hat{\mathrm{x}}_{k-1} + \mathrm{e}_{k-1}, \;\; \mathrm{e}_{k-1} \sim N(0, \mathrm{P}_{k-1})$$

System equation: $\mathrm{x}_k = \mathrm{A}_k \mathrm{x}_{k-1} + \mathrm{w}_k, \quad \mathrm{w}_k \sim N(0, \mathrm{W}_k)$

$$\Rightarrow \quad \mathrm{x}_k = \mathrm{A}_k \hat{\mathrm{x}}_{k-1} + \mathrm{A}_k \mathrm{e}_{k-1} + \mathbf{w}_k$$

Use properties i) and ii):

$$\mathrm{A}_k \mathrm{e}_{k\text{-}1} + \mathrm{w}_k \sim N(0, \mathrm{A}_k \mathrm{P}_{k-1} \mathrm{A}_k^T + \mathrm{W}_k)$$

Let

$$\hat{\mathrm{x}}_k^- = \mathrm{A}_k \hat{\mathrm{x}}_{k\text{-}1}, \quad \mathrm{P}_k^- = \mathrm{A}_k \mathrm{P}_{k-1} \mathrm{A}_k^T + \mathrm{W}_k,$$

then,

$$\mathrm{x}_k \mid \mathrm{Y}_{k-1} \sim N(\hat{\mathrm{x}}_k^-, \mathrm{P}_k^-)$$

Step II: Measurement Update

Time update:

$$p(\mathrm{x}_k \mid \mathrm{Y}_{k-1}) \propto \exp(-(\mathrm{x}_k - \hat{\mathrm{x}}_k^-)^T (\mathrm{P}_k^-)^{-1} (\mathrm{x}_k - \hat{\mathrm{x}}_k^-) / 2)$$

Measurement equation:

$$p(\mathrm{y}_k | \mathrm{x}_k) \propto \exp(-(\mathrm{y}_k - \mathrm{H}_k \mathrm{x}_k)^T \mathrm{Q}_k^{-1} (\mathrm{y}_k - \mathrm{H}_k \mathrm{x}_k) / 2)$$

Recursive update:

(details omitted)

$$p(\mathrm{x}_k \mid \mathrm{Y}_k) \propto p(\mathrm{y}_k \mid \mathrm{x}_k) p(\mathrm{x}_k \mid \mathrm{Y}_{k-1})$$

$$\propto \exp(-\frac{1}{2}(\mathrm{x}_k - \hat{\mathrm{x}}_k)^T (\mathrm{P}_k)^{-1} (\mathrm{x}_k - \hat{\mathrm{x}}_k))$$

where, $\hat{\mathrm{x}}_k = \hat{\mathrm{x}}_k^- + \mathrm{K}_k(\mathrm{y}_k - \mathrm{H}_k \hat{\mathrm{x}}_k^-)$, $\mathrm{P}_k = (\mathrm{I} - \mathrm{K}_k \mathrm{H}_k) \mathrm{P}_k^-$,

$$\mathrm{K}_k = \mathrm{P}_k^- \mathrm{H}_k^T (\mathrm{H}_k \mathrm{P}_k^- \mathrm{H}_k^T + \mathrm{Q}_k)^{-1}.$$

That is,

$$\mathrm{x}_k \mid \mathrm{Y}_k \sim N(\hat{\mathrm{x}}_k, \mathrm{P}_k)$$

Kalman Filter Algorithm

Time Update

Prior estimate

$$\hat{x}_k^- = A_k \hat{x}_{k-1}$$

Error covariance

$$P_k^- = A_k P_{k-1} A_k^T + W_k$$

Measurement Update

Posterior estimate

$$\hat{x}_k = \hat{x}_k^- + K_k (y_k - H_k \hat{x}_k^-)$$

Error covariance

$$P_k = (I - K_k H_k) P_k^-$$

Kalman gain

$$K_k = P_k^- H_k^T (H_k P_k^- H_k^T + Q_k)^{-1}$$

previous estimate of $\hat{x}_{k-1}$ and P_{k-1}

(initially: we can let $\hat{x}_1 = 0, P_1 = 0$)

Welch & Bishop, *An Introduction to the Kalman Filter*, 2006

Estimation Accuracy

- R^2 Error is commonly-used as a criterion to measure the estimation accuracy:

  Let x_k denote the true state and $\hat{x}_k$ denote the estimate. Then

$$R^2 = 1 - \frac{\sum_k \| x_k - \hat{x}_k \|^2}{\sum_k \| x_k - \bar{x} \|^2}$$

- R^2 Error can also be measured component-wise. That is, for the i-th component, we have

$$R_i^2 = 1 - \frac{\sum_k (x_{k,i} - \hat{x}_{k,i})^2}{\sum_k (x_{k,i} - \bar{x}_{\cdot,i})^2}$$