Kalman Filter: the Basics

Gang Liu

August 13, 2002

A few important concepts

- *Filtering* refers to estimating the state vector at the current time, based upon all past measurements.
- *Prediction* refers to estimating the state at a future time.
- *Smoothing* means estimating the value of the state at some prior time, based on all measurements taken up to the current time.
- An *estimate*, $\hat{\mathbf{x}}$, is the computed value of a quantity, $\mathbf{x}$, based upon a set of measurements, $\mathbf{y}$.

Desired properties

- An *unbiased* estimate is one whose expected value is the same as that of the quantity being estimated.
- A *minimum variance* (unbiased) estimate has the property that its error variance is less than or equal to that of any other unbiased estimate.
- A *consistent* estimate is one which converges to the true value of $\mathbf{x}$, as the number of measurements increases.

Assumptions to start with

- The measurement process is linear and stationary;
- Noise in measurement is random and additive.

Thus the measurement $\mathbf{y}$ can be expressed in terms of the state variable $\mathbf{x}$ and observation noise $\mathbf{v}$ as the following.

$$\mathbf{y} = H\mathbf{x} + \mathbf{v} \tag{1}$$

where H is called the observation matrix.

Which approach to choose?

There are many ways of estimating the true state variable $\mathbf{x}$ from noise corrupted measurement $\mathbf{y}$. Depending on the knowledge we have about system state $\mathbf{x}$ and noise $\mathbf{v}$, we can choose different approaches to calculate $\hat{\mathbf{x}}$, the estimate of $\mathbf{x}$.

- *Least-square-error estimation.* Suppose that we know nothing about the statistical properties of the system state $\mathbf{x}$ and the noise $\mathbf{v}$, one can select $\hat{\mathbf{x}}$ such that it minimizes the sum of squares of the deviations, $y_i - \hat{y}_i$, i.e., minimizes the quantity,

$$J = (\mathbf{y} - H\hat{\mathbf{x}})^T(\mathbf{y} - H\hat{\mathbf{x}}) \tag{2}$$

  The resulting least-square estimation can be found by setting $\partial J/\partial\hat{\mathbf{x}} = 0$.

$$\hat{\mathbf{x}} = (H^T H)^{-1}H^T\mathbf{y} \tag{3}$$

  Alternatively, one can choose to minimize the weighted sum of squares of deviations,

$$J = (\mathbf{y} - H\hat{\mathbf{x}})^T W(\mathbf{y} - H\hat{\mathbf{x}}) \tag{4}$$

  where W is a symmetric, positive definite weighting matrix. The *weighted-least-squares* estimate is,

$$\hat{\mathbf{x}} = (H^T W H)^{-1}H^T W\mathbf{y} \tag{5}$$

  Neither the plain nor the weighted least-squares estimation has any direct probabilistic interpretation. They are deterministic approaches.

- *Maximum liklihood estimation.* Suppose that we now have some knowledge about the measurement noise $\mathbf{v}$, say, the probability distribution of $\mathbf{v}$ is known. Then we can calculate $\hat{\mathbf{x}}$ such that it maximizes the probability of the measurements $\mathbf{y}$ that actually occured, taking into account known statistical properties of $\mathbf{v}$. But at this time, *there is still no statistical model assumed for system state* $\mathbf{x}$.

  In our sample problem, the conditional probability density function for $\mathbf{y}$, conditioned on a given value for $\mathbf{x}$, is just the density for $\mathbf{v}$ centered around $H\mathbf{x}$. If we assume the $\mathbf{v}$ is a zero mean Gaussian noise with covariance matrix P, we have,

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{(2\pi)^{k/2}|R|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{y} - H\mathbf{x})^T R^{-1}(\mathbf{y} - H\mathbf{x})\right] \tag{6}$$

  To maximize $p(\mathbf{y}|\mathbf{x})$ is the same as to minimize the exponent in the brackets. This is equivalent to minimizing the cost function in Eq (4), although now the matrix R in Eq (6) has a probabilistic meaning now.

- *Bayesian estimation* comes in when there is knowledge about both the system state $\mathbf{x}$ and the measurement noise $\mathbf{v}$. In this case, one seeks to compute $\hat{\mathbf{x}}$ from $p(\mathbf{x}|\mathbf{y})$, the *a posteriori* condition density function, which contains all the statistical information of interest. By Bayesian's theorem,

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{y})} \tag{7}$$

  where $p(\mathbf{x})$ is the *a priori* probability density function of $\mathbf{x}$, and $p(\mathbf{y})$ is the probability density function of the measurements. Depending upon the criterion of optimality, one

can compute $\hat{\mathbf{x}}$ from $p(\mathbf{x}|\mathbf{y})$. For example, if the object is to find a generalized *minimum variance* Bayes' estimate, i.e., to minimize the cost functional,

$$J = \int \cdots \int (\hat{\mathbf{x}} - \mathbf{x})^T S(\hat{\mathbf{x}} - \mathbf{x}) p(\mathbf{x}|\mathbf{y}) dx_1 dx_2 \cdots dx_n \tag{8}$$

where S is an arbitrary, positive semidefinite matrix. We simply set $\partial J/\partial \hat{\mathbf{x}} = \mathbf{0}$ to find, independent of S, that,

$$\hat{\mathbf{x}} = \int \cdots \int \mathbf{x} p(\mathbf{x}|\mathbf{y}) dx_1 dx_2 \cdots dx_n = \mathsf{E}[\mathbf{x}|\mathbf{y}] \tag{9}$$

which is the *conditional mean estimate*. Eq (8) has the characteristic structure,

$$J = \int \cdots \int L(\hat{\mathbf{x}} - \mathbf{x}) p(\mathbf{x}|\mathbf{y}) dx_1 dx_2 \cdots dx_n \tag{10}$$

where $L(\cdot)$ is a scalar "loss function" of the estimation error $\hat{\mathbf{x}} - \mathbf{x}$. Assuming Gaussian distribution for $\mathbf{x}$ and $\mathbf{v}$, the result of evaluating $\mathsf{E}[\mathbf{x}|\mathbf{y}]$ in Eq (9) is,

$$\hat{\mathbf{x}} = (P_0^{-1} + H^T R^{-1} H)^{-1} H^T R^{-1} \mathbf{y} \tag{11}$$

where P_0 is the *a priori* covariance matrix of $\mathbf{x}$.

To see the relationship between the various estimation methods, we note that if there is little or no *a priori* information, P_0^{-1} is very small and Eq (11) becomes Eq (5). And if we assume that all errors are uncorrelated (R is a diagonal matrix), and all errors have equal variance ($R = \sigma^2 I$), Eq (5) reduces to Eq (3).

Discrete Kalman filter

Extended Kalman filter

Error Propagation

Intuitive concepts

The Wiener filter

Appendix: Matrix gradient operations

- The *gradient* or derivative of a scalar function z, with respect to a vector $\mathbf{x}$, is the vector,

$$\frac{\partial z}{\partial \mathbf{x}} = \mathbf{a}, \quad \text{where } a_i = \frac{\partial z}{\partial x_i} \tag{12}$$

  Vector gradient of the inner product:

$$\frac{\partial}{\partial \mathbf{x}}(\mathbf{y}^T\mathbf{x}) = \mathbf{y} \tag{13}$$

$$\frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^T\mathbf{y}) = \mathbf{y} \tag{14}$$

  An exercise: $\frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^T A\mathbf{x}) = ?$

- *Hessian*: The second partial derivative of a scalar z, with respect to a vector $\mathbf{x}$, is a matrix denoted by,

$$\frac{\partial^2 z}{\partial \mathbf{x}^2} = A, \quad \text{with } a_{ij} = \frac{\partial^2 z}{\partial x_i \partial x_j} \tag{15}$$

  The determinant of A is called the *Hessian* of z.

- *Jacobian*: In general, the vector gradient of a vector is the matrix defined by,

$$\frac{\partial \mathbf{y}^T}{\partial \mathbf{x}} = A, \quad \text{with } a_{ij} = \frac{\partial y_j}{\partial x_i} \tag{16}$$

  If $\mathbf{y}$ and $\mathbf{x}$ are of equal dimension, the determinant of A can be found and is called the *Jacobian* of $\mathbf{y}$.

- The matrix gradient of a scalar z is defined by,

$$\frac{\partial z}{\partial A} = B, \quad \text{with } b_{ij} = \frac{\partial z}{\partial a_{ij}} \tag{17}$$

  Two scalar functions worth noting are the matrix trace and determinant. Some interesting results for square matrices A, B, and C are given below.

$$\frac{\partial}{\partial A}\text{trace}[A] = I \tag{18}$$

$$\frac{\partial}{\partial A}\text{trace}[BAC] = B^T C^T \tag{19}$$

$$\frac{\partial}{\partial A}\text{trace}[ABA^T] = A(B + B^T) \tag{20}$$

$$\frac{\partial}{\partial A}\text{trace}[e^A] = e^{A^T} \tag{21}$$

$$\frac{\partial}{\partial A}|BAC| = |BAC|(A^{-1})^T \tag{22}$$