Problems in Stochastic Estimation – The Bayesian Solution

Benjamin K. ESHUN

ben@benjamineshun.de

TABLE OF CONTENTS

ACKNOWLEDGMENT

I would like to thank Professor Francois Desbouvries of Institut National de Telecommunication, France.

1.0 INTRODUCTION

Filter is usually used to refer to system that is designed to extract information about a prescribed quantity of interest from noisy data. In particular, estimate the value taken by some random variable given the observation of some other random variable. With this aim, for instance Kalman filter, finds applications in many diverse fields: industrial processes, communications (Digital communication and signal processing), radar, sonar, navigation, spacecraft, seismology, biomedical engineering, and financial engineering among others.

The Kalman filter is implementable in the form of an algorithm for a digital computer, which was replacing analog circuitry for estimation and control at the time that Kalman filter was introduced. This implementation may be slower, but it is capable of much greater accuracy than had been achievable with analog filter. The Kalman filter does not require that the deterministic dynamic or the random process have stationary properties, and many application of importance include non-stationary stochastic processes. It is compatible with the state-space formulation of optimal controller for dynamic systems. It caters for the dual properties of estimation and control for systems. The Kalman filter provides the necessary information for mathematically sound, statistically-based decision methods for detecting and rejecting anomalous measurement.

In this paper we discuss the basic methodology for solving state space system. In particular, we discuss the Kalman filter in the context of Bayesian technique.

2.0 BAYESIAN TECHNIQUE FOR ESTIMATION AND CONTROL PROBLEM

Formulation of solution for the estimation problem is outlined which forms the necessary fundamentals and procedure for approaching a state system problem of recursive nature.

We need to obtain the state estimate $\hat{x}$ of x on which the optimal measurements will be based. This estimate is based on the available observations D= (z_1, z_2, ……., z_k) and initial information of association between this observations and state of nature or signal to be estimated x considering noise.

2.1.0 Single stage estimation problem

Consider an estimation problem in a model in which the physical relationship is given by,

$$Z= g(x ,v) \qquad (1)$$

Where z is the measurement vector, x is the state vector *(nx1)*, and v is the noise (error) vector *(qx1)*.

To estimate $\hat{x}$ of *x*, we adopt the Bayesian approach. The parameter, which is attempted to estimate is viewed as a realization of the random variable *x*, as such, the data are described by the joint probability density function (PDF).

$$p(z, x)=p(z/x)p(x) \qquad (2)$$

Where *p(x)* is the prior PDF , summarizing the knowledge about x before any data are observed , and p(z/x) is a conditional PDF , summarizing the knowledge provided by data z conditioned on knowing x.

In order to evaluate *p(z, x)* , we have to evaluate *p(z/x)* , and Bayesian approach provides a formalism which allows the a prior known information regarding the parameters of interest to be included in terms of their associated probability density functions. Thus it is useful to consider the Bayesian approach as fundamental way which leads to optimal estimation such as the maximization with respect to *x* of *p(x/z)* where,

$$p(x/z)=p(z/x)p(x)/p(z) \qquad (3) \qquad \text{that is,} \quad Posterior = \frac{Likelihood}{Evidence}\Pr ior$$

The posteriori density function can also be calculated from,

$$p(x/z)=p\,(x,z)/p(z) \qquad (4)$$

Where

$$p(x,z)=p(x,v=g^*(x,z))J\,, \qquad (5) \qquad J = \det\left[\frac{\partial g^*(x,z)}{\partial z}\right] \;,$$

if the dimensions of z and v are the same. However, it may be difficult to obtain in general, since g* may not exist either because of the non-linear nature of g. And *p(z)* is evidence density function.

Three main criterion functions for estimation are used which depend on Bayesian approach.

1) Minimize mean square error (MMSE)

According to (1), so that $MMSE(\hat{x}) = \int\left[\int\left(x-\hat{x}\right)?p(x/z)dx\right]p(z)dx$ (6)

Now since *p(z)* ≥ 0 for all z, if the integral in brackets can be minimized for each z, then the Bayesian MSE will be minimized.

$$\frac{\partial}{\partial \hat{x}}\int(x-\hat{x})?p(x/z)dx = \int\frac{\partial}{\partial \hat{x}}(x-\hat{x})?p(x/z)dx = -2\int xp(x/z)dx + 2\hat{x}\int p(x/z)dx \quad (7)$$

which when set equal to zero results in

$$\hat{x} = \int xp(x/z)dx = E(x/z) \quad (8)$$

It is seen that the optimal estimator in terms of minimizing the Bayesian MSE is the mean of the posterior PDF *p(x/z)*.

2) Maximize a posterior estimator (MAP)

The MAP estimation approach chooses $\hat{x}$ to maximize the posterior PDF

$$\hat{x} = max\ p(x/z) \quad (9)$$

in finding the maximum of *p(x/z)* we observe that

$$p(x/z)=p(z/x)p(x)/p(z)$$

So an equivalent maximization of *p(z/x)p(x).*

3) Minimize Maximum $|x-\hat{x}|$

$\hat{x}$ =medium of *p(x/z)* (10)

If we have minimum a prior information regarding the parameter *x*, then maximization of *p(x/z)* is equivalent to the maximization of *p(z/x)* , which is the Maximum Likelihood (ML) estimator in the case of Gaussian Distribution.

2.1.1 Special case for Kalman filter (single stage)

We consider the a special case for Kalman filter in which $D = (z_1, z_2, \ldots, z_{k-1})$ is a set of measurements, therefore the physical relationship can be described as follows:

$$z = Hx + v$$

in order to obtain the estimatied value of x , we compute

$$p(x/z) = \frac{p(x)p(v)}{p(z)} \qquad (11)$$

given, *p(x), p(v)* to be Gaussian distributions it follows that *p(z)* is also Gaussian.

Gaussian distribution is a probability density function, which commonly used in stochastic system models and diverse fields. It is classified into Univariate and Multivariate Guassian probability distributions respectively.

The Univariate Gaussian probability distiribution is denoted by N(mean, variance) with density function,

$$p(y) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{1}{2} \frac{(y - \bar{y})}{\sigma^2} \right] \qquad (12)$$

, where $mean = \bar{y}$ and $variance = \sigma^2$

The Multivariate Gaussian probability distribution is denoted by N(mean vector, symmetric positive-definite covariance matrix). The density function is, for instance n-dimensional with $mean = \bar{y}$ (*n*-vector) and *covariance*= *L* (*nxn* symmetric positive –definite matrix) ,

$$p(y) = \frac{1}{\sqrt{(2\pi)^n |L|}} \exp\left[-\frac{1}{2} (y - \bar{y})^T L^{-1} (y - \bar{y}) \right] \qquad (13)$$

Hence given , $x \cdots N(\bar{x}, P_0)$, $v \cdots N(0, R)$

Where,

$E(x) = \bar{x}$, $cov(x) = P_0$ and $E(v)=0$, $cov(v)=R$

Also, $z=Hx+v \Rightarrow E(z)=HE(x)+E(v)$, $E(z)=H\bar{x}$, $cov(z)=Hcov(x)H^T+R=HP_0H^T+R$

$$\Rightarrow z..........N(H\bar{x}, HP_0H^T + R)$$

therefore we can calculate the *p(x/z)*:

$$p(x/z) = \rho \exp\left\{-\frac{1}{2}\left[\left(x-\bar{x}\right)^T p_0^{-1}\left(x-\bar{x}\right) + \left(z-Hx\right)^T R^{-1}\left(z-Hx\right) - \left(z-H\bar{x}\right)^T \left(HP_0H^T+R\right)^{-1}\left(z-H\bar{x}\right)\right]\right\}$$

$$\rho = \frac{\left|HP_0H^T + R\right|^{\frac{1}{2}}}{(2\pi)^{\frac{n-1}{2}}\left|P_0\right|^{\frac{1}{2}}\left|R\right|^{\frac{1}{2}}}$$

$$\Rightarrow p(x/z) = \rho \exp\left\{-\frac{1}{2}\left[\left(x-\hat{x}\right)^T p^{-1}\left(x-\hat{x}\right)\right]\right\} \qquad (14)$$

$$p^{-1} = p_0^{-1} + H^T R^{-1} H\ ,$$
$$P = P_0 - P_0 H^T (HP_0H^T + R)^{-1} HP_0 \qquad (15)$$

$$\hat{x} = \bar{x} + PH^T R^{-1}(z - H\bar{x}) \qquad (16)$$

The difference $(z - H\bar{x})$ is called *the measurement innovation or the residual*. The residual reflects the discrepancy between the predicted measurement $H\bar{x}$ and the actual measurement z. A residual of zero means that the two are in complete agreement.

$PH^T R^{-1}$ is chosen to be the *gain or blending factor* that minimizes the a posteriori error covariance.

Since *p(x/z)* is Gaussian , the most probable, conditional mean and minimax estimate all coincide and is given by $\hat{x}$.

2.2.0 Multistage estimation problem

Consider a system having a state x_k at time step *k*, which is observed by some process generating an observation z_k. The state transition and observation equations can be written as follows:

$$x_k = f(x_{k-1}, w_k)$$
$$z_k = h(x_k, v_k)$$

where w_k and v_k are system and observation noise respectively. Given a set of observations $D = \{z_1, z_2, \ldots z_k\}$, we wish to determine $p(x_k / D_k)$, the distribution over the state at the current time. Using Bayes rule we can write the following:

$$p(x_k / D_k) = p(x_k / z_k, D_{k-1}) \propto p(z_k / x_k) p(x_k / D_{k-1}) \qquad (17)$$

Also, given the Markov structure of the problem, we have:

$$p(x_k, x_{k-1} / D_{k-1}) = p(x_k / x_{k-1}) p(x_{k-1} / D_{k-1}) \qquad (18)$$

From equations. (17) and (18) we can derive the following recursive state estimation equations:

$$p(x_k / D_{k-1}) = \int p(x_k / x_{k-1}) p(x_{k-1} / D_{k-1}) dx_{k-1} \Rightarrow prediction \qquad (19)$$
$$p(x_k / D_k) = c_k p(z_k / x_k) p(x_k / D_{k-1}) \Rightarrow filter \quad update \qquad (20)$$
$$c_k = \int p(z_k / x_k) p(x_k / D_{k-1})$$

In general these equations are difficult to evaluate. For special cases of state transition and observation functions, and noise distributions however, these equations become exactly solvable.

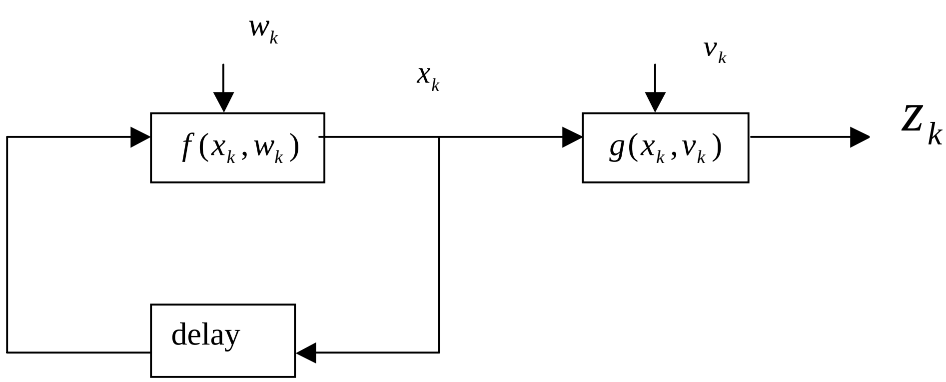

Figure 1 : Generic system diagram

2.2.1 Special case of Kalman filter (Multistage)

The Kalman filter essentially solves the prediction and filter update equations for the special case in which the system transition and observation functions are linear in the state and noise, and the noise is Gaussian. In this case since the noise is Gaussian, the integrals are analytically tractable, and the linearity of the transition and observation functions ensure that the state and observation distributions retain their Gaussian form at each step. We can write the state transition and observation equations as follows:

$$x_k = \phi x_{k-1} + \Gamma w_k$$
$$z_k = H x_k + v_k$$

Where we assume $w_k \sim N(0,Q)$, $v_k \sim N(0,R)$.We define the following statistics for the prior and posterior distribution

$$a_{k/k-1} = E[x_k / D_{k-1}] \Rightarrow prior \quad mean$$
$$P_{k/k-1} = Cov(x_k / D_{k-1}) \Rightarrow prior \quad covariance$$

$$a_k = E[x_k / D_k] \Rightarrow posterior \quad mean$$
$$P_k = Cov(x_k / D_k) \Rightarrow posterior \quad covariance$$

We can deduce from the linear equations of x_k and z_k that

$$x_k / x_{k-1} --- N(\phi x_{k-1}, \Gamma Q \Gamma^T)$$
$$z_k / x_k --- N(H x_k, R)$$

- **Prediction**

Equation (19) allows us to create a prior distribution of the current state x_k with only knowledge of past observations D_{k-1}. The reason this is a prior distribution is because we have yet to make an observation z_k in this state. Once we make the observation, we will need to integrate it with this distribution to generate a posterior distribution.

$$
\begin{aligned}
p(x_k / D_{k-1}) &= \int p(x_k / x_{k-1}) p(x_{k-1} / D_{k-1}) dx_{k-1} \\
&= \int N(\phi x_{k-1}, \Gamma Q \Gamma^T) N(a_{k-1}, P_{k-1}) dx_{k-1} \\
&= \beta \int e^{-\frac{1}{2}A} dx_{k-1}
\end{aligned} \qquad (21)
$$

The term A inside the exponent can be expanded as follows:

$$
\begin{aligned}
A &= (x_k - \phi x_{k-1})^T (\Gamma Q \Gamma^T)^{-1} (x_k - \phi x_{k-1}) + (x_{k-1} - a_{k-1})^T P_{k-1}^{-1} (x_{k-1} - a_{k-1}) \\
&= x_{k-1}{}^T \underbrace{(\phi^T (\Gamma Q \Gamma^T)^{-1} \phi + P_{k-1}^{-1})}_{B} x_{k-1} - 2 x_{k-1}^T \underbrace{(\phi^T (\Gamma Q \Gamma^T)^{-1} x_k + P_{k-1}^{-1} a_{k-1})}_{C} + x_k^T (\Gamma Q \Gamma^T)^{-1} x_k + a_{k-1}^T P_{k-1}^{-1} a_{k-1}
\end{aligned}
$$

This is a quadratic in x_{k-1}, which can be written as

$$
A = (x_{k-1} - \mu)^T \sum{}^{-1} (x_{k-1} - \mu) - C^T B C + x_k{}^T (\Gamma Q \Gamma^T)^{-1} x_k + a_{k-1}^T P_{k-1}^{-1} a_{k-1}
$$

Where $\sum = B^{-1}$ and $\mu = \sum C$. Substituting this result back into the exponent of (21) we get

$$
\begin{aligned}
p(x_k / D_{k-1}) &= \beta \int e^{-\frac{1}{2}A} dx_{k-1} \\
&= \beta \exp\left\{ -\frac{1}{2} \left[x_k{}^T (\Gamma Q \Gamma^T)^{-1} x_k + a_{k-1}^T P_{k-1}^{-1} a_{k-1} - C^T B C \right] \right\} \bullet \int \exp\left\{ -\frac{1}{2} (x_{k-1} - \mu)^T \sum{}^{-1} (x_{k-1} - \mu) \right\} dx_{k-1} \\
&= \beta (2\pi)^{\frac{d}{2}} \left| \sum \right|^{\frac{1}{2}} \exp\left\{ -\frac{1}{2} \underbrace{\left[x_k{}^T (\Gamma Q \Gamma^T)^{-1} x_k + a_{k-1}^T P_{k-1}^{-1} a_{k-1} - C^T B C \right]}_{D} \right\}
\end{aligned}
$$

This distribution has a quadratic inside an exponential term implying that $p(x_k / D_{k-1})$ has a Gaussian form. We can expand the quadratic term D as follows:

$$
D = -(\phi^T (\Gamma Q \Gamma^T)^{-1} x_k + P_{k-1}^{-1} a_{k-1})^T (\phi^T (\Gamma Q \Gamma^T)^{-1} \phi + P_{k-1}^{-1})^{-1} (\phi^T (\Gamma Q \Gamma^T)^{-1} x_k + P_{k-1}^{-1} a_{k-1}) + x_k{}^T (\Gamma Q \Gamma^T)^{-1} x_k + a_{k-1}^T P_{k-1}^{-1} a_{k-1}
$$

Consider terms in D that are quadratic in x_k :

$$
\begin{aligned}
D_2 &= x_k{}^T (\Gamma Q \Gamma^T)^{-1} x_k - x_k{}^T (\Gamma Q \Gamma^T)^{-1} \phi (\phi^T (\Gamma Q \Gamma^T)^{-1} \phi + P_{k-1}^{-1})^{-1} \phi^T (\Gamma Q \Gamma^T)^{-1} x_k \\
&= x_k{}^T (\phi P_{k-1} \phi^T + \Gamma Q \Gamma^T)^{-1} x_k
\end{aligned}
$$

Since this is the only term in D that is quadratic in x_k , we can infer the covariance of the distribution as follows:

$$
\Rightarrow P_{k/k-1} = \phi P_{k-1} \phi^T + \Gamma Q \Gamma^T \qquad (22)
$$

Consider terms in D that are linear in x_k

$$
\begin{aligned}
D_1 &= -2x_k(\Gamma Q\Gamma^T)^{-1}\phi(\phi^T(\Gamma Q\Gamma^T)^{-1}\phi + P_{k-1}^{-1})^{-1}P_{k-1}^{-1}a_{k-1} \\
&= -2x_k(\Gamma Q\Gamma^T)^{-1}\phi[P_{k-1} - P_{k-1}\phi^T(\Gamma Q\Gamma^T + \phi P_{k-1}\phi^T)^{-1}\phi P_{k-1}]P_{k-1}^{-1}a_{k-1} \\
&= -2x_k(\Gamma Q\Gamma^T)^{-1}[\phi - \phi P_{k-1}\phi^T(\Gamma Q\Gamma^T + \phi P_{k-1}\phi^T)^{-1}\phi]a_{k-1} \\
&= -2x_k(\Gamma Q\Gamma^T + \phi P_{k-1}\phi^T)^{-1}\phi a_{k-1} \\
&= -2x_k p_{k/k-1}^{-1}\phi a_{k-1}
\end{aligned}
$$

$$
\Rightarrow a_{k/k-1} = \phi a_{k-1} \qquad (23)
$$

Therefore, the prior distribution $p(x_k / D_{k-1})$ is Gaussian , with the mean and variance given by

$$
\begin{aligned}
a_{k/k-1} &= \phi a_{k-1} \\
P_{k/k-1} &= \phi P_{k-1}\phi^T + \Gamma Q\Gamma^T
\end{aligned}
$$

- **Filter update**

$$
p(x_k / D_k) \propto p(z_k / x_k)\, p(x_k / D_{k-1})
$$

$$
\propto \exp\left\{-\frac{1}{2}\Big[\underbrace{(z_k - Hx_k)^T R^{-1}(z_k - Hx_k) + (x_k - a_{k/k-1})^T P_{k/k-1}^{-1}(x_k - a_{k/k-1})}_{E}\Big]\right\}
$$

Now

$$
\begin{aligned}
E &= (z_k - Hx_k)^T R^{-1}(z_k - Hx_k) + (x_k - a_{k/k-1})^T P_{k/k-1}^{-1}(x_k - a_{k/k-1}) \\
&= x_k^{\,T}(H^T R^{-1}H + P_{k/k-1}^{-1})x_k - 2x_k^{\,T}(H^T R^{-1}z_k + P_{k/k-1}^{-1}a_{k/k-1}) + k
\end{aligned}
$$

From which we can deduce:

$$
\begin{aligned}
P_k &= (H^T R^{-1}H + P_{k/k-1}^{-1})^{-1} \\
&= P_{k/k-1} - P_{k/k-1}H^T(R + H^T P_{k/k-1}H)^{-1}HP_{k/k-1} \\
&= P_{k/k-1} - P_{k/k-1}H^T K^{-1}HP_{k/k-1} \qquad (24)
\end{aligned}
$$

Where ,

$$
K = R + H^T P_{k/k-1}H
$$

Now,

$$a_k = P_k(H^T R^{-1} z_k + P_{k/k-1}^{-1} a_{k/k-1})$$
$$= (P_{k/k-1} - P_{k/k-1} H^T K^{-1} H P_{k/k-1})(H^T R^{-1} z_k + P_{k/k-1}^{-1} a_{k/k-1})$$
$$= a_{k/k-1} - P_{k/k-1} H^T K^{-1} H a_{k/k-1} + P_{k/k-1} H^T \underbrace{[I - K^{-1} H P_{k/k-1} H^T]}_{F} R^{-1} z_k$$

Now,

$$F = I - K^{-1} H P_{k/k-1} H^T$$
$$= I - K^{-1} H P_{k/k-1} H^T - K^{-1} R + K^{-1} R$$
$$= K^{-1} R$$

Hence,

$$a_k = a_{k/k-1} + P_{k/k-1} H^T K^{-1} (z_k - H a_{k/k-1}) \qquad (25)$$

This allows us to characterize the posterior distribution $p(x_k / D_k)$ as a Gaussian with the following mean and variance:

$$a_k = a_{k/k-1} + P_{k/k-1} H^T K^{-1} (z_k - H a_{k/k-1})$$
$$P_k = P_{k/k-1} - P_{k/k-1} H^T K^{-1} H P_{k/k-1} \qquad (26)$$

Where

$$K = R + H^T P_{k/k-1} H$$
$$e_k = z_k - H a_{k/k-1} \qquad (27)$$

We identify that the mean of the posterior distribution is the regression function of a_k and e_k. According to the structure of sequential Bayesian estimation, a new posterior distribution occurs with each consecutive data. Kalman filter can be seen as the development of series of regression functions of a_k on e_k, at times 1,2….*k-1,k*, each having a potentially different intercept and regression coefficient.

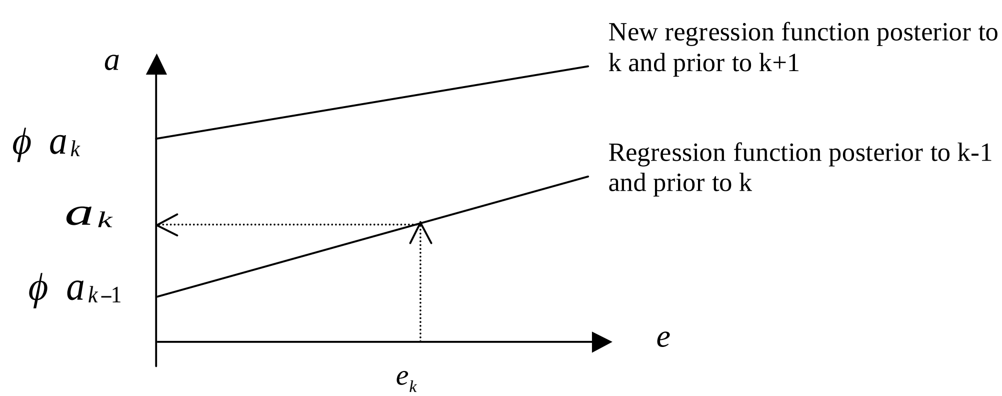

Fig 2: Regression of a_k on e_k

3.0 CONCLUSION

Posterior probability density function of the state given all the measurement is derived, which can be termed a complete solution to the estimation problem because all available information are used, from the probability density function, an optimal estimation can be theoretically be found for any criterion. However, the problems involve specifically the computation of the a posterior conditional density function; the relationship between state of nature and measurement variables must be linear or scalar, the posterior conditional density function must be in analytical form(integrals should be tractable), furthermore, the density functions need to be from the same distribution (Gaussian distribution). Under these conditions, an optimal solution exists using Kalman filter.

4.0 REFERENCES

1.0 R. J.Meinhold and N. D. Singpurwalla (1983), “Understanding the Kalman Filter”,American Statistician Vol.37,No. 2

2.0 S. Hanykin(2002) , “Adaptive Filter Theory” ,Prentice Hall

3.0 R. E. Kalman(1960) “A new approach to linear filtering and prediction problems”, Transactions of the ASME -Journal of Basic Engineering,

4.0 M. S. Grewal and A. P. Andrews (2001), “Kalman Filtering Theory and Practice using MATLAB “, John Wiley&Sons

5.0 Y.C.HO and R.C.K LEE(1964),”A Bayesian Approach to Problems in Stochastic Estimation and Control”, IEEE Trans.Automat.Control ,Vol.AC-9,pp.333-339