Problem set 8
The Kalman filter.
November 11, 2014, due Nov 18

In the lecture on the Fourier transform we briefly discussed the Shannon sampling theorem which lies at the basis of modern digital technology. Another tool, about just as important in theoretical engineering, is the Kalman filter.

The Kalman filter was invented by old friend Rudy Kalman some sixty years ago. At the time, we were both working in a research laboratory of an airplane manufacturing company in Baltimore. Its use for updating information as it is received has been ubiquitous. It is used in a multitude of applications varying from gps to cad to the space program, etc. A google search under "Kalman filter" will yield over a three quarters of a million hits!

Kalman's approach at the time was via Hilbert space, much as in Problem set **1**, so it would be natural for me to present the theory this way. But to be perverse, I will do things via conditional densities and conditional expectations.

I will also be extremely pedantic in discussing vector valued random variables, especially since I have not seen the way I do (co-)variance via tensor products elsewhere in the literature. We will spend some time on Gaussian measures.

Contents

1 Generalities about expectation, variance, and characteristic functions.

1.1 Expectation and variance of a vector valued random variable.

Let V be a vector space (say over the reals and finite dimensional). Let X be a V-valued random variable. That is, we have some measure space $(M, \mathcal{F}, \mu)$ (which will be fixed and hidden throughout), where μ is a probability measure on M (meaning that $\mu(M) = 1$), and $X : M \to V$ is a measurable function.

If X is integrable, then

$$E(X) := \int_M X d\mu$$

is called the **expectation** of X and is an element of V.

The function $X \otimes X$ is a $V \otimes V$ valued function, and if it is also integrable, then

$$\mathrm{Var}(X) := E(X \otimes X) - E(X) \otimes E(X) = E\left(X - E(X)\right) \otimes \left((X - E(X))\right)$$

is called the **variance** of X and is an element of $V \otimes V$. It is by its definition a symmetric tensor, and so can be thought of as a quadratic form on V^*.

1.2 Pushforward under a linear map.

If $A : V \to W$ is a linear map, then AX is a W valued random variable, and

$$E(AX) = AE(X), \quad \mathrm{Var}(AX) = (A \otimes A)\,\mathrm{Var}(X) \tag{1}$$

assuming that $E(X)$ and $\mathrm{Var}(X)$ exist.

We can also write this last equation as

$$\mathrm{Var}(AX)(\eta) = \mathrm{Var}(X)(A^*\eta), \quad \eta \in W^* \tag{2}$$

if we think of the variance as quadratic function on the dual space.

1.2.1 Pushforward to $\mathbb{R}$.

To get a feeling for the equation

$$\mathrm{Var}(AX) = (A \otimes A)\,\mathrm{Var}(X)$$

consider the case where $A = \xi$ is a linear map from V to $\mathbb{R}$. Then $\mathrm{Var}(X)(\xi) = \mathrm{Var}(\xi \cdot X)$ is the usual variance of the scalar valued random variable $\xi \cdot X$. Thus we see that $\mathrm{Var}(X)(\xi) \geq 0$, so $\mathrm{Var}(X)$ is non-negative definite symmetric bilinear form on V^*.

The variance of a scalar valued random variable vanishes if and only if it is a constant. Thus $\mathrm{Var}(X)$ is positive definite unless X is concentrated on hyperplane.

1.2.2 Pushforward under an isomorphism.

Suppose that $A : V \to W$ is an isomorphism, and that $X_*\mu$ is absolutely continuous with respect to Lebesgue measure, so

$$X_*\mu = \rho dv$$

where ρ is some function on V (called the probability density of X) and dv is Lebesgue measure.

Then $(AX)_*\mu$ is absolutely continuous with respect to Lebesgue measure on W and its density σ is given by

$$\sigma(w) = \rho(A^{-1}w)|\det A|^{-1} \tag{3}$$

as follows from the change of variables formula for multiple integrals.

1.2.3 Pushforward under a projection.

Suppose that $V = U \oplus W$ and A is projection onto the second component. Then AX is a W valued random variable called the **marginal** random variable.

If $V = \mathbb{R}^N = \mathbb{R}^{N-k} \oplus \mathbb{R}^k$ with Lebesgue measure $dv = dudw$ according to this decomposition and both X and AX are absolutely continuous with respect to their Lebesgue measures so that

$$X_*\mu = \rho_V(u, w)dudw \quad \text{and} \quad (AX)_*\mu = \rho_W(w)dw$$

then ρ_W is called the **marginal probability density** and, assuming that ρ_W is everywhere positive, the function

$$u \mapsto \frac{\rho_V(u, w)}{\rho_W(w)}$$

is called the **conditional probability density** of u given w. (Check back to the definition of conditional probability to see why this is a reasonable definition.)

For a future computation it is useful to rewrite this definition of conditional probability in exponential form: that is, if $\rho_V(u, v) > 0$ so that

$$\rho_V(u, v) = e^{\tau_V(u,v)} \quad \text{and} \quad \rho_W(w) = e^{\tau_W(w)}$$

then the conditional probability density is given by

$$\rho_{U|w}(u) = e^{\tau_V(u,w) - \tau_W(w)}. \tag{4}$$

1.3 The characteristic function.

The function on V^* given by

$$\xi \mapsto E(e^{i\xi \cdot X})$$

is called the **characteristic function** associated to X and is denoted by ϕ_X. Here we have used the notation $\xi \cdot v$ to denote the value of $\xi \in V^*$ on $v \in V$. It is a version of the Fourier transform (with the conventions used by the probabilists).

More precisely, let $X_*\mu$ denote the push forward of the measure μ by the map X, so that $X_*\mu$ is a probability measure on V. Then ϕ_X is the Fourier transform of this measure except that there are no powers of 2π in front of the integral and a plus rather than a minus sign is before the i in the exponent. These are the conventions of the probabilists.

What is important for us is the fact that the Fourier transform determines the measure, i.e. ϕ_X determines $X_*\mu$. The probabilists would say that the *law* of the random variable (meaning $X_*\mu$) is determined by its characteristic function.

2 Gaussian random variables.

2.1 The unit Gaussian.

Let d be a positive integer. We say that N is a **unit (d-dimensional) Gaussian random variable** if N is a random variable with values in $\mathbb{R}^d$ with density

$$(2\pi)^{-d/2}e^{-(x_1^2\cdots+x_d^2)/2}.$$

It is clear that $E(N) = 0$ and, since

$$(2\pi)^{-d/2}\int x_i x_j e^{-(x_1^2\cdots+x_d^2)/2}dx = \delta_{ij},$$

that

$$\mathrm{Var}(N) = \sum_i \delta_i \otimes \delta_i \tag{5}$$

where $\delta_1, \ldots, \delta_d$ is the standard basis of $\mathbb{R}^d$. We will sometimes denote this tensor by I_d.

2.1.1 The variance as a "matrix".

In general we have the identification $V \otimes V$ with $\mathrm{Hom}(V^*, V)$, so we can think of the $\mathrm{Var}(X)$ as an element of $\mathrm{Hom}(V^*, V)$ if X is a V-valued random variable.

If we identify $\mathbb{R}^d$ with its dual space using the standard basis, then the variance can be thought of as a matrix.

In particular, I_d can be thought of as the identity matrix.

2.1.2 The characteristic function of the unit Gaussian.

We can compute the characteristic function of N by reducing the computation to a product of one dimensional integrals yielding

$$\phi_N(t_1, \ldots, t_d) = e^{-(t_1^2+\cdots+t_d^2)/2}. \tag{6}$$

2.2 The general Gaussian random variable.

A V-valued random variable X is called **Gaussian** if it is equal in law to a random variable of the form

$$AN + a$$

where $A : \mathbb{R}^d \to V$ is a linear map, where $a \in V$, and where N is a unit Gaussian random variable.

Clearly

$$E(X) = a,$$

$$\mathrm{Var}(X) = (A \otimes A)(I_d)$$

or, put another way,
$$\text{Var}(X)(\xi) = I_d(A^*\xi)$$

and hence

$$\phi_X(\xi) = \phi_N(A^*\xi)e^{i\xi\cdot a} = e^{-\frac{1}{2}I_d(A^*\xi)}e^{i\xi\cdot a}$$

or

$$\phi_X(\xi) = e^{-\text{Var}(X)(\xi)/2 + i\xi\cdot E(X)}. \tag{7}$$

2.2.1 Centered Gaussians.

It is a bit of a nuisance to carry along the $E(X)$ in all the computations, so we shall frequently restrict ourselves to **centered Gaussian** random variables meaning that $E(X) = 0$. Thus for a centered Gaussian random variable we have

$$\phi_X(\xi) = e^{-\text{Var}(X)(\xi)/2}. \tag{8}$$

2.2.2 Characterizing Gaussians.

Conversely, suppose that X is a V valued random variable whose characteristic function is of the form

$$\phi_X(\xi) = e^{-Q(\xi)/2},$$

where Q is a quadratic form.

Since $|\phi_X(\xi)| \le 1$ we see that Q must be non-negative definite. Suppose that we have chosen a basis of V so that V is identified with $\mathbb{R}^q$ where $q = \dim V$. By the principal axis theorem we can always find an orthogonal transformation (c_{ij}) which brings Q to diagonal form. In other words, if we set

$$\eta_j := \sum_i c_{ij}\xi_i \quad \text{then} \quad Q(\xi) = \sum_j \lambda_j \eta_j^2.$$

The λ_j are all non-negative since Q is non-negative definite. So if we set

$$a_{ij} := \lambda_j^{\frac{1}{2}} c_{ij}, \text{ and } A = (a_{ij})$$

we find that $Q(\xi) = I_q(A^*\xi)$. Hence X has the same characteristic function as a Gaussian random variable hence must be Gaussian.

2.2.3 The variance of a Gaussian with density.

In our definition of a (centered) Gaussian random variable we were careful not to demand that the map A be an isomorphism. For example, if A were the zero map then we would end up with the δ function (at the origin for centered Gaussians) which (for reasons of passing to the limit) we want to consider as a Gaussian random variable.

But suppose that A *is* an isomorphism. Then by (3), X will have a density which is proportional to

$$e^{-S(v)/2}$$

where S is the quadratic form on V given by

$$S(v) = J_d(A^{-1}v)$$

and J_d is the unit quadratic form on $\mathbb{R}^d$:

$$J_d(x) = x_1^2 \cdots + x_d^2$$

or, in terms of the basis $\{\delta_i^*\}$ of the dual space to $\mathbb{R}^d$,

$$J_d = \sum_i \delta_i^* \otimes \delta_i^*.$$

Here $J_d \in (\mathbb{R}^d)^* \otimes (\mathbb{R}^d)^* = \mathrm{Hom}(\mathbb{R}^d, (\mathbb{R}^d)^*)$. It is the inverse of the map I_d. We can regard S as belonging to $\mathrm{Hom}(V, V^*)$ while we also regard $\mathrm{Var}(X) = (A \otimes A) \circ I_d$ as an element of $\mathrm{Hom}(V^*, V)$. I claim that $\mathrm{Var}(X)$ and S are inverses to one another. Indeed, dropping the subscript d which is fixed in this computation, $\mathrm{Var}(X)(\xi, \eta) = I(A^*\xi, A^*\eta) = \eta \cdot (A \circ I \circ A^*)\xi$ when thought of as a bilinear form on $V^* \otimes V^*$, and hence

$$\mathrm{Var}(X) = A \circ I \circ A^*$$

when thought of as an element of $\mathrm{Hom}(V^*, V)$. Similarly thinking of S as a bilinear form on V we have $S(v, w) = J(A^{-1}v, A^{-1}w) = J(A^{-1}v) \cdot A^{-1}w$ so

$$S = A^{-1*} \circ J \circ A^{-1}$$

when S is thought of as an element of $\mathrm{Hom}(V, V^*)$. Since I and J are inverses of one another, the two above displayed expressions for S and $\mathrm{Var}(X)$ show that these are inverses on one another.

This has the following very important computational consequences:

Suppose we are given a random variable X with (whose law has) a density proportional to $e^{-S(v)/2}$ where S is a quadratic form which is given as a "matrix" $S = (S_{ij})$ in terms of a basis of V^*. Then $\mathrm{Var}(X)$ is given by S^{-1} in terms of the dual basis of V.

If X is a Gaussian is a centered random variable with $\mathrm{Var}(X)$ positive definite, then X has a density with respect to Lebesgue measure which is proportional to $e^{-S(v)/2}$ where $S^{-1} = \mathrm{Var}(X)$.

If M is the matrix of $A : V \to W$ relative to the choice of a basis of each and P is the matrix of $\mathrm{Var}(X)$ relative to the choice of a basis of V and its dual basis, then

$$MPM^*$$

is the matrix of $\mathrm{Var}(AX)$ relative to the basis of W and its dual basis.

2.2.4 Example: The matrix of a projection.

Suppose that $V = U \oplus W$ and we have chosen a basis of U and W so together they form a basis of V so that P has the block decomposition

$$P = \begin{pmatrix} P_{(11)} & P_{(12)} \\ P_{(21)} & P_{(22)} \end{pmatrix}.$$

The matrix of π, the projection onto the second factor has the block decomposition $(0, I)$ where $I = I_{(22)}$ is the $k \times k$ identity matrix. So the matrix of $\mathrm{Var}(\pi X)$ is

$$\begin{pmatrix} 0 & I \end{pmatrix} \begin{pmatrix} P_{(11)} & P_{(12)} \\ P_{(21)} & P_{(22)} \end{pmatrix} \begin{pmatrix} 0 \\ I \end{pmatrix} = P_{(22)}.$$

So if $P_{(22)}$ is invertible, the marginal random variable πX is absolutely continuous with respect to Lebesgue measure and has a density proportional to $e^{-S_{(22)}(w)/2}$ where $S_{(22)}$ is the quadratic form whose matrix is $P_{(22)}^{-1}$.

3 The conditional probability density of a Gaussian is a Gaussian density.

It follows from the very definition that the pushforward of a Gaussian is a Gaussian, in particular the marginal distribution of a Gaussian is a Gaussian. It is not immediately obvious that the conditional probability of a Gaussian is a Gaussian (when it exists). This is well known, but not very well known. It is the key result for the computations needed for the Kalman filter. We state it in terms of a basis, where we assume that $P_{(22)}$ is invertible.

Theorem 3.1 *If* $\begin{pmatrix} x \\ y \end{pmatrix}$ *is Gaussian with mean* $\begin{pmatrix} \mu \\ \nu \end{pmatrix}$ *and variance* P *then the conditional probability of x given y is Gaussian with mean*

$$\mu_{x|y} = \mu + P_{12} P_{22}^{-1}(y - \nu) \tag{9}$$

and variance

$$P_{x|y} = P_{11} - P_{12} P_{22}^{-1} P_{21}. \tag{10}$$

Without loss of generality we may assume that $\mu = 0$ and $\nu = 0$. Let

$$M := \begin{pmatrix} I_{(11)} & -P_{(12)} P_{(22)}^{-1} \\ 0 & I_{(22)} \end{pmatrix}$$

1. Using the fact that P is symmetric compute M^* and then $B := MPM^*$.

Assume that B is invertible and that P is invertible so that

$$P^{-1} = M^* B^{-1} M.$$

8

Let
$$z = x - P_{(12)}P_{(22)}^{-1}y.$$

2. Show that
$$\left(\begin{pmatrix} x \\ y \end{pmatrix}, P^{-1} \begin{pmatrix} x \\ y \end{pmatrix} \right) = \left(\begin{pmatrix} z \\ y \end{pmatrix}, B^{-1} \begin{pmatrix} z \\ y \end{pmatrix} \right)$$
and conclude the truth of the theorem. [Hint: Use(4).]

4 The static version of the Kalman filter.

We want to apply Therorem 3.1 to the following situation: $\mathbf{x}$ is a Gaussian random variable with known mean $\hat{\mathbf{x}}^-$ and variance $\mathbf{P}^-$. (The meaning of the superscript minus is that these are the values "before we make a measurement".) We make a measurement $\mathbf{z}$ which is related to $\mathbf{x}$ by

$$\mathbf{z} = \mathbf{Hx} + \mathbf{v}$$

where $\mathbf{H}$ is a known matrix and $\mathbf{v}$ is the corruption of the measurement and is assumed to be Gaussian with mean zero and known (co)variance $\mathbf{R}$ and assumed to be independent of $\mathbf{x}$. So $\begin{pmatrix} \mathbf{x} \\ \mathbf{v} \end{pmatrix}$ is normally distributed with mean

$$\begin{pmatrix} \hat{\mathbf{x}}^- \\ \mathbf{0} \end{pmatrix}$$

and variance

$$\begin{pmatrix} \hat{\mathbf{P}}^- \\ \mathbf{R} \end{pmatrix}.$$

3. Compute the mean and variance of $\begin{pmatrix} \mathbf{x} \\ \mathbf{z} \end{pmatrix}$.

We now make the observation and observe z. So we want to compute the (new) conditional mean and variance of $\mathbf{x}$ conditioned upon having made the observation z. Define the **gain matrix** K by

$$\mathbf{K} := \mathbf{P}^-\mathbf{H}^T(\mathbf{HP}^-\mathbf{H}^T + \mathbf{R})^{-1} \tag{11}$$

then

Theorem 4.1 *After making the observation z, the variable $\mathbf{x}$ is normally distributed with mean*
$$\hat{\mathbf{x}}^+ = \hat{\mathbf{x}}^- + \mathbf{K}(z - \mathbf{H}\hat{\mathbf{x}}^-) \tag{12}$$

and variance
$$\mathbf{P}^+ = \mathbf{P}^- - \mathbf{KHP}^-. \tag{13}$$

4. Prove this theorem.

4.1 An alternative version for large initial variances.

There is an alternative form of these equations which is useful when $\mathbf{P}^-$ is very large, i.e. when we have very little confidence in our initial estimate. I claim that

$$\mathbf{P}^+ = [(\mathbf{P}^-)^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}]^{-1}. \tag{14}$$

5. Prove this. In oher word show that

$$[(\mathbf{P}^-)^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}][\mathbf{P}^- - \mathbf{P}^-\mathbf{H}^T(\mathbf{H}\mathbf{P}^-\mathbf{H}^T + \mathbf{R})^{-1}\mathbf{H}\mathbf{P}^-] = I.$$

[Hint: multiply out to get $I + \{\ \}\mathbf{H}\mathbf{P}^-$ where $\{\ \} =$

$$\mathbf{H}^T\mathbf{R}^{-1} - \mathbf{H}^T\left(I + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\mathbf{P}^-\mathbf{H}^T\right)\left(\mathbf{H}\mathbf{P}^-\mathbf{H}^T + \mathbf{R}\right)^{-1}.$$

Multiply through on the right by $(\mathbf{H}\mathbf{P}^-\mathbf{H}^T + \mathbf{R})$.]

4.1.1 The case of no initial confidence.

So, for example, in the case of no initial confidence, so "$\mathbf{P}^{-1} = 0$", the new variance is given by

$$\mathbf{P}^+ = \left(\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\right)^{-1}. \tag{15}$$

4.1.2 An alternative expression for K.

We also have an alternative expression for $\mathbf{K}$, namely

$$\mathbf{K} = \mathbf{P}^+\mathbf{H}^T\mathbf{R}^{-1}. \tag{16}$$

Proof. We must show that

$$\mathbf{P}^+\mathbf{H}^T\mathbf{R}^{-1} = \mathbf{P}^-\mathbf{H}^T(\mathbf{H}\mathbf{P}^-\mathbf{H}^T + \mathbf{R})^{-1}.$$

Multiplying on the right by $(\mathbf{H}\mathbf{P}^-\mathbf{H}^T + \mathbf{R})$ and on the left by $(\mathbf{P}^+)^{-1}$ we see that we must show that

$$(\mathbf{P}^+)^{-1}\mathbf{P}^-\mathbf{H}^T = (\mathbf{R}^{-1}\mathbf{H}\mathbf{P}^- + \mathbf{P}^+)\mathbf{H}^T.$$

Using the value of $\mathbf{P}^+$ given in (14) gives the result. □

4.1.3 An alternative formula for the mean.

It also follows directly from (14) and (16) that $I - \mathbf{K}\mathbf{H} = \mathbf{P}^+(\mathbf{P}^-)^{-1}$ and hence (12) becomes

$$\hat{\mathbf{x}}^+ = \mathbf{P}^+(\mathbf{P}^-)^{-1}\hat{\mathbf{x}}^- + [\mathbf{P}^+\mathbf{H}^T\mathbf{R}^{-1}]z. \tag{17}$$

Of course, in the case of zero prior information the first term disappears.

4.1.4 An elementary example.

As an example, consider the case where x is a scalar, and we make two independent measurements with variances σ_1^2 and σ_2^2. So

$$\mathbf{z} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} x + \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$$

so

$$\mathbf{H} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \mathbf{R} = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$$

so (15) gives

$$\mathbf{P}^+ = \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} 1/\sigma_1^2 & 0 \\ 0 & 1/\sigma_2^2 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \frac{1}{(1/\sigma_1^2) + (1/\sigma_2^2)},$$

while

$$\hat{\mathbf{x}}^+ = \mathbf{P}^+\mathbf{H^T}\mathbf{R^{-1}}z = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}z_1 + \frac{\sigma_2^2}{\sigma_2^2 + \sigma_2^2}z_2.$$

Intuitively this says that we take the weighted average of the two measurements, weighted according to the variance. The equation for $\mathbf{P}^+$ says that the variance for the combined measurement is less than the variance of each measurement separately.

5   The full Kalman filter.

The Kalman filter addresses the general problem of trying to estimate the state of $\mathbf{x} \in \mathbb{R}^n$ governed by a stochastic difference equation of the form

$$\mathbf{x}_k = A\mathbf{x}_{k-1} + Bu_{k-1} + \mathbf{w}_{k-1}, \tag{18}$$

with a measurement

$$\mathbf{z}_k = H\mathbf{x}_k + \mathbf{v}_k. \tag{19}$$

The matrix A relates the state at the previous time step $k-1$ to the state at the current step, in the absence of either a driving function or process noise. The matrix B gives the effect of the control variable $=$ driving function and the random variable $\mathbf{w}_{k-1}$ is the process noise, assumed to be normal with mean zero and variance $\mathbf{Q}$. The measurement noise is assumed to be normal with mean zero and variance $\mathbf{R}$, as before. The random variables $\mathbf{w}$ and $\mathbf{v}$ are assumed to be independent of one another.

In practice A might change with each time step, but here we assume it is constant. Similarly for H.

The Kalman filter estimates a process by using a form of feedback control: the filter estimates the process state at some time and then obtains feedback in the form of (noisy) measurements. As such, the equations for the Kalman filter fall into two groups: time update equations and measurement update equations.

The time update equations are responsible for projecting forward (in time) the current state and error covariance estimates to obtain the a priori estimates for the next time step. The measurement update equations are responsible for the feedback, i.e., for incorporating a new measurement into the a priori estimate to obtain an improved a posteriori estimate. The time update equations can also be thought of as predictor equations, while the measurement update equations can be thought of as corrector equations.

The time update equations for the mean and variance of $\mathbf{x}_k$ prior to the measurement at time t_k are given by

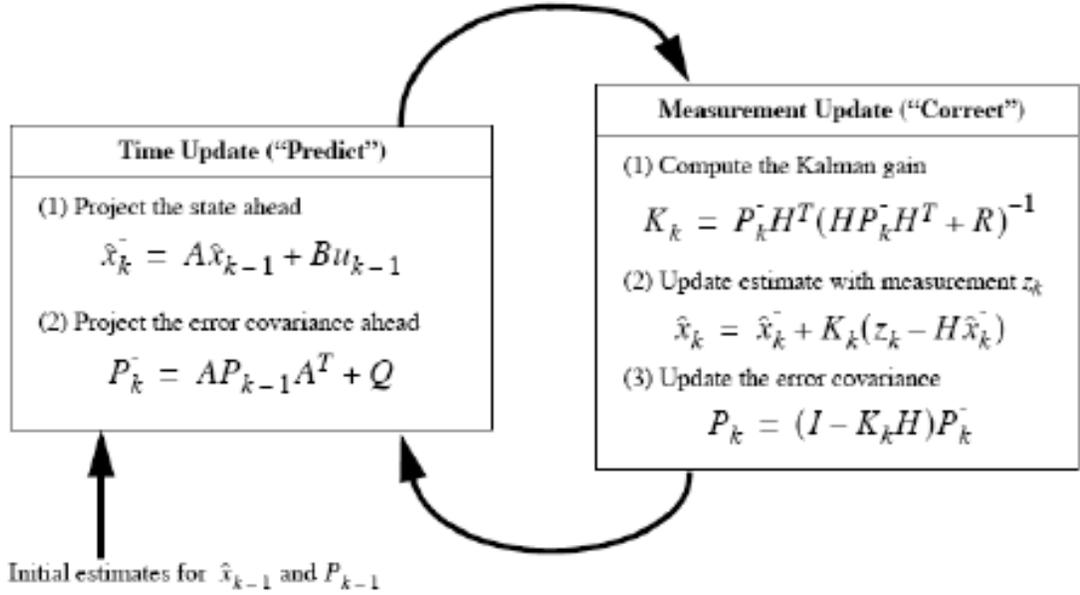$$\hat{\mathbf{x}}_k^- = A\hat{\mathbf{x}}_{k-1}^+ + Bu_{k-1} \tag{20}$$
$$\mathbf{P}_k^- = A\mathbf{P}_{k-1}^+ A^T + Q \tag{21}$$

The measurement update equations are as before:

$$\mathbf{K}_k = \mathbf{P}_k^- \mathbf{H}^T(\mathbf{H}\mathbf{P}_k^- \mathbf{H}^T + \mathbf{R})^{-1} \tag{22}$$
$$\hat{\mathbf{x}}_k^+ = \hat{\mathbf{x}}_k^- + \mathbf{K}_k(z_k - \mathbf{H}\hat{\mathbf{x}}_k^-) \tag{23}$$
$$\mathbf{P}_k^+ = (I - \mathbf{K}_k\mathbf{H})\mathbf{P}_k^-. \tag{24}$$

Time Update ("Predict")

(1) Project the state ahead
$$\hat{x}_k^- = A\hat{x}_{k-1} + Bu_{k-1}$$

(2) Project the error covariance ahead
$$P_k^- = AP_{k-1}A^T + Q$$

Measurement Update ("Correct")

(1) Compute the Kalman gain
$$K_k = P_k^- H^T(HP_k^- H^T + R)^{-1}$$

(2) Update estimate with measurement z_k
$$\hat{x}_k = \hat{x}_k^- + K_k(z_k - H\hat{x}_k^-)$$

(3) Update the error covariance
$$P_k = (I - K_k H)P_k^-$$

Initial estimates for $\hat{x}_{k-1}$ and P_{k-1}

After each time and measurement update pair, the process is repeated with the previous *a posteriori* estimates used to project or predict the new *a priori* estimates. This recursive nature is one of the very appealing features of the Kalman filter.

Here is a recent picture of Rudy Kalman. Of course, 50 years ago we were in our twenties.