Kalman Filtering: an Application of Generalized Least Squares
Bill Menke, inspired by a discussion with George Lu, April 18, 2022

I've always found Kalman Filtering rather mysterious. However, it turns out that it is just a simple application of Generalized Least Squares (least squares with prior information).

Kalman filtering combines two ideas:

The first idea is that the state of a dynamic system is represented by a vector, $\mathbf{m}^{(k)}$, of initially unknown model parameters. The goal of Kalman filtering is to estimate the state. It is known to evolve with time, $k\Delta t$, according to the linear rule:

$$\mathbf{m}^{(k)} = \mathbf{F}^{(k)}\mathbf{m}^{(k-1)} + \mathbf{g}^{(k)} \tag{1}$$

Here, $\mathbf{F}^{(k)}$is a known dynamics matrix, $\mathbf{g}^{(k)} = \mathbf{B}^{(k)}\mathbf{u}^{(k)} + \mathbf{w}^{(k)}$ is a forcing with known $\mathbf{B}^{(k)}$ and $\mathbf{u}^{(k)}$ and $\mathbf{w}^{(k)}$ is a vector of Normally-distributed noise with zero mean and covariance $[\operatorname{cov}\mathbf{g}^{(k)}]$. By rewriting the dynamical equation as

$$\frac{\mathbf{m}^{(k)} - \mathbf{m}^{(k-1)}}{\Delta t} = \frac{1}{\Delta t}\left[\left(\mathbf{F}^{(k)} - \mathbf{I}\right)\mathbf{m}^{(k-1)} + \mathbf{g}^{(k)}\right] \tag{2}$$

it is shown to be the finite difference approximation to a system of coupled first-order differential equations in $\mathbf{m}$. Such systems can represent a wide variety of dynamics.

The second idea is that the state, $\mathbf{m}^{(k)}$, is not directly observable. Instead, data are related to the state via the linear data equation:

$$\mathbf{d}^{(k)} = \mathbf{G}^{(k)}\mathbf{m}^{(k)} + \mathbf{n}^{(k)} \tag{3}$$

Here, $\mathbf{G}^{(k)}$is a known data kernel matrix, $\mathbf{m}^{(k)}$ are the unknown model parameters and $\mathbf{n}^{(k)}$ is a vector of Normally-distributed noise with zero mean and covariance $[\operatorname{cov}\mathbf{d}^{(k)}]$.

Suppose that prior information of the solution, $\langle\mathbf{m}^{(1)}\rangle$, and its covariance, $[\operatorname{cov}\mathbf{m}^{(1)}]_A$, is available for time $k = 1$. Kalman filtering estimates the state $\mathbf{m}^{(k)}$ for k>1 according to the following steps:

(1) Set $k = 1$ and take note of $\langle\mathbf{m}^{(k)}\rangle$ and $[\operatorname{cov}\mathbf{m}^{(k)}]_A$.

(2) Solve the data equation $\mathbf{d}^{(k)} = \mathbf{G}^{(k)}\mathbf{m}^{(k)}$ (covariance $[\operatorname{cov}\mathbf{d}^{(k)}]$) together with the prior information equation $\mathbf{m}^{(k)} = \langle\mathbf{m}^{(k)}\rangle$ (covariance $[\operatorname{cov}\mathbf{m}^{(k)}]_A$) using Generalized Least Squares to achieve a best-estimate, $\mathbf{m}^{(k)}$, of the state, together with its covariance, $[\operatorname{cov}\mathbf{m}^{(k)}]$:

$$\mathbf{m}^{(\mathrm{k})} = \langle\mathbf{m}^{(k)}\rangle + \mathbf{G}_k^{-g}\Delta\mathbf{d}^{(k)}$$

$$with\ \ \mathbf{G}_k^{-g} = [\operatorname{cov}\mathbf{m}^{(k)}]_A\mathbf{G}^{(k)T}\left[\mathbf{G}^{(k)}[\operatorname{cov}\mathbf{m}^{(k)}]_A\mathbf{G}^{(k)T} + [\operatorname{cov}\mathbf{d}^{(k)}]\right]^{-1}$$

$$\text{and} \quad \Delta\mathbf{d}^{(k)} = \mathbf{d}^{(k)} - \mathbf{G}^{(k)}\langle\mathbf{m}^{(k)}\rangle \tag{4}$$

$$[\text{cov}\,\mathbf{m}^{(k)}] = (\mathbf{I} - \mathbf{G}_{\boldsymbol{k}}^{-g}\mathbf{G}^{(k)})[\text{cov}\,\mathbf{m}^{(k)}]_A \tag{5}$$

(3) Presume that the dynamical equation provides prior information about the value of stae, $\langle\mathbf{m}^{(k+1)}\rangle$, at the next time step. The covariance, $[\text{cov}\,\mathbf{m}^{(k+1)}]_A$, is found using standard error propagation.

$$\langle\mathbf{m}^{(k+1)}\rangle = \mathbf{F}^{(k+1)}\mathbf{m}^{(k)} + \mathbf{g}^{(k+1)} \tag{6}$$

$$[\text{cov}\,\mathbf{m}^{(k+1)}]_A = \mathbf{F}^{(k+1)}[\text{cov}\,\mathbf{m}^{(k)}]\mathbf{F}^{(k+1)} + [\text{cov}\,\mathbf{g}^{(k+1)}] \tag{7}$$

(4) Increment k and repeat, starting with step 2.

My analysis proceeded by matching expressions in Wikipedia's Kalman Filtering article with the Generalized Least Squares (GLS) notation in Menke, Geophysical Data Analysis: Discrete Inverse Theory, Fourth Edition, 2018 (Table 1). One further derivation, provided below, is need to match the formulas for $[\text{cov}\,\mathbf{m}^{(k)}]$:

We start by noting that $\langle\mathbf{m}\rangle$ has covariance, $[\text{cov}\,\mathbf{m}]_A$, $\mathbf{d}$ has covariance, $[\text{cov}\,\mathbf{d}]$, and the Generalized Least Squares solution is:

$\mathbf{m}^{\text{est}} = \langle\mathbf{m}\rangle + \mathbf{G}^{-g}(\mathbf{d} - \mathbf{G}\langle\mathbf{m}\rangle) = \mathbf{G}^{-g}\mathbf{d}^{obs} - (\mathbf{I} - \mathbf{G}^{-g}\mathbf{G})\langle\mathbf{m}\rangle$

with generalized inverse:

$\mathbf{G}^{-g} = [\text{cov}\,\mathbf{m}]_A\mathbf{G}^{\mathrm{T}}\mathbf{A}^{-1} \quad \text{with} \quad \mathbf{A} = [\mathbf{G}[\text{cov}\,\mathbf{m}]_A\mathbf{G}^T + [\text{cov}\,\mathbf{d}]]$

and so that $\mathbf{G}^{-gT} = \mathbf{A}^{-1}\mathbf{G}[\text{cov}\,\mathbf{m}]_A$. Then, via standard error propagation:

$[\text{cov}\,\mathbf{m}]_{est} = \mathbf{G}^{-g}[\text{cov}\,\mathbf{d}]_A\mathbf{G}^{-gT} + (\mathbf{I} - \mathbf{G}^{-g}\mathbf{G})[\text{cov}\,\mathbf{m}]_A(\mathbf{I} - \mathbf{G}^{\mathrm{T}}\mathbf{G}^{-gT})$

$= [\text{cov}\,\mathbf{m}]_A + \mathbf{G}^{-g}[\text{cov}\,\mathbf{d}]_A\mathbf{G}^{-gT} + \mathbf{G}^{-g}\mathbf{G}[\text{cov}\,\mathbf{m}]_A\mathbf{G}^{\mathrm{T}}\mathbf{G}^{-gT}$

$-[\text{cov}\,\mathbf{m}]_A\mathbf{G}^{\mathrm{T}}\mathbf{G}^{-gT} - \mathbf{G}^{-g}\mathbf{G}[\text{cov}\,\mathbf{m}]_A =$

$= [\text{cov}\,\mathbf{m}]_A + \mathbf{G}^{-g}\left[\mathbf{G}[\text{cov}\,\mathbf{m}]_A\mathbf{G}^{\mathrm{T}} + [\text{cov}\,\mathbf{d}]\right]\mathbf{G}^{-gT}$

$-[\text{cov}\,\mathbf{m}]_A\mathbf{G}^{\mathrm{T}}\mathbf{G}^{-gT} - \mathbf{G}^{-g}\mathbf{G}[\text{cov}\,\mathbf{m}]_A =$

$= [\text{cov}\,\mathbf{m}]_A + [\text{cov}\,\mathbf{m}]_A\mathbf{G}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{A}\mathbf{A}^{-1}\mathbf{G}[\text{cov}\,\mathbf{m}]_A$

$-[\text{cov}\,\mathbf{m}]_A\mathbf{G}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{G}[\text{cov}\,\mathbf{m}]_A - [\text{cov}\,\mathbf{m}]_A\mathbf{G}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{G}[\text{cov}\,\mathbf{m}]_A =$

$= [\text{cov}\,\mathbf{m}]_A + [\text{cov}\,\mathbf{m}]_A\mathbf{G}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{G}[\text{cov}\,\mathbf{m}]_A$

$$-[\text{cov}\,\mathbf{m}]_A\mathbf{G}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{G}[\text{cov}\,\mathbf{m}]_A - [\text{cov}\,\mathbf{m}]_A\mathbf{G}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{G}[\text{cov}\,\mathbf{m}]_A =$$
$$= [\text{cov}\,\mathbf{m}]_A - [\text{cov}\,\mathbf{m}]_A\mathbf{G}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{G}[\text{cov}\,\mathbf{m}]_A =$$
$$= [\text{cov}\,\mathbf{m}]_A - \mathbf{G}^{-g}\mathbf{G}[\text{cov}\,\mathbf{m}]_A = [\mathbf{I} - \mathbf{G}^{-g}\mathbf{G}][\text{cov}\,\mathbf{m}]_A$$

(8)

An important question is whether Kalman filtering produces different estimates of $\mathbf{m}^{(k)}$ and $\left[\text{cov}\,\mathbf{m}^{(k)}\right]$ then if one were to solve the time-coupled generalized least squares problem.

In the problem, all the model parameters are concatenated into a single vector, $\mathbf{m}^{(all)} = [\mathbf{m}^{(1)};\ \mathbf{m}^{(2)};\ \mathbf{m}^{(3)}; \cdots]$. The data equation is $\mathbf{d}^{(all)} = \mathbf{G}^{(all)}\mathbf{m}^{(all)}$, with covariance $\left[\text{cov}\,\mathbf{d}^{(all)}\right] = \text{diag}\left(\left[\text{cov}\,\mathbf{d}^{(1)}\right], \left[\text{cov}\,\mathbf{d}^{(2)}\right], \cdots\right)$, represents all the data. The prior information equation represents all the prior information, and has the form $\mathbf{h}^{(all)} = \mathbf{H}^{(all)}\mathbf{m}^{(all)}$. It includes initial conditions, boundary conditions, and the recursion equation, $\mathbf{F}^{(k)}\mathbf{m}^{(k-1)} - \mathbf{m}^{(k)} = -\mathbf{g}^{(k)}$ as its rows. The variances of these equations and

$$\left[\text{cov}\,\mathbf{h}^{(all)}\right] = \text{diag}\left([\text{cov}\,ic], [\text{cov bc}], \left[\text{cov}\,\mathbf{g}^{(1)}\right], \left[\text{cov}\,\mathbf{g}^{(2)}\right], \cdots\right)$$

(9)

where $[\text{cov}\,ic]$ and $[\text{cov bc}]$ are the covariances of the initial and boundary condition, respectively.

Example.

The diffusion equation,

$$\frac{\partial}{\partial t}m(t,x) = c_0\frac{\partial^2}{\partial x^2}m(t,x) + g(t,x)$$

(10)

corresponds to the recursion equation

$$\frac{m(t_{i+1},x_j) - m(t_i,x_j)}{\Delta t} = c_0\left(m(t_i,x_{j-1}) - 2m(t_i,x_j) + m(t_i,x_{j+1})\right) + g(t_i,x_j)$$

(11)

which when rewritten as

$$m(t_{i+1},x_j) = m(t_i,x_j) + c_0\Delta t\left(m(t_i,x_{j-1}) - 2m(t_i,x_j) + m(t_i,x_{j+1})\right) + \Delta t g(t_i,x_j)$$

(12)

has the form of Equation (1). We presume that there are M model parameters, with uniform sampling Δx, and N time steps, with uniform sampling Δt We also introduce the initial condition $m(t_1,x_j) = 0$ and the boundary conditions $m(t,x_1) = m(t,x_M) = 0$.

We introduce a data equation, $\mathbf{d}^{(k)} = \mathbf{G}^{(k)}\mathbf{m}^{(k)}$, corresponding to making K measurements of $m(t_k, x)$ at randomly chosen positions, x. Each row of the data kernel contains a single entry of unity, with all other entries being zero.

The source, $g(t_i, x_j)$, is taken to be impulsive in time and Gaussian in position:

$$g(t_i, x_j) = \Delta t \delta(t_i - t_1) \exp\left(-\left(x_j - ½x_M\right)^2 / 2\sigma_x^2\right)$$

The results are shown in Figure 1.

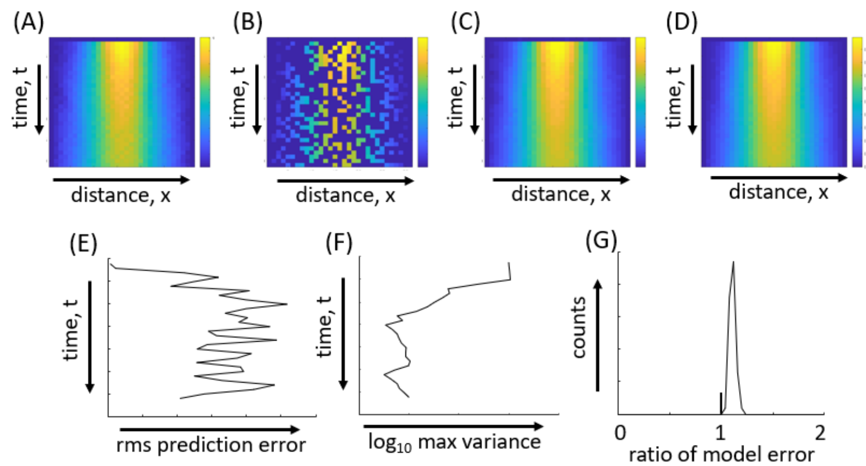

Figure 1. (A) True model with $M = 31$, $\Delta t = \Delta x = 1$, $\sigma_x = 5$, $c_0 = 0.4$, $K = 10$, $\sigma_d = 0.01$, $\sigma_g = 0.01$, $\langle \mathbf{m}^{(1)} \rangle = 0$, $\left[\text{cov}\, \mathbf{m}^{(1)}\right]_A = \sigma_h^2 \mathbf{I}$ and $\sigma_h = 0.1$. (B) Observed data. (C) Estimated model by Kalman filtering. (D) Estimated model by fully-coupled generalized Least Squares. (E) Time evolution of root mean squared (r.m.s) data prediction error for Kalman filtering. (F) Time evolution of the logarithm of maximum model variance for Kalman filtering. (G) Histogram of the ratio of r.m.s. model error for Kalman filtering and model error for fully-coupled Generalized Least Squares, for 200 realizations of the data.

Fully-coupled Generalized Least Squares seems to slightly outperform Kalman filtering, by the criterion of model error. However, there are so many tunable parameters that I am unsure whether the difference is fundamental.

However, Kalman filtering cannot be exactly equivalent to the fully-coupled Generalized Least Squares. The reason is that the Gram matrix for Fully-coupled Generalized Least Squares, when formulated in terms of single-time-step submatrices, is tridiagonal, and it is known that a tridiagonal system cannot be solved by a process that, like Kalman filtering, is forward-in-time, only. The standard algorithm for solving a tridiagonal system consists of two steps, the first forward-in-time, and the second backward-in-time. Consequently, future measurements affect past estimated values of the GLS solution, whereas in Kalman filtering, they do not.

A question that I've not worked on (yet): Is Kalman Filtering really "filtering"; that is, can the formula for $\mathbf{m}^{(k)}$ be manipulated into the form of an FIR filter, $a(t) * m(t) =.b(t) * g(t)$. If so, what are the filters $a(t)$, $b(t)$, and $a^{-1}(t) * b(t)$?

Table 1. Correspondence between the variables in Wikipedia's Kalman Filtering article and the variables in Menke, Geophysical Data Analysis: Discrete Inverse Theory, 4th Edition, 2018.

No.	Wikipedia	Generalized Least Squares	Comment
1	k	k	time k
2	$\mathbf{x}_k$	$\mathbf{m}^{(k)}$	state variables / model parameters
3	$\mathbf{u}_k$	$\mathbf{u}^{(k)}$	
4	$\mathbf{B}_k\mathbf{u}_k$	$\mathbf{g}^{(k)} = \mathbf{B}^{(k)}\mathbf{u}^{(k)}$	forcing
5	$\mathbf{Q}_k$	$\left[\text{cov}\,\mathbf{g}^{(k)}\right]$	covariance of dynamics
4	$\mathbf{x}_k = \mathbf{F}_k\mathbf{x}_{k-1} + \mathbf{B}_k\mathbf{u}_k$	$\langle\mathbf{m}^{(k)}\rangle = \mathbf{F}^{(k)}\mathbf{m}^{(k-1)} + \mathbf{g}^{(k)}$	dynamics recursion rule
6	$\mathbf{P}_{k\|k-1} = \mathbf{F}_k\mathbf{P}_{k-1\|k-1}\mathbf{F}_k^T + \mathbf{Q}_k$	$\left[\text{cov}\,\mathbf{m}^{(k)}\right]_A = \mathbf{F}^{(k)}\left[\text{cov}\,\mathbf{m}^{(k-1)}\right]_A\mathbf{F}^{(k)} + \left[\text{cov}\,\mathbf{g}^{(k)}\right]$	prior covariance recursion rule
7	$\mathbf{z}_k$	$\mathbf{d}^{(k)}$	data
8	$\mathbf{H}_k$	$\mathbf{G}^{(k)}$	data kernel
9	$\mathbf{z}_k = \mathbf{H}_k\mathbf{x}_k$	$\mathbf{d}^{(k)} = \mathbf{G}^{(k)}\mathbf{m}^{(k)}$	data equation
10	$\mathbf{R}_k$	$\left[\text{cov}\,\mathbf{d}^{(k)}\right]$	covariance of the data
11	$\mathbf{S}_k = \mathbf{H}_k\mathbf{P}_{k\|k-1}\mathbf{H}_k^T + \mathbf{R}_k$	$\mathbf{A}^{(k)} = \mathbf{G}^{(k)}\left[\text{cov}\,\mathbf{m}^{(k)}\right]_A\mathbf{G}^{(k)T} + \left[\text{cov}\,\mathbf{d}^{(k)}\right]$	
12	$\mathbf{K}_k = \mathbf{P}_{k\|k-1}\mathbf{H}_k^T\mathbf{S}_k^{-1}$	$\mathbf{G}_k^{-g} = \left[\text{cov}\,\mathbf{m}^{(k)}\right]_A\mathbf{G}^{(k)T}\left[\mathbf{A}^{(k)}\right]^{-1}$	generalized inverse
13	$\hat{\mathbf{x}}_{k\|k-1}$	$\langle\mathbf{m}^{(k)}\rangle$	prior model parameters
14	$\tilde{\mathbf{y}}_k = \mathbf{z}_k - \mathbf{H}_k\hat{\mathbf{x}}_{k\|k-1}$	$\Delta\mathbf{d}^{(k)} = \mathbf{d}^{(k)} - \mathbf{G}^{(k)}\langle\mathbf{m}^{(k)}\rangle$	data deviation
15	$\hat{\mathbf{x}}_{k\|k} = \hat{\mathbf{x}}_{k\|k-1} + \mathbf{K}_k\tilde{\mathbf{y}}_k$	$\mathbf{m}^{(\mathrm{k})} = \langle\mathbf{m}^{(k)}\rangle + \mathbf{G}_k^{-g}\Delta\mathbf{d}^{(k)}$	GLS solution
16	$\tilde{\mathbf{y}}_{k\|k} = \mathbf{z}_k - \mathbf{H}_k\hat{\mathbf{x}}_{k\|k}$	$\mathbf{e}^{(k)} = \mathbf{d}^{(k)} - \mathbf{G}^{(k)}\mathbf{m}^{(\mathrm{k})}$	prediction error
17		$\left[\text{cov}\,\mathbf{m}^{(k)}\right] = \left(\mathbf{I} - \mathbf{G}_k^{-g}\mathbf{G}^{(k)}\right)\left[\text{cov}\,\mathbf{m}^{(k)}\right]_A$	posterior covariance of model parameters (see proof, below)
18	$\hat{\mathbf{x}}_{k+1\|k} = \hat{\mathbf{x}}_{k\|k-1}$	$\langle\mathbf{m}^{(\mathrm{k+1})}\rangle = \mathbf{m}^{(\mathrm{k})}$	model recursion rule
19	$\mathbf{P}_{k\|k} = (\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)\mathbf{P}_{k\|k-1}$	$\left[\text{cov}\,\mathbf{m}^{(k+1)}\right]_A = \left[\text{cov}\,\mathbf{m}^{(k)}\right]$	covariance recursion rule