Kalman Filtering

Another of the Alzheimer Notes Series[1]

09 September 1996

1 Introduction

This set of notes surveys a number of estimation techniques as background for Kalman filtering. It is decidedly applied in its orientation and supplies sufficient mathematics to demonstrate relations among the components. The review begins with linear least squares estimation of a fixed set of data and builds from there the components required to describe a dynamic system with a Kalman Filter. Intellectual stubs supporting this line of development are consigned to appendices.

In the sense used here, filtering is the estimation of the current (actual) value of a parameter or set of parameters, $\mathbf{x}_k$, given all of the data points up to and including $\mathbf{x}_k$. Smoothing and prediction are kissing cousins to filtering wherein given the same set of data, past and future values of $\mathbf{x}_k$ are estimated, respectively.

One warning should be emphasized up front. With all of these least squares techniques, one should make certain that the minimization of *squared* error makes sense in the application. If alternative metrics such as exponential or absolute value weighting of the error is appropriate then other (perhaps more computationally costly) methods should be used. Consider as an example, the case where you want to minimize the errors in disbursement of compensation. Is the overpayment of one individual by \$10 twice or four times as bad as underpaying another by \$5?

2 Notation

Constants or scalar variables are represented by lower-case letters (x) and functions by upper-case letters (X).

Variables that are vectors[2] are denoted by bold lower-case letters ($\mathbf{x}$). Means obtained from samples of these vector variables are presented in bold with an overbar ($\bar{\mathbf{x}}$). Estimated vector variables are presented in bold with a hat ($\widehat{\mathbf{x}}$).

$E[\mathbf{x}]$ is the representation for the expected value of the vector $\mathbf{x}$; $E[\mathbf{X}]$ for the matrix $\mathbf{X}$.

Matrices are denoted in bold by upper-case letters ($\mathbf{X}$). Two special matrices are the variance matrix ($\mathbf{V}$) of the vector $\mathbf{x}$ and the covariance matrix ($\mathbf{C}$) for the vectors $\mathbf{x}$ and $\mathbf{y}$ (the arguments are presented as subscripts):

[1] This exoresource was created so that if I blinked, the critical issues of justification and implementation would not be lost. The content has been plundered (sometimes from other plunderers) with no or very little attribution. It was thrashed out in discussions with Jeff Mendelsohn. This note began with him and any and all errors are his. - Brian Madden

[2] In this notation, vectors are $n \times 1$ column vectors.

$$\mathbf{V_x} = \begin{bmatrix} \sigma^2_{y_1,y_1} & \sigma^2_{y_1,y_2} & \cdots & \sigma^2_{y_1,y_n} \\ \sigma^2_{y_2,y_1} & \sigma^2_{y_2,y_2} & \cdots & \sigma^2_{y_2,y_n} \\ \vdots & \vdots & & \vdots \\ \sigma^2_{y_n,y_1} & \sigma^2_{y_n,y_2} & \cdots & \sigma^2_{y_n,y_n} \end{bmatrix}; \qquad \mathbf{C_{x,y}} = \begin{bmatrix} \sigma^2_{x_1,y_1} & \sigma^2_{x_1,y_2} & \cdots & \sigma^2_{x_1,y_n} \\ \sigma^2_{x_2,y_1} & \sigma^2_{x_2,y_2} & \cdots & \sigma^2_{x_2,y_n} \\ \vdots & \vdots & & \vdots \\ \sigma^2_{x_n,y_1} & \sigma^2_{x_n,y_2} & \cdots & \sigma^2_{x_n,y_n} \end{bmatrix}$$

where the statistics[3] are computed over a population sample of m n-dimensional vectors ($\mathbf{V_x}$) or m vector pairs ($\mathbf{C_{x,y}}$):

$$\sigma^2_{x_i,y_j} = \sum_{k=1}^{m} \frac{(x_{ik} - \bar{x})(y_{jk} - \bar{y})}{m-1}$$

3 Linear Least Squares Estimation (LLSE)

Let $\exists$ a (not necessarily linear) mapping from an n-dimensional vector space into itself ($\mathcal{A} : \mathbf{y} \mapsto \mathbf{x}$). Given m samples of this mapping $(\mathbf{y}, \mathbf{x})$ obtain the best (minimum squared-error)[4] linear mapping, $\mathbf{A} : \mathbf{y} \mapsto \widehat{\mathbf{x}}$. The linear mapping is an approximation, $\mathbf{x} = \widehat{\mathbf{x}} + \mathbf{e}$. The LLSE $\widehat{\mathbf{x}}$ is the estimate which minimizes the dot product $E\left[(\mathbf{x} - \widehat{\mathbf{x}})^T(\mathbf{x} - \widehat{\mathbf{x}})\right]$ (thereby minimizing the algebraic sum of the variance of the individual vector elements) over the class of *linear* functions of $\mathbf{y}$: $\widehat{\mathbf{x}} = \mathbf{A}\mathbf{y} + \mathbf{b}$. The LLSE estimate is given by:

$$\widehat{\mathbf{x}} \;=\; \bar{\mathbf{x}} + \mathbf{C_{x,y}} \mathbf{V_y}^{-1}(\mathbf{y} - \bar{\mathbf{y}})$$

This estimate is based entirely on the obtained measurements and does not include any model of the process. The only assumptions are that $\mathbf{x}$ and $\mathbf{y}$ are linearly related and that any deviations ($\mathbf{e}$) from this assumption have a zero-mean[5]. Note that in the computation of $\widehat{\mathbf{x}}$, neither $\bar{\mathbf{y}}$ nor $\mathbf{V_y}$ are updated with the new measurement $\mathbf{y}$ since there is no corresponding measurement $\mathbf{x}$.[6]

From the assumptions that $\widehat{\mathbf{x}}$ is a linear function of $\mathbf{y}$, $\mathbf{x} - \widehat{\mathbf{x}}$ is zero-mean, and $\mathbf{x} - \widehat{\mathbf{x}}$ is uncorrelated[7] with $\mathbf{y}$:

---

[3] An unbiased estimate of the variance is used.

[4] As noted above, implicit in LLSE is the relevance of squared error to the problem at hand. If, however, it is more appropriate that other measures such as exponential weighting or the linear sum (absolute value) of the error, then in these circumstances the computational and theoretical convenience of squared error techniques may be diminished.

[5] The source of the deviation may come from fitting a linear model to data that are in fact generated by a higher order function, by errors in the measurement procedure, or from nonstationarities in the process over the time the measurements were acquired.

[6] Why not? The best estimate of $\mathbf{y}$ should be independent of the estimation of $\mathbf{x}$. The answer is that such modifications go beyond the assumptions required for an optimal linear mapping based on a given set of $(\mathbf{y}, \mathbf{x})$ pairs and that additional constraints need to be assessed before adaptive models can be employed with assurance.

[7] Although no linear relation may exist, it is still possible for higher order interactions to exist in the measurement noise or in the actual activity of the plant.

$$
\begin{aligned}
\widehat{\mathbf{x}} &= \mathbf{A}\mathbf{y} + \mathbf{b} \text{ ; } \textit{the estimate is a linear function of the measurement} \\
E\left[\mathbf{x} - \widehat{\mathbf{x}}\right] &= \mathbf{0} \text{ ; } \textit{the estimate has a zero-mean error distribution} \\
E\left[(\mathbf{x} - \widehat{\mathbf{x}})\mathbf{y}^T\right] &= \mathbf{0} \text{ ; } \textit{the error in the estimate is uncorrelated with the measurement}
\end{aligned}
$$

In practice, difficulties arise because the noise does not have a zero mean or it is correlated with $\mathbf{y}$, or both. The need for a Gaussian noise distribution stems from the conditions required to make LLSE a maximum likelihood estimator (MLE). This relation (the guarantee that the minimum of a given function of the data coincides with the most probable answer) is what makes LLSE useful. The use of LLSE with probability density functions (pdfs) that do not support this relation is one of the two most common errors in the application of this class of techniques (the other being the use of squared-error metrics where they are not appropriate) (see Appendix A).

By looking at the zero-mean equation above, the offset in the linear equation ($\mathbf{b} = \bar{\mathbf{x}} - \mathbf{A}\bar{\mathbf{y}}$) is obtained:

$$
\begin{aligned}
\mathbf{0} &= E\left[\mathbf{x} - \widehat{\mathbf{x}}\right] \\
&= E\left[\mathbf{x} - \mathbf{A}\mathbf{y} - \mathbf{b}\right] \\
&= \bar{\mathbf{x}} - \mathbf{A}\bar{\mathbf{y}} - \mathbf{b}
\end{aligned}
$$

From the uncorrelated condition, the proportionality matrix ($\mathbf{A} = \mathbf{C}_{\mathbf{x},\mathbf{y}}\mathbf{V}_{\mathbf{y}}^{-1}$) is obtained:

$$
\begin{aligned}
\mathbf{0} &= E\left[(\mathbf{x} - \widehat{\mathbf{x}})\mathbf{y}^T\right] \\
&= E\left[(\mathbf{x} - \bar{\mathbf{x}} - \mathbf{A}\mathbf{y} + \mathbf{A}\bar{\mathbf{y}})\mathbf{y}^T\right] \text{ ; } \textit{substituting for } \widehat{\mathbf{x}} \textit{ and then for } \mathbf{b} \\
&= E\left[\mathbf{x}\mathbf{y}^T\right] - \bar{\mathbf{x}}\bar{\mathbf{y}}^T - \mathbf{A}E\left[\mathbf{y}\mathbf{y}^T\right] + \mathbf{A}\bar{\mathbf{y}}\bar{\mathbf{y}}^T \text{ ; } \textit{factoring out constants from expected values} \\
&= \mathbf{C}_{\mathbf{x},\mathbf{y}} - \mathbf{A}\mathbf{V}_{\mathbf{y}}
\end{aligned}
$$

Therefore, substituting for $\mathbf{A}$ and for $\mathbf{b}$ from the above, we get $\widehat{\mathbf{x}} = \bar{\mathbf{x}} + \mathbf{C}_{\mathbf{x},\mathbf{y}}\mathbf{V}_{\mathbf{y}}^{-1}(\mathbf{y} - \bar{\mathbf{y}})$ by:

$$
\begin{aligned}
\widehat{\mathbf{x}} &= \mathbf{A}\mathbf{y} + \mathbf{b} \\
&= \mathbf{C}_{\mathbf{x},\mathbf{y}}\mathbf{V}_{\mathbf{y}}^{-1}\mathbf{y} + \bar{\mathbf{x}} - \mathbf{C}_{\mathbf{x},\mathbf{y}}\mathbf{V}_{\mathbf{y}}^{-1}\bar{\mathbf{y}} \\
&= \bar{\mathbf{x}} + \mathbf{C}_{\mathbf{x},\mathbf{y}}\mathbf{V}_{\mathbf{y}}^{-1}(\mathbf{y} - \bar{\mathbf{y}})
\end{aligned}
$$

Note that the expected value of $\widehat{\mathbf{x}}$ is $\bar{\mathbf{x}}$.

What does it mean functionally that $\mathbf{V}_{\mathbf{y}}^{-1}$ exists, i.e., $\mathbf{V}_{\mathbf{y}}$ is nonsingular? None of the elements of the measurement vector may be replaced by a linear combination of the others. It is required that all of the parameters in the linear description make a unique contribution to the prediction of $\mathbf{x}$ given $\mathbf{y}$. $\mathbf{V}_{\mathbf{y}}$ represents the variability of the individual elements of the measurement vector as well as their interrelatedness (the major and minor axes of the measurement uncertainty ellipses). But what does $\mathbf{V}_{\mathbf{y}}^{-1}$ *mean*? It certainly doesn't transform the measurements into unique properties in the way an eigenvalue decomposition would. Sometimes a matrix is a useful linear transform and sometimes it is just a matrix. Now does the transformation $\mathbf{C}_{\mathbf{x},\mathbf{y}}\mathbf{V}_{\mathbf{y}}^{-1}$ have any meaning? Is it more *meaningful* in the sense of the ratio of two uncertainty measures or is it better described as the product of two linear transforms of the n-dimensional space onto itself?

The variance of the error in the estimate (a description of the multidimensional distribution) is expressible as:

$$\mathbf{V}_{\mathbf{x}-\widehat{\mathbf{x}}} = \mathbf{V}_{\mathbf{x}} - \mathbf{C}_{\mathbf{x},\mathbf{y}}\mathbf{V}_{\mathbf{y}}^{-1}\mathbf{C}_{\mathbf{x},\mathbf{y}}^{T}$$

This relation follows from:

$$\begin{aligned}
\mathbf{V}_{\mathbf{x}-\widehat{\mathbf{x}}} &= E\left[(\mathbf{x}-\widehat{\mathbf{x}})(\mathbf{x}-\widehat{\mathbf{x}})^T\right] \\
&= E\left[(\mathbf{x}-\mathbf{A}\mathbf{y}-\mathbf{b})(\mathbf{x}-\mathbf{A}\mathbf{y}-\mathbf{b})^T\right] \ ; \ \textit{substitute for } \widehat{\mathbf{x}} \\
&= E\left[((\mathbf{x}-\bar{\mathbf{x}})-\mathbf{A}(\mathbf{y}-\bar{\mathbf{y}}))((\mathbf{x}-\bar{\mathbf{x}})-\mathbf{A}(\mathbf{y}-\bar{\mathbf{y}}))^T\right] \ ; \ \textit{substitute for } \mathbf{b} \\
&= \mathbf{V}_{\mathbf{x}} - E\left[(\mathbf{x}-\bar{\mathbf{x}})(\mathbf{y}-\bar{\mathbf{y}})^T\right]\mathbf{A}^T - \mathbf{A}E\left[(\mathbf{y}-\bar{\mathbf{y}})(\mathbf{x}-\bar{\mathbf{x}})^T\right] + \mathbf{A}\mathbf{V}_{\mathbf{y}}\mathbf{A}^T \ ; \ \textit{expand product} \\
&= \mathbf{V}_{\mathbf{x}} - \mathbf{C}_{\mathbf{x},\mathbf{y}}\mathbf{A}^T - \mathbf{A}\mathbf{C}_{\mathbf{y},\mathbf{x}} + \mathbf{A}\mathbf{V}_{\mathbf{y}}\mathbf{A}^T \\
&= \mathbf{V}_{\mathbf{x}} - \mathbf{C}_{\mathbf{x},\mathbf{y}}\mathbf{V}_{\mathbf{y}}^{-1}\mathbf{C}_{\mathbf{x},\mathbf{y}}^T - \mathbf{C}_{\mathbf{x},\mathbf{y}}\mathbf{V}_{\mathbf{y}}^{-1}\mathbf{C}_{\mathbf{x},\mathbf{y}}^T + \mathbf{C}_{\mathbf{x},\mathbf{y}}\mathbf{V}_{\mathbf{y}}^{-1}\mathbf{V}_{\mathbf{y}}(\mathbf{C}_{\mathbf{x},\mathbf{y}}\mathbf{V}_{\mathbf{y}}^{-1})^T \\
&= \mathbf{V}_{\mathbf{x}} - \mathbf{C}_{\mathbf{x},\mathbf{y}}\mathbf{V}_{\mathbf{y}}^{-1}\mathbf{C}_{\mathbf{x},\mathbf{y}}^T \ ; \ \textit{from noting } (\mathbf{V}_{\mathbf{y}}^{-1})^T = \mathbf{V}_{\mathbf{y}}^{-1} \textit{ and } \mathbf{C}_{\mathbf{y},\mathbf{x}} = \mathbf{C}_{\mathbf{x},\mathbf{y}}^T
\end{aligned}$$

In as similar vein as above, what is the functional interpretation of $\mathbf{C}_{\mathbf{x},\mathbf{y}}\mathbf{V}_{\mathbf{y}}^{-1}\mathbf{C}_{\mathbf{x},\mathbf{y}}^T$? Why does the second term reduce the variance of the error of the estimate below that which is the inherent covariance of the estimated variables themselves? [NOTE: See CBs TR; it covers some of this.]

3.1 LLSE Example

You have the percentages obtained from equal-sized samples from 10 cities indicating the proportion of smokers and drivers as well as the percentage of cases of emphysema and cancer. Given the percentage of smokers and drivers in a sample taken from another city, you want to predict the likely proportion of cases of emphysema and cancer in that sample. [NB: With more detailed (unmerged) information, one could certainly do better.]

$$y = \begin{bmatrix} \%\ \textit{drivers} \\ \%\ \textit{smokers} \end{bmatrix} = \begin{bmatrix} 73 & 82 & 32 & 55 & 29 & 88 & 56 & 61 & 49 & 53 \\ 33 & 47 & 30 & 50 & 16 & 34 & 27 & 22 & 41 & 33 \end{bmatrix}$$

$$x = \begin{bmatrix} \%\ \textit{emphysema} \\ \%\ \textit{cancer} \end{bmatrix} = \begin{bmatrix} 11 & 13 & 8 & 14 & 6 & 11 & 15 & 8 & 17 & 10 \\ 9 & 11 & 8 & 20 & 5 & 13 & 10 & 12 & 15 & 11 \end{bmatrix}$$

$$\hat{\mathbf{x}} = \begin{bmatrix} \% \widehat{emphysema} \\ \% \widehat{cancer} \end{bmatrix} = \overbrace{\begin{bmatrix} 11.3 \\ 11.4 \end{bmatrix}}^{\bar{\mathbf{x}}} + \overbrace{\begin{bmatrix} 21.8 & 25.5 \\ 25.08 & 32.4 \end{bmatrix}}^{\mathbf{C}_{\mathbf{x},\mathbf{y}}} \overbrace{\begin{bmatrix} 0.0033 & -0.0026 \\ -0.0026 & 0.00110 \end{bmatrix}}^{\mathbf{V}_{\mathbf{y}}^{-1}} \overbrace{\begin{bmatrix} \%\ drivers_{obs} & - & 57.8 \\ \%\ smokers_{obs} & - & 33.3 \end{bmatrix}}^{\mathbf{y}-\bar{\mathbf{y}}}$$

$$= \begin{bmatrix} 11.3 \\ 11.4 \end{bmatrix} + \begin{bmatrix} 0.006 & 0.224 \\ -0.002 & 0.292 \end{bmatrix} \begin{bmatrix} \%\ drivers_{obs} & - & 57.8 \\ \%\ smokers_{obs} & - & 33.3 \end{bmatrix}$$

Examination of the results of this (contrived) example show that for this data set, the percentage of drivers is a poor predictor of the percentage of either emphysema or cancer (0.006 and -0.002); however, there is an appreciable relation between the percentage of smokers and these diseases (0.224 and 0.292).

$$\mathbf{V}_{\mathbf{x}-\hat{\mathbf{x}}} = error\ variance = \overbrace{\begin{bmatrix} 12.01 & 9.09 \\ 9.09 & 16.71 \end{bmatrix}}^{\mathbf{V}_{\mathbf{x}}} - \overbrace{\begin{bmatrix} 21.84 & 25.46 \\ 24.98 & 32.42 \end{bmatrix}}^{\mathbf{C}_{\mathbf{x},\mathbf{y}}} \overbrace{\begin{bmatrix} 0.0033 & -0.0026 \\ -0.0026 & 0.00110 \end{bmatrix}}^{\mathbf{V}_{\mathbf{y}}^{-1}} \overbrace{\begin{bmatrix} 21.84 & 24.98 \\ 25.46 & 32.42 \end{bmatrix}}^{\mathbf{C}_{\mathbf{x},\mathbf{y}}^{T}}$$

$$= \begin{bmatrix} 6.19 & 1.70 \\ 1.70 & 7.29 \end{bmatrix}$$

The standard deviation of the estimation is 2.5% (emphysema) and 2.7% (cancer) with a moderate correlation (0.25) between the errors of the two estimates.

3.2 LLSE Summary

Given the values for $\bar{\mathbf{x}}$, $\bar{\mathbf{y}}$, $\mathbf{V}_{\mathbf{y}}$, $\mathbf{C}_{\mathbf{x},\mathbf{y}}$ and the observation $\mathbf{y}$, the linear least squares estimate of $\mathbf{x}$ and the variance of the associated error are given by:

$$\begin{aligned} \hat{\mathbf{x}} &= \bar{\mathbf{x}} + \mathbf{C}_{\mathbf{x},\mathbf{y}}\mathbf{V}_{\mathbf{y}}^{-1}(\mathbf{y} - \bar{\mathbf{y}})\ ; && estimation\ equation \\ \mathbf{V}_{\mathbf{x}-\hat{\mathbf{x}}} &= \mathbf{V}_{\mathbf{x}} - \mathbf{C}_{\mathbf{x},\mathbf{y}}\mathbf{V}_{\mathbf{y}}^{-1}\mathbf{C}_{\mathbf{x},\mathbf{y}}^{T}\ ; && distribution\ equation \end{aligned}$$

4 Obtaining $\hat{\mathbf{x}}$ Given $\mathbf{y} = \mathbf{D}\mathbf{x} + \mathbf{b} + \mathbf{v}$ and $\mathbf{y}$

With LLSE in hand, we now construct one of the components that make up the Kalman filter, the measurement equation. Consider the measurement model $\mathbf{y} = \mathbf{D}\mathbf{x} + \mathbf{b} + \mathbf{v}$ where $\mathbf{b}$ is a known constant[8], $\mathbf{v}$ is zero-mean noise

[8] In addition to providing the offset for the linear plant relation, $\mathbf{b}$ is also used to absorb/cancel any known nonzero-mean of the noise distribution.

with variance $\mathbf{V_v}$ and is uncorrelated with $\mathbf{x}$. Given the relationship $\mathbf{Dx}+\mathbf{b}+\mathbf{v}$ and the measurement $\mathbf{y}$, estimate $\mathbf{x}$:

$$\widehat{\mathbf{x}} = \bar{\mathbf{x}} + \mathbf{V_x D}^T \left[\mathbf{D V_x D}^T + \mathbf{V_v}\right]^{-1} (\mathbf{y} - \mathbf{D}\bar{\mathbf{x}} - \mathbf{b})$$

This follows primarily from the previous section's results. First, note that $\bar{\mathbf{y}} = \mathbf{D}\bar{\mathbf{x}} + \mathbf{b}$:

$$\begin{aligned} \bar{\mathbf{y}} &= E[\mathbf{y}] \\ &= E[\mathbf{Dx}+\mathbf{b}+\mathbf{v}] \\ &= \mathbf{D}\bar{\mathbf{x}} + \mathbf{b} \end{aligned}$$

Then obtain $\mathbf{V_y} = \mathbf{D V_x D}^T + \mathbf{V_v}$ from the given equations:

$$\begin{aligned} \mathbf{V_y} &= E\left[(\mathbf{y}-\bar{\mathbf{y}})(\mathbf{y}-\bar{\mathbf{y}})^T\right] \\ &= E\left[(\mathbf{D}(\mathbf{x}-\bar{\mathbf{x}})+\mathbf{v})(\mathbf{D}(\mathbf{x}-\bar{\mathbf{x}})+\mathbf{v})^T\right] \ ; && \textit{from the above expressions for } \mathbf{y} \textit{ and } \bar{\mathbf{y}} \\ &= \mathbf{D V_x D}^T + \mathbf{D}E\left[(\mathbf{x}-\bar{\mathbf{x}})\mathbf{v}^T\right] + E\left[\mathbf{v}(\mathbf{x}-\bar{\mathbf{x}})^T\right]\mathbf{D}^T + E\left[\mathbf{v}\mathbf{v}^T\right] \ ; && \textit{from } (\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T \\ &= \mathbf{D V_x D}^T + E\left[\mathbf{v}\mathbf{v}^T\right] \ ; && \textit{cross terms eliminated due to} \\ &= \mathbf{D V_x D}^T + \mathbf{V_v} && \textit{uncorrelated noise} \end{aligned}$$

Obtain $\mathbf{C_{x,y}} = \mathbf{V_x D}^T$ by direct computation:

$$\begin{aligned} \mathbf{C_{x,y}} &= E\left[(\mathbf{x}-\bar{\mathbf{x}})(\mathbf{y}-\bar{\mathbf{y}})^T\right] \\ &= E\left[(\mathbf{x}-\bar{\mathbf{x}})(\mathbf{D}(\mathbf{x}-\bar{\mathbf{x}})+\mathbf{v})^T\right] \\ &= \mathbf{V_x D}^T + E\left[(\mathbf{x}-\bar{\mathbf{x}})\mathbf{v}^T\right] \\ &= \mathbf{V_x D}^T \end{aligned}$$

Finally:

$$\begin{aligned} \widehat{\mathbf{x}} &= \bar{\mathbf{x}} + \mathbf{C_{x,y}} \mathbf{V_y}^{-1}(\mathbf{y}-\bar{\mathbf{y}}) \\ &= \bar{\mathbf{x}} + \mathbf{V_x D}^T \left[\mathbf{D V_x D}^T + \mathbf{V_v}\right]^{-1} (\mathbf{y} - \mathbf{D}\bar{\mathbf{x}} - \mathbf{b}) \end{aligned}$$

Using the above result, the variance result from the previous section and the fact that $\mathbf{V_x}^T = \mathbf{V_x}$:

$$\begin{aligned} \mathbf{V_{x-\widehat{x}}} &= \mathbf{V_x} - \mathbf{C_{x,y}} \mathbf{V_y}^{-1} \mathbf{C_{x,y}}^T \\ &= \mathbf{V_x} - \mathbf{V_x D}^T \left[\mathbf{D V_x D}^T + \mathbf{V_v}\right]^{-1} \mathbf{D V_x} \end{aligned}$$

4.1 Summary

Given the model $\mathbf{y} = \mathbf{Dx}+\mathbf{b}+\mathbf{v}$ where $\mathbf{v}$ is a zero mean vector which is uncorrelated with $\mathbf{x}$:

$$
\begin{aligned}
\widehat{\mathbf{x}} &= \bar{\mathbf{x}} + \mathbf{V}_{\mathbf{x}}\mathbf{D}^T\left[\mathbf{D}\mathbf{V}_{\mathbf{x}}\mathbf{D}^T + \mathbf{V}_{\mathbf{v}}\right]^{-1}(\mathbf{y} - \mathbf{D}\bar{\mathbf{x}} - \mathbf{b})\ ; \quad \textit{estimation equation} \\
\mathbf{V}_{\mathbf{x}-\widehat{\mathbf{x}}} &= \mathbf{V}_{\mathbf{x}} - \mathbf{V}_{\mathbf{x}}\mathbf{D}^T\left[\mathbf{D}\mathbf{V}_{\mathbf{x}}\mathbf{D}^T + \mathbf{V}_{\mathbf{v}}\right]^{-1}\mathbf{D}\mathbf{V}_{\mathbf{x}}\ ; \quad \textit{distribution equation}
\end{aligned}
$$

5 Obtaining $\hat{\mathbf{z}}$ Given $\mathbf{z} = \mathbf{Fx} + \mathbf{Hw} + \mathbf{Bu}$ and $\mathbf{y}$

This section shows that if two variables are linearly related and are perturbed by uncorrelated zero-mean noise, then their estimates are also linearly related. This result provides the second component of the Kalman filter, the state prediction equation. Consider the state prediction model $\mathbf{z} = \mathbf{Fx} + \mathbf{Hw} + \mathbf{Bu}$, where $\mathbf{u}$ is a known constant vector (e.g., fixed offsets, programmed changes) and $\mathbf{w}$ is zero mean noise with variance $\mathbf{V}_{\mathbf{w}}$ and is uncorrelated with *both* $\mathbf{x}$ and $\mathbf{y}$. The noise transform, $\mathbf{H}$, and the state constant transform, $\mathbf{B}$, are assumed known and their form depends on where the noise is introduced in the plant relative to the output. In this case, there is a known relationship $\mathbf{F}$ between $\mathbf{z}$ and $\mathbf{x}$. Both $\mathbf{z}$ and $\mathbf{x}$ depend linearly on $\mathbf{y}$. Given a measurement $\mathbf{y}$, $\widehat{\mathbf{x}}$ is obtained and [NOTE: expand and explain], in turn, $\widehat{\mathbf{z}}$.

$$
\bar{\mathbf{z}} = \mathbf{F}\bar{\mathbf{x}} + \mathbf{Bu}
$$

$$
\begin{aligned}
\mathbf{C}_{\mathbf{z},\mathbf{y}} &= E\left[(\mathbf{z} - \bar{\mathbf{z}})(\mathbf{y} - \bar{\mathbf{y}})^T\right] \\
&= E\left[(\mathbf{F}(\mathbf{x} - \bar{\mathbf{x}}) + \mathbf{Hw})(\mathbf{y} - \bar{\mathbf{y}})^T\right] \\
&= \mathbf{F}\mathbf{C}_{\mathbf{x},\mathbf{y}}
\end{aligned}
$$

$$
\begin{aligned}
\widehat{\mathbf{z}} &= \bar{\mathbf{z}} + \mathbf{C}_{\mathbf{z},\mathbf{y}}\mathbf{V}_{\mathbf{y}}^{-1}(\mathbf{y} - \bar{\mathbf{y}})\ ; \quad \textit{LLSE estimation equation} \\
&= \mathbf{F}\bar{\mathbf{x}} + \mathbf{Bu} + \mathbf{F}\mathbf{C}_{\mathbf{x},\mathbf{y}}\mathbf{V}_{\mathbf{y}}^{-1}(\mathbf{y} - \bar{\mathbf{y}})\ ; \quad \textit{substitute for first two variables} \\
&= \mathbf{F}\left[\bar{\mathbf{x}} + \mathbf{C}_{\mathbf{x},\mathbf{y}}\mathbf{V}_{\mathbf{y}}^{-1}(\mathbf{y} - \bar{\mathbf{y}})\right] + \mathbf{Bu} \\
&= \mathbf{F}\widehat{\mathbf{x}} + \mathbf{Bu}
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{V}_{\mathbf{z}} &= E\left[(\mathbf{Fx} + \mathbf{Hw} + \mathbf{Bu} - \bar{\mathbf{z}})(\mathbf{Fx} + \mathbf{Hw} + \mathbf{Bu} - \bar{\mathbf{z}})^T\right] \\
&= E\left[(\mathbf{F}(\mathbf{x} - \bar{\mathbf{x}}) + \mathbf{Hw})(\mathbf{F}(\mathbf{x} - \bar{\mathbf{x}}) + \mathbf{Hw})^T\right]\ ; \quad \textit{substitute for } \bar{\mathbf{z}} \\
&= \mathbf{F}\mathbf{V}_{\mathbf{x}}\mathbf{F}^T + \mathbf{F}E\left[(\mathbf{x} - \bar{\mathbf{x}})\mathbf{w}^T\right] \supset mH^T + \mathbf{H}E\left[\mathbf{w}(\mathbf{x} - \bar{\mathbf{x}})^T\right]\mathbf{F}^T + \mathbf{H}E\left[\mathbf{w}\mathbf{w}^T\right]\mathbf{H}^T \\
&= \mathbf{F}\mathbf{V}_{\mathbf{x}}\mathbf{F}^T + \mathbf{F}E\left[(\mathbf{x} - \bar{\mathbf{x}})\mathbf{w}^T\right] \supset mH^T + \mathbf{H}E\left[\mathbf{w}(\mathbf{x} - \bar{\mathbf{x}})^T\right]\mathbf{F}^T + \mathbf{H}\mathbf{V}_{\mathbf{w}}\mathbf{H}^T \\
&= \mathbf{F}\mathbf{V}_{\mathbf{x}}\mathbf{F}^T + \mathbf{H}\mathbf{V}_{\mathbf{w}}\mathbf{H}^T
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{V}_{\mathbf{z}-\widehat{\mathbf{z}}} &= \mathbf{V}_{\mathbf{z}} - \mathbf{C}_{\mathbf{z},\mathbf{y}}\mathbf{V}_{\mathbf{y}}^{-1}\mathbf{C}^T_{jvz,\mathbf{y}}\ ; \quad \textit{LLSE distribution equation} \\
&= \mathbf{F}\mathbf{V}_{\mathbf{x}}\mathbf{F}^T + \mathbf{H}\mathbf{V}_{\mathbf{w}}\mathbf{H}^T - \mathbf{F}\mathbf{C}_{\supset vx,\mathbf{y}}\mathbf{V}_{\mathbf{y}}^{-1}(\mathbf{F}\mathbf{C}_{\mathbf{x},\mathbf{y}})^T \\
&= \mathbf{F}(\mathbf{V}_{\mathbf{x}} - \mathbf{C}_{\mathbf{x},\mathbf{y}}\mathbf{V}_{\mathbf{y}}^{-1}\mathbf{C}^T_{\mathbf{x},\mathbf{y}})\mathbf{F}^T + \mathbf{H}\mathbf{V}_{\mathbf{w}}\mathbf{H}^T \\
&= \mathbf{F}\mathbf{V}_{\mathbf{x}-\widehat{\mathbf{x}}}\mathbf{F}^T + \mathbf{H}\mathbf{V}_{\mathbf{w}}\mathbf{H}^T
\end{aligned}
$$

5.1 Summary

$$\begin{aligned} \widehat{\mathbf{z}} &= \mathbf{F}\widehat{\mathbf{x}} + \mathbf{B}\mathbf{u} \; ; \; \textit{estimation equation} && (Eq1) \\ \mathbf{V}_{\mathbf{z}-\widehat{\mathbf{z}}} &= \mathbf{F}\mathbf{V}_{\mathbf{x}-\widehat{\mathbf{x}}}\mathbf{F}^T + \mathbf{H}\mathbf{V}_{\mathbf{w}}\mathbf{H}^T \; ; \; \textit{distribution equation} && (Eq2) \end{aligned}$$

6 Discrete Time Kalman Filter

Combining the results of the previous two sections, we obtain a time varying version of linear least squares estimation, the Kalman filter. Consider the following dynamic system:

$$\begin{aligned} \mathbf{x}_{k+1} &= \mathbf{F}_k\mathbf{x}_k + \mathbf{B}_k\mathbf{u}_k + \mathbf{H}_k\mathbf{w}_k \; ; \textit{state prediction equation} && (Eq3) \\ \mathbf{y}_{k+1} &= \mathbf{D}_{k+1}\mathbf{x}_{k+1} + \mathbf{b}_{k+1} + \mathbf{v}_{k+1} \; ; \textit{measurement equation} && (Eq4) \end{aligned}$$

where $\mathbf{x}$ is the estimated variable, $\mathbf{u}$ is a constant, $\mathbf{y}$ is the measurement, and $\mathbf{w}$ and $\mathbf{v}$ are zero-mean Gaussian noise. $\mathbf{F}$, $\mathbf{B}$, $\mathbf{D}$, $\mathbf{H}$, $\mathbf{b}$ and the variance of the noise can change over time as long as the values are known *a priori*.

Eq3 provides a means for obtaining the estimate $\widehat{\mathbf{x}}_{k+1}$ from the first k observations; this estimate will be denoted by $\widehat{\mathbf{x}}_{k+1|k}$. The following equation provides a means for obtaining $\widehat{\mathbf{x}}_{k+1}$ from the $k^{th}+1$ observation; this estimate will be denoted by $\widehat{\mathbf{x}}_{k+1|k+1}$.

The goal is to develop a recursive procedure for optimal filtering; for estimating $\mathbf{x}_k$ based on $\mathbf{y}_\tau$ where $\tau \leq k$.

Eq1[9] through *Eq3* provide the following:

$$\begin{aligned} \widehat{\mathbf{x}}_{k+1|k} &= \mathbf{F}_k\widehat{\mathbf{x}}_{k|k} + \mathbf{B}_k\mathbf{u}_k \; ; && (Eq5) \\ \mathbf{V}_{\mathbf{x}_{k+1}-\widehat{\mathbf{x}}_{k+1|k}} &= \mathbf{F}_k\mathbf{V}_{\mathbf{x}_k-\widehat{\mathbf{x}}_{k|k}}\mathbf{F}_k^T + \mathbf{H}_k\mathbf{V}_{\mathbf{w}_k}\mathbf{H}_k^T \; ; && (Eq6) \end{aligned}$$

$Eq6$ is the crux of the ability of the Kalman filter to dynamically alter the relative weighting given to the predictions due to the measurements and the predictions due to the model.

Define the following two vectors:

$$\begin{aligned} \mathbf{e}_{k+1} &= \mathbf{x}_{k+1} - \widehat{\mathbf{x}}_{k+1|k} \; ; \quad \textit{error of the predicted estimate} \\ \mathbf{n}_{k+1} &= \mathbf{y}_{k+1} - \widehat{\mathbf{y}}_{k+1} \; ; \quad \textit{error of the predicted measurement, the innovation vector} \end{aligned}$$

Note that they have the following properties:

[9] The random variable, $\widehat{\mathbf{z}}$, in *Eq1* is the estimate of the next state vector, $\widehat{\mathbf{x}}_{k+1}$

$$
\begin{aligned}
\widehat{\mathbf{e}}_{k+1} &= \widehat{\mathbf{x}}_{k+1|k+1} - \widehat{\mathbf{x}}_{k+1|k} \;;\quad \textit{from } Eq1
\end{aligned}
$$

$$
\begin{aligned}
\bar{\mathbf{e}}_{k+1} &= E\left[\mathbf{x}_{k+1} - \widehat{\mathbf{x}}_{k+1|k}\right] \\
&= \bar{\mathbf{x}}_{k+1} - \mathbf{F}_k\bar{\mathbf{x}}_k - \mathbf{B}_k\mathbf{u}_k \;;\quad \textit{by taking } E[Eq3] \\
&= \mathbf{0} \\
\mathbf{V}_{\mathbf{e}_{k+1}} &= E\left[(\mathbf{x}_{k+1} - \widehat{\mathbf{x}}_{k+1|k})(\mathbf{x}_{k+1} - \widehat{\mathbf{x}}_{k+1|k})^T\right] \\
&= \mathbf{V}_{\mathbf{x}_{k+1} - \widehat{\mathbf{x}}_{k+1|k}} \;;\quad \textit{since } \bar{\mathbf{e}}_{k+1} = 0
\end{aligned}
$$

$$
\begin{aligned}
\bar{\mathbf{n}}_{k+1} &= E\left[\mathbf{y}_{k+1} - \widehat{\mathbf{y}}_{k+1}\right] \\
&= \mathbf{0} \;;\quad \textit{since } \mathbf{y}_{k+1} = \widehat{\mathbf{y}}_{k+1},\ \textit{from } E[\widehat{\mathbf{x}}] = \bar{\mathbf{x}}
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{V}_{\mathbf{n}_{k+1}} &= E\left[(\mathbf{y}_{k+1} - \widehat{\mathbf{y}}_{k+1})(\mathbf{y}_{k+1} - \widehat{\mathbf{y}}_{k+1})^T\right] \\
&= E\left[(\mathbf{D}_{k+1}(\mathbf{x}_{k+1} - \widehat{\mathbf{x}}_{k+1|k}) + \mathbf{v}_{k+1})(\mathbf{D}_{k+1}(\mathbf{x}_{k+1} - \widehat{\mathbf{x}}_{k+1|k}) + \mathbf{v}_{k+1})^T\right] \;;\quad \textit{from } Eq4 \\
&= \mathbf{D}_{k+1}\mathbf{V}_{\mathbf{x}_{k+1} - \widehat{\mathbf{x}}_{k+1|k}}\mathbf{D}_{k+1}^T + \mathbf{V}_{\mathbf{v}_{k+1}}
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{C}_{\mathbf{e}_{k+1},\mathbf{n}_{k+1}} &= E\left[(\mathbf{e}_{k+1})(\mathbf{n}_{k+1})^T\right] \;;\quad \textit{since both are zero mean} \\
&= E\left[(\mathbf{x}_{k+1} - \widehat{\mathbf{x}}_{k+1|k})(\mathbf{D}_{k+1}(\mathbf{x}_{k+1} - \widehat{\mathbf{x}}_{k+1|k}) + \mathbf{v}_{k+1})^T\right] \;;\quad \textit{def}^n \textit{ of } \mathbf{e} \textit{ and } \mathbf{n} \\
&= \mathbf{V}_{\mathbf{x}_{k+1} - \widehat{\mathbf{x}}_{k+1|k}}\mathbf{D}_{k+1}^T + E\left[(\mathbf{x}_{k+1})(\mathbf{v}_{k+1})^T\right] - E\left[(\widehat{\mathbf{x}}_{k+1|k})(\mathbf{v}_{k+1})^T\right] \\
&= \mathbf{V}_{\mathbf{x}_{k+1} - \widehat{\mathbf{x}}_{k+1|k}}\mathbf{D}_{k+1}^T - E\left[(\widehat{\mathbf{x}}_{k+1|k})(\mathbf{v}_{k+1})^T\right] \;;\quad \textit{since } \mathbf{x} \textit{ and } \mathbf{v} \textit{ are uncorrelated} \\
&= \mathbf{V}_{\mathbf{x}_{k+1} - \widehat{\mathbf{x}}_{k+1|k}}\mathbf{D}_{k+1}^T \;;\quad \textit{since } E[\mathbf{v}] = 0
\end{aligned}
$$

Now we relate the change in the state vector to the unexpected part of the measurement vector. The LLSE of $\mathbf{e}_k$ from $\mathbf{n}_k$, the improvement of the estimation of $\mathbf{x}$ due to the new information at time $k+1$, and its error variance is given by:

$$
\begin{aligned}
\widehat{\mathbf{e}}_{k+1} &= \bar{\mathbf{e}}_{k+1} + \mathbf{C}_{\mathbf{e}_{k+1},\mathbf{n}_{k+1}}\mathbf{V}_{\mathbf{n}_{k+1}}^{-1}(\mathbf{n}_{k+1} - \widehat{\mathbf{n}}_{k+1}) \;;\quad \textit{from LLSE estimation equation} \\
&= \mathbf{V}_{\mathbf{x}_{k+1} - \widehat{\mathbf{x}}_{k+1|k}}\mathbf{D}_{k+1}^T\left[\mathbf{D}_{k+1}\mathbf{V}_{\mathbf{x}_{k+1} - \widehat{\mathbf{x}}_{k+1|k}}\mathbf{D}_{k+1}^T + \mathbf{V}_{\mathbf{v}_{k+1}}\right]^{-1}(\mathbf{y}_{k+1} - \mathbf{D}_{k+1}\widehat{\mathbf{x}}_{k+1|k}) \qquad (Eq7)
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{V}_{\mathbf{e}_{k+1} - \widehat{\mathbf{e}}_{k+1}} &= \mathbf{V}_{\mathbf{e}_{k+1}} - \mathbf{C}_{\mathbf{e}_{k+1},\mathbf{n}_{k+1}}\mathbf{V}_{\mathbf{n}_{k+1}}^{-1}\mathbf{C}_{\mathbf{e}_{k+1},\mathbf{n}_{k+1}}^T \;;\quad \textit{from LLSE distribution equation} \\
&= \mathbf{V}_{\mathbf{x}_{k+1} - \widehat{\mathbf{x}}_{k+1|k}} \\
&\quad - \mathbf{V}_{\mathbf{x}_{k+1} - \widehat{\mathbf{x}}_{k+1|k}}\mathbf{D}_{k+1}^T\left[\mathbf{D}_{k+1}\mathbf{V}_{\mathbf{x}_{k+1} - \widehat{\mathbf{x}}_{k+1|k}}\mathbf{D}_{k+1}^T + \mathbf{V}_{\mathbf{v}_{k+1}}\right]^{-1}\mathbf{D}_{k+1}\mathbf{V}_{\mathbf{x}_{k+1} - \widehat{\mathbf{x}}_{k+1|k}}
\end{aligned}
$$

In answer to the question: *What do you trust, the data or the prediction model?:*

Define the Kalman gain matrix (a relative weighting of the predictions obtained from the data and from the model based on their variability):

$$
\begin{aligned}
\mathbf{K}_{k+1} &= \mathbf{V}_{\mathbf{x}_{k+1}-\widehat{\mathbf{x}}_{k+1|k}} \mathbf{D}_{k+1}^T \left[\mathbf{D}_{k+1} \mathbf{V}_{\mathbf{x}_{k+1}-\widehat{\mathbf{x}}_{k+1|k}} \mathbf{D}_{k+1}^T + \mathbf{V}_{\mathbf{v}_{k+1}}\right]^{-1}
\end{aligned}
$$

$$
\begin{aligned}
\widehat{\mathbf{x}}_{k+1|k+1} &= \widehat{\mathbf{x}}_{k+1|k} + \widehat{\mathbf{e}}_{k+1} \ ; \ \textit{from Eq7 and substituting in } \mathbf{K} \\
&= \underbrace{\widehat{\mathbf{x}}_{k+1|k}}_{model} + \mathbf{K}_{k+1} \underbrace{(\mathbf{y}_{k+1} - \mathbf{D}_{k+1}\widehat{\mathbf{x}}_{k+1|k})}_{data}
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{V}_{\mathbf{x}_{k+1}-\widehat{\mathbf{x}}_{k+1|k+1}} &= \mathbf{V}_{\mathbf{e}_{k+1}-\widehat{\mathbf{e}}_{k+1}} \\
&= \mathbf{V}_{\mathbf{x}_{k+1}-\widehat{\mathbf{x}}_{k+1|k}} - \mathbf{K}_{k+1} \mathbf{D}_{k+1} \mathbf{V}_{\mathbf{x}_{k+1}-\widehat{\mathbf{x}}_{k+1|k}} \qquad (Eq8)
\end{aligned}
$$

How can we iterate the process?:

Assume[10] we have $\widehat{\mathbf{x}}_{k|k}$ and $\mathbf{V}_{\mathbf{x}_k-\widehat{\mathbf{x}}_{k|k}}$.

Apply *Eq5* to get $\widehat{\mathbf{x}}_{k+1|k}$.

Apply *Eq6* to get $\mathbf{V}_{\mathbf{x}_{k+1}-\widehat{\mathbf{x}}_{k+1|k}}$.

Assume we now have $\mathbf{y}_{k+1}$.

In theory, we could apply *Eq4* to get $\widehat{\mathbf{x}}_{k+1|k+1}$
but instead we use the innovation process and apply *Eq7*:

$$
\begin{aligned}
\widehat{\mathbf{x}}_{k+1|k+1} &= \widehat{\mathbf{x}}_{k+1|k} + \widehat{\mathbf{e}}_{k+1} \\
& \quad \textit{where the first term is known and the second} \\
& \quad \textit{is estimated from the innovation vector.}
\end{aligned}
$$

Finally, apply $Eq8$ to get $\mathbf{V}_{\mathbf{x}_{k+1}-\widehat{\mathbf{x}}_{k+1|k+1}}$

Iterate, iterate, iterate, ...

6.1 Constant Velocity Example

In this problem we have an object we wish to track that can independently change position in two dimensions and we assume that it maintains a constant velocity. What follows is the heart of the Kalman filtering technique. We have two relations that involve the variable we want to predict, $\mathbf{x}_{k+1}$, the state of the system at time $k+1$: the x and y components of the position and velocity of the object.

The first relation is the state prediction equation ($Eq3$):

[10] When $k = 0$, these values arise from calibration or revelation.

$$\overbrace{\begin{bmatrix} pos^x_{k+1} \\ pos^y_{k+1} \\ spd^x_{k+1} \\ spd^y_{k+1} \end{bmatrix}}^{\mathbf{x}_{k+1}} = \overbrace{\begin{bmatrix} 1 & 0 & \Delta t & 0 \\ 0 & 1 & 0 & \Delta t \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}}^{\mathbf{F}_k} \overbrace{\begin{bmatrix} pos^x_k \\ pos^y_k \\ spd^x_k \\ spd^y_k \end{bmatrix}}^{\mathbf{x}_k} + \overbrace{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}}^{\mathbf{B}_k} \overbrace{\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}}^{\mathbf{u}_k} + \overbrace{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}}^{\mathbf{H}_k} \overbrace{\begin{bmatrix} w_{pos^x_k} \\ w_{pos^y_k} \\ w_{spd^x_k} \\ w_{spd^y_k} \end{bmatrix}}^{\mathbf{w}_k}$$

In this simple example, the x and y position estimates are updated at each step by the change in position due to the respective speed components times the interval between samples, Δt. The constant vector $\mathbf{u}_k$ reflects any controlled (*a priori*) changes in the state vector (e.g., a programmed change in the velocity of the system). No changes are induced in this manner in this example. The last term in the noise vector and its associated transform. The vector, $\mathbf{w}_k$, reflects the prediction noise in the model. In the current example, these might be errors due to air currents acting on the object or roundoff error in the computations.

The second relation is the measurement equation($Eq4$):

$$\overbrace{\begin{bmatrix} pos^x_{k+1} \\ pos^y_{k+1} \end{bmatrix}}^{\mathbf{y}_{k+1}} = \overbrace{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}}^{\mathbf{D}_{k+1}} \overbrace{\begin{bmatrix} pos^x_{k+1} \\ pos^y_{k+1} \\ spd^x_{k+1} \\ spd^y_{k+1} \end{bmatrix}}^{\mathbf{x}_{k+1}} + \overbrace{\begin{bmatrix} 0 \\ 0 \end{bmatrix}}^{\mathbf{b}_{k+1}} + \overbrace{\begin{bmatrix} v_{pos^x_k} \\ v_{pos^y_k} \end{bmatrix}}^{\mathbf{v}_{k+1}}$$

The linear transform $\mathbf{D}_{k+1}$ established the relation between the new measurement $\mathbf{y}_{k+1}$ and the state vector $\mathbf{x}_{k+1}$. The constant vector $\mathbf{b}_{k+1}$, here set to zero, could be used to incorporate any discrepancies between the measurement and state domains or to cancel the mean of the noise distribution. Finally, additive noise $\mathbf{v}_{k+1}$ is assumed to be injected into the measurement.

The initial state vector $\widehat{\mathbf{x}}_{0|0}$ and $\mathbf{V}_{\mathbf{x}_0-\widehat{\mathbf{x}}_{0|0}}$ are assumed known. In practice, the initial vector is assigned a theoretically consistent value, a value obtained by running test trials or a calibration, or simply a good guess.

$$\widehat{\mathbf{x}}_{0|0} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}$$

In Sections 4 & 5 we developed estimation procedures for each of these prediction equations. For the state prediction equation (from $Eq5$ and with $\Delta t = 1$):

$$\widehat{\mathbf{x}}_{k+1} = \mathbf{F}_k \widehat{\mathbf{x}}_{k|k} + \mathbf{B}_k \mathbf{u}_k$$

$$\widehat{\mathbf{x}}_{1|0} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

How does $\mathbf{y}_1$ change $\widehat{\mathbf{x}}_{1|0}$ into $\widehat{\mathbf{x}}_{1|1}$?

Let the new measurement be:

$$\mathbf{y}_1 = \begin{bmatrix} 0.9 \\ 1.1 \end{bmatrix}$$

$$\mathbf{n}_1 = \mathbf{y}_1 - \widehat{\mathbf{y}}_1 = \mathbf{y}_1 - \mathbf{D}_{x+1}\widehat{\mathbf{x}}_{1|0} = \begin{bmatrix} -0.1 \\ 0.1 \end{bmatrix}$$

$$\widehat{\mathbf{x}}_{1|1} = \widehat{\mathbf{x}}_{1|0} + \widehat{\mathbf{e}}_1$$

Another little-mentioned sleazy fact in all of this is how the error covariance estimates are obtained in practice. In our case, we will assume it was possible to obtain 100 samples of the measurement noise vector, $\mathbf{v}$. The covariance matrix from 100 2 by 1 sample vectors taken from a zero-mean Gaussian distribution with a standard deviation of 0.1 is:

$$\mathbf{V}_{\mathbf{v}} = \begin{bmatrix} 0.0083 & 0.0001 \\ 0.0001 & 0.0116 \end{bmatrix}$$

In our example, it is presumed that the measurement noise is stationary throughout the tracking period. If there were *a priori* knowledge to the contrary (e.g., sensor characteristics changing as they warm up, or age), then this matrix could be altered to reflect these changes.

A little more difficult to come by are estimates of the estimation noise, $\mathbf{w}$, and the initial state estimation variance itself, $\mathbf{V}_{\mathbf{x}_0 - \widehat{\mathbf{x}}_{0|0}}$. Any physical knowledge about the operation of the plant and its corresponding precision would be helpful here (e.g., consequences of backlash, resolution of encoders, forces exerted by swinging camera wires). In this example, we will assume we have access to a number of test trials where we ran the system on a calibration target having a known sequence of positions. Here as well, we assume the state noise is stationary and is zero-mean Gaussian with a standard deviation of 0.1:

$$\mathbf{V}_{\mathbf{w}} = \begin{bmatrix} 0.0091 & -0.0002 & 0.0004 & -0.0006 \\ -0.0002 & 0.0113 & 0.0000 & 0.0016 \\ 0.0004 & 0.0000 & 0.0118 & 0.0007 \\ -0.0006 & 0.0016 & 0.0007 & 0.0099 \end{bmatrix}$$

We also assume that calibration data yielded a distribution of estimation error that was zero-mean Gaussian with a standard deviation of 1.0:

$$\mathbf{V}_{\mathbf{x}_{0}-\widehat{\mathbf{x}}_{0|0}} = \begin{bmatrix} 0.9225 & 0.0019 & -0.0642 & 0.0188 \\ 0.0019 & 1.0515 & -0.0286 & 0.0200 \\ -0.0642 & -0.0286 & 1.1324 & 0.0858 \\ 0.0188 & 0.0200 & 0.0858 & 0.8843 \end{bmatrix}$$

Now from the your-mileage-may-vary category of advice: If you are unsure of your model, you may want to inflate $\mathbf{V}_{\mathbf{x}_{0}-\widehat{\mathbf{x}}_{0|0}}$ so that the filter will initially rely on the measurement data; however, if you are confident in your model and have a good estimate of the initial conditions (i.e., you have "homed" the system), you may want to use a smaller estimate of the initial estimation variance.

To get the ball rolling, we create an update of the estimation variance from initial conditions (*Eq6*):

$$\begin{aligned} \mathbf{V}_{\mathbf{x}_{k+1}-\widehat{\mathbf{x}}_{k+1|k}} &= \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0.9225 & 0.0019 & -0.0642 & 0.0188 \\ 0.0019 & 1.0515 & -0.0286 & 0.0200 \\ -0.0642 & -0.0286 & 1.1324 & 0.0858 \\ 0.0188 & 0.0200 & 0.0858 & 0.8843 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} + \\ & \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0.0091 & -0.0002 & 0.0004 & -0.0006 \\ -0.0002 & 0.0113 & 0.0000 & 0.0016 \\ 0.0004 & 0.0000 & 0.0118 & 0.0007 \\ -0.0006 & 0.0016 & 0.0007 & 0.0099 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 1.9356 & 0.0777 & 1.0686 & 0.1040 \\ 0.0777 & 1.9871 & 0.0572 & 0.9059 \\ 1.0686 & 0.0572 & 1.1442 & 0.0865 \\ 0.1040 & 0.9059 & 0.0865 & 0.8942 \end{bmatrix} \end{aligned}$$

And now computing the error term from *Eq7*:

$$
\widehat{\mathbf{e}}_1 = \begin{bmatrix} 1.9356 & 0.0777 & 1.0686 & 0.1040 \\ 0.0777 & 1.9871 & 0.0572 & 0.9059 \\ 1.0686 & 0.0572 & 1.1442 & 0.0865 \\ 0.1040 & 0.9059 & 0.0865 & 0.8942 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \times
$$

$$
\left[\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1.9356 & 0.0777 & 1.0686 & 0.1040 \\ 0.0777 & 1.9871 & 0.0572 & 0.9059 \\ 1.0686 & 0.0572 & 1.1442 & 0.0865 \\ 0.1040 & 0.9059 & 0.0865 & 0.8942 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0.0083 & 0.0001 \\ 0.0001 & 0.0116 \end{bmatrix} \right]^{-1} \begin{bmatrix} -0.1 \\ 0.1 \end{bmatrix}
$$

$$
= \begin{bmatrix} -0.0996 \\ 0.0994 \\ -0.0542 \\ 0.0416 \end{bmatrix}
$$

$$
\widehat{\mathbf{x}}_{1|1} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} -0.0996 \\ 0.0994 \\ -0.0542 \\ 0.0416 \end{bmatrix} = \begin{bmatrix} 0.9004 \\ 1.0994 \\ 0.9458 \\ 1.0416 \end{bmatrix}
$$

Now apply $Eq8$ to get $\mathbf{V}_{\mathbf{x}_{k+1} - \widehat{\mathbf{x}}_{k+1|k+1}}$ and you are ready for the next iteration.

As would be expected given the relative magnitude of the estimation variance as compared to the measurement variance, the prediction is heavily weighted by the data. If, however, the estimation variance is assumed equal to that of the measurement variance ($\sigma = 0.1$), a more balanced influence is observed:

$$
\widehat{\mathbf{x}}_{1|1} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} -0.0528 \\ 0.0478 \\ 0.0021 \\ -0.0002 \end{bmatrix} = \begin{bmatrix} 0.9472 \\ 1.0478 \\ 1.0021 \\ 0.9998 \end{bmatrix}
$$

6.2 Time-Invariant Kalman Filters

If you assume that the plant is fixed and you allow all actual variations to be absorbed by the noise term, there follows a number of computational advantages. [Make explicit what is kept constant and that only position (not velocity) measurements are used.] Obtaining the optimal estimate is simpler than in the time-varying case. The state estimation covariance and filter gain matrices have steady-state values that may be computed in advance. Two time-invariant variants are constant-velocity and constant acceleration systems.

In the constant-velocity case, the initial velocity of the target is altered over time through the effect of random accelerations included in the process noise. These random changes are assumed constant during each sampling interval but independent across intervals. Tracking an object that is, in fact, undergoing acceleration is accommodated by specifying high process noise.

Consider the 2D tracking example. The velocity and position in the X- and Y-dimensions change independently and the measurements of these parameters are subject to constant covariance. As before, the state of the target, $\mathbf{x}$ (the 2D position and velocity of the target), is updated by:

$$\mathbf{x}_{k+1} = \mathbf{F}_k \mathbf{x}_k + \mathbf{B}_k \mathbf{u}_k + \mathbf{H}_k \mathbf{w}_k$$

where $\mathbf{B}_k$ and $\mathbf{H}_k$ are identity matrices and there are no programmed changes ($\mathbf{u}_k = \mathbf{0}$).

The $\alpha - \beta$ (constant-velocity) state prediction equation produces an updated estimate of the state vector given the new measurement:

$$\widehat{\mathbf{x}}_{k+1|k+1} = \widehat{\mathbf{x}}_{k+1|k} + \begin{bmatrix} \alpha \\ \frac{\beta}{\Delta t} \end{bmatrix} \left[\mathbf{y}_{k+1} - \widehat{\mathbf{y}}_{k+1|k} \right]$$

This new state estimate is the weighted sum of the state predicted from the last estimate ($\mathbf{F}_k \mathbf{x}_{k|k}$) and the innovation term, the difference between the predicted and actual measurements. The optimal weights can be derived [NOTE: see Bar-Shalom] and are a function of solely of the ratio of the standard deviations of the process and measurement noises (the maneuvering index, λ):

$$\alpha = \frac{\lambda^2 + 8\lambda - (\lambda + 4)\sqrt{\lambda^2 + 8\lambda}}{8}$$

and

$$\beta = \frac{\lambda^2 + 4\lambda - \lambda\sqrt{\lambda^2 + 8\lambda}}{4}$$

The state estimation covariances may also be found in closed form. [NOTE: Also get from Bar-Shalom.]

In the constant-acceleration case, an additional weighting factor is computed (γ). This $\alpha - \beta - \gamma$ filter exhibits improved tracking of targets with smoothly changing velocities but at the cost of increased sensitivity to noise. In this model it is assumed that the target undergoes small random perturbations in acceleration (random jerk). The state prediction equation is:

$$\widehat{\mathbf{x}}_{k+1|k+1} = \widehat{\mathbf{x}}_{k+1|k} + \begin{bmatrix} \alpha \\ \frac{\beta}{\Delta t} \\ \frac{\gamma}{\Delta t^2} \end{bmatrix} \left[\mathbf{y}_{k+1} - \widehat{\mathbf{y}}_{k+1|k} \right]$$

where α and β are as above and:

$$\gamma = \frac{\beta^2}{\alpha}$$

6.3 Extended Kalman Filters

Another variation of the Kalman Filter technique is the extension to applications where the relation between the measurement and the estimation is nonlinear. In these cases, an attempt is made to apply piecewise linear solutions. [NOTE: find a good reference for this and extend the prior worked example.]

Appendices

A When is Least Squares a Maximum Likelihood Estimator?

Least squares provides a convenient computational form of estimation. Under what conditions is the obtained estimate the most likely estimate? What population distribution is most likely to have given rise to the obtained set of observations? The observed values can be converted into probabilities for a given probability density function and the joint probability of the occurrence of the set of measurements is the product of these probabilities. If the measurements are assumed to be corrupted by zero-mean Gaussian noise, then the computationally expensive products can be simplified.

Find the population mean that maximizes the probability of obtaining a given set of measurements (i.e., a set of values, $\mathbf{x}_i$):

$$\max\left(\prod \exp^{\frac{-(x_i-\mu)^2}{\sigma^2}}\right)$$

where σ^2 is the sample variance.[11]

Since the natural logarithm is a monotonically increasing function, the above expression is equivalent to:

$$\begin{aligned}
&= \max\left(\ln\left(\prod \exp^{\frac{-(x_i-\mu)^2}{\sigma^2}}\right)\right)\\
&= \max\left(\sum\left(\ln \exp^{\frac{-(x_i-\mu)^2}{\sigma^2}}\right)\right)\\
&= \max\left(\sum \frac{-(x_i-\mu)^2}{\sigma^2}\right)\\
&= \min\left(\sum \frac{(x_i-\mu)^2}{\sigma^2}\right)\\
&= \min\left(\sum (x_i-\mu)^2\right)
\end{aligned}$$

This last expression is equivalent to a least squares estimate for observations perturbed by Gaussian noise. The practical question remains as to what distortions of the Gaussian requirement can be sustained while still returning an estimate that is "close" to optimal. What other forms of exponential pdfs both guarantee that LLSE is an MLE and provide the computational convenience of a simple addition to find the minimum? In modeling practical systems is unimodality sufficient to achieve a good approximation or are there features of the distribution (such as heavy

[11] It is possible in many cases to have more detailed knowledge of the pattern of variability in the data (e.g., images were acquired from an array of cameras, some more noisy than others). In these cases, it is appropriate to partition the data and divide by the associated subset variance.

tails) that will disrupt the least squares approximation process? Are there sampling and measurements techniques that through their application, the Central Limit Theorem would predict that deviant data would approach the desired distribution? Alternatively, there are other computationally efficient estimation techniques such as Robust Estimation that don't require a Gaussian noise distribution (but may impose their own constraints).

B What is the Relation between Maximum A Posteriori and Optimal Linear Estimation?

In the Bayesian approach, the parameter being estimated is a random variable. There is no single true answer but, instead, there exists a pdf, $p(x)$. If the prior pdf is known then the posterior pdf may be computed (Bayes Rule):

$$p(x|Z^k) \;=\; \frac{p(Z^k|x)\;p(x)}{p(Z^k)}$$

The most likely value of the parameter, x, is the maximum a posteriori (MAP) estimate:

$$\hat{\mathbf{x}}_k^{MAP} \;\doteq\; argmax_x p(x|Z^k) \;=\; argmax_x p(Z^k|x)\;p(x)$$

Unlike the case with the maximum likelihood estimator (Appendix A), it is assumed here that there is some prior idea as to what the answer is. The relative influence the measurement and the prior information have on the final estimate depend on their respective variances. In the event that nothing is initially known about the parameter, the variance of the prior pdf is allowed to go to infinity (diffuse prior). Under these conditions, MAP becomes MLE.

C Useful Operators and their Transforms

Or: You can't tell the players without a scorecard.

Covariance, correlation, convolution - each is a way to express the interaction between two functions. The maxima of these relations can be derived from the fact that the arithmetic mean is always greater than or equal to the geometric mean of a set of positive numbers which, after a few manipulations leads to:

$$a^2 + b^2 \geq 2ab$$

The upshot of this is that the integrated product of two signals with similar profiles is maximized when the similar regions are brought into alignment.

Let a stochastic process define a family of real-valued functions, $\mathbf{a}$, defined in a probability space and sampled at discrete points in space (or time), then the sample at each position (or time) $\mathbf{a}_k$ is a random variable with a mean value at position (time) k:

$$E\left[\mathbf{a}_k\right] = \eta_k$$

The *autocorrelation* of the process, $\mathbf{a}$, is:

$$E\left[\mathbf{a}_i\ \mathbf{a}_j\right] = R_{aa}(i,j)$$

The *cross-correlation* of two processes, $\mathbf{a}$ and $\mathbf{b}$ is the expected value sof the product, ab:

$$E\left[\mathbf{a}_i\ \mathbf{b}_j\right] = R_{ab}(i,j)$$

The *autocovariance* of the process, $\mathbf{a}$, is:

$$E\left[(\mathbf{a}_i - \eta_i)\ (\mathbf{a}_j - \eta_j)\right] = C_{aa}(i,j)$$

The *cross-covariance* of two processes, $\mathbf{a}$ and $\mathbf{b}$ is:

$$E\left[(\mathbf{a}_i - \eta_i)\ (\mathbf{b}_j - \eta_j)\right] = C_{ab}(i,j)$$

A stochastic process is considered *wide-sense stationary* if its mean is constant and its autocorrelation depends only on the relative position of the sampling ($\tau = i - j$):

$$E\left[\mathbf{a}_k\right] = \eta = \ \ constant$$

$$R_{aa}(i,j) = R_{aa}(\tau) = E\left[(\mathbf{a}_i)\ (\mathbf{a}_i)\right]$$

The *average power* of a process, $\mathbf{a}$, is:

$$E\left[|\mathbf{a}_k|^2\right] = R_{kk}$$

In frequency terms, the *power spectrum* of a stationary process, $\mathbf{a}$, is the Fourier transform of its autocorrelation:

$$S_{aa}(\omega) = \int_{-\infty}^{\infty} R_{aa}(\tau)\ e^{-j\omega\tau}\ d\tau$$

The area of $S(\omega)$ is equal to the average power of $\mathbf{a}$:

$$E\left[|\mathbf{a}_k|^2\right] = R(0) = \frac{1}{2\pi}\int_{-\infty}^{\infty} S(\omega)\ d\omega$$

The *cross-power spectrum* of two jointly stationary processes, $\mathbf{a}$ and $\mathbf{b}$, is:

$$S_{ab}(\omega) = \int_{-\infty}^{\infty} R_{ab}(\tau)\ e^{-j\omega\tau}\ d\tau$$

And finally, for completeness [NOTE: compare and contrast cross-correlation and convolution; what doe the difference in sign reflect?], we include the definition of the convolution of two functions, **a** and **b**:

$$G(t) = \int_{-\infty}^{\infty} a(\tau)b(t-\tau)d\tau$$

C.1 What do you mean, "What do you mean?"?

Once again, beware of the crafty square! Power measures make sense when considering the energy dissipated in a resistor by the current flowing through it or the strength of a broadcast signal. In vision, it makes sense when talking about the initial transform of the amplitude of the electromagnetic wave to intensity (and very little else subsequent to that point). The square is either related to a metric whose computational necessity is increasingly in question or it is a gross hack wherein issues of phase are side-stepped.

D What is the Relation between Weiner-Hopf Filters and Optimal Linear Estimation?

The problem addressed here is the removal of additive Gaussian noise in a way that best reconstructs the original signal. The information that is available on the signal and the noise is statistical[12]. What is the best linear operation that can be applied to minimize the expected value of the squared difference between the original and obtained signal?

The take-home message is: If you know the power spectrum of the signal and of the noise, then you could attenuate the spectrum of the obtained signal by a function of the ratio of the signal to noise at each frequency. How do you show that this is the optimal reconstruction of the original signal that produced the input? Horn (*Robot Vision*, pp. 132-137) starts with a least squares error measure of the difference between the reconstructed signal (r) and the original signal (c):

$$E = \int_{-\infty}^{\infty} (r(x) - c(x))^2 \, dx$$

The input image is corrupted by additive Gaussian noise:

$$i(x) = c(x) + n(x)$$

The search for reconstruction functions is limited to linear functions that are convolved with the input:

[12] The treatment as presented here ignores the issues that actually make any such technique workable. How characterizations of the signal and the noise are obtained in practice is left as an exercise ...

$$r(x) = i(x) \otimes h(x)$$

By expanding the expression for E above:

$$E = \int_{-\infty}^{\infty} (r^2(x) - 2r(x)c(x) + c^2(x))dx$$

Expansion of r^2 into a product of convolutions yields (1^{st} term):

$$r^2(x) = \int_{-\infty}^{\infty} i(x-\alpha)h(\alpha)d\alpha \times \int_{-\infty}^{\infty} i(x-\beta)h(\beta)d\beta$$

Integrating r^2 (and an appropriate substitution of variables ($x \leftarrow (x-\beta)$) and an additional integration (over x), results in the two image (i) terms being converted into the autocorrelation of the image:

$$\int_{-\infty}^{\infty} r^2(x)dx = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \phi_{ii}(\alpha-\beta)h(\alpha)h(\beta)d\alpha d\beta$$

where $\phi_{ii}(x)$ is the autocorrelation of $i(x)$.

In addition (2^{nd} term):

$$r(x)c(x) = \int_{-\infty}^{\infty} i(x-\alpha)h(\alpha)c(x)d\alpha$$

In similar fashion as above:

$$\int_{-\infty}^{\infty} r(x)c(x)dx = \int_{-\infty}^{\infty} \phi_{ir}(\alpha)h(\alpha)d\alpha$$

where ϕ_{ir} is the cross-correlation of $i(x)$ and $c(x)$.

Finally (3^{rd} term),

$$\int_{-\infty}^{\infty} c^2(x)dx = \phi_{cc}(0)$$

where ϕ_{cc} is the autocorrelation of $c(x)$.

Having come to terms with the three components of the expansion of the least squares error metric for the difference between the obtained image and the original image, we can now write an expression for the error composed of the integration of the auto- and cross-correlation of $i(x)$ and $c(x)$:

$$E = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \phi_{ii}(\alpha-\beta)h(\alpha)h(\beta)d\alpha d\beta - 2\int_{-\infty}^{\infty} \phi_{ic}(\alpha)h(\alpha)d\alpha + \phi_{cc}(0)$$

The problem now is to find the transfer function $h(x)$ that minimizes this error. This feat is accomplished with the calculus of variations and a little magic. If we assume that the transfer function we have, $h(x)$, yields the minimum value of E and we let $\delta h(x)$ be an arbitrary function that modifies $h(x)$, then $h(x) + \epsilon\ \delta h(x)$ cannot be less than E, $\forall\ \delta h(x)$. Let the increased value of the error due to the perturbation of the transfer function be represented as $E + \delta E$, then if $h(x)$ truly yields a minimum:

$$\lim_{\epsilon \to 0} \frac{\partial}{\partial \epsilon}(E + \delta E) \ = \ 0 \qquad \forall\ \delta h(x)$$

Now replacing $h(x)$ with $h(x) + \epsilon\ \delta h(x)$, expanding the product of the two $h(x)$ terms and bringing the partial derivative wrt ϵ inside the integration, results in a two component expression that is composed solely of the cross terms (hence the 2 out front) of the $h(x)$ expansion in the first component and the simple derivative of the $h(x)$ expansion in the second:

$$\lim_{\epsilon \to 0} \frac{\partial}{\partial \epsilon}(E + \delta E) \ = \ 2 \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \phi_{ii}(\alpha - \beta) h(\alpha) \delta h(\beta) d\alpha d\beta \ - 2 \int_{-\infty}^{\infty} \phi_{ic}(\alpha) \delta h(\alpha) d\alpha$$

Then collecting terms by integration variable:

$$\lim_{\epsilon \to 0} \frac{\partial}{\partial \epsilon}(E + \delta E) \ = \ -2 \int_{-\infty}^{\infty} \left[\phi_{ic}(\alpha) \ - \ \int_{-\infty}^{\infty} \phi_{ii}(\alpha - \beta) h(\alpha) d\beta \right] \ \delta h(\alpha) d\alpha$$

Now noting that in order for the above expression on the right to be zero for any and all $\delta h(x)$, then (this is the magic [13]) the bracketed expression itself must be zero. Rearranging terms, this illumination results in:

$$\phi_{ic}(\alpha) \ = \ \int_{-\infty}^{\infty} \phi_{ii}(\alpha - \beta) h(\alpha) d\alpha$$

And, inserting a definition or two, we end up with an expression containing the transfer function we seek, $h(x)$, that will guarantee that our output signal is as close (in the least squares sense) as possible to the desired signal:

$$\phi_{ic} \ = \ \phi_{ii} \otimes h$$

A solution for $h(x)$ can be had by taking the Fourier transform of the above expression:

$$\Phi_{ic} \ = \ H \ \Phi_{ii}$$

where Φ_{ic} and Φ_{ii} are the power spectra.

Therefore, if one knows the power spectrum of the corrupted image as well as the power spectrum of the corrupted and original images, then one has all he needs to know to create the optimal reconstruction (given that one is restricted to linear operators and corruption by additive Gaussian noise[14]).

[13] This bit of legerdemain come from the fact that if the integral of the product of two continuous functions is zero over some interval and one of the functions is arbitrary, then the other function must be zero over the interval.

[14] It is also assumed here that in the reconstruction, an overshoot of two quantization levels is four times as bad as an undershoot of one quantization level.

The power spectrum of the corrupted and the original images is the Fourier transform of the cross-correlation of the the original image with itself plus the noise:

$$\Phi_{ic} = \Phi_{cc} + \Phi_{cn}$$

And, similarly, the power spectrum of the corrupted image is its autocorrelation:

$$\Phi_{ii} = \Phi_{cc} + \Phi_{cn} + \Phi_{nc} + \Phi_{nn}$$

By noting that the noise is uncorrelated with the image, we can set the $\Phi_{cn} + \Phi_{nc}$ to zero. Rearranging the above equation to solve for H yields:

$$H = \frac{\Phi_{ic}}{\Phi_{ii}} = \frac{\Phi_{cc}}{\Phi_{cc} + \Phi_{nn}} = \frac{1}{1 + \frac{\Phi_{nn}}{\Phi_{cc}}}$$

From the last equation we can see that the optimal (least-squares) filter forms the reconstruction by attenuating those frequencies in the image where the signal to noise ratio is small and retaining the amplitudes of those frequencies where the signal to noise ratio is high.

This annotated version of Horn's Weiner-Hopf derivation is helpful in that it not only demonstrates the relation between the transfer function of the system and the importance and meaning of the ratio of the covariance to variance matrices, it also demonstrates the optimality of the estimate given the least squares error metric.

[NOTE: Also check out Pratt's version of the proof.]

E Matlab code for constant velocity Kalman filter

```
% convergence is a Matlab program designed to test a constant velocity
% model Kalman filter. The two versions of Y (the measurement) are for
% the case where there is only sensor noise and the plant is perfect and
% the case where both the sensor and the plant are faulty. To get the
% "feel" of the system, one should vary the relative standard deviation
% of the various noise components (i.e., alter the divisor of any of the
% randn vectors).


% covariance of measurement noise
VV=cov(randn(100,2)/1);
```

```
% covariance of estimation noise
VW=cov(randn(100,4)/10);


% covariance of estimation error
VX=cov(randn(100,4));


% state estimate at k|k or k+1|k+1 (depending on context)
Xnew=zeros(4,100);

% state estimate at k+1|k
Xtrans=zeros(4,100);

% measurement vector
Y=zeros(2,100);

% error vector
E=zeros(4,100);

% variance of the state vector variables
Var=zeros(4,100);
 Var(:,1)   =   [   VX(1,1)
                    VX(2,2)
                    VX(3,3)
                    VX(4,4)   ];
% initial guess at state vector X_{0|0}
Xnew(:,1)=[0 0 1 1]';

% plant matrix
 A   =   [   1 0 1 0
             0 1 0 1
             0 0 1 0
             0 0 0 1   ];
% state/measurement matrix
 D   =   [   1 0 0 0
             0 1 0 0   ];
% iterate estimation for 100 steps
for j=2:100

        Xtrans(:,j)=A*Xnew(:,j-1);
```

```
        % Y(:,j)  =  [     j-1+randn/10
        %                  j-1+randn/10     ];
        Y(:,j)    =  [  Y(1,j-1)+1+randn/1
                        Y(2,j-1)+1+randn/1  ];
       VX=A*VX*A'+VW;
       E(:,j)=VX*D'*inv(D*VX*D'+VV)*(Y(:,j)-D*Xtrans(:,j));
       Xnew(:,j)=Xtrans(:,j)+E(:,j);
       VX=VX-(VX*D'*inv(D*VX*D'+VV))*D*VX;
       Var(:,j)  =  [  VX(1,1)
                       VX(2,2)
                       VX(3,3)
                       VX(4,4)  ];
end
% plot the individual X and Y measurements and the associated estimates
figure
plot(Y(1,:),'+')
hold on
set(gca,'xlabel',text(0,0,'time'));
set(gca,'ylabel',text(0,0,'position'));
plot(Xnew(1,:),'r+')
plot(Y(2,:))
plot(Xnew(2,:),'r')
axisL=axis;
text(15,0.95*axisL(4),'plus = X dimension');
text(15,0.90*axisL(4),'solid = Y dimension');
text(15,0.85*axisL(4),'yellow = measurement');
text(15,0.80*axisL(4),'red = estimate');
% plot the estimation variance of the state variables
figure
plot(Var(1,1:20),'+')
hold on
set(gca,'xlabel',text(0,0,'time'));
set(gca,'ylabel',text(0,0,'variance'));
```

```
plot(Var(1,1:20))
plot(Var(2,1:20))
plot(Var(3,1:20),'r+')
plot(Var(3,1:20),'r')
plot(Var(4,1:20),'r')
axisL=axis;
text(.65*axisL(2),0.95*axisL(4),'plus = X dimension');
text(.65*axisL(2),0.90*axisL(4),'solid = Y dimension');
text(.65*axisL(2),0.85*axisL(4),'yellow = position');
text(.65*axisL(2),0.80*axisL(4),'red = velocity');
```

% plot the error of the X and Y positions

```
figure
plot(E(1,1:100),'+')
hold on
set(gca,'xlabel',text(0,0,'time'));
set(gca,'ylabel',text(0,0,'error'));
plot(E(1,1:100))
plot(E(2,1:100),'r')
axisL=axis;
text(.65*axisL(2),0.90*axisL(4),'plus = X dimension');
text(.65*axisL(2),0.80*axisL(4),'solid = Y dimension');
```

E.1 Example Plots

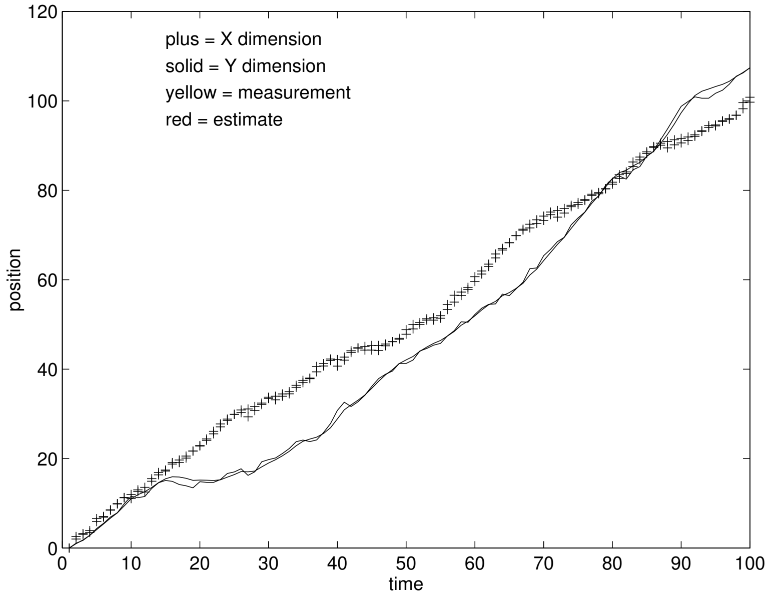

Figure 1: **Estimated and Measured X and Y Positions.** In this example both the plant and measurement processes are corrupted by Gaussian noise.

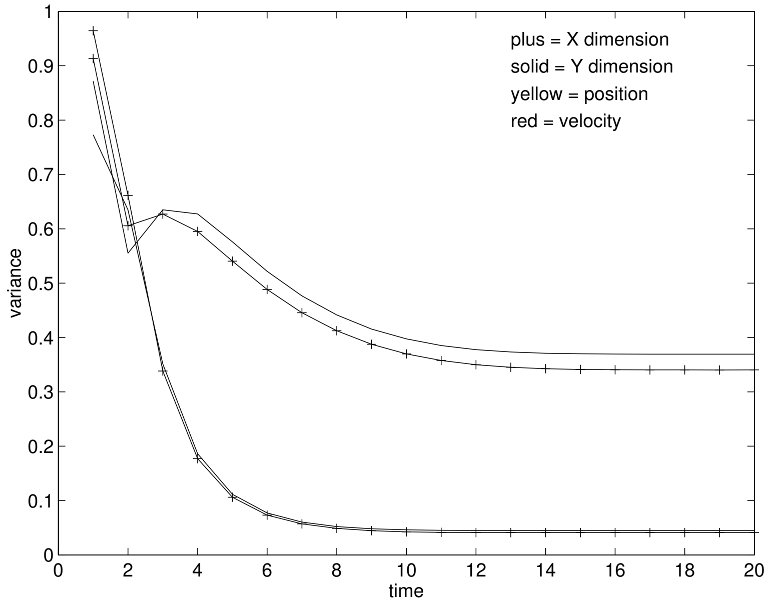

Figure 2: **Time Course of X and Y Position and Velocity Estimation Variance.** Note the asymptotic values and compare them to the variance of the given noise distributions. Also note the rapidity of convergence in this case. Alter the initial noise distributions ands attempt to modify the convergence rate.

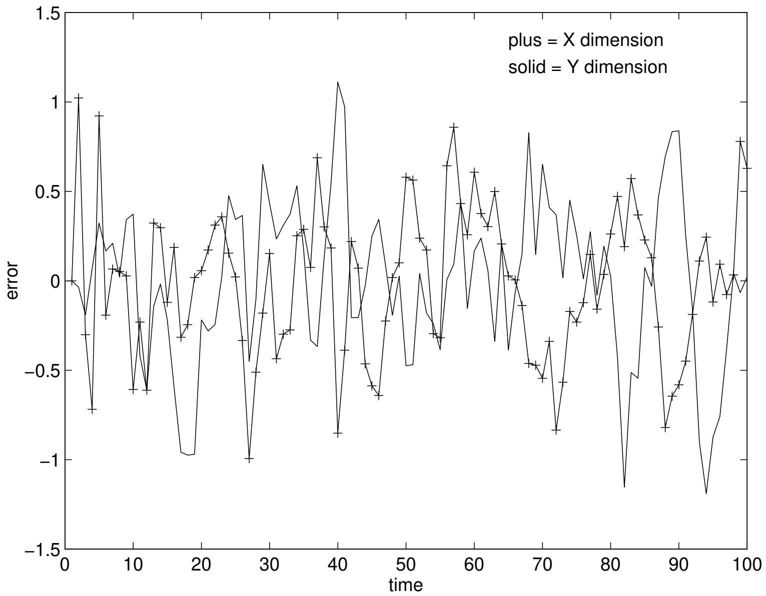

Figure 3: **Error in Estimated X and Y Position.** The range of position errors in this example is appreciable due to the fact that the injected noise ($\sigma = 1$) is on the order of the programmed change; however, the standard deviation of the error of the estimate is less than half that.