EL2320 - Gaussian Estimation

John Folkesson

Gaussian Estimation and Kalman Filter
(Chap 3 in Thrun)

The Gaussian Distribution

 Carl Friedrich Gauss invented the normal distribution in 1809 to help explain the method of least squares.
The scalar X is a Gaussian or normal variable if its pdf is of the form:

$$p(x) = G(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

(Laplace figured out the $\frac{1}{\sqrt{2\pi\sigma^2}}$)

where the mean of X is
$\mu = E[x]$

and the variance of X is $\sigma^2 = E[(x-\mu)^2]$

We say X is $N(\mu, \sigma^2)$ or $x \sim N(\mu, \sigma^2)$

Gaussian in many dimensions

The vector $\mathbf{X}$ is a Gaussian or normal variable if its pdf is of the form

$$p(\mathbf{x}) = G(\mathbf{x}, \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n det(\Sigma)}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1}(\mathbf{x}-\mu)}$$

where the mean of $\mathbf{X}$ is
$\mu = E[\mathbf{x}]$

and the covariance of $\mathbf{X}$ is $\Sigma = E[(\mathbf{x} - \mu)(\mathbf{x} - \mu)^T]$

We say $\mathbf{X}$ is $N(\mu, \Sigma)$

If we chose a basis for $\mathbf{X}$ where Σ is diagonal then the pdf becomes a product of n scalar Gaussians.

If we change variables to $\mathbf{y} = \Sigma^{-\frac{1}{2}}(\mathbf{x} - \mu)$ then $\mathbf{y} \sim N(0, I)$

Gaussians

x and **y** are Gaussian vector variables and A, B and C are matricies.

$$p(\mathbf{x},\mathbf{y}) = G\left(\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}, \begin{pmatrix}\mu_x\\ \mu_y\end{pmatrix}, \begin{pmatrix}A & C\\ C^T & B\end{pmatrix}\right)$$

$$p(\mathbf{y}) = \int_{-\infty}^{\infty} p(\mathbf{x},\mathbf{y})d\mathbf{x} = G(\mathbf{y},\mu_y,B)$$

$$p(\mathbf{x}|\mathbf{y}) = p(\mathbf{x},\mathbf{y})/p(\mathbf{y}) = G(\mathbf{x},\mu_x + CB^{-1}(\mathbf{y}-\mu_y), A - CB^{-1}C^T)$$

Proofs involve: Matrix algebra and 'completing the square'

$\alpha x^2 + \beta xy + \gamma = \alpha(x + \frac{\beta}{2\alpha}y)^2 + \gamma - (\frac{\beta}{2\alpha}y)^2$

Changing varibles, $\mathbf{x} \sim N(\mu,\Sigma)$ and $\mathbf{u} = A\mathbf{x} + \mathbf{b}$ then

$\mathbf{u} \sim N(A\mu + \mathbf{b}, A\Sigma A^T)$ and

$\int_{-\infty}^{\infty} G(\mathbf{u},\mu_u,S)d\mathbf{u} = 1.$

Gaussian Multiplication

We can multiply Gaussians together.

$$N(\mu_a, A)N(\mu_b, B) = z_c N(\mu_c, C)$$

$$C = (A^{-1} + B^{-1})^{-1} = A(A + B)^{-1}B = B(A + B)^{-1}A$$

$$C = A - A(A + B)^{-1}A = B - B(A + B)^{-1}B \text{ (Woodbury)}$$

$$\mu_c = CA^{-1}\mu_a + CB^{-1}\mu_b = B(A + B)^{-1}\mu_a + A(A + B)^{-1}\mu_b$$

When might we do this?

Gaussian Multiplication

For independent Gaussian measurements z of quantity x:
$p(z_2, z_1|x) = p(z_2|x)p(z_1|x) = G(x, z_2, \sigma_2^2)G(x, z_1, \sigma_1^2)$

But also in our Bayesian iterative inference formula:

$$p(x|z_2, z_1) = \frac{p(z_2,|x,z_1)p(x|z_1)}{p(z_2|z_1)} \propto p(z_2|x)p(x|z_1) = p(z_2|x)p(z_1|x)$$

We use the prior $p(x) = constant \Rightarrow p(x|z_1) = p(z_1|x)$.
We can ignore the denominator $p(z_2|z_1)$ because:

$$\int_{-\infty}^{\infty} p(x|z_2, z_1)dx = 1 => p(x|z_2, z_1) = G(x, \mu, \sigma^2)$$

$$\sigma^2 = (\sigma_1^{-2} + \sigma_2^{-2})^{-1} \text{ and } \mu = \sigma^2(\sigma_1^{-2}\mu_1 + \sigma_2^{-2}\mu_2)$$

Its a weighted average: $w_1 = \kappa/\sigma_1^2$ and $w_2 = \kappa/\sigma_2^2$ and κ turns out to make $w_1 + w_2 = 1$.

Gaussian Multiplication

You have two thermometers with measurement models for τ_i:

$\tau_i \sim N(T, \sigma_i^2)$

where T is the 'true' temperature and i is 1 or 2.

(That is it has a pdf that goes like $e^{-\frac{(\tau_i - T)^2}{2\sigma_i^2}}$) $\sigma_1^2 = 1$; and $\sigma_2^2 = 2$;

What is $bel(T)$: $p(T|\tau_1 = 20, \tau_2 = 19) =$?

What is the 'Maximum Likelihood Estimate' MLE for T?

Kalman Filter Update Preview

If we have independent Gaussian measurements z_i with variances $\sigma_i^2 = r_i$ our iterative formula

$$p(x|z_1, \ldots, z_n) \propto p(z_n|x)p(x|z_1, \ldots, z_{n-1})$$

leads to a series of Gaussians:

$$p(x|z_1, \ldots, z_n) = G(x, \mu_n, s_n)$$

$$s_n = (r_n^{-1} + s_{n-1}^{-1})^{-1} = s_{n-1}(r_n + s_{n-1})^{-1} r_n;$$

$$\mu_n = s_n(r_n^{-1} z_n + s_{n-1}^{-1}\mu_{n-1})$$
$$\mu_n = s_{n-1}(r_n + s_{n-1})^{-1} z_n + r_n(r_n + s_{n-1})^{-1}\mu_{n-1}$$

Just for fun define: $k_n = s_{n-1}(r_n + s_{n-1})^{-1}$;

$$s_n = s_{n-1} - k_n s_{n-1} \text{ and } \mu_n = \mu_{n-1} + k_n(z_n - \mu_{n-1})$$

$(z_n - \mu_{n-1})$ is called the 'innovation' and k_n the 'Kalman gain'.

Kalman Filter Update Preview

$k_n = s_{n-1}(s_{n-1} + r_n)^{-1} =>$

$s_n = s_{n-1} - k_n s_{n-1}$ and $\mu_n = \mu_{n-1} + k_n(z_n - \mu_{n-1})$

An example of using this might be anytime one measures the same quantity many times with differing accuracy.

So measuring the mass of the electron, diameter of the earth, height of a building,...

Soon we will see how different sorts of measurements, eg. compass, gps, inertial sensors,.. can be combined in 'data fusion'.

Information Filter Update Preview

$s_n = (r_n^{-1} + s_{n-1}^{-1})^{-1}$ and $\mu_n = s_n(r_n^{-1} z_n + s_{n-1}^{-1}\mu_{n-1})$

Just for more fun define:
$\omega_n = s_n^{-1}$,
$\zeta_n = \omega_n \mu_n$,

Then:
$\omega_n = \omega_{n-1} + r_n^{-1}$
$\zeta_n = \zeta_{n-1} + r_n^{-1} z_n$

ω_n is called the information 'matrix' (for now 1D).
ζ_n is called the information 'vector' (for now 1D).

The Observer

Up to now we have assumed we have a simple scalar as our state and we measure it directly.

We can generalize this to multidimensions and indirect measurements by introducing a vector state $\mathbf{x}$, a vector measurement $\mathbf{z}$, and an observer:

$$\hat{\mathbf{z}} = \mathbf{h}(\mathbf{x})$$

Think of $\hat{\mathbf{z}}$ as the expected measurement given the state $\mathbf{x}$. We then form an innovation:

$$\eta = \mathbf{z} - \hat{\mathbf{z}}$$

Think of η as the noise in the actual measurement. We will typically model the distribution $p(\eta|x)$ and try to infer an a posteriori distribution $p(x|\eta)$

The Observer

Often we assume a linear observer:

$\hat{\mathbf{z}} = \mathbf{h}(\mathbf{x}) = \bar{\mathbf{z}} + C\mathbf{x}$

The innovation is now linear: $\eta = \mathbf{z} - \bar{\mathbf{z}} - C\mathbf{x}$

Connect to linear filters:
$\mathbf{y} = \mathbf{z} - \bar{\mathbf{z}} = C\mathbf{x} + \eta$

The Observer

If we have Gaussian measurements then

$$p(\mathbf{z}|\mathbf{x}) = G(\mathbf{z}, \bar{\mathbf{z}} + C\mathbf{x}, R) = G(\eta, 0, R).$$

This can lead to some problems in that there may not be enough dimensions in $\mathbf{z}$.
The state may not be observable.
We can still find the state if we make multiple independent and different measurements $\mathbf{z}_n$ with cooresponding C_i and R_i. We need:

$$\sum C_i^T R_i^{-1} C_i$$

to be full rank. Lets see why.

The Observer

$$p(\mathbf{x}|\{\mathbf{z}_i\}) \propto p(\{\mathbf{z}_i\}|\mathbf{x}) = \prod_i \frac{1}{\sqrt{(2\pi)^n det R_i}} e^{-\frac{1}{2}\eta_i{}^T R_i^{-1} \eta_i}$$

$$\eta_i = \mathbf{z}_i - \bar{\mathbf{z}}_\mathbf{i} - C_i\mathbf{x}$$

Any multipling part that does not depend on x can be ignored.
Why?
We have a product of exponentials. So we can sum the exponents.
In the exponent we have -1/2 times a sum of terms:

$$[\mathbf{x}^T C_i^T R_i^{-1} C_i \mathbf{x} - [x^T C_i^T R_i^{-1}(\mathbf{z}_i - \bar{\mathbf{z}}_\mathbf{i}) + transpose] + \ldots] =$$

$$(\mathbf{x} - \mu_n)^T \Sigma_n^{-1} (\mathbf{x} - \mu_n) + \ldots \text{ (x independent stuff)}$$

$$\Sigma_n^{-1} = \sum_{i=1}^{n} C_i^T R_i^{-1} C_i = \Sigma_{n-1}^{-1} + C_n^T R_n^{-1} C_n$$

The linear in x terms define the $\mu_\mathbf{n}$ while the quadratic terms limit the uncertainty in x.

The Observer

$$(\mathbf{x} - \mu_n)^T \Sigma_n^{-1} (\mathbf{x} - \mu_n) + \ldots \text{ (x independent stuff)}$$

$$\Sigma_n^{-1} = \sum_{i=1}^{n} C_i^T R_i^{-1} C_i = \Sigma_{n-1}^{-1} + C_n^T R_n^{-1} C_n$$

We define can $\Omega_n = \Sigma_n^{-1}$the information matrix:

$$\Omega_n = \Omega_{n-1} + C_n^T R_n^{-1} C_n$$

This is the thing I said needs to be full rank for the system to be observable. That is for us to compute μ_n from the $\mathbf{z}_i$. Still need to show how.

The Observer

The MLE is where the exponent has a minimum (or zero gradient wrt x). This is also when $\Omega_n(\mathbf{x} - \mu_n) = 0$. Doing some math (starting from previous page) we can see:

$$[\Omega_n \mathbf{x} - \sum_{i=1}^{n} C_i^T R_i^{-1}(\mathbf{z}_i - \bar{\mathbf{z}}_\mathbf{i})] \mid_{\mathbf{x}=\mu_n} = 0$$

$$\Omega_n \mu_n \equiv \zeta_n = \zeta_{n-1} + C_n^T R_n^{-1}(\mathbf{z}_n - \bar{\mathbf{z}}_\mathbf{n}) \quad \rightarrow$$
$$\mu_\mathbf{n} = \mu_{\mathbf{n-1}} + \Omega_n^{-1} C_n^T R_n^{-1}(\mathbf{z}_n - \bar{\mathbf{z}}_\mathbf{n} - C_n \mu_{\mathbf{n-1}})$$

So we can solve for the mean/MLE of the state if the information matrix can be inverted.
In practical systems we often initailize by creating a first measurement with $C_1 = I$ and a large but finite R_1. This is our a priori state distribution. Everthing then looks the same but we call our estimates MAP estimates.

We have nearly an information filter which is basically the same as a Kalman filter. We still need to add a processes, i.e. dynamics.

The Markov Property

A process is a first order Markov process if, given that we know the present, the past has no influence on the future, i.e. if

for times
$t_0 < t_1 < t_2 < ... < t_n$

and cooresponding states of our process
$\{x_0, x_1, x_2, ..., x_n\}$

we have that

$$p(x_n|x_1, x_2, ..., x_{n-1}) = p(x_n|x_{n-1})$$

We typical call it Markov and skip the first order

Back to Gaussians: Growing States

Imagine that our state vector at time t grows.

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_{t-1} \\ \mathbf{x}_t \end{pmatrix}$$

We then add a number of independent measurements z_i as before only now our observers depend on the augmented state and its C_i matricies have more columns.
As long as the new state rows are 'observed' in the new C we can maintain observability in the system. Our iterative 'update' formula still works:

$$\Omega_n = \Omega_{n-1} + C_n^T R_n^{-1} C_n$$

$$\Omega_n \mathbf{x} = \zeta_n = \zeta_{n-1} + C_i^T R_i^{-1} (\mathbf{z}_i - \bar{\mathbf{z}}_\mathbf{i})$$

where we augment Ω_{n-1} by adding 0's to for the new state rows/cols. If we are working from the Covaraince Matrix then we must invert, then augment, then update, then invert back again.

Gaussians: Shrinking State

Now image we no longer care about some of the rows of out state vector:

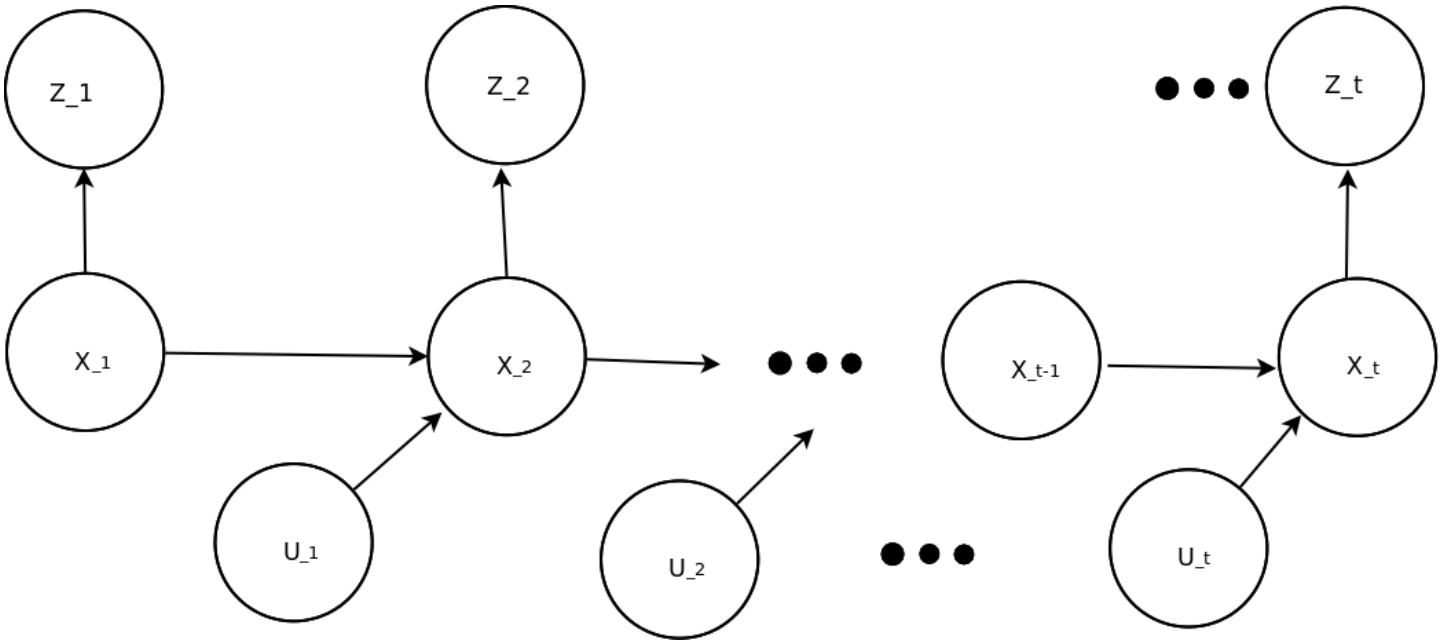
$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_{t-1} \\ \mathbf{x}_t \end{pmatrix} \Rightarrow (\mathbf{x}_t)$$

This is accomplished by marginalization.

Remember marginalizing part of a Gaussian system means simply restricting the covariance to the remaining rows and columns. The mean does not change.

When working with the Information Matrix we must first invert, then restrict, then invert the smaller matrix.

Bayes Filter - Recursive State Estimation



Predict: $p(x_t|z_1, \ldots, z_{t-1}, u_1, \ldots, u_t) =$

$$\int p(x_t|x_{t-1}, u_t)p(x_{t-1}|z_1, \ldots, z_{t-1}, , u_1, \ldots, u_{t-1})dx_{t-1}$$

Update: $p(x_t|z_1, \ldots, z_n) \propto p(z_n|x)p(x_t|z_1, \ldots, z_{n-1})$

Gaussian: Prediction

Now we introduce a concept of control measurement or control data.
We want to do a sequence of grow and shrink on our state where we first augument the state at time t-1 by adding the state at time t. Then we shrink it by removing the state at time t-1.
We know that we can shrink our system if we know the covaraince.
We assume that we have seperated out a set of measurement data as a seperate class which we call 'control data', $\mathbf{u}_t$.
Control data is assumed known with any noise or uncertainty absorbed into the process noise.
Instead of our observer we use a 'predictor'.
$\mathbf{x}_t = g(\mathbf{u}_t, \mathbf{x}_{t-1}) + \varepsilon_t$
where ε_t is 'process noise and disturbances'

Prediction

For Linear systems the predictor is typically written as:

$\hat{\mathbf{x}}_t = A\mathbf{x}_{t-1} + B\mathbf{u}_t$

With the noise and disturbances modeled in terms of:

$\varepsilon_t = (\mathbf{x}_t - A\mathbf{x}_{t-1} - B\mathbf{u}_t)$

For Gaussian processes this is assumed $N(0, Q_t)$

$$p(\mathbf{x}_t, \mathbf{x}_{t-1}|\{\mathbf{u}_j\}, \{\mathbf{z}_i\}) = p(\mathbf{x}_t|\mathbf{x}_{t-1}, \{\mathbf{u}_j\}, \{\mathbf{z}_i\})p(\mathbf{x}_{t-1}|\{\mathbf{u}_j\}, \{\mathbf{z}_i\})$$
$$= G(\varepsilon_t, 0, Q_t)G(\mathbf{x}_{t-1}, \mu_{\mathbf{t-1}}, \Sigma_{t-1})$$

By collecting the quadratic terms, the Information Matrix after the update becomes (hats are to distinguish from the shrunk system in the next step):

$$\hat{\Omega}_t = \hat{\Sigma}_t^{-1} = \begin{pmatrix} \Sigma_{t-1}^{-1} + A^T Q^{-1} A & -A^T Q^{-1} \\ -Q^{-1} A & Q^{-1} \end{pmatrix}$$

and the new mean its inverse is:

$$\hat{\mu}_t = \begin{pmatrix} \mu_{t-1} \\ A\mu_{t-1} + B\mathbf{u}_t \end{pmatrix} \text{ and } \hat{\Sigma}_t = \begin{pmatrix} \Sigma_{t-1} & \Sigma_{t-1}A^T \\ A\Sigma_{t-1} & Q + A\Sigma_{t-1}A^T \end{pmatrix}$$

Prediction

$$\hat{\mu}_t = \begin{pmatrix} \mu_{t-1} \\ A\mu_{t-1} + B\mathbf{u}_t \end{pmatrix} \text{ and } \hat{\Sigma}_t = \begin{pmatrix} \Sigma_{t-1} & \Sigma_{t-1}A^T \\ A\Sigma_{t-1} & Q + A\Sigma_{t-1}A^T \end{pmatrix}$$

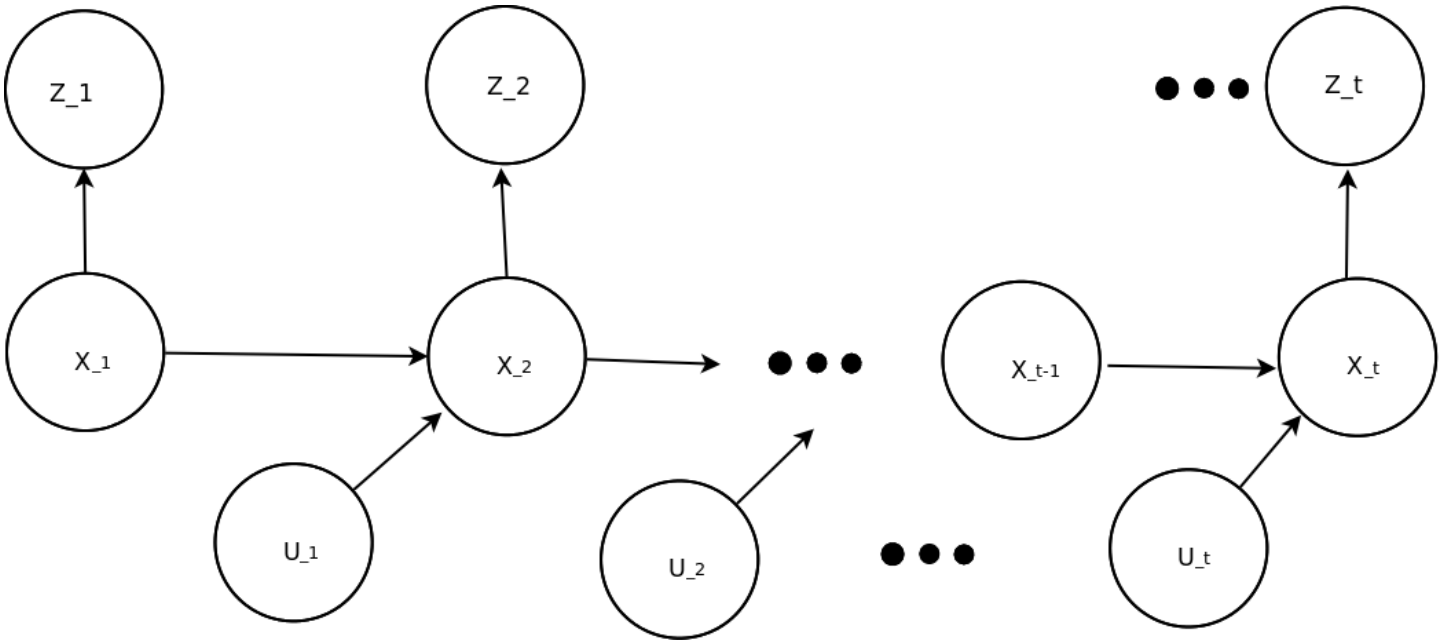
Finally shrinking

$\mu_t = A\mu_{t-1} + B\mathbf{u}_t$ and

$\Sigma_t = Q + A\Sigma_{t-1}A^T$

$\Omega_t = \Sigma_t^{-1}$

Bayes Filter - Recursive State Estimation



Predict: $p(x_t|z_1, \ldots, z_{t-1}, u_1, \ldots, u_t) =$

$$\int p(x_t|x_{t-1}, u_t)p(x_{t-1}|z_1, \ldots, z_{t-1}, , u_1, \ldots, u_{t-1})dx_{t-1}$$

Update: $p(x_t|z_1, \ldots, z_n) \propto p(z_n|x)p(x_t|z_1, \ldots, z_{n-1})$

Kalman Filter - Prediction and Observation

To allow both prediction and observation at times t we introduce a bar to denote the values after prediction. (warning the book switches R and Q) So for the Kalman Filter we define:
$K_t = \bar{\Sigma}_t C_t^T (C_t \bar{\Sigma}_t C_t^T + R_t)^{-1}$ and $\mathbf{y}_t = \mathbf{z}_t - \bar{\mathbf{z}}_t - C_t \bar{\mu}_t$
Predict phase:
$\bar{\Sigma}_t = Q_t + A_t \Sigma_{t-1} A_t^T$ uncertainty is often growing
$\bar{\mu}_\mathbf{t} = A_t \mu_{t-1} + B_t \mathbf{u}_t$
and update:
$\Sigma_t = \bar{\Sigma}_t - K_t C_t \bar{\Sigma}_t$ Use the Woodbury matrix identity
$\mu_t = \bar{\mu}_\mathbf{t} + K_t \mathbf{y}_t$

Information filter Prediction:
$\bar{\Omega}_t = (Q_t + A_t \Omega_{t-1}^{-1} A_t^T)^{-1}$
$\bar{\zeta}_t = \bar{\Omega}_t [A_t \Omega_{t-1}^{-1} \zeta_{\mathbf{t-1}} + B_t \mathbf{u}_t]$
and Update:
$\Omega_t = \bar{\Omega}_t + C_t^T R_t^{-1} C_t$ information is growing
$\zeta_\mathbf{t} = \bar{\zeta}_\mathbf{t} + C_t^T R_t^{-1} \mathbf{y}_t.$

Kalman Filter - Assumptions and Properties

Assumption, the system is linear and discribed by:

$\mathbf{x}_t = A_t\mathbf{x}_{t-1} + B_t\mathbf{u}_t + \varepsilon_t$; $\quad\quad$ ε_t is $N(0, Q_t)$.

$\mathbf{y_t} = C_t\mathbf{x}_t + \delta_t$; $\quad\quad$ δ_t is $N(0, R_t)$.

This implies then that:

- the a posteriori distribution is Gaussian. Why?
- the system is Markov. Why?
- the Kalman Filter is optimal in the sense of having the minimum expected variance from the state estimate. Why?

Fact: Kalman Filter is the optimal linear filter even if the noise is not Gaussian, so long as it is zero mean and independent.

Kalman Filter - Other reading as needed

For those of you that are still not completely clear on the Kalamn filter.
There are a clear presentations of the Kalman Filter in:

An Introduction to the Kalman Filter, Welch et al. (2006),

and the classic reference on Kalman Filters:

Stochastic models, estimation, and control, Maybeck 1979.

Kalman Filter - Hello World

We have a house with two rooms, one thermometer, and a heater in each room.

$$\text{State: } \mathbf{x} = \begin{pmatrix} temperature - In - Room - 1 \\ temperature - In - Room - 2 \\ outside - Temperature \end{pmatrix}$$

- The rate of change of the temperature in each room is proportional to the rate of energy (watts) entering or leaving the room.
- The rate of energy passing from room 1 to room 2 is proportional to the temperature difference. And similarly for the outside walls.
- The identical heaters produce heating watts proportional to some control signal plus Gaussian noise.
- The thermometer is in room 1 and is subject to Gaussian noise.

Kalman Filter - Hello World

$$\mathbf{x}_t = \begin{pmatrix} 1-a-b & a & b \\ d & 1-d-c & c \\ 0 & 0 & 1 \end{pmatrix} \mathbf{x}_{t-1} + \begin{pmatrix} g & 0 \\ 0 & g \\ 0 & 0 \end{pmatrix} \mathbf{u}_t + \varepsilon_t$$

$$\mathbf{y}_t = x_t^1 + \delta_t \qquad 1 >> a, b, c, d, g > 0$$

What are the physical meaning of a,b,c,d and g?
How would you check if this system observable?
What if the thermometer was outside?
What happens if I enter the house by opening and closing the door?
What would change if we added a second identical house but no more themometers?

A bonus questions not part of this course:
How would you check controllablity?