ME 7247: Advanced Control Systems **Fall 2022–23**

Lecture 10: Joint distribution and estimation

Tuesday October 11, 2022

Lecturer: Laurent Lessard *Scribe: Tony Smoragiewicz*

In lecture today, we covered how the expected value of an estimate is affected when we have joint distributions. In particular, we learned how errors in our measurement are reflected and modeled in our estimate. Of particular importance were the different formulations of the expectation and covariance of a conditional distribution. Eqs. (1) and (2) show the different formulations.

We have two ways to write the conditional expectation $\mathbf{E}(x \mid y)$:

$$\begin{aligned}\mathbf{E}(x \mid y) &= \mu_x + \Sigma_x A^\mathsf{T}(A\Sigma_x A^\mathsf{T} + \Sigma_w)^{-1}(y - A\mu_x) \\ &= \mu_x + (A^\mathsf{T}\Sigma_w^{-1}A + \Sigma_x^{-1})^{-1}A^\mathsf{T}\Sigma_w^{-1}(y - A\mu_x)\end{aligned} \tag{1}$$

We have two ways to write the conditional covariance $\mathbf{Cov}(x \mid y)$:

$$\begin{aligned}\mathbf{Cov}(x \mid y) &= \Sigma_x - \Sigma_x A^\mathsf{T}(A\Sigma_x A^\mathsf{T} + \Sigma_w)^{-1}A\Sigma_x \\ &= (\Sigma_x^{-1} + A^\mathsf{T}\Sigma_w^{-1}A)^{-1}\end{aligned} \tag{2}$$

The first version is called the *covariance* formula, and the second is the *information* formula. The covariance is a proxy for the error. So smaller covariance means less error. The inverse of the covariance (the *information*) is the opposite; more information means less error. When we add a new measurement, the covariance decreases and the information increases.

1 Joint Distributions

Recall from last time our picture of two random variables with positive correlation. Correlated variables means that the off-diagonal values of the covariance matrix are non-zero. If the values were uncorrelated, the ellipsoid would change to a circle. Fig. 1 below shows the two positively correlated random variables, x_1 and x_2.

If x and y are jointly Gaussian, we can write

$$\begin{bmatrix} x \\ y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_y \end{bmatrix}\right). \tag{3}$$

Prior to measuring y, the probability distribution (the *prior*) is the marginal distribution for x:

$$x \sim \mathcal{N}(\mu_x, \Sigma_x).$$

Once we measure y, the distribution of x conditioned on this measurement is the *conditional distribution*, also known as the *posterior distribution*, given by

$$(x \mid y) \sim \mathcal{N}\left(\mu_x + \Sigma_{xy}\Sigma_y^{-1}(y - \mu_y), \Sigma_x - \Sigma_{xy}\Sigma_y^{-1}\Sigma_{yx}\right). \tag{4}$$

Through this improved estimate, the mean of x has shifted from μ_x to $\mu_x + \Sigma_{xy}\Sigma_y^{-1}(y - \mu_y)$ and the variance has decreased from Σ_x to $\Sigma_x - \Sigma_{xy}\Sigma_y^{-1}\Sigma_{yx}$.

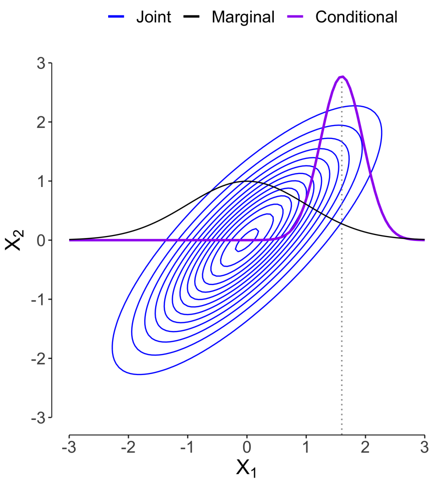

Figure 1: Joint, Marginal, and Conditional Distributions

Problem: Assume we have a prior on x and a linear measurement model for y that includes additive Gaussian noise.

$$y = Ax + w, \quad \text{where}$$
$$x \sim \mathcal{N}(\mu_x, \Sigma_x) \quad \text{and} \quad w \sim \mathcal{N}(0, \Sigma_w)$$

What is the best estimate of x given y? Here, *best* is something we need to define. Just like there are many ways to make a vector *small* (we discussed different notions of norm), there are many ways to make the estimation error small.

Side Note: It's possible to have noise with non-zero mean but in this case it is usually modeled as part of the system itself. Similarly, if the noise has fixed but unknown bias (non-zero mean), we can augment the state x to include the unknown mean μ_w:

$$y = \begin{bmatrix} A & I \end{bmatrix} \begin{bmatrix} x \\ \mu_w \end{bmatrix} + w$$

2 Mean Squared Error (MSE)

If $\hat{x}$ is our estimate of x, the *mean squared error* (MSE) is defined as

$$\text{MSE} = \mathbf{E}\big(\|x - \hat{x}\|^2 \mid y\big). \tag{5}$$

This is an intuitive way to characterize the magnitude of the error in our estimate. We can expand Eq. (5) as follows:

$$\begin{aligned} \text{MSE} &= \mathbf{E}\big((x - \hat{x})^\mathsf{T}(x - \hat{x}) \mid y\big) \\ &= \mathbf{E}\big(\|x\|^2 - 2\hat{x}^\mathsf{T}x + \|\hat{x}\|^2 \mid y\big) \\ &= \mathbf{E}\big(\|x\|^2 \mid y\big) - 2\hat{x}^\mathsf{T}\mathbf{E}(x \mid y) + \|\hat{x}\|^2. \end{aligned} \tag{6}$$

We seek to find an $\hat{x}$ that minimizes our mean squared error. This is called the *minimum mean squared error* (MMSE) estimate. We can find this by taking the gradient of Eq. (6) with respect to $\hat{x}$ and setting it equal to zero.

$$\text{MMSE} \quad = \quad \underset{\hat{x}}{\text{minimize}} \quad \mathbf{E}(\|x\|^2 \mid y) - 2\hat{x}^\mathsf{T}\mathbf{E}(x \mid y) + \|\hat{x}\|^2$$

We have $\nabla_{\hat{x}} = -2\mathbf{E}(x \mid y) + 2\hat{x}$, which leads to $\hat{x} = \mathbf{E}(x \mid y)$. So the MMSE estimator is precisely the conditional expectation.

If we substitute this value in, we find that the MMSE is:

$$\begin{aligned}
\text{MMSE} &= \mathbf{E}\left((x - \mathbf{E}\left(x \mid y\right))^\mathsf{T}\left(x - \mathbf{E}\left(x \mid y\right)\right) \,\middle|\, y\right) \\
&= \mathbf{E}\left(\mathbf{tr}\left((x - \mathbf{E}\left(x \mid y\right))\left(x - \mathbf{E}\left(x \mid y\right)\right)^\mathsf{T}\right) \,\middle|\, y\right) \\
&= \mathbf{tr}\left(\mathbf{E}\left((x - \mathbf{E}\left(x \mid y\right))\left(x - \mathbf{E}\left(x \mid y\right)\right)^\mathsf{T} \,\middle|\, y\right)\right) \\
&= \mathbf{tr}\left(\mathbf{Cov}\left(x \mid y\right)\right)
\end{aligned}$$

So the minimum mean squared error is achieved by using the conditional mean as the estimate, and the associated MSE is the trace of the conditional covariance.

We can now return to our original problem of minimizing the MSE, and instead just seek the conditional distribution of x given y. Recall from last time:

$$\text{if} \quad x \sim \mathcal{N}(\mu_x, \Sigma_x) \quad \text{then} \quad Ax + b \sim \mathcal{N}(A\mu_x + b, A\Sigma_x A^\mathsf{T}).$$

Given our measurement and noise model, we have

$$y = Ax + w \qquad \text{and} \qquad \begin{bmatrix} x \\ w \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_x \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma_x & 0 \\ 0 & \Sigma_w \end{bmatrix}\right)$$

To find the joint distribution of (x, y), we can write (x, y) as a linear transformation of (x, w).

$$\begin{aligned}
\begin{bmatrix} x \\ y \end{bmatrix} &= \begin{bmatrix} I & 0 \\ A & I \end{bmatrix} \begin{bmatrix} x \\ w \end{bmatrix} \\
\implies \begin{bmatrix} x \\ y \end{bmatrix} &\sim \mathcal{N}\left(\begin{bmatrix} I & 0 \\ A & I \end{bmatrix} \begin{bmatrix} \mu_x \\ 0 \end{bmatrix}, \begin{bmatrix} I & 0 \\ A & I \end{bmatrix} \begin{bmatrix} \Sigma_x & 0 \\ 0 & \Sigma_w \end{bmatrix} \begin{bmatrix} I & 0 \\ A & I \end{bmatrix}^\mathsf{T}\right) \\
&= \mathcal{N}\left(\begin{bmatrix} \mu_x \\ A\mu_x \end{bmatrix}, \begin{bmatrix} \Sigma_x & \Sigma_x A^\mathsf{T} \\ A\Sigma_x & A\Sigma_x A^\mathsf{T} + \Sigma_w \end{bmatrix}\right)
\end{aligned}$$

Applying Eq. (4), we can find the conditional distribution:

$$(x \mid y) \sim \mathcal{N}\Big(\underbrace{\mu_x + \Sigma_x A^\mathsf{T}(A\Sigma_x A^\mathsf{T} + \Sigma_w)^{-1}(y - A\mu_x)}_{\mathbf{E}(x|y)},\ \underbrace{\Sigma_x - \Sigma_x A^\mathsf{T}(A\Sigma_x A^\mathsf{T} + \Sigma_w)^{-1}A\Sigma_x}_{\mathbf{Cov}(x|y)}\Big)$$

Recall: Matrix Inversion Lemma (MIL)

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

We choose a mapping $(A, B, C, D) \to (\Sigma_x^{-1}, -A^\mathsf{T}, A, \Sigma_w)$ so that we can use the MIL to rewrite the conditional covariance in a different way.

$$\begin{aligned}\mathbf{Cov}(x \mid y) &= \Sigma_x - \Sigma_x A^\mathsf{T}(A\Sigma_x A^\mathsf{T} + \Sigma_w)^{-1}A\Sigma_x \\ &= (\Sigma_x^{-1} + A^\mathsf{T}\Sigma_w^{-1}A)^{-1}\end{aligned}$$

Another way of writing this is:

$$\mathbf{Cov}(x \mid y)^{-1} = \Sigma_x^{-1} + A^\mathsf{T}\Sigma_w^{-1}A$$

The covariance is a proxy for error. Smaller covariance means smaller error. In the original covariance formula, we saw that covariance decreased (in the semidefinite sense) when we observe a measurement.

The inverse of the covariance is called the *information*. In this new formula, we see that observing a measurement increases our information (again in the semidefinite sense).

Recall: Push through identity

$$A(BA + I)^{-1} = (AB + I)^{-1}A$$

We will use the push through identity to simplify the conditional mean $\mathbf{E}(x \mid y)$

$$\mathbf{E}(x \mid y) = \mu_x + \underbrace{\Sigma_x A^\mathsf{T}(A\Sigma_x A^\mathsf{T} + \Sigma_w)^{-1}}_{\text{simplified below}}(y - A\mu_x)$$

$$\begin{aligned}\Sigma_x A^\mathsf{T}(A\Sigma_x A^\mathsf{T} + \Sigma_w)^{-1} &= \Sigma_x A^\mathsf{T}[\Sigma_w(\Sigma_w^{-1}A\Sigma_x A^\mathsf{T} + I)]^{-1} \\ &= \Sigma_x A^\mathsf{T}(\Sigma_w^{-1}A\Sigma_x A^\mathsf{T} + I)^{-1}\Sigma_w^{-1} \\ &= \Sigma_x(A^\mathsf{T}\Sigma_w^{-1}A\Sigma_x + I)^{-1}A^\mathsf{T}\Sigma_w^{-1} \\ &= (A^\mathsf{T}\Sigma_w^{-1}A + \Sigma_x^{-1})^{-1}A^\mathsf{T}\Sigma_w^{-1}\end{aligned}$$

Therefore, we have the new formula

$$\mathbf{E}(x \mid y) = \mu_x + (A^\mathsf{T}\Sigma_w^{-1}A + \Sigma_x^{-1})^{-1}A^\mathsf{T}\Sigma_w^{-1}(y - A\mu_x)$$

Both ways of writing the conditional mean and covariance are useful. From a computational standpoint, the covariance formulation requires inverting a matrix of the size of Σ_w (size of y), while the information formulation requires inverting a matrix of the size of Σ_x (size of x). So if x is much larger than y or vice versa, using one formulation over the other might be faster.

3 MAP and MMSE Estimates

There are two common estimators we will use, Minimum mean squared error (MMSE) and Maximum a posteriori probability (MAP). MAP finds the x that maximizes the posterior pdf $f_{x|y}(x, y)$, while the MMSE picks the mean of the posterior pdf. If the distribution is normal, then these two estimates are equal. However, in the case where the distribution is not symmetric, such as in Fig. 2, the values are not equal.

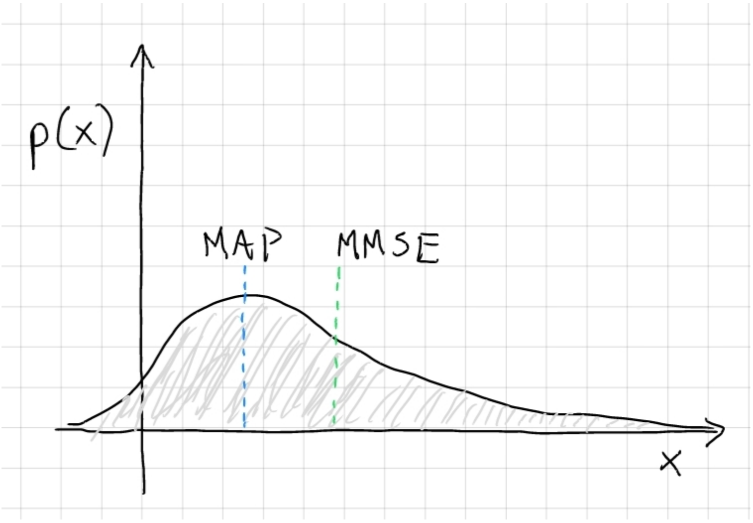

Figure 2: MAP and MMSE estimates on a non normal distribution.

4 Least-squares Approach

The pdf of our joint measurement model is:

$$f(x,y) = (\text{const}) \cdot \exp\left(-\frac{1}{2} \begin{bmatrix} x-\mu_x \\ y - A\mu_x \end{bmatrix}^\mathsf{T} \begin{bmatrix} \Sigma_x & \Sigma_x A^\mathsf{T} \\ A\Sigma_x & A\Sigma_x A^\mathsf{T} + \Sigma_w \end{bmatrix}^{-1} \begin{bmatrix} x-\mu_x \\ y - A\mu_x \end{bmatrix} \right)$$

We seek to find x that maximizes $f_{x|y}(x,y)$ for some given measurement y. Since $f_{x|y}(x,y) = \frac{f(x,y)}{f_y(y)}$ from Bayes' rule, maximizing the posterior pdf is the same as maximizing the joint pdf. We can maximize $e^{-g(x)}$ by minimizing $g(x)$. So our goal is to

$$\underset{x}{\text{minimize}} \quad \begin{bmatrix} x-\mu_x \\ y - A\mu_x \end{bmatrix}^\mathsf{T} \begin{bmatrix} \Sigma_x & \Sigma_x A^\mathsf{T} \\ A\Sigma_x & A\Sigma_x A^\mathsf{T} + \Sigma_w \end{bmatrix}^{-1} \begin{bmatrix} x-\mu_x \\ y - A\mu_x \end{bmatrix}$$

To do this, we will use the factorization

$$\begin{bmatrix} I & 0 \\ A & I \end{bmatrix} \begin{bmatrix} \Sigma_x & 0 \\ 0 & \Sigma_w \end{bmatrix} \begin{bmatrix} I & A^\mathsf{T} \\ 0 & I \end{bmatrix}$$

Substituting this in, we obtain:

$$\begin{aligned} & \begin{bmatrix} x-\mu_x \\ y - A\mu_x \end{bmatrix}^\mathsf{T} \begin{bmatrix} \Sigma_x & \Sigma_x A^\mathsf{T} \\ A\Sigma_x & A\Sigma_x A^\mathsf{T} + \Sigma_w \end{bmatrix}^{-1} \begin{bmatrix} x-\mu_x \\ y - A\mu_x \end{bmatrix} \\ &= \begin{bmatrix} x-\mu_x \\ y - A\mu_x \end{bmatrix}^\mathsf{T} \begin{bmatrix} I & -A^\mathsf{T} \\ 0 & I \end{bmatrix} \begin{bmatrix} \Sigma_x^{-1} & 0 \\ 0 & \Sigma_w^{-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ A & I \end{bmatrix} \begin{bmatrix} x-\mu_x \\ y - A\mu_x \end{bmatrix} \\ &= \begin{bmatrix} x-\mu_x \\ y - Ax \end{bmatrix}^\mathsf{T} \begin{bmatrix} \Sigma_x^{-1} & 0 \\ 0 & \Sigma_w^{-1} \end{bmatrix} \begin{bmatrix} x-\mu_x \\ y - Ax \end{bmatrix} \end{aligned}$$

Expanding this out, we can write

$$\begin{aligned}&= \underbrace{(x-\mu_x)^\mathsf{T}\Sigma_x^{-1}(x-\mu_x)}_{\text{prior on } x} + \underbrace{(y-Ax)\Sigma_w^{-1}(y-Ax)}_{\text{noise}} \\ &= \|x-\mu_x\|^2_{\Sigma_x^{-1}} + \|y-Ax\|^2_{\Sigma_w^{-1}}\end{aligned}$$

In the last step, we used the definition $\|x\|_Q^2 := x^\mathsf{T} Q x$, which is called a *weighted 2-norm*. So finding the MAP estimator (which is the same as the MMSE in this Gaussian setting) amounts to solving a multi-objective least squares problem! The goal is to find an x that is simultaneously close to its prior mean μ_x and also for which the noise $w = y - Ax$ is small. The relative weight we give to these two objectives is determined by the covariance matrices Σ_x and Σ_w.

4.1 Test cases

Case 1. We have an uninformative prior and the measurement noise is a ball (covariance is a multiple of the identity). In this case, the optimization problem simplifies to

$$\left.\begin{aligned}\Sigma_x &\to \infty \\ \Sigma_w &= \sigma^2 I\end{aligned}\right\} \to \min_x \frac{1}{\sigma^2}\|y-Ax\|^2$$

In other words, the problem reduces to standard least squares. Geometrically, we pick the largest density (likeliest noise) that intersects range(A). When the density contours are circles as in the case $\Sigma_w = \sigma^2 I$, the optimal point is forms a right angle with y, so it's the same as projecting y onto range(A) (i.e. it's the least squares solution). If the ellipsoid contours are not circles, then the optimal point will still be the point of tangency, but it may no longer be orthogonal. In this case, the optimal point corresponds to an *oblique projection.* See Fig. 3.

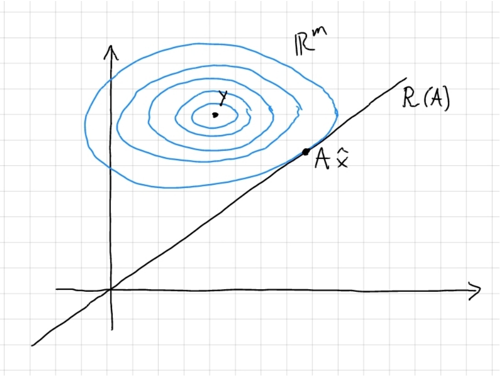

Figure 3: Case 1

Case 2. We have an informative prior, so $\Sigma_x \prec \infty$. Here, we can plot the space $\mathbb{R}^n$ (space of x values). Depending on the value of Σ_x, we will either obtain a solution that is close to μ_x (the prior mean), or close to $\boldsymbol{A}^\dagger y$ (the least-squares/least-norm solution). The optimal solution is when the ellipsoids of equal confidence are tangent.

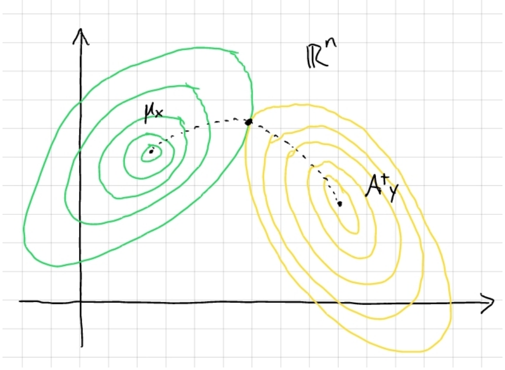

Figure 4: Case 2

5 Recursive Least-Squares

Imagine a series of measurements $y_1, \ldots, y_m$. In addition, we will assume that the noise in each measurement is independent of the noise in all other measurements. This means that all off-diagonal elements in the noise covariance matrix are zero.

$$\begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} A_1 \\ \vdots \\ A_m \end{bmatrix} x + \begin{bmatrix} w_1 \\ \vdots \\ w_m \end{bmatrix}, \qquad w \sim \mathcal{N}\left(0, \begin{bmatrix} \Sigma_{w_1} & & 0 \\ & \ddots & \\ 0 & & \Sigma_{w_m} \end{bmatrix}\right)$$

In this scenario, we have a prior on x and use y to update our distribution for x. One way to do this is to solve a new problem each time we get a new measurement:

$$\begin{aligned} \hat{x}_1 &= x \mid y_1 \\ \hat{x}_2 &= x \mid y_1, y_2 \\ &\vdots \\ \hat{x}_m &= x \mid y_1, \ldots, y_m \end{aligned}$$

This is obviously inefficient as the number of measurements grows larger and we have to solve ever-growing problems at each step. We also have to store all the measurements! To increase efficiency, we can simply use our posterior as our new prior.

$$\begin{aligned} \hat{x}_1 &= x \mid y_1 \\ \hat{x}_2 &= \hat{x}_1 \mid y_2 \\ &\vdots \\ \hat{x}_m &= \hat{x}_{m-1} \mid y_m \end{aligned}$$

This works as long as the different measurements are *conditionally independent given* x. In other words, it's fine that y_i and y_j are correlated (of course, they will be, since changing x will affect both). But if we are given x, then y_i and y_j become independent because w_i and w_j are independent by assumption.

Define the partial conditional expectations and covariances as follows:

$$\begin{aligned}
\hat{x}_k &= \mathbf{E}(x \mid y_1, \ldots, y_k) \\
\hat{\Sigma}_k &= \mathbf{Cov}(x \mid y_1, \ldots, y_k) \\
\hat{x}_0 &= \mu_x \\
\hat{\Sigma}_0 &= \Sigma_x
\end{aligned}$$

Substituting the above in Eq. (1) and Eq. (2), we can write

$$\boxed{\begin{aligned}
\hat{x}_{k+1} &= \hat{x}_k + \hat{\Sigma}_k A_{k+1}^{\mathsf{T}} (A_{k+1} \hat{\Sigma}_k A_{k+1}^{\mathsf{T}} + \Sigma_{w_{k+1}})^{-1} (y_{k+1} - A_{k+1} \hat{x}_k) && (7) \\
\hat{\Sigma}_{k+1} &= \hat{\Sigma}_k - \hat{\Sigma}_k A_{k+1}^{\mathsf{T}} (A_{k+1} \hat{\Sigma}_k A_{k+1}^{\mathsf{T}} + \Sigma_{w_{k+1}})^{-1} A_{k+1} \hat{\Sigma}_k && (8) \\
\hat{\Sigma}_{k+1}^{-1} &= \hat{\Sigma}_k^{-1} + A_{k+1}^{\mathsf{T}} \hat{\Sigma}_{w_{k+1}}^{-1} A_{k+1} && (9)
\end{aligned}}$$

As we get new measurements, we can incrementally get more confident in our estimate. Note that Eqs. (8) and (9) are two alternative formulas that say the same thing. In the former, we see that $\hat{\Sigma}$ is updated by subtracting something positive definite from it. So the error covariance shrinks. In the information formulation, as we gather more measurements, the inverse covariance gets larger; we get more information.

This comes up a lot in estimating/triangulating position. A practical example is the map applications on our phones. Initially, the position estimate is very crude. Within a few milliseconds, the information from more satellites and relative positions are incorporated and the error shrinks.