12.3 Multivariate Gaussian and Weighted Least Squares

The normal probability density $p(x)$ (the Gaussian) depends on only two numbers:

$$\textbf{Mean } m \textbf{ and variance } \sigma^2 \qquad \boldsymbol{p(x) = \frac{1}{\sqrt{2\pi}\,\sigma}\, e^{-(x-m)^2/2\sigma^2}}. \tag{1}$$

The graph of $p(x)$ is a bell-shaped curve centered at $x = m$. The continuous variable x can be anywhere between $-\infty$ and ∞. With probability close to $\frac{2}{3}$, that random x will lie between $m - \sigma$ and $m + \sigma$ (less than one standard deviation σ from its mean value m).

$$\int_{-\infty}^{\infty} p(x)\,dx = 1 \quad \text{and} \quad \int_{\boldsymbol{m-\sigma}}^{\boldsymbol{m+\sigma}} p(x)\,dx = \frac{1}{\sqrt{2\pi}} \int_{\mathbf{-1}}^{\mathbf{1}} e^{-X^2/2}\,dX \approx \frac{\mathbf{2}}{\mathbf{3}}. \tag{2}$$

That integral has a change of variables from x to $X = (x-m)/\sigma$. This simplifies the exponent to $-X^2/2$ and it simplifies the limits of integration to -1 and 1. Even the $1/\sigma$ from p disappears outside the integral because dX equals dx/σ. Every Gaussian turns into a **standard Gaussian** $p(X)$ with mean $m = 0$ and variance $\sigma^2 = 1$. Just call it $p(x)$:

$$\textbf{The standard normal distribution } \boldsymbol{N(0,1)} \quad \textbf{has} \quad \boldsymbol{p(x) = \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}}. \tag{3}$$

Integrating $p(x)$ from $-\infty$ to x gives the cumulative distribution $F(x)$: the probability that a random sample is below x. That probability will be $F = \frac{1}{2}$ at $x = 0$ (the mean).

Two-dimensional Gaussians

Now we have $M = 2$ Gaussian random variables x and y. They have means m_1 and m_2. They have variances σ_1^2 and σ_2^2. If they are *independent*, then their probability density $p(x, y)$ is just $\boldsymbol{p_1(x)}$ **times** $\boldsymbol{p_2(y)}$. Multiply probabilities when variables are independent:

$$\textbf{Independent } \boldsymbol{x} \textbf{ and } \boldsymbol{y} \quad p(x,y) = \frac{1}{2\pi\sigma_1\sigma_2}\; e^{-(x-m_1)^2/2\sigma_1^2}\, e^{-(y-m_2)^2/2\sigma_2^2} \tag{4}$$

The covariance of x and y will be $\boldsymbol{\sigma_{12} = 0}$. The covariance matrix V will be *diagonal*. The variances σ_1^2 and σ_2^2 are always on the main diagonal of V. The exponent in $p(x, y)$ is just the sum of the x-exponent and the y-exponent. Good to notice that the two exponents can be combined into $-\frac{1}{2}\,(\boldsymbol{x}-\boldsymbol{m})^{\mathrm{T}}\, V^{-1}\,(\boldsymbol{x}-\boldsymbol{m})$ with V^{-1} in the middle:

$$-\frac{(x-m_1)^2}{2\sigma_1^2} - \frac{(y-m_2)^2}{2\sigma_2^2} = -\frac{1}{2}\begin{bmatrix} x-m_1 & y-m_2 \end{bmatrix} \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}^{-1} \begin{bmatrix} x-m_1 \\ y-m_2 \end{bmatrix} \tag{5}$$

Non-independent x and y

We are ready to give up independence. The exponent (5) with V^{-1} is still correct when V is no longer a diagonal matrix. **Now the Gaussian depends on a vector $\boldsymbol{m}$ and a matrix $\boldsymbol{V}$.**

When $M = 2$, the first variable x may give partial information about the second variable y (and vice versa). Maybe part of y is decided by x and part is truly independent. It is the M by M covariance matrix V that accounts for dependencies between the M variables $\boldsymbol{x} = x_1, \ldots, x_M$. Its inverse V^{-1} goes into $p(\boldsymbol{x})$:

Multivariate Gaussian probability distribution

$$\boxed{\boldsymbol{p(x)} = \frac{1}{(\sqrt{2\pi})^M\sqrt{\det V}}\ \boldsymbol{e^{-(x-m)^{\mathrm{T}}V^{-1}(x-m)/2}}} \quad (6)$$

The vectors $\boldsymbol{x} = (x_1, \ldots, x_M)$ and $\boldsymbol{m} = (m_1, \ldots, m_M)$ contain the random variables and their means. The M square roots of 2π and the determinant of V are included to make the total probability equal to 1. Let me check that by linear algebra. I use the eigenvalues λ and orthonormal eigenvectors $\boldsymbol{q}$ of the symmetric matrix $V = Q\Lambda Q^{\mathrm{T}}$. So $\boldsymbol{V^{-1} = Q\Lambda^{-1}Q^{\mathrm{T}}}$:

$$\boldsymbol{X} = \boldsymbol{x} - \boldsymbol{m} \qquad (\boldsymbol{x} - \boldsymbol{m})^{\mathrm{T}}\boldsymbol{V^{-1}}(\boldsymbol{x} - \boldsymbol{m}) = \boldsymbol{X^{\mathrm{T}}Q\Lambda^{-1}Q^{\mathrm{T}}X = Y^{\mathrm{T}}\Lambda^{-1}Y}$$

Notice! The combinations $\boldsymbol{Y = Q^{\mathrm{T}}X = Q^{\mathrm{T}}(x - m)}$ are statistically independent. *Their covariance matrix Λ is diagonal.*

This step of diagonalizing V by its eigenvector matrix Q is the same as "uncorrelating" the random variables. Covariances are zero for the new variables $X_1, \ldots X_m$. This is the point where linear algebra helps calculus to compute multidimensional integrals.

The integral of $p(\boldsymbol{x})$ is not changed when we center the variable $\boldsymbol{x}$ by subtracting $\boldsymbol{m}$ to reach $\boldsymbol{X}$, and rotate that variable to reach $\boldsymbol{Y = Q^{\mathrm{T}}X}$. The matrix $\boldsymbol{\Lambda}$ is diagonal! So the integral we want splits into M separate one-dimensional integrals that we know:

$$\int \cdots \int \boldsymbol{e^{-Y^{\mathrm{T}}\Lambda^{-1}Y/2}\,dY} = \left(\int_{-\infty}^{\infty} e^{-y_1^2/2\lambda_1}\,dy_1\right) \cdots \left(\int_{-\infty}^{\infty} e^{-y_M^2/2\lambda_M}\,dy_M\right)$$

$$= \left(\sqrt{2\pi\lambda_1}\right) \cdots \left(\sqrt{2\pi\lambda_M}\right) = \left(\sqrt{2\pi}\right)^M \sqrt{\det V}. \quad (7)$$

The determinant of V (also the determinant of Λ) is the product $(\lambda_1) \ldots (\lambda_M)$ of the eigenvalues. Then (7) gives the correct number to divide by so that $p(x_1, \ldots, x_M)$ in equation (6) has integral $= 1$ as desired.

The mean and variance of $p(\boldsymbol{x})$ are also M-dimensional integrals. The same idea of diagonalizing V by its eigenvectors and introducing $\boldsymbol{Y = Q^{\mathrm{T}}X}$ will find those integrals:

Vector $\boldsymbol{m}$ of means
$$\int \cdots \int \boldsymbol{x}\, p(\boldsymbol{x})\, \boldsymbol{dx} = (m_1, m_2, \ldots) = \boldsymbol{m} \quad (8)$$

Covariance matrix $\boldsymbol{V}$
$$\int \cdots \int (\boldsymbol{x} - \boldsymbol{m})\, p(\boldsymbol{x})(\boldsymbol{x} - \boldsymbol{m})^{\mathrm{T}}\, \boldsymbol{dx} = \boldsymbol{V}. \quad (9)$$

Conclusion: Formula (6) for the probability density $p(\boldsymbol{x})$ has all the properties we want.

Weighted Least Squares

In Chapter 4, least squares started from an unsolvable system $A\boldsymbol{x} = \boldsymbol{b}$. We chose $\widehat{\boldsymbol{x}}$ to minimize the error $||\boldsymbol{b} - A\boldsymbol{x}||^2$. That led us to the least squares equation $A^{\mathrm{T}}A\widehat{\boldsymbol{x}} = A^{\mathrm{T}}\boldsymbol{b}$. The best $A\widehat{\boldsymbol{x}}$ is the projection of $\boldsymbol{b}$ onto the column space of A. But is this squared distance $E = ||\boldsymbol{b} - A\boldsymbol{x}||^2$ the right error measure to minimize ?

If the measurement errors in $\boldsymbol{b}$ are independent random variables, with mean $m = 0$ and variance $\sigma^2 = 1$ and a normal distribution, Gauss would say **yes** : *Use least squares*. If the errors are not independent or their variances are not equal. Gauss would say **no** : *Use **weighted** least squares*. This section will show that the good measure of error is $\boldsymbol{E} = (\boldsymbol{b} - A\boldsymbol{x})^{\mathrm{T}}\boldsymbol{V^{-1}}(\boldsymbol{b} - A\boldsymbol{x})$. The equation for the best $\widehat{\boldsymbol{x}}$ uses the covariance matrix V :

Weighted least squares $$\boldsymbol{A^{\mathrm{T}}V^{-1}A\widehat{x} = A^{\mathrm{T}}V^{-1}b}. \tag{10}$$

The most important examples have m *independent* errors in $\boldsymbol{b}$. Those errors have variances $\sigma_1^2, \ldots, \sigma_m^2$. By independence, V is a diagonal matrix. The good weights $1/\sigma_1^2, \ldots, 1/\sigma_m^2$ come from V^{-1}. *We are weighting the errors in* $\boldsymbol{b}$ *to have* **variance** $= \mathbf{1}$:

Weighted least squares
Independent errors in *b*

$$\boxed{\text{Minimize} \quad E = \sum_{i=1}^{m} \frac{(\boldsymbol{b} - A\boldsymbol{x})_i^2}{\sigma_i^2}} \tag{11}$$

By weighting the errors, we are "whitening" the noise. **White noise** is a quick description of independent errors based on the standard Gaussian $\mathbf{N}(0, 1)$ with mean zero and $\sigma^2 = 1$.

Let me write down the steps to equations (10) and (11) for the best $\widehat{\boldsymbol{x}}$:

Start with $A\boldsymbol{x} = \boldsymbol{b}$ (m equations, n unknowns, $m > n$, no solution)

Each right side b_i has mean zero and variance σ_i^2. The b_i are independent.

Divide the ith equation by σ_i to have variance $= 1$ for every b_i/σ_i

That division turns $A\boldsymbol{x} = \boldsymbol{b}$ into $V^{-1/2}A\boldsymbol{x} = V^{-1/2}\boldsymbol{b}$ with $V^{-1/2} = \operatorname{diag}(1/\sigma_1, \ldots, 1/\sigma_m)$

Ordinary least squares on those weighted equations has $A \to V^{-1/2}A$ and $\boldsymbol{b} \to V^{-1/2}\boldsymbol{b}$

$$(V^{-1/2}A)^{\mathrm{T}}(V^{-1/2}A)\widehat{\boldsymbol{x}} = (V^{-1/2}A)^{\mathrm{T}}V^{-1/2}\,\boldsymbol{b} \quad \text{is} \quad \boldsymbol{A^{\mathrm{T}}V^{-1}A\widehat{x} = A^{\mathrm{T}}V^{-1}b}. \tag{12}$$

Because of $1/\sigma^2$ in V^{-1}, more reliable equations (*smaller* σ) get heavier weights. This is the main point of weighted least squares.

Those diagonal weightings (uncoupled equations) are the most frequent and the simplest. They apply to *independent errors in the* b_i. When these measurement errors are not independent, V is no longer diagonal—but (12) is still the correct weighted equation.

In practice, finding all the covariances can be serious work. Diagonal V is simpler.

The Variance in the Estimated $\widehat{\boldsymbol{x}}$

One more point: Often the important question is not the best $\widehat{\boldsymbol{x}}$ for one particular set of measurements $\boldsymbol{b}$. This is only one sample! The real goal is to know the reliability of the whole experiment. That is measured (as reliability always is) by the **variance in the estimate** $\widehat{\boldsymbol{x}}$. First, zero mean in $\boldsymbol{b}$ gives zero mean in $\widehat{\boldsymbol{x}}$. Then the formula connecting variance $\boldsymbol{V}$ in the inputs $\boldsymbol{b}$ to variance $\boldsymbol{W}$ in the outputs $\widehat{\boldsymbol{x}}$ turns out to be beautiful:

Variance-covariance matrix $\boldsymbol{W}$ for $\widehat{\boldsymbol{x}}$ $\quad \mathrm{E}[(\widehat{\boldsymbol{x}}-\boldsymbol{x})(\widehat{\boldsymbol{x}}-\boldsymbol{x})^{\mathrm{T}}] = \boldsymbol{(A^{\mathrm{T}}V^{-1}A)^{-1}}$.(13)

That smallest possible variance comes from the best possible weighting, which is V^{-1}.

This key formula is a perfect application of Section 12.2. **If $\boldsymbol{b}$ has covariance matrix** V**, then** $\widehat{\boldsymbol{x}} = \boldsymbol{Lb}$ **has covariance matrix $\boldsymbol{LVL^{\mathrm{T}}}$.** Equation (12) above tells us that L is $(A^{\mathrm{T}}V^{-1}A)^{-1}A^{\mathrm{T}}V^{-1}$. Now substitute this into LVL^{T} and watch equation (13) appear:

$$\boldsymbol{LVL^{\mathrm{T}}} = (A^{\mathrm{T}}V^{-1}A)^{-1}A^{\mathrm{T}}V^{-1} \quad V \quad V^{-1}A\,(A^{\mathrm{T}}V^{-1}A)^{-1} = \boldsymbol{(A^{\mathrm{T}}V^{-1}A)^{-1}}.$$

This is the covariance W of the output, our best estimate $\widehat{\boldsymbol{x}}$. It is time for examples.

Example 1 Suppose a doctor measures your heart rate x three times $(m=3, n=1)$:

$$\begin{matrix} x = b_1 \\ x = b_2 \\ x = b_3 \end{matrix} \quad \text{is} \quad A\boldsymbol{x} = \boldsymbol{b} \quad \text{with} \quad A = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \quad \text{and} \quad V = \begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{bmatrix}$$

The variances could be $\sigma_1^2 = 1/9$ and $\sigma_2^2 = 1/4$ and $\sigma_3^2 = \mathbf{1}$. You are getting more nervous as measurements are taken: b_3 is less reliable than b_2 and b_1. All three measurements contain some information, so they all go into the best (weighted) estimate $\widehat{\boldsymbol{x}}$:

$$V^{-1/2}A\widehat{\boldsymbol{x}} = V^{-1/2}\boldsymbol{b} \quad \text{is} \quad \begin{matrix} 3x = 3b_1 \\ 2x = 2b_2 \\ 1x = 1b_3 \end{matrix} \quad \text{leading to} \quad \boldsymbol{A^{\mathrm{T}}V^{-1}A\widehat{x} = A^{\mathrm{T}}V^{-1}b}$$

$$\begin{bmatrix} 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 9 & & \\ & 4 & \\ & & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \widehat{\boldsymbol{x}} = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 9 & & \\ & 4 & \\ & & 1 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$$

$$\boxed{\widehat{x} = \frac{\boldsymbol{9b_1 + 4b_2 + b_3}}{\boldsymbol{14}} \quad \textbf{is a weighted average of } \boldsymbol{b_1, b_2, b_3}}$$

Most weight is on b_1 since its variance σ_1 is smallest. The variance of $\widehat{\boldsymbol{x}}$ has the beautiful formula $W = (A^{\mathrm{T}}V^{-1}A)^{-1} = 1/14$:

$$\textbf{Variance of } \widehat{\boldsymbol{x}} \qquad \left(\begin{bmatrix} 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 9 & & \\ & 4 & \\ & & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \right)^{-1} = \frac{1}{14} \quad \text{is smaller than} \quad \frac{1}{9}$$

The BLUE theorem of Gauss (proved on the website) says that our $\widehat{\boldsymbol{x}} = L\boldsymbol{b}$ is the best linear unbiased estimate of the solution to $A\boldsymbol{x} = \boldsymbol{b}$. Any other unbiased choice $\boldsymbol{x}^* = L^*\boldsymbol{b}$ has greater variance than $\widehat{\boldsymbol{x}}$. All unbiased choices have $L^*A = I$ so that an exact $A\boldsymbol{x} = \boldsymbol{b}$ will produce the right answer $\boldsymbol{x} = L^*\boldsymbol{b} = L^*A\boldsymbol{x}$.

Note. I must add that there are reasons not to minimize squared errors in the first place. One reason: This $\widehat{\boldsymbol{x}}$ often has many small components. The squares of small numbers are very small, and they appear when we minimize. It is easier to make sense of *sparse* vectors—only a few nonzeros. Statisticians often prefer to minimize **unsquared errors**: **the sum of** $|(\boldsymbol{b} - A\boldsymbol{x})_i|$. *This error measure is* L^1 *instead of* L^2. Because of the absolute values, the equation for $\widehat{\boldsymbol{x}}$ becomes nonlinear (it is actually piecewise linear).

Fast new algorithms are computing a sparse $\widehat{\boldsymbol{x}}$ quickly and the future may belong to L^1.

The Kalman Filter

The "Kalman filter" is the great algorithm in dynamic least squares. That word *dynamic* means that new measurements $\boldsymbol{b}_k$ keep coming. So the best estimate $\widehat{\boldsymbol{x}}_k$ keeps changing (based on all of $\boldsymbol{b}_0, \ldots, \boldsymbol{b}_k$). More than that, the matrix A is also changing. So $\widehat{\boldsymbol{x}}_2$ will be our best least squares estimate of the latest solution $\boldsymbol{x}_k$ to the **whole history of observation equations and update equations (state equations) up to time 2**:

$$A_0\boldsymbol{x}_0 = \boldsymbol{b}_0 \qquad \boldsymbol{x}_1 = F_0\boldsymbol{x}_0 \qquad A_1\boldsymbol{x}_1 = \boldsymbol{b}_1 \qquad \boldsymbol{x}_2 = F_1\boldsymbol{x}_1 \qquad A_2\boldsymbol{x}_2 = \boldsymbol{b}_2 \qquad (14)$$

The Kalman idea is to introduce one equation at a time. There will be errors in each equation. With every new equation, we update the best estimate $\widehat{\boldsymbol{x}}_k$ for the current $\boldsymbol{x}_k$. But history is not forgotten! This new estimate $\widehat{\boldsymbol{x}}_k$ uses all the past observations $\boldsymbol{b}_0$ to $\boldsymbol{b}_{k-1}$ and all the state equations $\boldsymbol{x}_{\text{new}} = F_{\text{old}}\,\boldsymbol{x}_{\text{old}}$. A large and growing least squares problem.

One more important point. Each least squares equation is **weighted** using the covariance matrix V_k for the error in b_k. There is even a covariance matrix C_k for errors in the update equations $\boldsymbol{x}_{k+1} = F_k x_k$. The best $\widehat{\boldsymbol{x}}_2$ then depends on $\boldsymbol{b}_0, \boldsymbol{b}_1, \boldsymbol{b}_2$ and V_0, V_1, V_2 and C_1, C_2. The good way to write $\widehat{\boldsymbol{x}}_k$ is as an update to the previous $\widehat{\boldsymbol{x}}_{k-1}$.

Let me concentrate on a simplified problem, without the matrices F_k and the covariances C_k. We are estimating the same true $\boldsymbol{x}$ at every step. How do we get $\widehat{\boldsymbol{x}}_1$ from $\widehat{\boldsymbol{x}}_0$?

$$\textbf{OLD} \quad A_0\,\boldsymbol{x}_0 = \boldsymbol{b}_0 \text{ leads to the weighted equation } A_0^{\mathrm{T}}\,V_0^{-1}\,A_0\,\widehat{\boldsymbol{x}}_0 = A_0^{\mathrm{T}}\,V_0^{-1}\,\boldsymbol{b}_0. \qquad (15)$$

$$\textbf{NEW} \quad \begin{bmatrix} A_0 \\ A_1 \end{bmatrix} \widehat{\boldsymbol{x}}_1 = \begin{bmatrix} \boldsymbol{b}_0 \\ \boldsymbol{b}_1 \end{bmatrix} \text{ leads to the following weighted equation for } \widehat{\boldsymbol{x}}_{\boldsymbol{1}}:$$

$$\begin{bmatrix} A_0^{\mathrm{T}} & A_1^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} V_0^{-1} & \\ & V_1^{-1} \end{bmatrix} \begin{bmatrix} A_0 \\ A_1 \end{bmatrix} \widehat{\boldsymbol{x}}_1 = \begin{bmatrix} A_0^{\mathrm{T}} & A_1^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} V_0^{-1} & \\ & V_1^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{b}_0 \\ \boldsymbol{b}_1 \end{bmatrix}. \quad (16)$$

Yes, we could just solve that new problem and forget the old one. But the old solution $\widehat{\boldsymbol{x}}_0$ needed work that we hope to reuse in $\widehat{\boldsymbol{x}}_1$. What we look for is **an update to $\widehat{\boldsymbol{x}}_0$** :

Kalman update gives $\widehat{\boldsymbol{x}}_1$ from $\widehat{\boldsymbol{x}}_0$ $\quad \widehat{\boldsymbol{x}}_1 = \widehat{\boldsymbol{x}}_0 + K_1(\boldsymbol{b}_1 - A_1\,\widehat{\boldsymbol{x}}_0). \quad (17)$

The update correction is the mismatch $\boldsymbol{b}_1 - A_1\widehat{\boldsymbol{x}}_0$ between the old state $\widehat{\boldsymbol{x}}_0$ and the new measurements $\boldsymbol{b}_1$—multiplied by the *Kalman gain matrix* K_1. The formula for K_1 comes from comparing the solutions $\widehat{\boldsymbol{x}}_1$ and $\widehat{\boldsymbol{x}}_0$ to (15) and (16). And when we update $\widehat{\boldsymbol{x}}_0$ to $\widehat{\boldsymbol{x}}_1$ based on new data $\boldsymbol{b}_1$, **we also update the covariance matrix $\boldsymbol{W_0}$ to $\boldsymbol{W_1}$**. Remember $W_0 = (A_0^{\mathrm{T}}\, V_0^{-1}\, A_0)^{-1}$ from equation (13). Update its inverse to W_1^{-1} :

Covariance $\boldsymbol{W_1}$ of errors in $\widehat{\boldsymbol{x}}_1$ $\quad \boldsymbol{W_1^{-1} = W_0^{-1} + A_1^{\mathrm{T}}\, V_1^{-1}\, A_1} \quad (18)$

Kalman gain matrix $\boldsymbol{K_1}$ $\quad \boldsymbol{K_1 = W_1\, A_1^{\mathrm{T}}\, V_1^{-1}} \quad (19)$

This is the heart of the Kalman filter. Notice the importance of the W_k. Those matrices measure the reliability of the whole process, where the vector $\widehat{\boldsymbol{x}}_k$ estimates the current state based on the particular measurements $\boldsymbol{b}_0$ to $\boldsymbol{b}_k$.

Whole chapters and whole books are written to explain the dynamic Kalman filter, when the states $\boldsymbol{x}_k$ are also changing (based on the matrices F_k). There is a *prediction* of $\boldsymbol{x}_k$ using F, followed by a *correction* using the new data $\boldsymbol{b}$. Perhaps best to stop here.

This page was about **recursive least squares**: adding new data b_k and updating both $\widehat{\boldsymbol{x}}$ and W : the best current estimate based on all the data, and its covariance matrix.

Problem Set 12.3

1 Two measurements of the same variable x give two equations $x = b_1$ and $x = b_2$. Suppose the means are zero and the variances are σ_1^2 and σ_2^2, with independent errors: V is diagonal with entries σ_1^2 and σ_2^2. Write the two equations as $A\boldsymbol{x} = \boldsymbol{b}$ (A is 2 by 1). As in the text Example 1, find this best estimate $\widehat{\boldsymbol{x}}$ based on b_1 and b_2 :

$$\widehat{\boldsymbol{x}} = \frac{b_1/\sigma_1^2 + b_2/\sigma_2^2}{1/\sigma_1^2 + 1/\sigma_2^2} \qquad \mathrm{E}\left[\widehat{\boldsymbol{x}}\,\widehat{\boldsymbol{x}}^{\mathrm{T}}\right] = \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right)^{-1}.$$

2 (a) In Problem 1, suppose the second measurement b_2 becomes super-exact and its variance $\sigma_2 \to 0$. What is the best estimate $\widehat{\boldsymbol{x}}$ when σ_2 reaches zero?

(b) The opposite case has $\sigma_2 \to \infty$ and no information in b_2. What is now the best estimate $\widehat{\boldsymbol{x}}$ based on b_1 and b_2 ?

3 If x and y are independent with probabilities $p_1(x)$ and $p_2(y)$, then $p(x,y) = p_1(x)\,p_2(y)$. By separating double integrals into products of single integrals ($-\infty$ to ∞) show that

$$\iint p(x,y)\,dx\,dy = \mathbf{1} \qquad \text{and} \qquad \iint (x+y)\,p(x,y)\,dx\,dy = \boldsymbol{m_1} + \boldsymbol{m_2}.$$

4 Continue Problem 3 for independent x, y to show that $p(x,y) = p_1(x)\,p_2(y)$ has

$$\iint (x-m_1)^2\,p(x,y)\,dx\,dy = \boldsymbol{\sigma_1^2} \qquad \iint (x-m_1)(y-m_2)\,p(x,y)\,dx\,dy = \mathbf{0}.$$

So the 2 by 2 covariance matrix V is diagonal and its entries are _____.

5 Show that the inverse of a 2 by 2 covariance matrix V is

$$V^{-1} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}^{-1} = \frac{1}{1-\rho^2}\begin{bmatrix} 1/\sigma_1^2 & -\rho/\sigma_1\sigma_2 \\ -\rho/\sigma_1\sigma_2 & 1/\sigma_2^2 \end{bmatrix} \quad \begin{matrix}\text{with correlation} \\ \rho = \sigma_{12}/\sigma_1\sigma_2.\end{matrix}$$

This produces the exponent $-(\boldsymbol{x}-\boldsymbol{m})^{\mathrm{T}}\,\boldsymbol{V}^{-1}(\boldsymbol{x}-\boldsymbol{m})$ in a 2-variable Gaussian.

6 Suppose $\widehat{x}_k$ is the average of $b_1, \ldots, b_k$. A new measurement b_{k+1} arrives and we want the new average $\widehat{x}_{k+1}$. The Kalman update equation (17) is

$$\textbf{New average} \qquad \widehat{x}_{k+1} = \widehat{x}_k + \frac{1}{k+1}\left(b_{k+1} - \widehat{x}_k\right).$$

Verify that $\widehat{x}_{k+1}$ *is the correct average of* $b_1 \ldots, b_{k+1}$.

7 Also check the update equation (18) for the variance $W_{k+1} = \sigma^2/(k+1)$ of this average $\widehat{x}$ assuming that $W_k = \sigma^2/k$ and b_{k+1} has variance $V = \sigma^2$.

8 (**Steady model**) Problems 6–7 were *static* least squares. All the sample averages $\widehat{x}_k$ were estimates of the same x. To make the Kalman filter *dynamic*, include also a *state equation* $x_{k+1} = Fx_k$ *with its own error variance* s^2. The dynamic least squares problem allows x to "drift" as k increases:

$$\begin{bmatrix} 1 & \\ -F & 1 \\ & 1 \end{bmatrix}\begin{bmatrix} x_0 \\ x_1 \end{bmatrix} = \begin{bmatrix} b_0 \\ 0 \\ b_1 \end{bmatrix} \text{ with variances } \begin{bmatrix} \sigma^2 \\ s^2 \\ \sigma^2 \end{bmatrix}.$$

With $F = 1$, divide both sides of those three equations by σ, s, and σ. Find $\widehat{x_0}$ and $\widehat{x_1}$ by least squares, which gives more weight to the recent b_1. The Kalman filter is developed in *Algorithms for Global Positioning* (Borre and Strang, Wellesley-Cambridge Press).

Change in A^{-1} from a Change in A

This final page connects the beginning of the book (inverses and rank one matrices) with the end of the book (dynamic least squares and filters). Begin with this basic formula:

$$\boxed{\textbf{The inverse of } \boldsymbol{M = I - uv}^{\mathrm{T}} \textbf{ is } \boldsymbol{M^{-1} = I} + \frac{\boldsymbol{uv}^{\mathrm{T}}}{\boldsymbol{1 - v}^{\mathrm{T}}\boldsymbol{u}}}$$

The quickest proof is $MM^{-1} = I - \boldsymbol{uv^{\mathrm{T}}} + (1 - \boldsymbol{uv}^{\mathrm{T}})\dfrac{\boldsymbol{uv}^{\mathrm{T}}}{1 - \boldsymbol{v}^{\mathrm{T}}\boldsymbol{u}} = I - \boldsymbol{uv}^{\mathrm{T}} + \boldsymbol{uv}^{\mathrm{T}} = I$.

M is not invertible if $\boldsymbol{v}^{\mathrm{T}}\boldsymbol{u} = 1$ (then $M\boldsymbol{u} = \boldsymbol{0}$). Here $\boldsymbol{v}^{\mathrm{T}} = \boldsymbol{u}^{\mathrm{T}} = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}$:

Example The inverse of $M = I - \begin{bmatrix} 1\,1\,1 \\ 1\,1\,1 \\ 1\,1\,1 \end{bmatrix}$ is $M^{-1} = I + \dfrac{1}{\boldsymbol{1-3}} \begin{bmatrix} 1\,1\,1 \\ 1\,1\,1 \\ 1\,1\,1 \end{bmatrix}$

But we don't always start from the identity matrix. Many applications need to invert $\boldsymbol{M = A - uv^{\mathrm{T}}}$. After we solve $A\boldsymbol{x} = \boldsymbol{b}$ we expect a rank one change to give $M\boldsymbol{y} = \boldsymbol{b}$. The division by $1 - \boldsymbol{v}^{\mathrm{T}}\boldsymbol{u}$ above will become a division by $c = 1 - \boldsymbol{v}^{\mathrm{T}}A^{-1}\boldsymbol{u} = 1 - \boldsymbol{v}^{\mathrm{T}}\boldsymbol{z}$.

Step 1 Solve $A\boldsymbol{z} = \boldsymbol{u}$ and compute $c = 1 - \boldsymbol{v}^{\mathrm{T}}\boldsymbol{z}$.

Step 2 If $c \neq 0$ then $M^{-1}\boldsymbol{b}$ is $\boldsymbol{y} = \boldsymbol{x} + \dfrac{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{x}}{c}\,\boldsymbol{z}$.

Suppose A is easy to work with. A might already be factored into LU by elimination. Then this Sherman-Woodbury-Morrison formula is the fast way to solve $M\boldsymbol{y} = \boldsymbol{b}$. Here are three problems to end the book!

9 Take Steps 1–2 to find $\boldsymbol{y}$ when $A = I$ and $\boldsymbol{u}^{\mathrm{T}} = \boldsymbol{v}^{\mathrm{T}} = [1\ \ 2\ \ 3]$ and $\boldsymbol{b}^{\mathrm{T}} = [2\ \ 1\ \ 4]$.

10 Step 2 in this "update formula" claims that $M\boldsymbol{y} = \left(A - \boldsymbol{uv}^{\mathrm{T}}\right)\left(\boldsymbol{x} + \dfrac{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{x}}{c}\,\boldsymbol{z}\right) = \boldsymbol{b}$.

Simplify this to $\dfrac{\boldsymbol{uv}^{\mathrm{T}}\boldsymbol{x}}{c}\,[1 - c - \boldsymbol{v}^{\mathrm{T}}\boldsymbol{z}] = \boldsymbol{0}$. This is true since $c = 1 - \boldsymbol{v}^{\mathrm{T}}\boldsymbol{z}$.

11 When A has a new row $\boldsymbol{v}^{\mathrm{T}}$, $A^{\mathrm{T}}A$ in the least squares equation changes to M:

$$M = \begin{bmatrix} A^{\mathrm{T}} & \boldsymbol{v} \end{bmatrix} \begin{bmatrix} A \\ \boldsymbol{v}^{\mathrm{T}} \end{bmatrix} = A^{\mathrm{T}}A + \boldsymbol{vv}^{\mathrm{T}} = \text{ rank one change in } A^{\mathrm{T}}A.$$

Why is that multiplication correct? The updated $\widehat{\boldsymbol{x}}_{\text{new}}$ comes from Steps 1 and 2.

For reference here are four formulas for M^{-1}. The first two were given above, when the change was $\boldsymbol{uv}^{\mathrm{T}}$. Formulas 3 and 4 go beyond rank one to allow matrices U, V, W.

1	$M = I - \boldsymbol{uv}^{\mathrm{T}}$	and	$M^{-1} = I + \boldsymbol{uv}^{\mathrm{T}}/(1 - \boldsymbol{v}^{\mathrm{T}}\boldsymbol{u})$ (*rank* 1 *change*)
2	$M = A - \boldsymbol{uv}^{\mathrm{T}}$	and	$M^{-1} = A^{-1} + A^{-1}\boldsymbol{uv}^{\mathrm{T}}A^{-1}/(1 - \boldsymbol{v}^{\mathrm{T}}A^{-1}\boldsymbol{u})$
3	$M = I - \boldsymbol{UV}$	and	$M^{-1} = I_n + U(\boldsymbol{I_m} - \boldsymbol{VU})^{-1}V$
4	$M = A - \boldsymbol{UW^{-1}V}$	and	$M^{-1} = A^{-1} + A^{-1}U(\boldsymbol{W} - \boldsymbol{VA^{-1}U})^{-1}VA^{-1}$

Formula **4** is the "matrix inversion lemma" in engineering. Not seen until now! The Kalman filter for solving block tridiagonal systems uses formula **4** at each step.