

The linear model

The linear model

- ▶ the *linear model*

$$y = Ax + w$$

- ▶ x and w are independent
- ▶ The matrix A is $m \times n$
- ▶ $x \sim \mathcal{N}(0, \Sigma_x)$ and $w \sim \mathcal{N}(0, \Sigma_w)$
- ▶ We measure $y = y_{\text{meas}}$ and would like to estimate x

The linear map

$$y = Ax + b \rightarrow \mu_y = A\mu_x + b$$
$$\Sigma_y = A\Sigma_x A^T$$

- since $y = Ax + w$, we have

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} I & 0 \\ A & I \end{bmatrix} \begin{bmatrix} x \\ w \end{bmatrix}$$

- We have (x, y) is jointly Gaussian, with covariance

$$\begin{bmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_y \end{bmatrix} = \text{cov} \begin{bmatrix} x \\ y \end{bmatrix}$$
$$= \begin{bmatrix} I & 0 \\ A & I \end{bmatrix} \begin{bmatrix} \Sigma_x & 0 \\ 0 & \Sigma_w \end{bmatrix} \begin{bmatrix} I & 0 \\ A & I \end{bmatrix}^T$$
$$= \begin{bmatrix} \Sigma_x & \Sigma_x A^T \\ A\Sigma_x & A\Sigma_x A^T + \Sigma_w \end{bmatrix}$$

- we have $\text{cov}(y) = A\Sigma_x A^T + \Sigma_w$ where

- $A\Sigma_x A^T$ is 'signal covariance'
- Σ_w is 'noise covariance'

X, Y indep.

$$\mathbb{E}[f(x)g(y)]$$
$$= \mathbb{E}(f(x)) \cdot \mathbb{E}(g(y))$$

$$\text{Cov}(X, Y) =$$

$$\mathbb{E}((X - \mu_x)(Y - \mu_y))$$

$$= \underbrace{\mathbb{E}(X - \mu_x)}_0 \underbrace{\mathbb{E}(Y - \mu_y)}_0$$

$$= 0$$

Linear measurements with Gaussian noise

$$\mu_1 = \mu_x + \Sigma_{xy} \Sigma_y^{-1} (y_{\text{meas}} - \mu_y)$$

$$\Sigma_1 = \Sigma_x - \Sigma_{xy} \Sigma_y^{-1} \Sigma_{xy}^T$$

$$\begin{bmatrix} \Sigma_x & \Sigma_x A^T \\ A \Sigma_x & A \Sigma_x A^T + \Sigma_w \end{bmatrix}$$

- The MMSE estimate of x given $y = y_{\text{meas}}$ is

L

$$\hat{x}_{\text{mmse}} = \Sigma_x A^T (A \Sigma_x A^T + \Sigma_w)^{-1} y_{\text{meas}}$$

- because we know $\hat{x}_{\text{mmse}} = \Sigma_{xy} \Sigma_y^{-1} y_{\text{meas}}$
- The matrix $L = \Sigma_x A^T (A \Sigma_x A^T + \Sigma_w)^{-1}$ is called the *estimator gain*

Example: linear measurements with Gaussian noise

$$y \approx 2x$$
$$\rightarrow \hat{x} = \frac{1}{2} y$$

Suppose $y = 2x + w$, with

- ▶ prior covariance $\text{cov}(x) = 1$
- ▶ noise covariance $\text{cov}(w) = 3$
- ▶ the estimator is

$$x_{\text{mmse}} = \frac{2y_{\text{meas}}}{7}$$

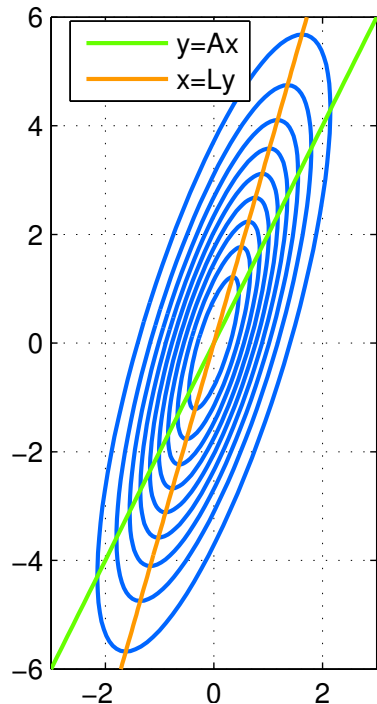
- ▶ The MMSE estimator gives a smaller answer than just inverting A ,

$$|x_{\text{mmse}}| \leq |A^{-1}y_{\text{meas}}|$$

since we have prior information about x

$$\hat{x}_{\text{mmse}} = \underbrace{\Sigma_x}_{1 \times 2} \underbrace{A^T}_{2 \times 1} \underbrace{(A \Sigma_x A^T + \Sigma_w)^{-1}}_{(2 \times 2 + 3)^{-1}} y_{\text{meas}} = \frac{2}{7} y$$

$$y \approx Ax$$
$$\hat{x} = A^{-1}y$$



Non-zero means

- ▶ Suppose $x \sim \mathcal{N}(\mu_x, \Sigma_x)$ and $w \sim \mathcal{N}(\mu_w, \Sigma_w)$.
- ▶ The MMSE estimate of x given $y = y_{\text{meas}}$ is

$$\hat{x}_{\text{mmse}} = \mu_x + \Sigma_x A^T (A \Sigma_x A^T + \Sigma_w)^{-1} (y_{\text{meas}} - A \mu_x - \mu_w)$$

The signal to noise ratio

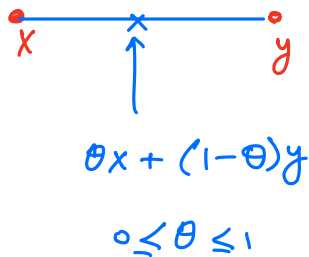
- Suppose where x , y and w are scalar, and $y = Ax + w$. The *signal-to-noise ratio* is

$$s = \frac{\sqrt{A^2 \Sigma_x}}{\sqrt{\Sigma_w}}$$

- Commonly used for scalar w, x, y ; no use in vector case
- In terms of s , the MMSE estimate is

$$\begin{aligned} x_{\text{mmse}} &= \mu_x + \frac{A \Sigma_x}{A^2 \Sigma_x + \Sigma_w} (y_{\text{meas}} - A \mu_x) \\ &= \frac{1}{1 + s^2} \mu_x + \frac{s^2}{1 + s^2} A^{-1} y_{\text{meas}} \end{aligned}$$

Scalar systems and the SNR



- The MMSE estimate is

$$x_{\text{mmse}} = \underbrace{\frac{1}{1+s^2}}_{\theta} \mu_x + \underbrace{\frac{s^2}{1+s^2}}_{1-\theta} A^{-1} y_{\text{meas}}$$

- let $\theta = \frac{1}{1+s^2}$, then $x_{\text{mmse}} = \theta \mu_x + (1 - \theta) A^{-1} y$
- a *convex linear combination* of the prior mean and the least-squares estimate
- when s is small, $x_{\text{mmse}} \approx \mu_x$, the *prior mean*
- when s is large, $x_{\text{mmse}} \approx A^{-1} y$, the *least-squares estimate* of y

Example: small noise

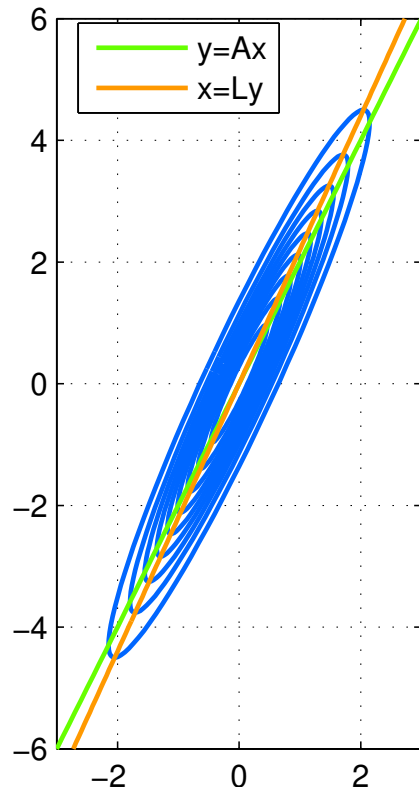
- ▶ Suppose $y = 2x + w$, with
 - ▶ prior covariance $\text{cov}(x) = 1$
 - ▶ noise covariance $\text{cov}(w) = 0.4$; signal is large compared to noise

- ▶ SNR $s = \frac{\sqrt{A^2 \Sigma_x}}{\sqrt{\Sigma_w}} \approx 3.2$

- ▶ Estimate is

$$\begin{aligned}x_{\text{mmse}} &= \frac{s^2}{1 + s^2} A^{-1} y_{\text{meas}} \\&\approx 0.9 A^{-1} y_{\text{meas}}\end{aligned}$$

i.e., close to $y_{\text{meas}}/2$



Example: large noise

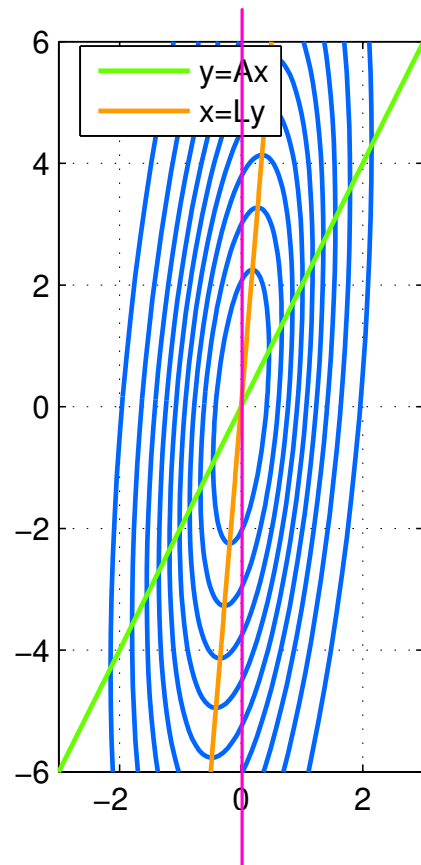
- ▶ Suppose $y = 2x + w$, with
 - ▶ prior covariance $\text{cov}(x) = 1$
 - ▶ noise covariance $\text{cov}(w) = 20$; signal is small compared to noise

- ▶ SNR $s = \frac{\sqrt{A^2 \Sigma_x}}{\sqrt{\Sigma_w}} \approx 0.45$

- ▶ Estimate is

$$\begin{aligned}x_{\text{mmse}} &= \frac{s^2}{1 + s^2} A^{-1} y_{\text{meas}} \\ &\approx 0.17 A^{-1} y_{\text{meas}}\end{aligned}$$

i.e., closer to 0 for all y_{meas}



The posterior covariance

- ▶ The posterior covariance of x given $y = y_{\text{meas}}$ is

$$\mathbf{cov}(x \mid y = y_{\text{meas}}) = \Sigma_x - \Sigma_x A^T (A \Sigma_x A^T + \Sigma_w)^{-1} A \Sigma_x$$

- ▶ above follows because

$$\mathbf{cov}(x \mid y = y_{\text{meas}}) = \Sigma_x - \Sigma_{xy} \Sigma_y^{-1} \Sigma_{yx}$$

- ▶ We can use this to compute the MSE since

$$\mathbf{E}(\|x - \hat{x}_{\text{mmse}}\|^2 \mid y = y_{\text{meas}}) = \mathbf{trace} \mathbf{cov}(x \mid y = y_{\text{meas}})$$

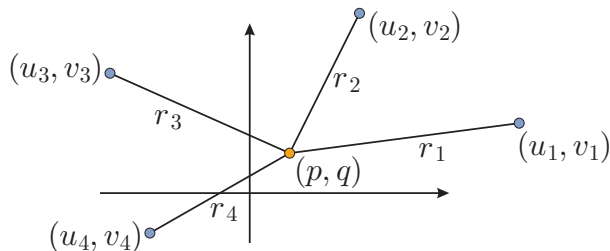
The posterior covariance and SNR

- For scalar problems, the posterior covariance of x given $y = y_{\text{meas}}$ is

$$\text{cov}(x \mid y = y_{\text{meas}}) = \frac{\Sigma_x}{1 + s^2}$$

- The *uncertainty* (covariance) in x is reduced by the factor $\frac{1}{1 + s^2}$ by measurement

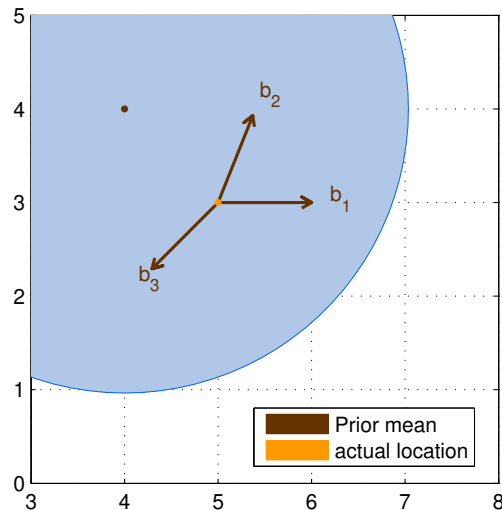
Example: navigation



- ▶ $x = \begin{bmatrix} p \\ q \end{bmatrix}$ our location, we measure distances r_i to m beacons at points (u_i, v_i)
- ▶ assume p, q are small compared to u_i, v_i . then, approximately $y = Ax$
- ▶ $A \in \mathbb{R}^{m \times 2}$, i th row of A is the transpose of unit vector in the direction of beacon i
- ▶ $y = \begin{bmatrix} \sqrt{u_1^2 + v_1^2} - r_1 \\ \vdots \\ \sqrt{u_m^2 + v_m^2} - r_m \end{bmatrix}$ measured vector of distances

Example: navigation

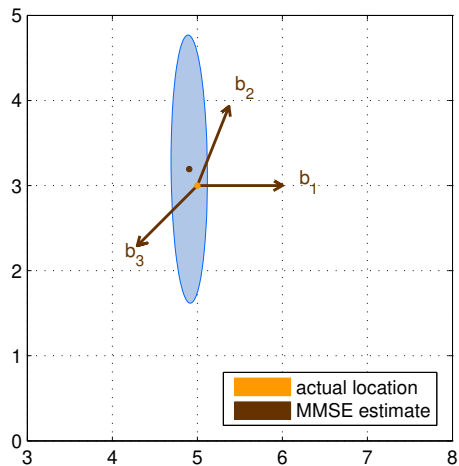
- ▶ here $A \in \mathbb{R}^{3 \times 2}$ with $A = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$
- ▶ $y = Ax$. Each b_i is a unit vector.
- ▶ Prior information is $x \sim \mathcal{N}\left(\begin{bmatrix} 4 \\ 4 \end{bmatrix}, \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}\right)$
- ▶ y is measured; y_i is range measurement in the direction b_i with noise w added
- ▶ beacons at $\begin{bmatrix} 50 \\ 0 \end{bmatrix}$, $\begin{bmatrix} 20 \\ 50 \end{bmatrix}$, $\begin{bmatrix} -50 \\ -50 \end{bmatrix}$
- ▶ figure shows prior 90% confidence ellipsoid



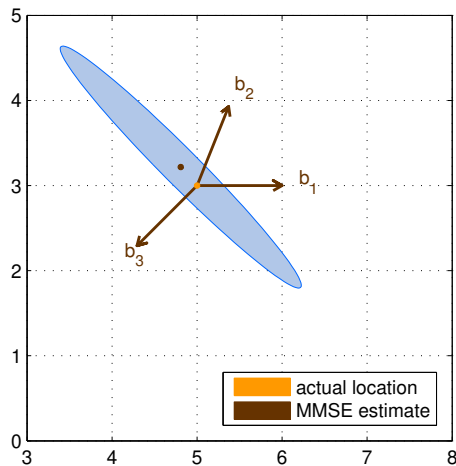
Example: posterior confidence ellipsoids

Posterior confidence ellipsoids for two different possible noise covariances.

$$\Sigma_w = \begin{bmatrix} 0.01 & & \\ & 1 & \\ & & 1 \end{bmatrix}$$



$$\Sigma_w = \begin{bmatrix} 1 & & \\ & 1 & \\ & & 0.01 \end{bmatrix}$$



Alternative formula

- ▶ There is another way to write the posterior covariance:

$$\mathbf{cov}(x | y = y_{\text{meas}}) = (\Sigma_x^{-1} + A^T \Sigma_w^{-1} A)^{-1}$$

- ▶ follows from the *Sherman-Morrison-Woodbury* formula

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

- ▶ This is very useful when we have fewer unknowns than measurements; i.e., Σ_x is smaller than $A\Sigma_x A^T$

Alternative formula

- There is also an alternative formula for the estimator gain

$$L = (\Sigma_x^{-1} + A^T \Sigma_w^{-1} A)^{-1} A^T \Sigma_w^{-1}$$

- Because

$$\begin{aligned} L &= \Sigma_x A^T (A \Sigma_x A^T + \Sigma_w)^{-1} \\ &= \Sigma_x A^T (\Sigma_w^{-1} A \Sigma_x A^T + I)^{-1} \Sigma_w^{-1} \\ &= (\Sigma_x A^T \Sigma_w^{-1} A + I)^{-1} \Sigma_x A^T \Sigma_w^{-1} && \text{by push-through identity} \\ &= (A^T \Sigma_w^{-1} A + \Sigma_x^{-1})^{-1} A^T \Sigma_w^{-1} \end{aligned}$$

Comparison with least-squares

- The least-squares approach minimizes

$$\|y - Ax\|^2 = \sum_{i=1}^m (y_i - a_i^T x)^2$$

where $A = \begin{bmatrix} a_1 & a_2 & \dots & a_m \end{bmatrix}^T$

- suppose instead we minimize

$$\sum_{i=1}^m w_i (y_i - a_i^T x)^2$$

where $w_1, w_2, \dots, w_m$ are positive *weights*

Weighted norms

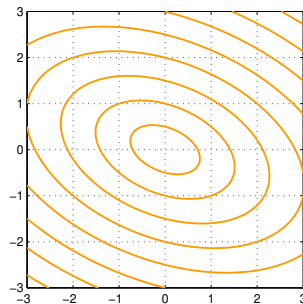
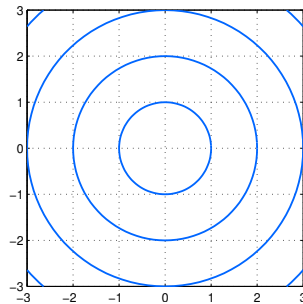
- ▶ More generally, let's look at *weighted norms*
- ▶ contours of the 2-norm
- ▶ contours of the *weighted-norm*

$$\|x\|_2 = \sqrt{x^T x}$$

$$\|x\|_W = \sqrt{x^T W x}$$

$$= \|W^{\frac{1}{2}} x\|_2$$

$$\text{where } W = \begin{bmatrix} 2 & 1 \\ 1 & 4 \end{bmatrix}$$



Weighted least squares

- ▶ the *weighted least-squares* problem; given $y_{\text{meas}} \in \mathbb{R}^m$,

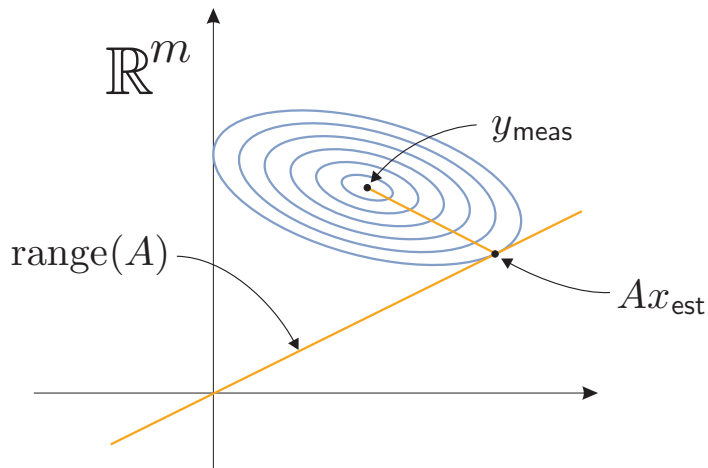
$$\text{minimize} \quad \|y_{\text{meas}} - Ax\|_W$$

- ▶ assume $A \in \mathbb{R}^{m \times n}$, skinny, full rank, and $W \in \mathbb{R}^{m \times m}$ and $W > 0$

- ▶ (by differentiating) the optimum x is

$$x_{\text{wls}} = (A^T W A)^{-1} A^T W y_{\text{meas}}$$

Weighted least squares



- ▶ if there is no noise, y lies in **range** A
- ▶ the weighted least-squares estimate x_{wls} minimizes

$$\|y_{\text{meas}} - Ax\|_W$$

- ▶ Ax_{wls} is the closest (in weighted-norm) point in **range** A to y_{meas}

MMSE and weighted least squares

- ▶ suppose we choose weight $W = \Sigma_w^{-1}$; then WLS solution is

$$x_{\text{wls}} = (A^T \Sigma_w^{-1} A)^{-1} A^T \Sigma_w^{-1} y_{\text{meas}}$$

- ▶ compare with MMSE estimate when $x \sim \mathcal{N}(0, \Sigma_x)$ and $w \sim \mathcal{N}(0, \Sigma_w)$

$$x_{\text{mmse}} = (\Sigma_x^{-1} + A^T \Sigma_w^{-1} A)^{-1} A^T \Sigma_w^{-1} y_{\text{meas}}$$

- ▶ as the prior covariance $\Sigma_x \rightarrow \infty$, the MMSE estimate tends to the WLS estimate
- ▶ if $\Sigma_w = I$ then MMSE tends to usual least-squares solution as $\Sigma_x \rightarrow \infty$
- ▶ the weighted norm heavily penalizes the residual $y - Ax$ in low-noise directions