17 - Bias and variance

- Bias of the MMSE
- Conditional bias
- Conditional covariance
- Conditional mean-variance decomposition
- Conditional MSE and estimator gain
- The conditional pdf of the estimate
- Decomposition of the MSE
- The bias-variance trade-off
- The linear model
- The Fisher estimator
- Example: navigation
- Example: convolution

Bias of the MMSE

Suppose x, y are Gaussian, with

$$\mathbf{E}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix} \qquad \mathbf{cov}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_y \end{bmatrix}$$

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

$$\phi(y_{\text{meas}}) = \mathbf{E}(x \,|\, y = y_{\text{meas}})$$

It is *unbiased*, since the *error* $z = \phi(y) - x$ satisfies

$$\mathbf{E}\, z = 0$$

Conditional error

Consider a *linear* estimator

$$\phi(y) = \mu_x + L(y - \mu_y)$$

Then the mean error conditioned on $x = a$ is

$$\boxed{\mathbf{E}(z \,|\, x = a) = (L\Sigma_{yx}\Sigma_x^{-1} - I)(a - \mu_x)}$$

because

$$\begin{aligned}
\mathbf{E}(z \,|\, x = a) &= \mathbf{E}(\mu_x - x + L(y - \mu_y) \,|\, x = a) \\
&= \mu_x - a - L\mu_y + L\,\mathbf{E}(y \,|\, x = a) \\
&= \mu_x - a - L\mu_y + L\big(\mu_y + \Sigma_{yx}\Sigma_x^{-1}(a - \mu_x)\big)
\end{aligned}$$

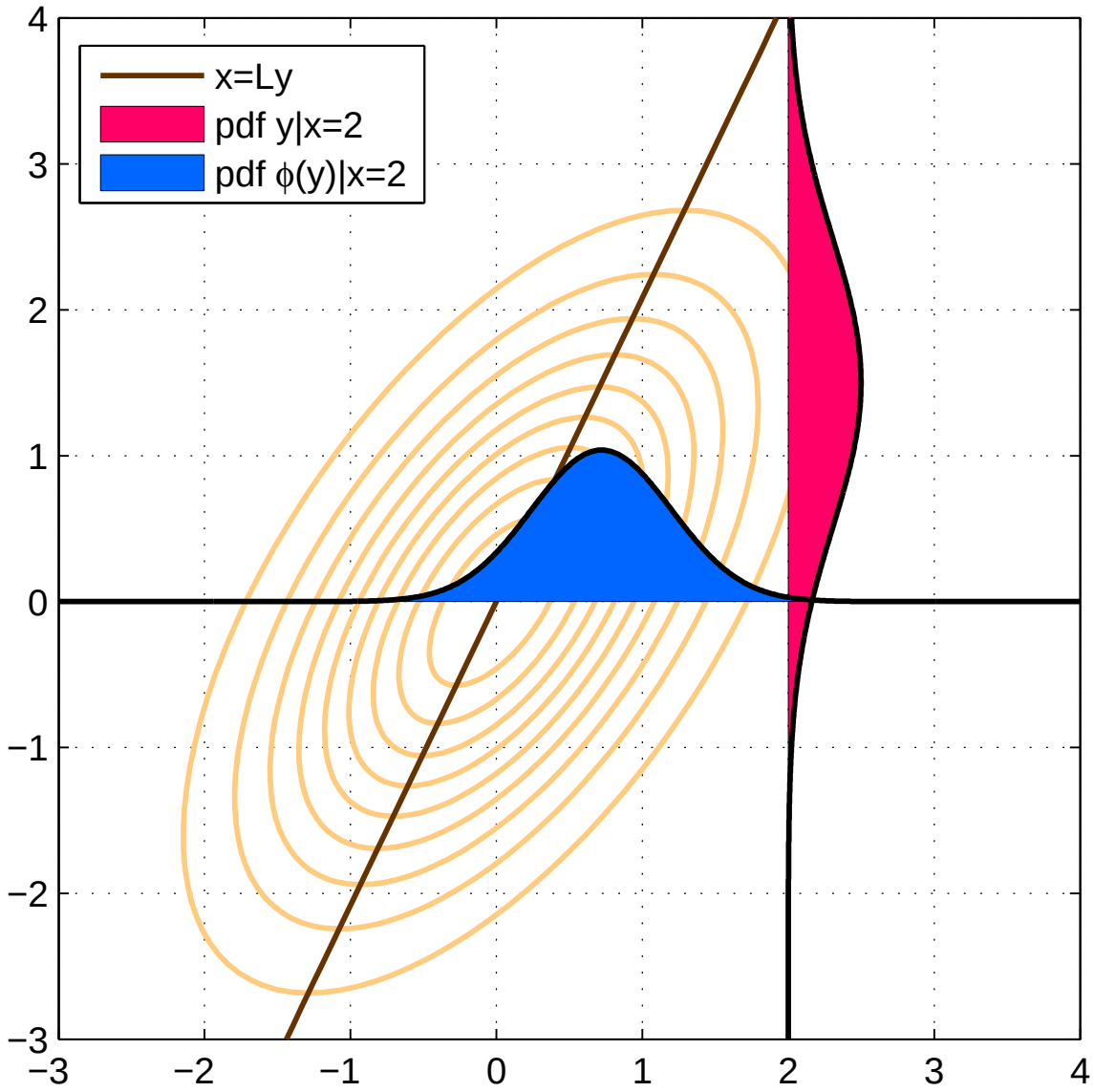
Conditional bias

- Given that $x = 2$,

  the mean estimate is not 2.

$$\mathbf{E}(\phi(y) \,|\, x = 2) \neq 2$$

- That is, the *conditional bias* is nonzero.



Covariance of the conditional error

The covariance of the error conditioned on $x = a$ is

$$\boxed{\mathbf{cov}(z \,|\, x = a) = L(\Sigma_y - \Sigma_{yx}\Sigma_x^{-1}\Sigma_{xy})L^T}$$

A measure of how much the *estimate* $\phi(y)$ varies, given $x = a$

because

$$\begin{aligned}
\mathbf{cov}(z \,|\, x = a) &= \mathbf{cov}\big(\mu_x - x + L(y - \mu_y) \,|\, x = a\big) \\
&= \mathbf{cov}\big(Ly \,|\, x = a\big) \\
&= L\,\mathbf{cov}\big(y \,|\, x = a\big)L^T \\
&= L(\Sigma_y - \Sigma_{yx}\Sigma_x^{-1}\Sigma_{xy})L^T
\end{aligned}$$

Conditional mean-variance decomposition

We can write the MSE conditioned on $x = a$ as

$$e_{\mathsf{given}x}(a) = \mathbf{E}\big(\|z\|^2 \,|\, x = a\big)$$

From the mean-variance decomposition, we have

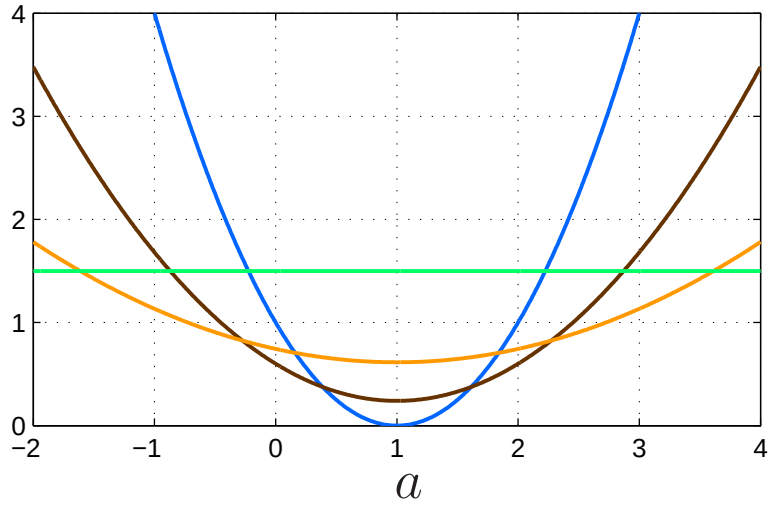
$$e_{\mathsf{given}x}(a) = \|\mathbf{E}(z \,|\, x = a)\|^2 + \mathbf{trace}\,\mathbf{cov}(z \,|\, x = a)$$

- The first term measures the *bias* of $\phi(y)$ conditioned on $x = a$
- The second term measures the variance of $\phi(y)$ conditioned on $x = a$

Conditional MSE and estimator gain

- $\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1.6 \end{bmatrix}$

$e_{\text{givenx}}(a)$



- When the estimator gain L is small

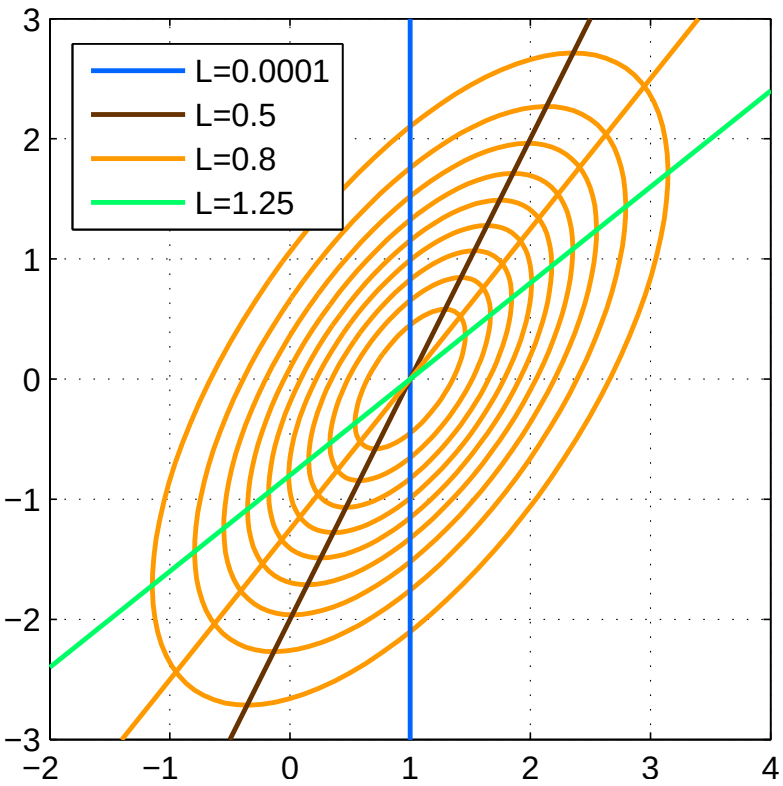
  the conditional MSE is small for a close to μ_x and large otherwise

- When the estimator gain L is increased

  the conditional MSE becomes flatter

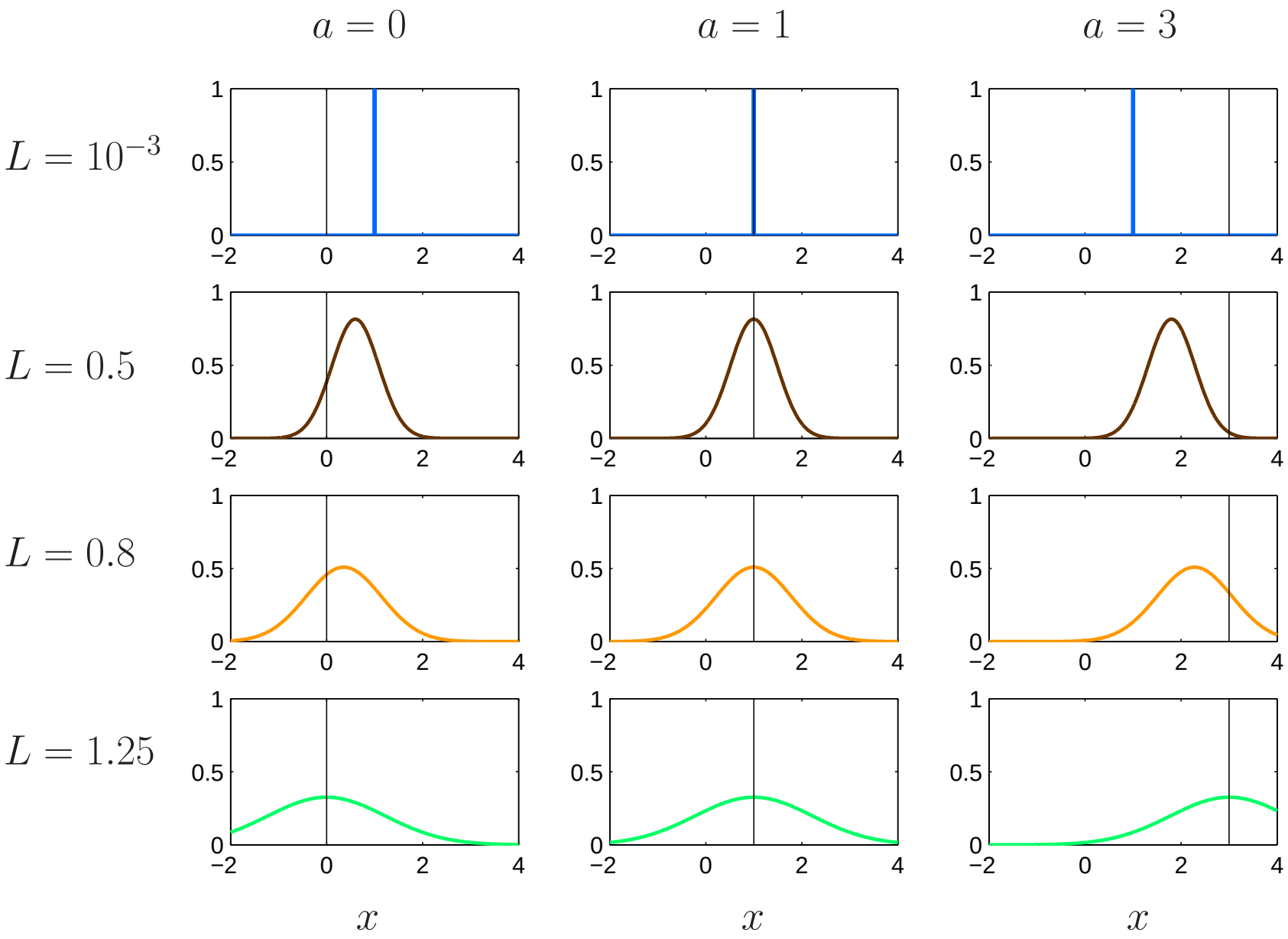
- When the gain satisfies $L^{-1} = \Sigma_{yx}\Sigma_x^{-1}$

  the conditional MSE is flat



The conditional pdf of the estimate

The pdf of $\phi(y) \,|\, x = a$ for various a and L values



Decomposition of the MSE

Define

$$b(a) = \|\mathbf{E}(z \,|\, x = a)\|^2 \qquad v(a) = \mathbf{trace}\,\mathbf{cov}(z \,|\, x = a)$$

The mean square error satisfies

$$\boxed{\mathbf{E}\big(\|\phi(y) - x\|^2\big) = \mathbf{E}\big(b(x)\big) + \mathbf{E}\big(v(x)\big)}$$

because

$$\mathbf{E}\big(\|\phi(y) - x\|^2\big) = \mathbf{E}\big(e_{\mathsf{givenx}}(x)\big)$$

The bias-variance trade-off

We'll minimize the multiobjective cost

$$J_1 + \mu J_2$$

Here

- $J_1 = \mathbf{E}\big(b(x)\big)$ is the *mean conditional bias*
- $J_2 = \mathbf{E}\big(v(x)\big)$ is the *mean conditional error variance*

The bias-variance trade-off

We have

$$J_1 = \mathbf{trace}\Big((L\Sigma_{yx}\Sigma_x^{-1} - I)\Sigma_x(L\Sigma_{yx}\Sigma_x^{-1} - I)^T\Big)$$

$$J_2 = \mathbf{trace}\Big(L(\Sigma_y - \Sigma_{yx}\Sigma_x^{-1}\Sigma_{xy})L^T\Big)$$

Hence

$$J_1 + \mu J_2 = \mathbf{trace}\left(\begin{bmatrix} L & I \end{bmatrix}\begin{bmatrix} \mu\Sigma_y + (1-\mu)\Sigma_{yx}\Sigma_x^{-1}\Sigma_{xy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_x \end{bmatrix}\begin{bmatrix} L^T \\ I \end{bmatrix}\right)$$

The bias-variance trade-off

Completing the square gives

$$L_{\mathsf{opt}} = \Sigma_{xy}\left(\mu\Sigma_y + (1-\mu)\Sigma_{yx}\Sigma_x^{-1}\Sigma_{xy}\right)^{-1}$$

The bias-variance trade-off

- As $\mu \to 0$ then L_{opt} becomes

$$L_{\text{opt}} = \lim_{\mu \to 0} \Sigma_{xy}\left(\mu\Sigma_y + (1-\mu)\Sigma_{yx}\Sigma_x^{-1}\Sigma_{xy}\right)^{-1}$$

when x, y are scalar, this is $L_{\text{opt}}^{-1} = \Sigma_{yx}\Sigma_x^{-1}$

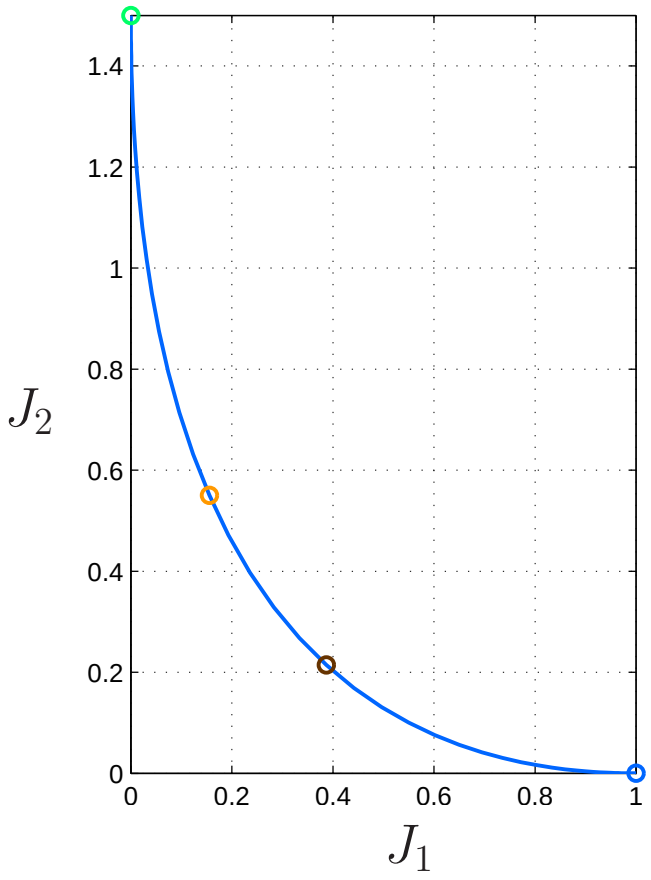
This estimator has $J_1 = 0$, so it is unbiased.

- If $\mu = 1$ then L_{opt} is the MMSE estimator

$$L_{\text{opt}} = \Sigma_{xy}\Sigma_y^{-1}$$

- As $\mu \to \infty$ we find $L \to 0$

This estimator has $J_2 = 0$, so it has zero conditional variance.



The linear model

Suppose $x \sim \mathcal{N}(\mu_x, \Sigma_x)$ and $w \sim \mathcal{N}(0, \Sigma_w)$, and

$$y = Ax + w$$

Then we have two formulae for L_{opt}

$$\begin{aligned} L_{\text{opt}} &= \Sigma_x A^T (A\Sigma_x A^T + \mu\Sigma_w)^{-1} \\ &= (\mu\Sigma_x^{-1} + A^T\Sigma_w^{-1}A)^{-1} A^T\Sigma_w^{-1} \end{aligned}$$

The Fisher estimator

Suppose A is skinny and full rank, i.e., more measurements than unknowns.

When $\mu = 0$, we have

$$L_{\text{opt}} = (A^T \Sigma_w^{-1} A)^{-1} A^T \Sigma_w^{-1}$$

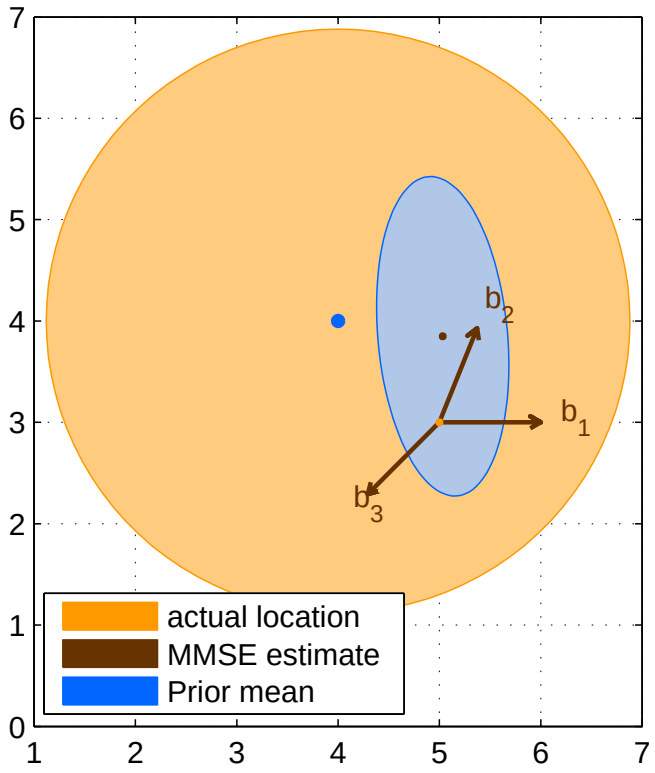
- Called the *Fisher* or *Gauss-Markov* estimator
- Does not depend on the prior covariance Σ_x
- Conditional mean of error $\mathbf{E}(z \mid x = a) = 0$ for all a
- L is a left inverse of A
- L is the pseudo-inverse of A when $\Sigma_w = I$

Example: navigation

- Prior information is

$$x \sim \mathcal{N}\left(\begin{bmatrix} 4 \\ 4 \end{bmatrix}, \begin{bmatrix} 1.8 & 0 \\ 0 & 1.8 \end{bmatrix} \right)$$

- beacons at $\begin{bmatrix} 50 \\ 0 \end{bmatrix}, \begin{bmatrix} 20 \\ 50 \end{bmatrix}, \begin{bmatrix} -50 \\ -50 \end{bmatrix}$

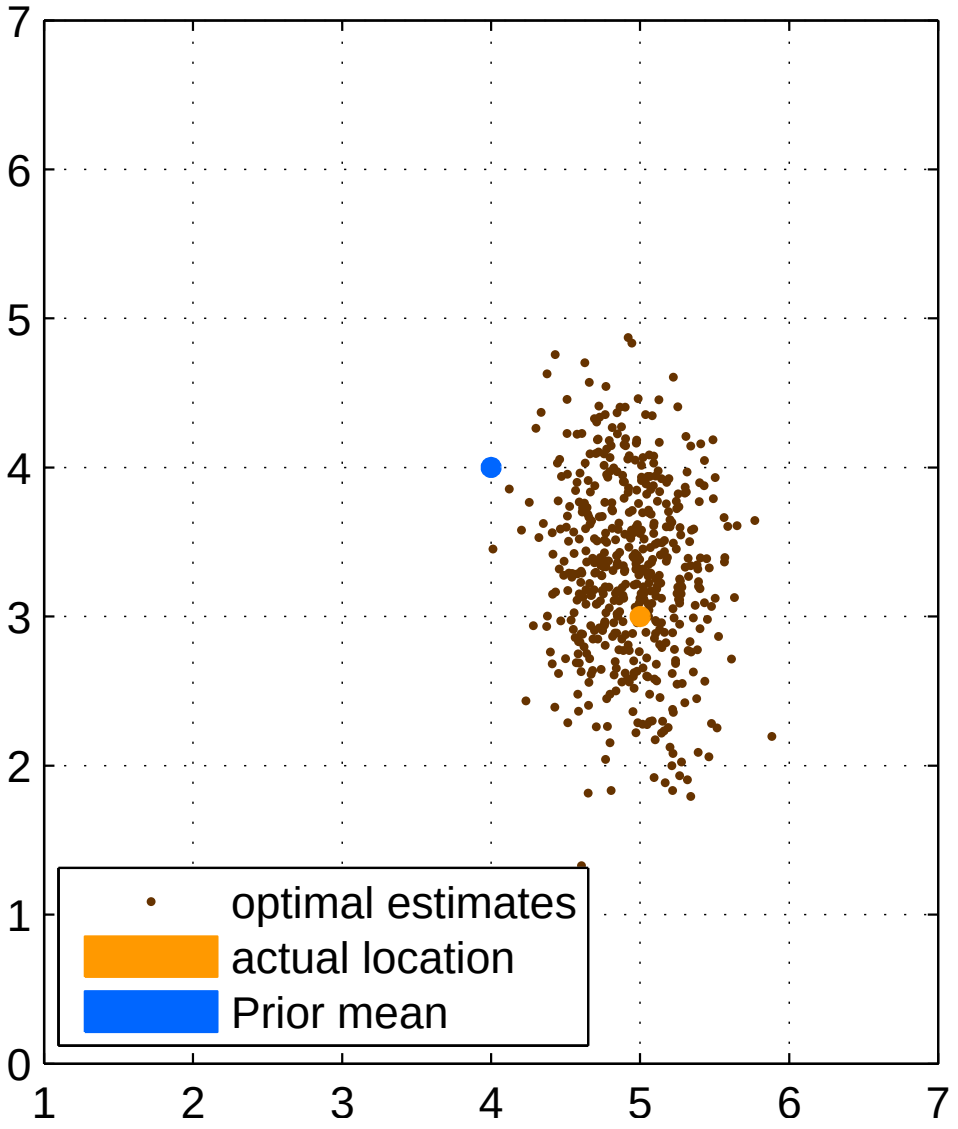


- figure shows prior 90% confidence ellipsoid
  and posterior 90% confidence ellipsoid using MMSE estimate

Example: MMSE estimates

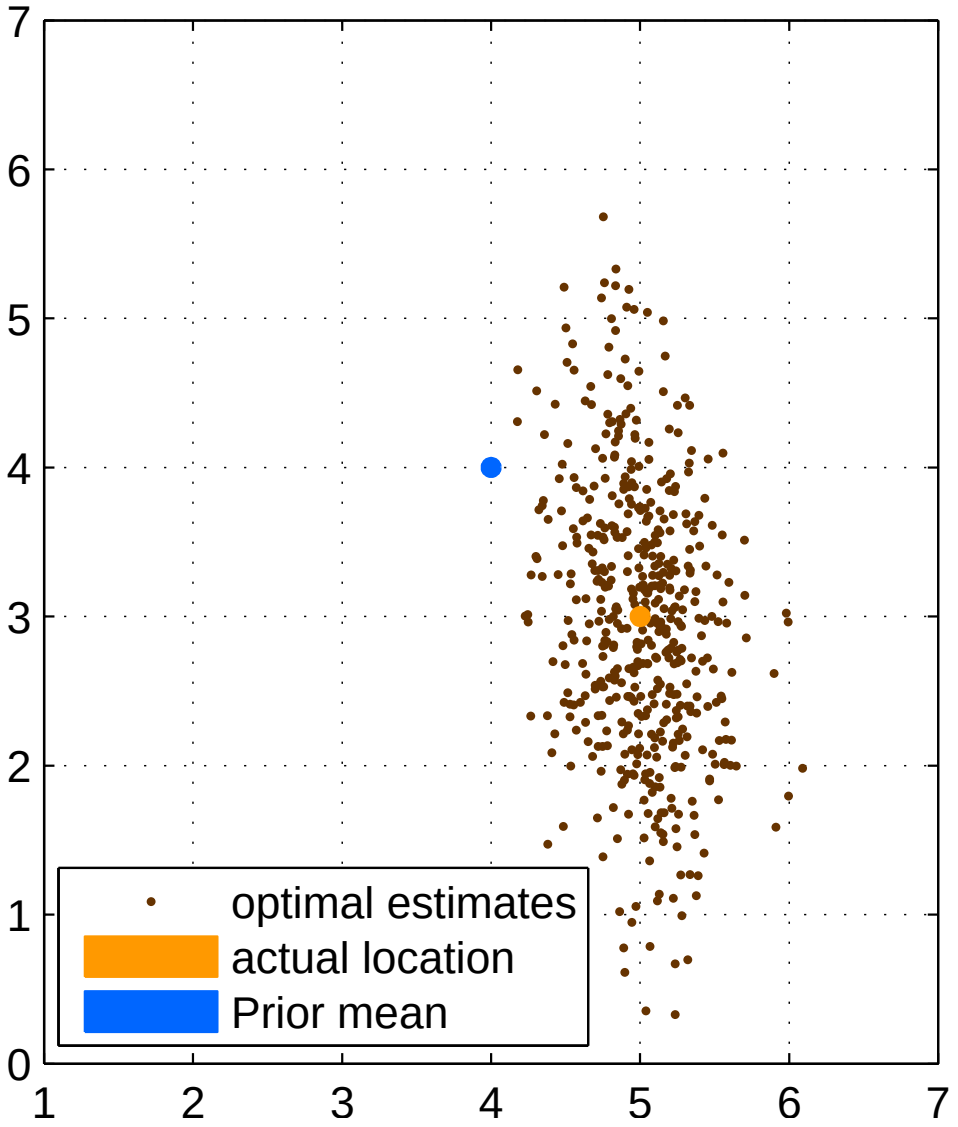
- 500 experiments
- Use MMSE estimator; i.e., $\mu = 1$
- In each, actual location is $\begin{bmatrix} 5 \\ 3 \end{bmatrix}$
- Prior mean is $\begin{bmatrix} 4 \\ 4 \end{bmatrix}$
- The center of the cloud of MMSE estimates is not at the true location

  It is biased towards the mean



Example: Fisher estimates

- Use Fisher estimator; i.e., $\mu = 0$
- The center of the cloud of estimates is at the true location
- But we see much more variance in the estimates

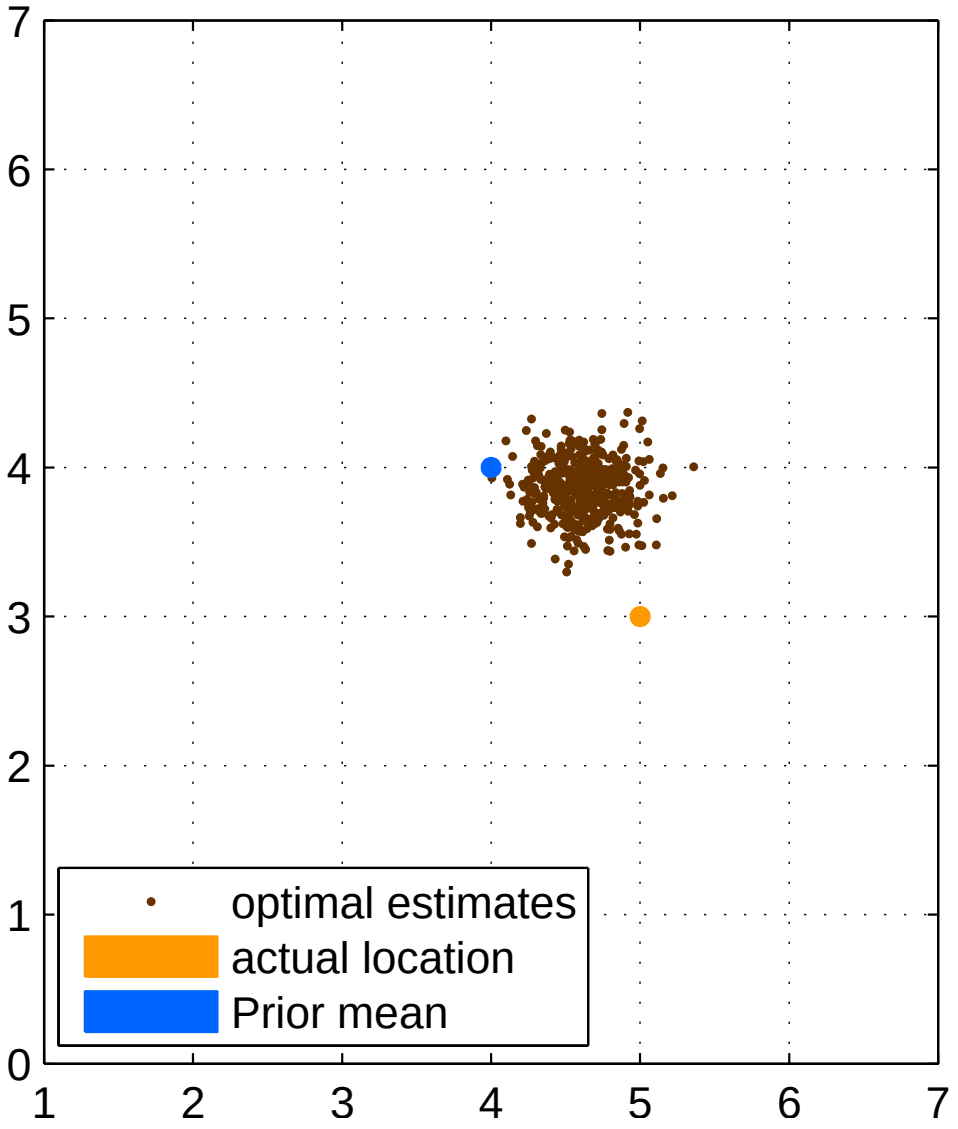


The Fisher estimate does not minimize the MMSE

But it also doesn't bias the answer towards prior mean

Example: another estimator

- Use $\mu = 10$
- The center of the cloud of estimates is very biased towards the mean
- But we see much less variance in the estimates

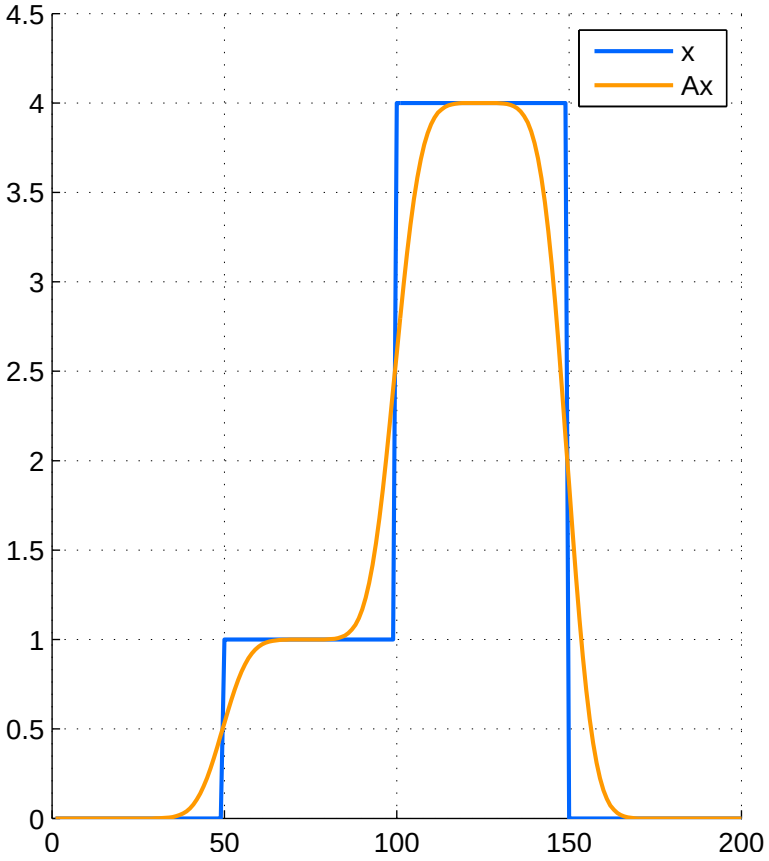


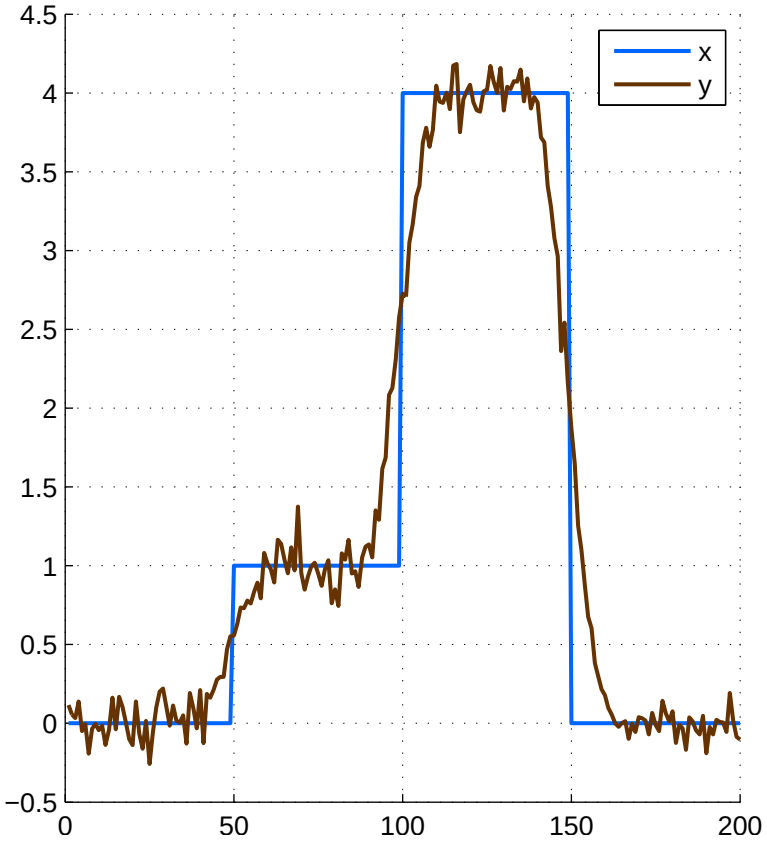
Example: convolution

Consider the convolution filter system

$$y = Ax + w$$

where A is a Toeplitz matrix $A_{ij} = c_{i-j}$





Example: convolution

For various μ, the regularized estimate is below.