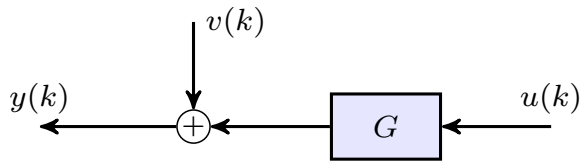
System Identification

Lecture 2: Data fitting and statistics

Roy Smith

Averaging estimates

Modeling framework



R experiments:

True system: $y_r = G_0\, u_r + v_r, \quad v_r \sim \mathcal{N}(0, \sigma_v^2), \quad v_r$ is i.i.d.

Each estimate comes from a single input-output pair: (y_r, u_r)

$$\hat{G}_r = \frac{y_r}{u_r} = G_0 + \frac{v_r}{u_r}, \qquad r = 1, \ldots, R.$$

Bias: $\mathcal{E}\Big\{\hat{G}_r - G_0\Big\} = 0$ Variance: $\mathcal{E}\Big\{(\hat{G}_r - G_0)^2\Big\} = \dfrac{\sigma_v^2}{|u_r|^2} =: \sigma_r^2.$

Averaging estimates

Averaging estimates (linear weighting)

Can we combine R unbiased estimates to get a better estimate?

$$\hat{G}_{\text{avg}} = \sum_{r=1}^{R} \alpha_r \hat{G}_r, \quad \alpha_r \geqslant 0, \qquad \text{(sample mean estimate: } \alpha_r = 1/R\text{)}$$

Bias

$$\text{Bias} = \mathcal{E}\left\{\hat{G}_{\text{avg}}\right\} - G_0 = \mathcal{E}\left\{\sum_{r=1}^{R} \alpha_r \hat{G}_r\right\} - G = \left(\sum_{r=1}^{R} \alpha_r - 1\right) G_0.$$

The averaged estimate, $\hat{G}_{\text{avg}}$, is unbiased if and only if $\sum_{r=1}^{R} \alpha_r = 1$.

Why do we care about an estimate being unbiased?

Averaging estimates

Averaging estimates (linear weighting)

$$\hat{G}_{\text{avg}} = \sum_{r=1}^{R} \alpha_r \hat{G}_r$$

Variance

$$\text{Variance} = \mathcal{E}\left\{\left(\hat{G}_{\text{avg}} - \mathcal{E}\left\{\hat{G}_{\text{avg}}\right\}\right)^2\right\} = \mathcal{E}\left\{\left|\sum_{r=1}^{R} \alpha_r \left(G + \frac{v_r}{u_r}\right) - \sum_{r=1}^{R} \alpha_r G\right|^2\right\}$$

$$= \mathcal{E}\left\{\left|\sum_{r=1}^{R} \alpha_r \frac{v_r}{u_r}\right|^2\right\} = \sum_{r=1}^{R} \alpha_r^2 \sigma_r^2$$

$$= \sum_{r=1}^{R} \alpha_r^2 \frac{\sigma_v^2}{|u_r|^2}.$$

Averaging estimates

Best Linear Unbiased Estimator (BLUE)

Choose the α_r to make $\hat{G}_{\text{avg}}$ unbiased and also have the minimum variance.

Optimisation framework: $\underset{\alpha_r}{\text{minimise}} \quad \sum_{r=1}^{R} \alpha_r^2 \sigma_r^2$

$$\text{subject to} \quad \left(1 - \sum_{r=1}^{R} \alpha_r\right) = 0.$$

Lagrangian: $L(\alpha_r, \lambda) = \sum_{r=1}^{R} \alpha_r^2 \, \sigma_r^2 + \lambda \left(1 - \sum_{r=1}^{R} \alpha_r\right).$

$$\frac{\partial L(\alpha_r, \lambda)}{\partial \alpha_r} = 2\alpha_r \sigma_r^2 - \lambda = 0, \quad \Longrightarrow \quad \alpha_r = \frac{\lambda}{2\sigma_r^2}, \quad r = 1, \ldots, R.$$

$$\frac{\partial L(\alpha_r, \lambda)}{\partial \lambda} = 0 \quad \Longrightarrow \quad \sum_{r=1}^{R} \alpha_r = 1 \quad \Longrightarrow \quad \lambda = \frac{2}{\sum_{r=1}^{R} \frac{1}{\sigma_r^2}}$$

Averaging estimates

Best Linear Unbiased Estimator (BLUE)

Solving for α_r gives,

$$\alpha_r = \frac{\frac{1}{\sigma_r^2}}{\sum_{r=1}^{R} \frac{1}{\sigma_r^2}} = \frac{|u_r|^2}{\sum_{r=1}^{R} |u_r|^2}.$$

The minimum variance weights are proportional to the inverse of the variance.

Low variance estimates have greater weight.

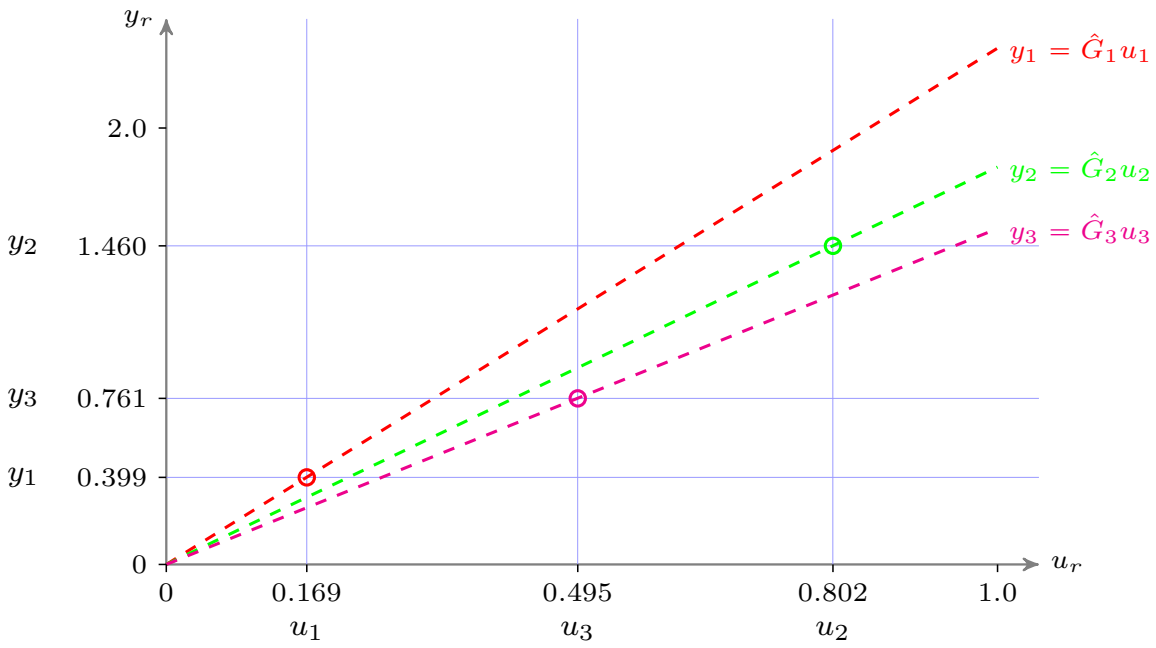
The resulting variance is:

$$\text{var}\left(\hat{G}_{\text{avg}}\right) = \frac{1}{\sum_{r=1}^{R} \frac{1}{\sigma_r^2}} = \frac{\sigma_v^2}{\sum_{r=1}^{R} |u_r|^2}.$$

The variance decreases as $R \longrightarrow \infty$, implying that the estimator is consistent.

Example

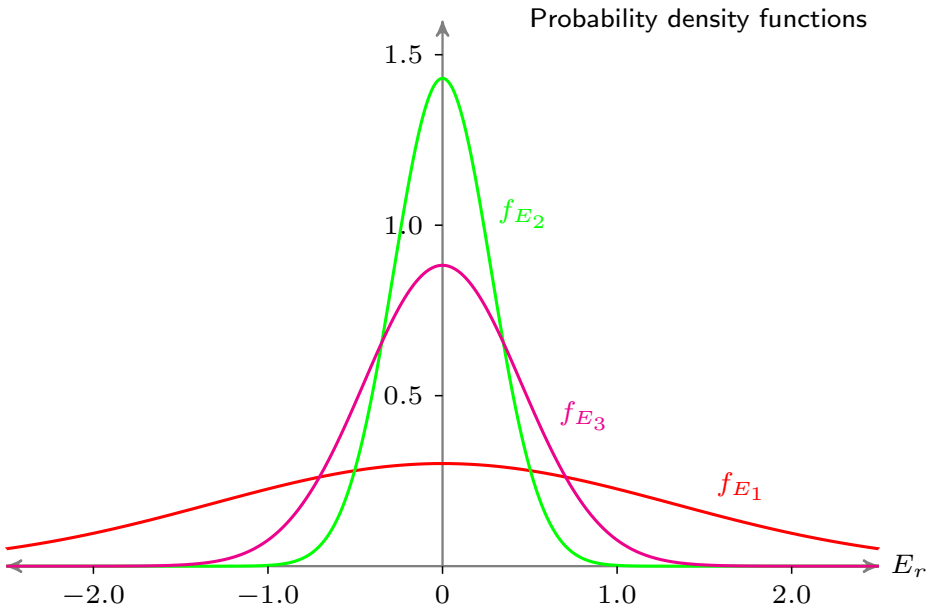
$$y = G_0\, u + v, \qquad v \sim \mathcal{N}(0, 0.05), \qquad (\sigma_v^2 = 0.05)$$



Example

Error distributions: $E_r = \hat{G}_r - G_0$

$$\mathcal{E}\{E_r\} = 0, \qquad \sigma_{E_r}^2 = \frac{\sigma_v^2}{|u_r|^2}, \quad r = 1, 2, 3.$$

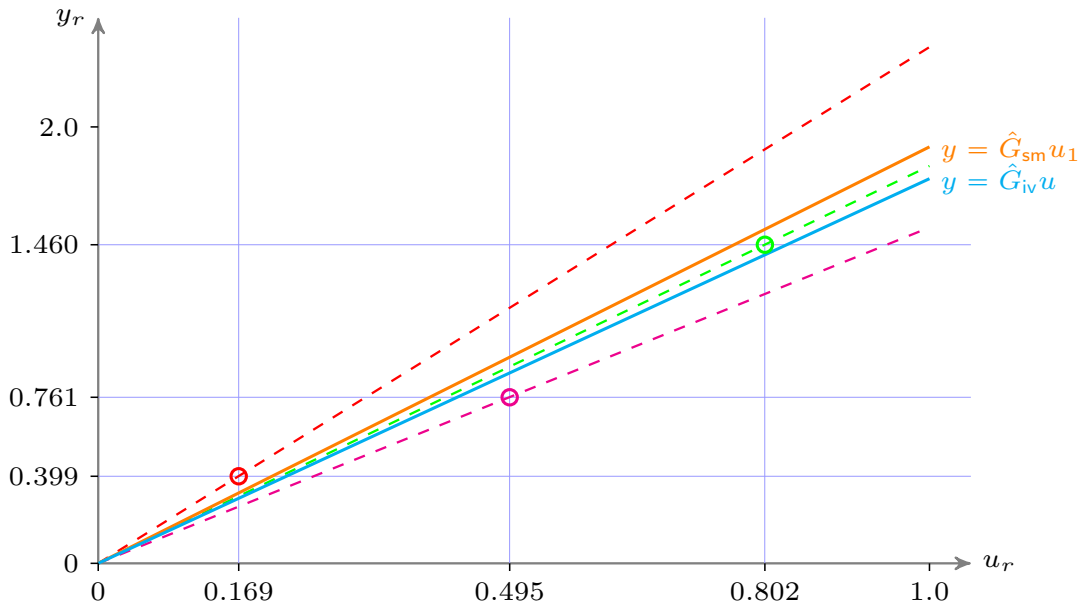


Example

Averaging: estimated models

Sample mean $\hat{G}_{\text{sm}} : \alpha_r = \dfrac{1}{R}$, Inverse variance $\hat{G}_{\text{iv}} : \alpha_r = \dfrac{1/\sigma_r^2}{\sum_{r=1}^{R} 1/\sigma_r^2}$.

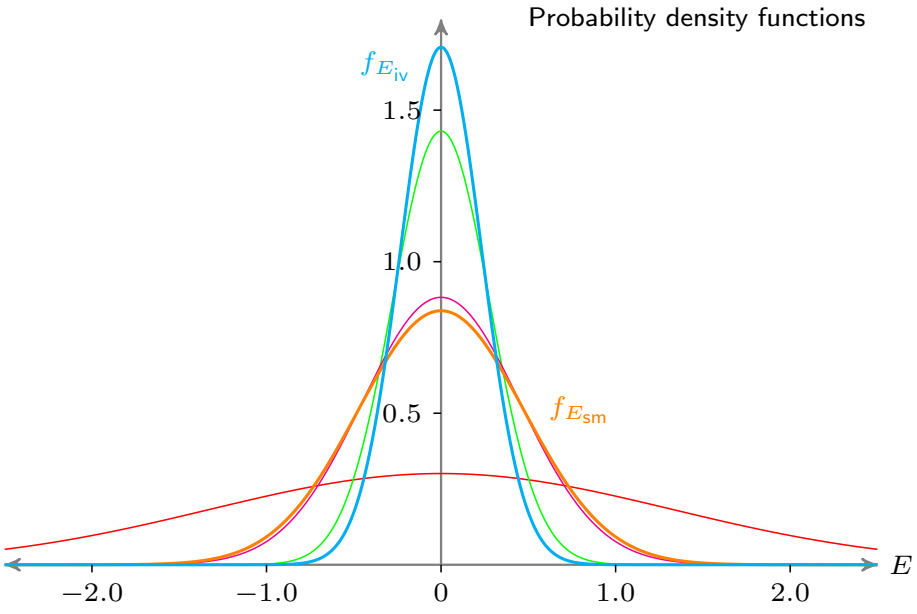
$\hat{G}_{\text{sm}} = 1.9087$ $\hat{G}_{\text{iv}} = 1.7627$



Example

Averaging: error variances

Sample mean $\hat{G}_{\text{sm}} : \alpha_r = \dfrac{1}{R}$, Inverse variance $\hat{G}_{\text{iv}} : \alpha_r = \dfrac{1/\sigma_r^2}{\sum_{r=1}^{R} 1/\sigma_r^2}$.



$\sigma_{\text{sm}}^2 = 0.2266$

$\sigma_{\text{iv}}^2 = 0.0546$

Averaging estimates

Minimising the Mean-Square Error (MSE)

Does unbiased and minimum variance imply minimum MSE?

$$\text{MSE}(\hat{G}_{\text{avg}}) \;=\; \text{bias}^2 \;+\; \text{variance} \;=\; \left(1 - \sum_{r=1}^{R} \alpha_r\right)^2 |G|^2 \;+\; \sum_{r=1}^{R} \alpha_r^2 \sigma_r^2.$$

Minimising this over $\alpha_r > 0$ gives,

$$\alpha_r \;=\; \frac{\dfrac{1}{\sigma_r^2}}{\dfrac{1}{|G|^2} + \displaystyle\sum_{r=1}^{R} \frac{1}{\sigma_r^2}} \;=\; \frac{|u_r|^2}{\dfrac{\sigma_v^2}{|G|^2} + \displaystyle\sum_{r=1}^{R} |u_r|^2}.$$

The optimal weighting, α_r, for minimising the MSE depends on σ_v^2 and $|G|^2$.

Averaging estimates

Minimising the Mean-Square Error (MSE)

Using the optimal α_r weighting gives the best possible MSE,

$$\text{MSE}(\hat{G}_{\text{avg}}) \;=:\; \frac{1}{\dfrac{1}{|G|^2} + \displaystyle\sum_{r=1}^{R} \frac{1}{\sigma_r^2}} \;=\; \frac{\sigma_v^2}{\dfrac{\sigma_v^2}{|G|^2} + \displaystyle\sum_{r=1}^{R} |u_r|^2}.$$

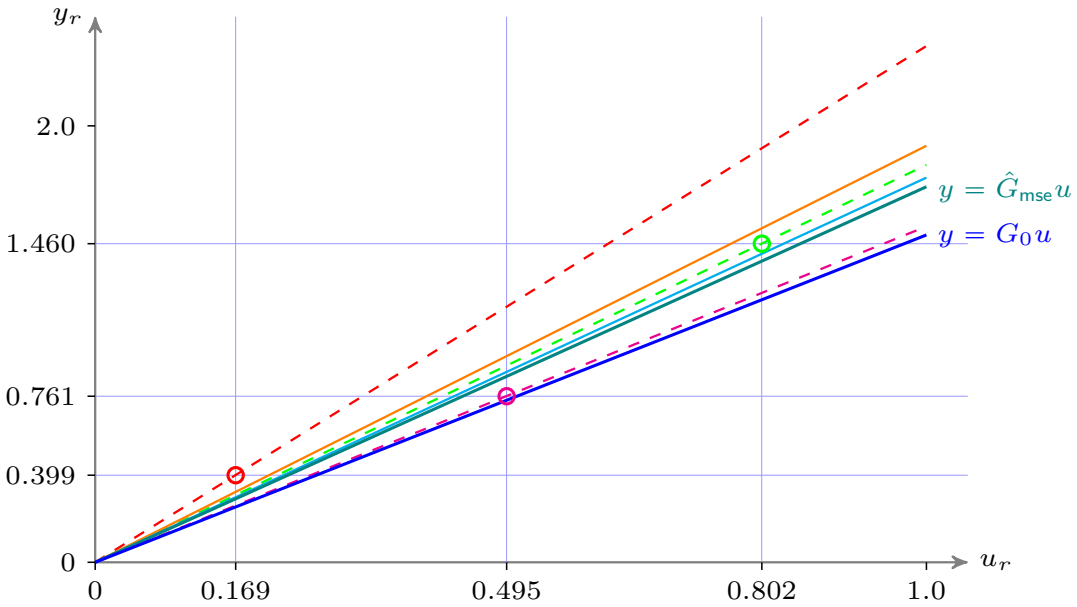
This is strictly less than the MSE for the minimum variance estimate.

Example

Averaging: error variances

Minimum MSE $\hat{G}_{\text{mse}}$: $\alpha_r = \dfrac{1/\sigma_r^2}{\sigma_v^2/|G_0|^2 + \sum_{r=1}^{R} 1/\sigma_r^2}.$

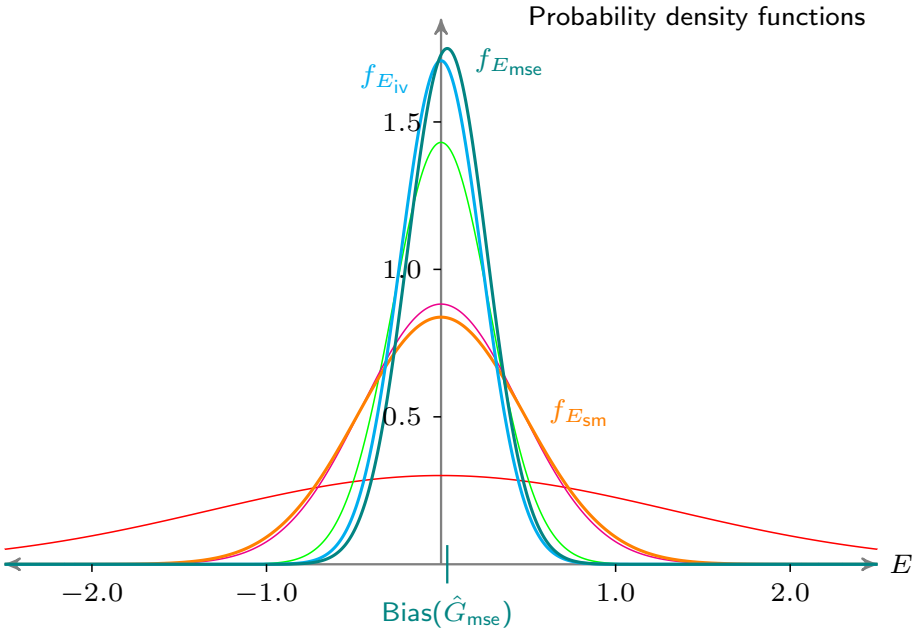
$\hat{G}_{\text{sm}} = 1.9087, \quad \hat{G}_{\text{iv}} = 1.7627, \quad \hat{G}_{\text{mse}} = 1.7209, \quad G_0 = 1.5$



Example

Averaging: error variances

Minimum MSE $\hat{G}_{\text{mse}}$: $\alpha_r = \dfrac{1/\sigma_r^2}{\sigma_v^2/|G_0|^2 + \sum_{r=1}^{R} 1/\sigma_r^2}.$



$\sigma_{\text{sm}}^2 = 0.2266$

$\sigma_{\text{iv}}^2 = 0.0546$

$\sigma_{\text{mse}}^2 = 0.0520$

$\text{Bias}(\hat{G}_{\text{mse}}) = 0.0355$

$\text{MSE}(\hat{G}_{\text{mse}}) = 0.0533$

Averaging estimates

Bias-variance trade-off

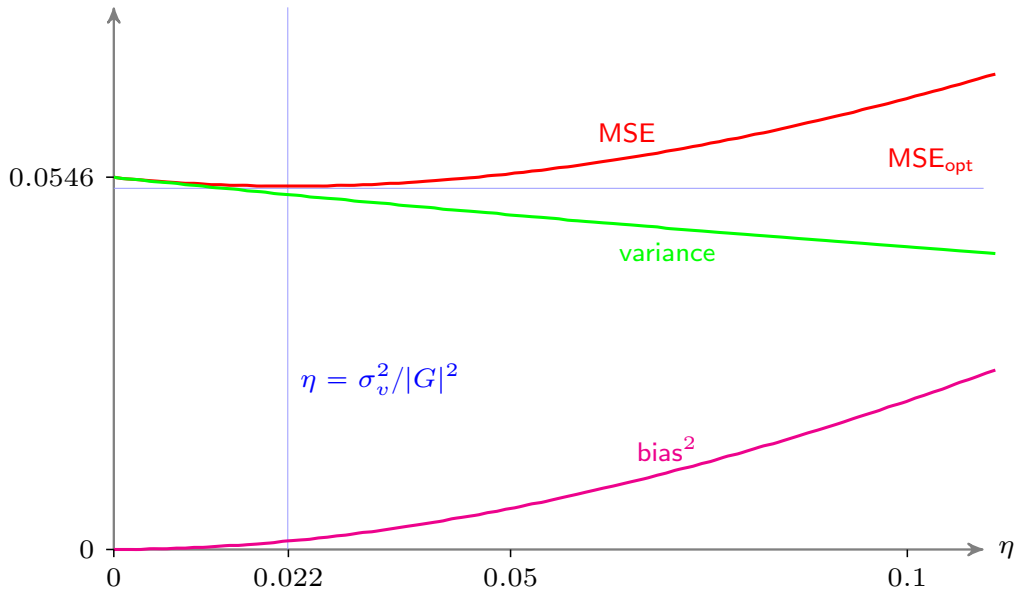
Parametrised family of estimates:

$$\alpha_r \;=\; \frac{|u_r|^2}{\eta \;+\; \sum_{r=1}^{R} |u_r|^2}, \qquad \eta \geqslant 0. \qquad \text{(} \eta \text{ is a hyperparameter)}$$

Ideally $\eta = \sigma_v^2/|G|^2$ and $\eta = 0$ corresponds to the minimum variance (unbiased) estimate.

Example

Bias-Variance trade-off



Averaging estimates

Picking η: validation

Averaging estimates: summary

$$\hat{G}_{\text{avg}} = \sum_{r=1}^{R} \alpha_r \hat{G}_r, \quad \alpha_r \geqslant 0,$$

Bias and Variance of $\hat{G}_{\text{avg}}$:

$$\text{Bias} = \left(\sum_{r=1}^{R} \alpha_r - 1 \right) G_0, \qquad \text{Variance} = \sum_{r=1}^{R} \alpha_r^2 \sigma_r^2.$$

Weights: α_r

Inverse variance weighting	Minimum MSE weighting
$\alpha_r = \frac{1}{\sigma_r^2} \left(\sum_{r=1}^{R} \frac{1}{\sigma_r^2} \right)^{-1}$	$\alpha_r = \frac{1}{\sigma_r^2} \left(\frac{1}{\lvert G \rvert^2} + \sum_{r=1}^{R} \frac{1}{\sigma_r^2} \right)^{-1}$
unbiased Best Linear Unbiased Estimator variance decrease order $1/R$	biased (finite R) asymptotically unbiased minimum MSE variance decrease order $1/R$

Parameter estimation statistics

Basic formulation

Model parameters: θ

True parameter: θ_0

Estimated parameter: $\hat{\theta}$

Maximum likelihood estimation

Basic formulation

Consider K observations, $z_1, \ldots, z_K$.

Each is a realisation of a random variable, with joint probability distribution,

$$f(\underbrace{x_1, \ldots, x_K}_{\text{random variables}}; \theta) \quad \longleftarrow \quad \text{family of distributions parametrised by } \theta.$$

Another common notation is,

$$f(x_1, \ldots, x_K \mid \theta) \quad \longleftarrow \quad \text{the pdf for } x_1, \ldots, x_K \text{ given } \theta.$$

For independent variables,

$$f(x_1, \ldots, x_K; \theta) = f_1(x_1; \theta) f_2(x_2; \theta) \cdots f_K(x_K; \theta) = \prod_{i=1}^{K} f_i(x_i; \theta)$$

Maximum likelihood estimation

Likelihood function

Substituting the observation, $Z_K = \{z_1, \ldots, z_K\}$, gives a function of θ,

$$\mathcal{L}(\theta) = f(x_1, \ldots, x_K; \theta)\Big|_{x_i = z_i, i=1,\ldots,K.} \qquad \text{(Likelihood function)}$$

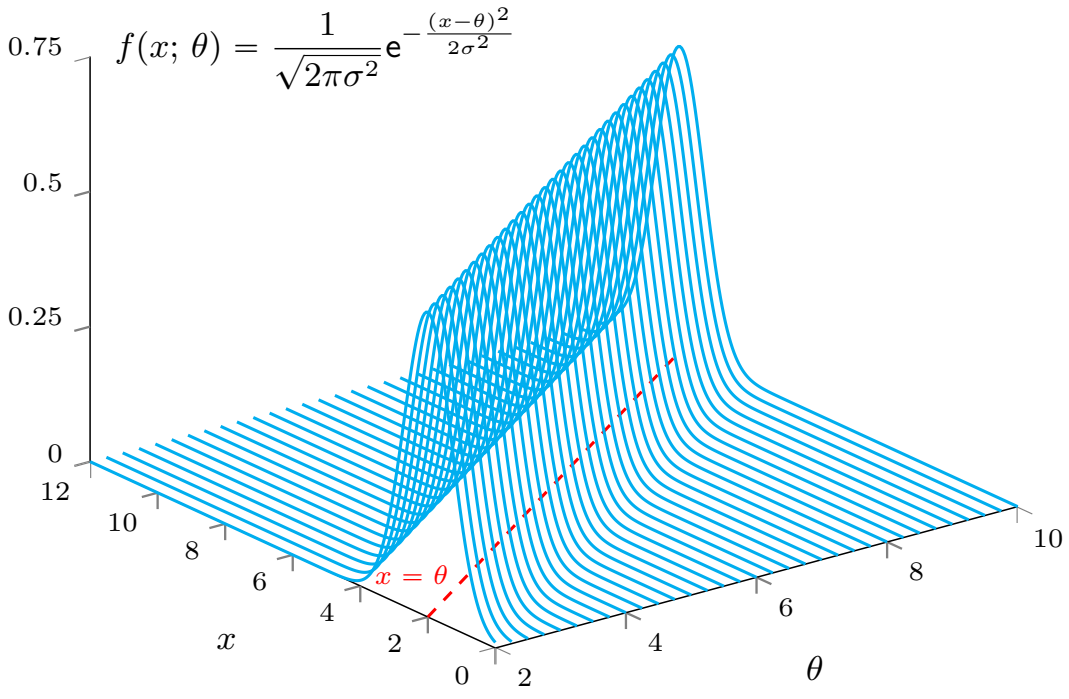
Maximum likelihood estimator:

$$\hat{\theta}_{\mathsf{ML}} = \operatorname*{argmax}_{\theta} \mathcal{L}(\theta).$$

The value chosen for θ is the one that gives the most "agreement" with the observation.

Maximum likelihood estimation

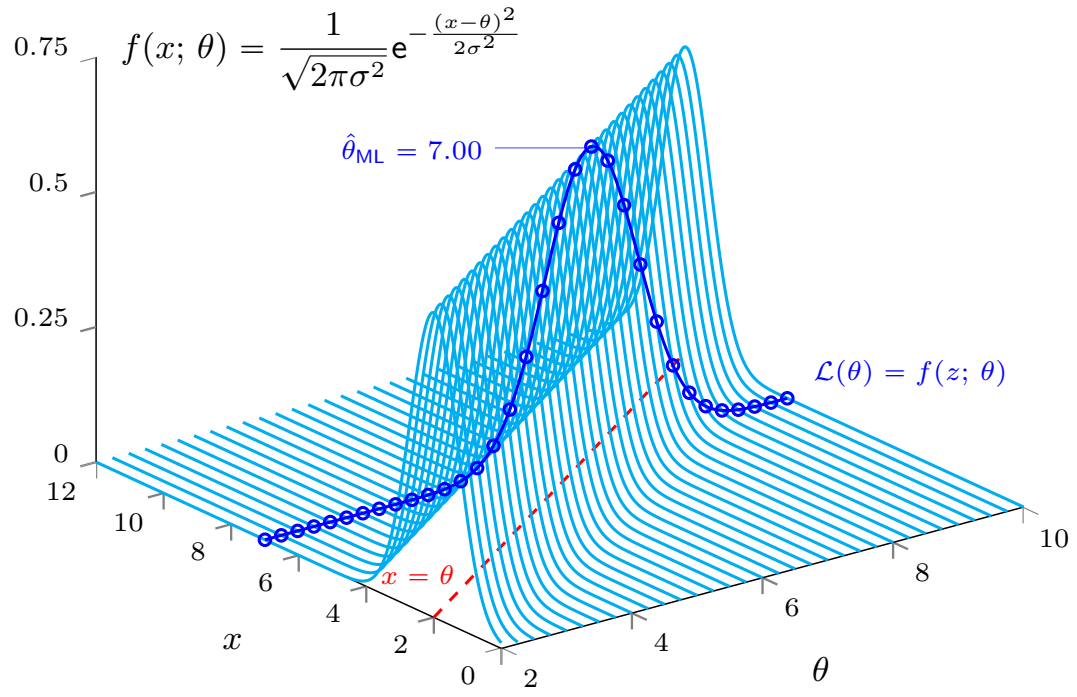
Estimating the mean of a Gaussian distribution ($\sigma^2 = 0.5$)



Maximum likelihood estimation

Estimating the mean of a Gaussian distribution ($\sigma^2 = 0.5$)

Datum: $z = 7.0$



Maximum likelihood estimation

Log-likelihood function

It is often mathematically easier to consider,

$$\hat{\theta}_{\mathsf{ML}} = \underset{\theta}{\operatorname{argmax}} \;\; \ln \mathcal{L}(\theta).$$

As the $\ln$ function is monotonic this gives the same θ.

This is typically the natural logarithm so as to be able to handle the exponentiation in typical pdfs.

Example

Estimation of the mean of a set of samples

$z_i, \quad i = 1, \ldots, K \qquad z_i \sim \mathcal{N}(\theta_0, \sigma_i^2).$ (note: different variances)

Sample mean estimate: $\hat{\theta}_{\text{SM}} = \dfrac{1}{K} \sum_{i=1}^{K} z_i$

Probability density functions (pdf): θ is the common mean of the distributions.

$$f_i(x_i;\, \theta) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(x_i - \theta)^2}{2\sigma_i^2} \right)$$

For independent samples the joint pdf is:

$$f(x_1, \ldots, x_K;\, \theta) = \prod_{i=1}^{K} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(x_i - \theta)^2}{2\sigma_i^2} \right)$$

Example

Estimation of the mean of a set of samples

$$\begin{aligned}
\theta_{\text{ML}} &= \underset{\theta}{\text{argmax}}\ \ln f(x_1, \ldots, x_K;\, \theta)\Big|_{x_i = z_i, i = 1, \ldots, K.} \\
&= \underset{\theta}{\text{argmax}}\ \ln \mathcal{L}(\theta) \\
&= \underset{\theta}{\text{argmax}} \left(-\frac{K}{2} \ln(2\pi) - \sum_{i=1}^{K} \frac{1}{2} \ln(\sigma_i^2) - \frac{1}{2} \sum_{i=1}^{K} \frac{(z_i - \theta)^2}{\sigma_i^2} \right)
\end{aligned}$$

This gives (differentiate and equate to zero),

$$\hat{\theta}_{\text{ML}} = \left(\frac{1}{\sum_{i=1}^{K} \frac{1}{\sigma_i^2}} \right) \sum_{i=1}^{K} \frac{z_i}{\sigma_i^2}$$

Bayesian approach

Random parameter framework

Thomas Bayes
1701–1761.

Consider θ to be a random variable with pdf: $f_\theta(x)$.

This is an a priori distribution (assumed before the experiment).

Conditional distribution (inference from the experiment)

Our model (plus assumptions) gives a conditional distribution,

$$f(x_1, \ldots, x_K \mid \theta)$$

On the basis of the experiment ($x_i = z_i$),

$$\text{Prob}(\theta \mid z_1, \ldots, z_K) = \frac{\text{Prob}(Z_K \mid \theta)\,\text{Prob}(\theta)}{\text{Prob}(Z_K)}$$

So,

$$\underset{\theta}{\text{argmax}}\, f(\theta \mid z_1, \ldots, z_K) \;=\; \underset{\theta}{\text{argmax}}\, f(Z_K \mid \theta) f_\theta(\theta)$$

Maximum a posteriori (MAP) estimation

Estimator

Given data, Z_K,

$$\hat{\theta}_{\text{MAP}} = \underset{\theta}{\text{argmax}}\; f(Z_K \mid \theta) f_\theta(\theta).$$

We can interpret the maximum likelihood estimator as,

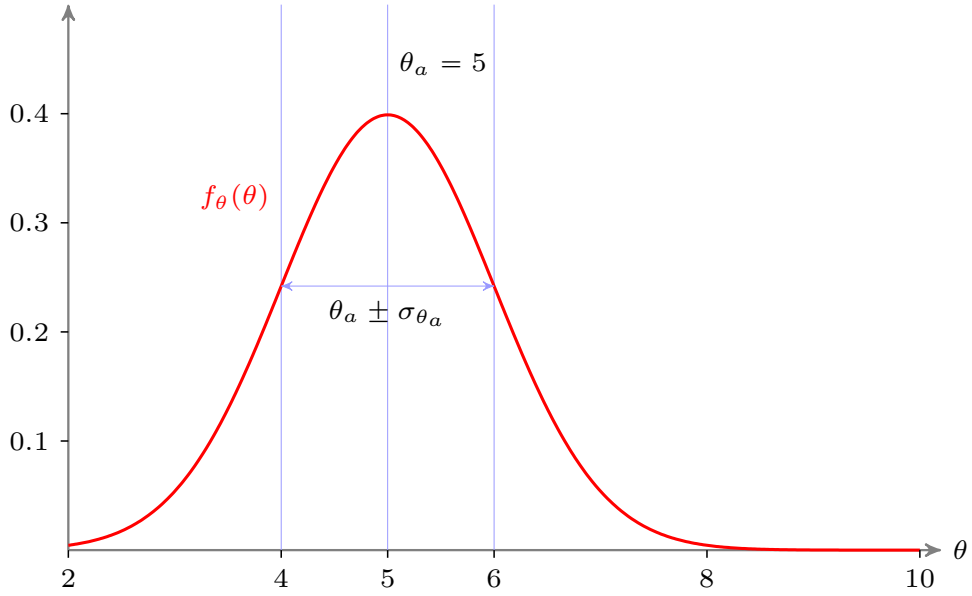
$$\begin{aligned}\theta_{\text{ML}} &= \underset{\theta}{\text{argmax}}\, f(x_1, \ldots, x_K; \theta)\Big|_{x_i = z_i, i=1,\ldots,K.} \\ &= \underset{\theta}{\text{argmax}}\; f(Z_K \mid \theta)\end{aligned}$$

These estimates coincide if we assume a uniform distribution for θ.

MAP estimation

A priori parameter distribution

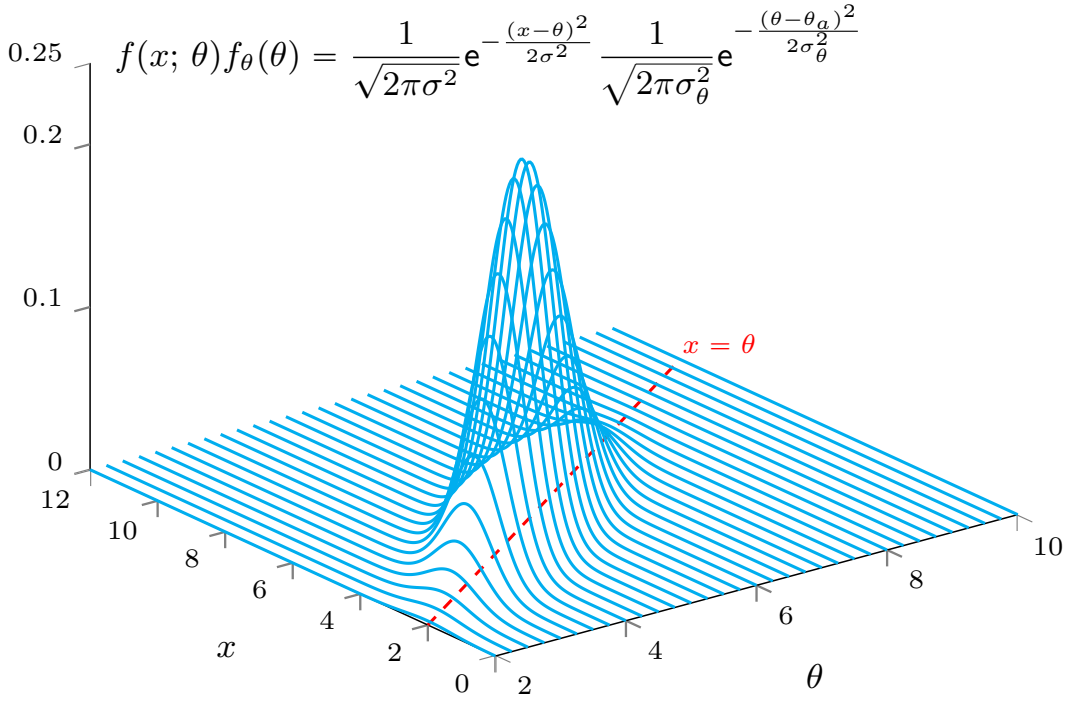
$$f_\theta(\theta) = \frac{1}{\sqrt{2\pi\sigma_\theta^2}} \mathrm{e}^{-\frac{(\theta-\theta_a)^2}{2\sigma_\theta^2}}, \qquad \theta_a = 5, \quad \sigma_\theta^2 = 1.$$



MAP estimation

Estimating the mean: Gaussian distribution ($\sigma^2 = 0.5$, $\theta_a = 5$, $\sigma_{\theta_a}^2 = 1$)

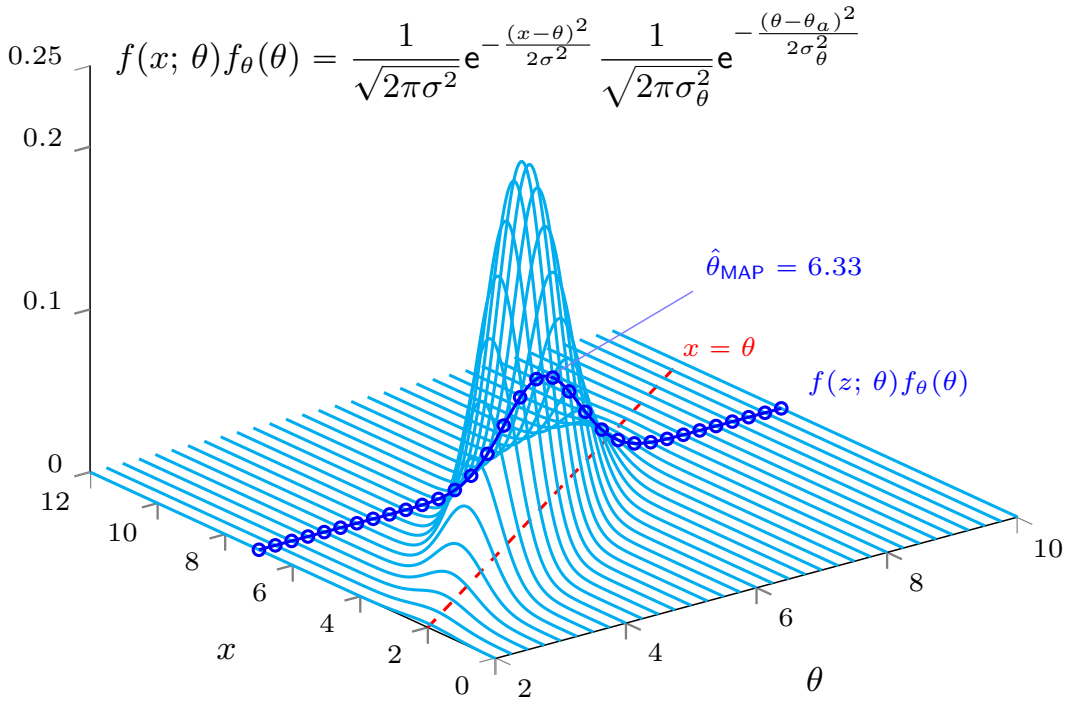
$$f(x;\theta) f_\theta(\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(x-\theta)^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi\sigma_\theta^2}} \mathrm{e}^{-\frac{(\theta-\theta_a)^2}{2\sigma_\theta^2}}$$



MAP estimation

Estimating the mean: Gaussian distribution ($\sigma^2 = 0.5$, $\theta_a = 5$, $\sigma_{\theta_a}^2 = 1$)

Datum: $z = 7.0$



Cramér-Rao bound

Mean-square error matrix

$$P = \mathcal{E}\left\{\left(\hat{\theta}(\mathcal{Z}_K) - \theta_0\right)\left(\hat{\theta}(\mathcal{Z}_K) - \theta_0\right)^T\right\}$$

Assume that the pdf for $\mathcal{Z}_K$ is $f(\mathcal{Z}_K;\,\theta)$.

Cramér-Rao inequality

Assume $\mathcal{E}\left\{\hat{\theta}(\mathcal{Z}_K)\right\} = \theta_0$, and $\mathcal{Z}_K \subset \mathcal{R}^K$.

Then, $P \geqslant M^{-1}$ (M is the Fisher Information Matrix)

$$M = \mathcal{E}\left\{\left(\frac{d}{d\theta}\ln f(\mathcal{Z}_K;\,\theta)\right)\left(\frac{d}{d\theta}\ln f(\mathcal{Z}_K;\,\theta)\right)^T\right\}\Bigg|_{\theta=\theta_0}$$

$$= -\,\mathcal{E}\left\{\frac{d^2}{d\theta^2}\ln f(\mathcal{Z}_K;\,\theta)\right\}\Bigg|_{\theta=\theta_0}$$

Asymptotic results for i.i.d. variables

Consider a parametrised family of pdfs,

$$f(x_1, \ldots, x_K;\, \theta) = \prod_{i=1}^{K} f_i(x_i;\, \theta).$$

Then,

$$\lim_{K \longrightarrow \infty} \hat{\theta}_{\mathsf{ML}} \xrightarrow{\text{w.p. } 1} \theta_0,$$

and

$$\lim_{K \longrightarrow \infty} \sqrt{K} \left(\hat{\theta}_{\mathsf{ML}}(\mathcal{Z}_K) - \theta_0 \right) \sim \mathcal{N}\left(0, M^{-1}\right).$$