Advanced Section #2
Model Selection & Information Criteria
Akaike Information Criterion

Marios Mattheakis and Pavlos Protopapas

CS109A Introduction to Data Science

Pavlos Protopapas and Kevin Rader

Outline

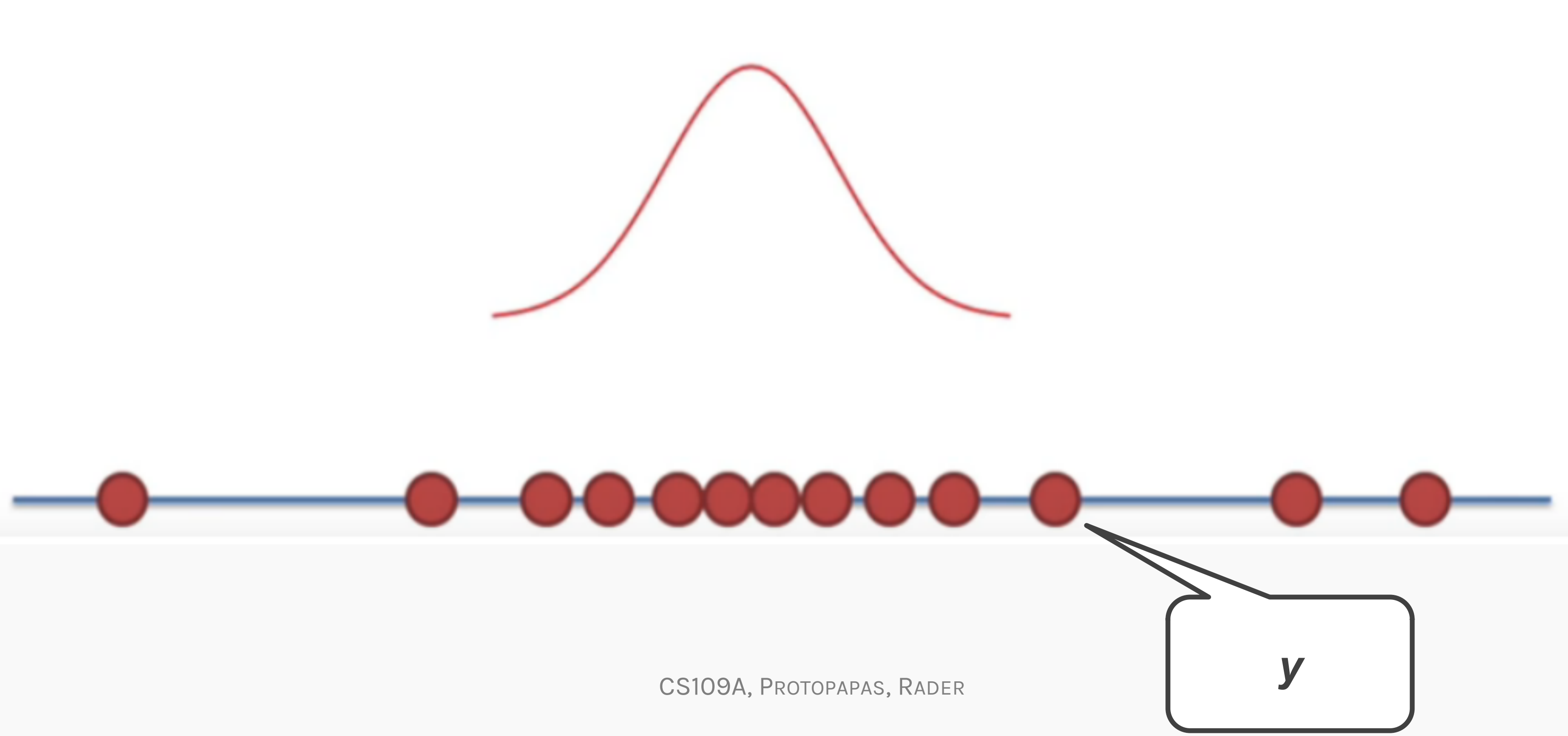
- Maximum Likelihood Estimation (MLE). Fit a distribution
  - Exponential distribution
  - Normal (Linear Regression Model)
- Model Selection & Information Criteria
  - KL divergence
  - MLE justification through KL divergence
  - Model Comparison
  - Akaike Information Criterion (AIC)

Maximum Likelihood Estimation (MLE)
& Parametric Models

Maximum Likelihood Estimation (MLE)

Fit your data with a parametric distribution $q(\mathbf{y}|\boldsymbol{\theta})$.

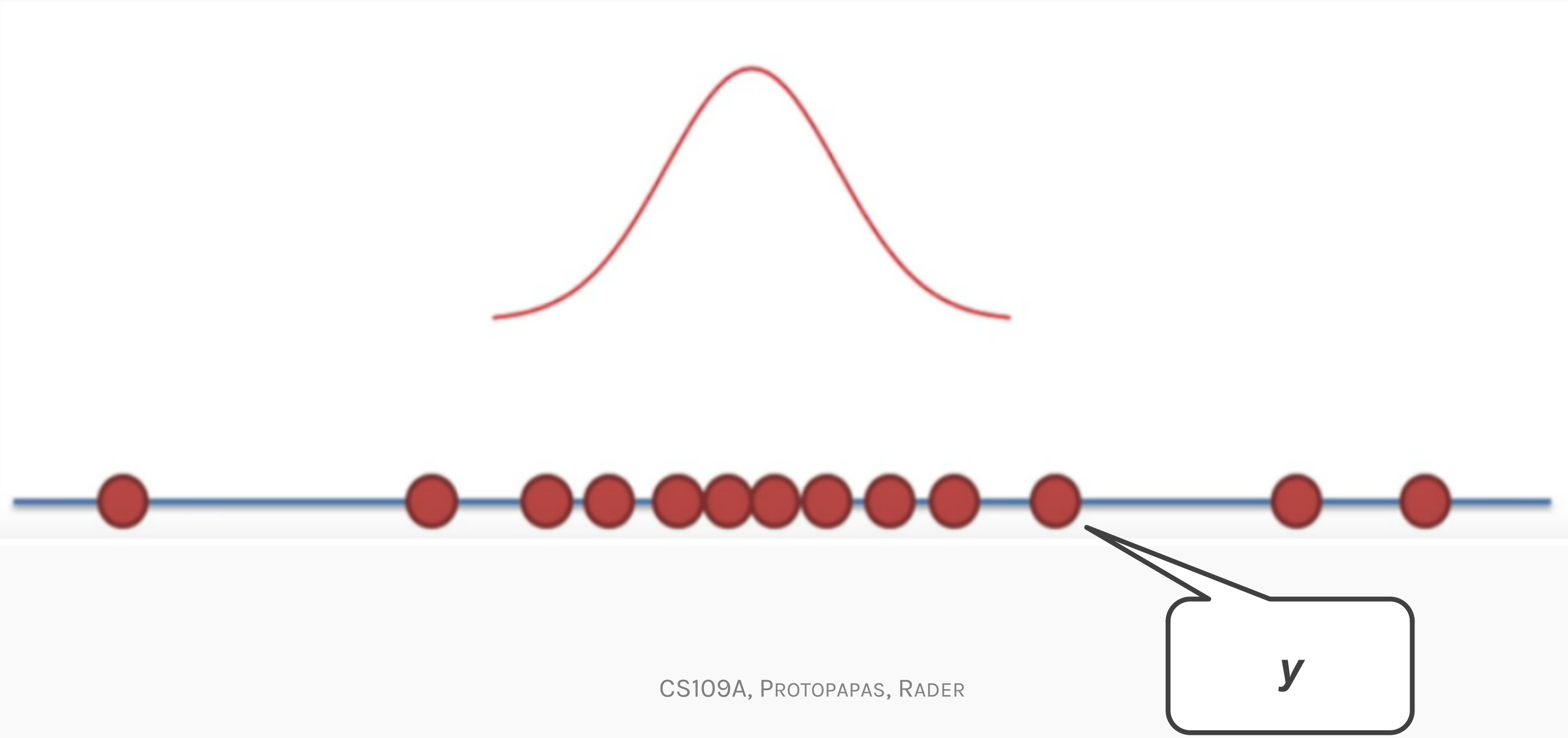
$\boldsymbol{\theta}=(\theta_1, \ldots, \theta_k)$ is a parameter set to be estimated.



Maximum Likelihood Estimation (MLE)

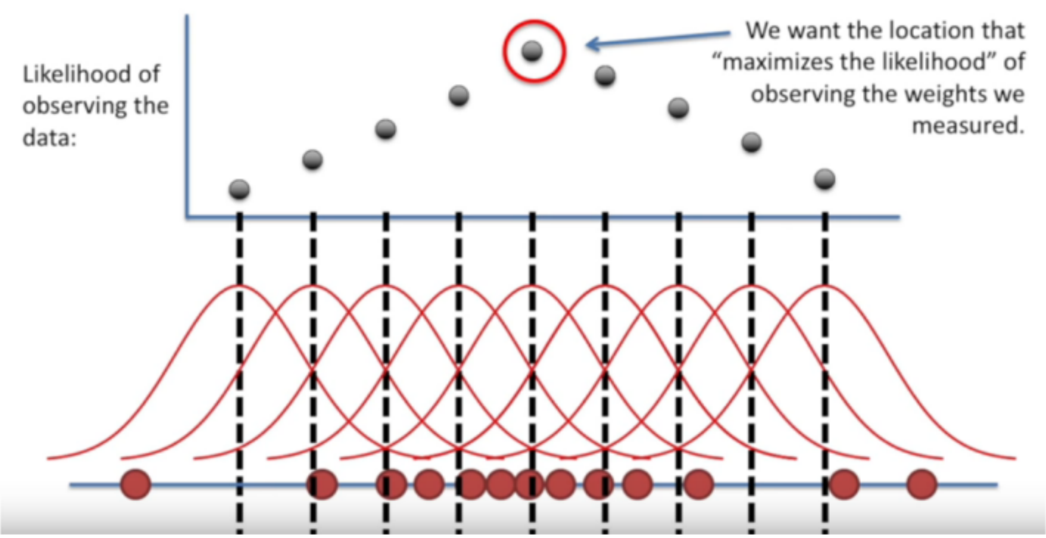
Fit your data with a parametric distribution $q(\mathbf{y}|\boldsymbol{\theta})$.

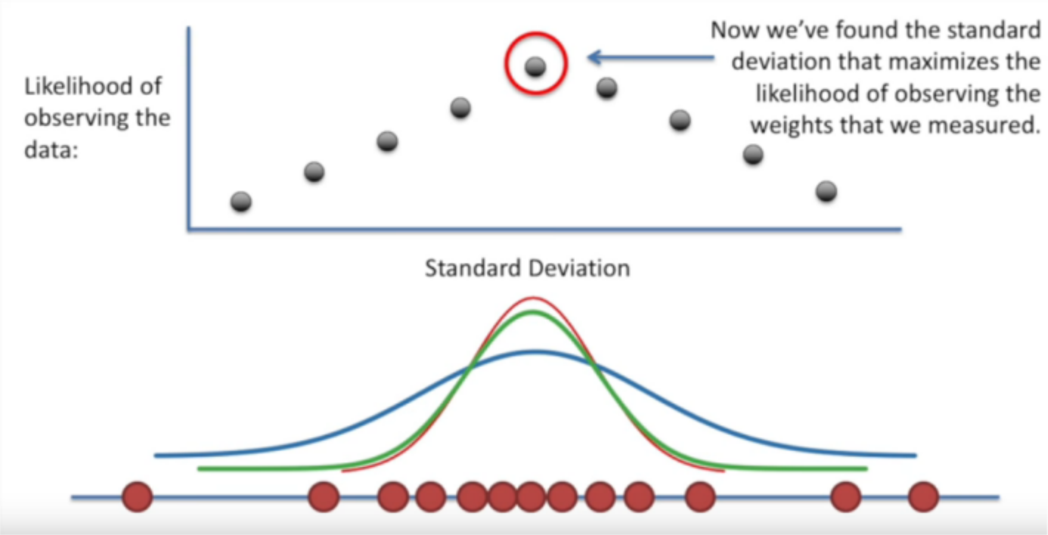
$\boldsymbol{\theta}=(\theta_1, \ldots , \theta_k)$ is a parameter set to be estimated.



Maximize the Likelihood L

Scanning over all the parameters until find the maximum L





…but this is a too time-consuming approach.

Maximum Likelihood Estimation (MLE)

A formal and efficient method is given by MLE

Observations: $\mathbf{y}=(y_1, \ldots, y_n)$

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} q(y_i|\boldsymbol{\theta}),$$

Easier and numerically more stable to work with log-likelihood

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log\left(q(y_i|\boldsymbol{\theta})\right)$$

Maximum Likelihood Estimation (MLE)

Easier and numerically more stable to work with log-likelihood

$$\frac{\partial \ell}{\partial \boldsymbol{\theta}} = \frac{\partial}{\partial \boldsymbol{\theta}} \log L = \frac{1}{L} \frac{\partial L}{\partial \boldsymbol{\theta}} \quad \Longrightarrow \quad \frac{\partial}{\partial \boldsymbol{\theta}} L(\boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\mathrm{MLE}}} = \frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\mathrm{MLE}}} = 0$$

Exponential distribution: A simple and useful example

A one parameter distribution: *rate parameter* λ

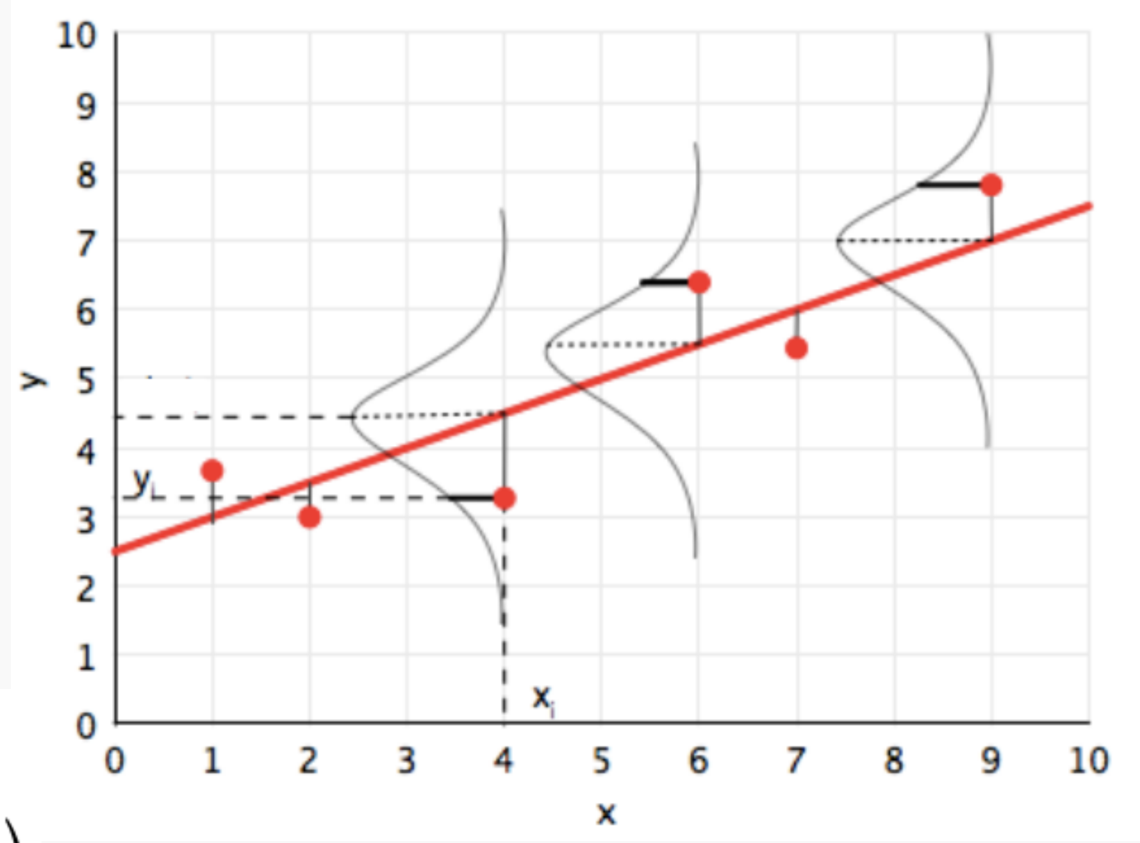
$$f(y_i|\lambda) = \begin{cases} \lambda e^{-\lambda y_i} & y_i \geq 0 \\ 0 & y_i < 0 \end{cases}$$

$$\ell(\lambda) = \sum_{i=1}^{n} \log\left(\lambda e^{-\lambda y_i}\right) = \sum_{i=1}^{n} \left(\log(\lambda) - \lambda y_i\right)$$

$$\lambda_{\text{MLE}} = \left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)^{-1}$$

Linear Regression Model with gaussian error

$$
\begin{aligned}
y_i &= \sum_{j=0}^{k} x_{ij}\beta_j + \epsilon_i \\
&= \mathbf{x}_i \cdot \boldsymbol{\beta} + \epsilon_i \\
&= \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i
\end{aligned}
$$



$$
\begin{aligned}
y_i = q(y_i|\mu_i, \sigma^2) &= \mathcal{N}(\mu_i, \sigma^2) = \mathcal{N}(\mathbf{x}_i^T \boldsymbol{\beta}, \sigma^2) \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2}{2\sigma^2}\right)
\end{aligned}
$$

Linear Regression Model through MLE

$$L(\boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2}{2\sigma^2}\right)$$

$$\begin{aligned}
\ell(\boldsymbol{\beta}, \sigma^2) &= \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2}{2\sigma^2}\right)\right) \\
&= -\sum_{i=1}^{n} \left(\frac{1}{2}\log(2\pi) + \frac{1}{2}\log(\sigma^2) + \frac{(y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2}{2\sigma^2}\right) \\
&= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \boxed{\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2}
\end{aligned}$$

Loss Function

Linear Regression Model: Standard Formulas

Minimize the loss essentially maximize the likelihood, and we get

$$\beta_{\text{MLE}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

$$\sigma^2_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \mathbf{x}_i^T\beta_{\text{MLE}}\right)^2$$

$\mathbf{X}$ is called *the design matrix*

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix} = \begin{pmatrix} 1 & \mathbf{x}_{11} & \cdots & \mathbf{x}_{1\nu} \\ 1 & \mathbf{x}_{21} & \cdots & \mathbf{x}_{2\nu} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \mathbf{x}_{n1} & \cdots & \mathbf{x}_{n\nu} \end{pmatrix}$$

Model Selection & Information Theory:
Akaike Information Criterion

Kullback-Leibler (KL) divergence (or relative entropy)

How good do we fit the data?

What additional uncertainty have we introduced?

$$\mathcal{D}_{\mathrm{KL}}(p \,\|\, q) = \sum_{i=1}^{n} p(y_i) \log\left(\frac{p(y_i)}{q(y_i|\boldsymbol{\theta})}\right)$$
$$= \int_{-\infty}^{\infty} p(\mathbf{y}) \log\left(\frac{p(\mathbf{y})}{q(\mathbf{y}|\boldsymbol{\theta})}\right) d\mathbf{y}$$

p is the *real* distribution

q is the model distribution

$$\mathcal{D}_{\mathrm{KL}}(p \,\|\, q) = \mathbb{E}_p\left[\log\left(\frac{p(\mathbf{y})}{q(\mathbf{y}|\boldsymbol{\theta})}\right)\right]$$
$$= \mathbb{E}_p\left[\log(p(\mathbf{y})) - \log(q(\mathbf{y}|\boldsymbol{\theta}))\right]$$

KL divergence

The KL divergence shows the *distance* between two distributions, hence it is a non-negative quantity.

With Jensen's inequality for convex functions $f(\boldsymbol{y})$, $\mathbb{E}[f(\boldsymbol{y})] \geq f(\mathbb{E}[\mathbf{y}])$:

$$\begin{aligned}\mathcal{D}_{\mathrm{KL}}(p \parallel q) &= \mathbb{E}_p\left[\log\left(\frac{p(\mathbf{y})}{q(\mathbf{y}|\boldsymbol{\theta})}\right)\right] \\ &= \mathbb{E}_p\left[-\log\left(\frac{q(\mathbf{y}|\boldsymbol{\theta})}{p(\mathbf{y})}\right)\right] \geq -\log\left(\mathbb{E}_p\left[\frac{q(\mathbf{y}|\boldsymbol{\theta})}{p(\mathbf{y})}\right]\right) = 0\end{aligned}$$

KL divergence is a non-symmetric quantity $\mathcal{D}_{\mathrm{KL}}(p \parallel q) \neq \mathcal{D}_{\mathrm{KL}}(q \parallel p)$

MLE justification through KL divergence

Empirical distribution

$$p(\mathbf{y}) \simeq \frac{1}{n}\sum_{i=1}^{n}\delta(\mathbf{y} - y_i),$$

Minimize KL divergence is the same with maximize likelihood (empirical distribution)

$$\begin{aligned}
\mathcal{D}_{\mathrm{KL}}(p \,\|\, q) &\simeq \int_{-\infty}^{\infty} p(\mathbf{y}) \log\left(\frac{p(\mathbf{y})}{q(\mathbf{y}|\boldsymbol{\theta})}\right) d\mathbf{y} \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty} \delta(\mathbf{y} - y_i) \log\left(\frac{p(\mathbf{y})}{q(\mathbf{y}|\boldsymbol{\theta})}\right) d\mathbf{y} = \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{p(y_i)}{q(y_i|\boldsymbol{\theta})}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\log p(y_i) - \boxed{\log q(y_i|\boldsymbol{\theta})}\right),
\end{aligned}$$

log-likelihood

Model Comparison

Consider to model distributions $q(\mathbf{y}|\boldsymbol{\theta})$ and $r(\mathbf{y}|\boldsymbol{\theta})$

$$\begin{aligned}\mathcal{D}_{\mathrm{KL}}(p \,\|\, q) - \mathcal{D}_{\mathrm{KL}}(p \,\|\, r) &= \mathbb{E}_p\left[\log(p(\mathbf{y})) - \log(q(\mathbf{y}|\boldsymbol{\theta}))\right] - \mathbb{E}_p\left[\log(p(\mathbf{y})) - \log(r(\mathbf{y}|\boldsymbol{\theta}))\right] \\ &= \mathbb{E}_p\left[\log(r(\mathbf{y}|\boldsymbol{\theta})) - \log(q(\mathbf{y}|\boldsymbol{\theta}))\right] = \mathbb{E}_p\left[\log\left(\frac{r(\mathbf{y}|\boldsymbol{\theta})}{q(\mathbf{y}|\boldsymbol{\theta})}\right)\right]\end{aligned}$$

By using the empirical distribution:

$$\mathcal{D}_{\mathrm{KL}}(p \,\|\, q) - \mathcal{D}_{\mathrm{KL}}(p \,\|\, r) = \frac{1}{n}\log\left(\frac{L_r(\mathbf{y}|\boldsymbol{\theta})}{L_q(\mathbf{y}|\boldsymbol{\theta})}\right)$$

p is eliminated.

Akaike Information Criterion (AIC)

AIC is a trade off between the number of parameters *k* and the error that is introduced (overfitting).

AIC is an asymptotic approximation of the KL-divergence $\mathcal{D}_{\mathrm{KL}}(p \parallel q)$

The data are being used twice: first for MLE and second for the KL-divergence estimation.

AIC estimates which is the optimal number of parameters k

Polynomial Regression Model Example

Suppose a polynomial regression model

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij},$$

Which is the optimal k?

For k smaller than the optimal: Underfitting

For k larger than the optimal: Overfitting

Minimizing real and empirical KL-divergence

Suppose many models indicated by index j

Work with the *j*-th model which has k_j parameters

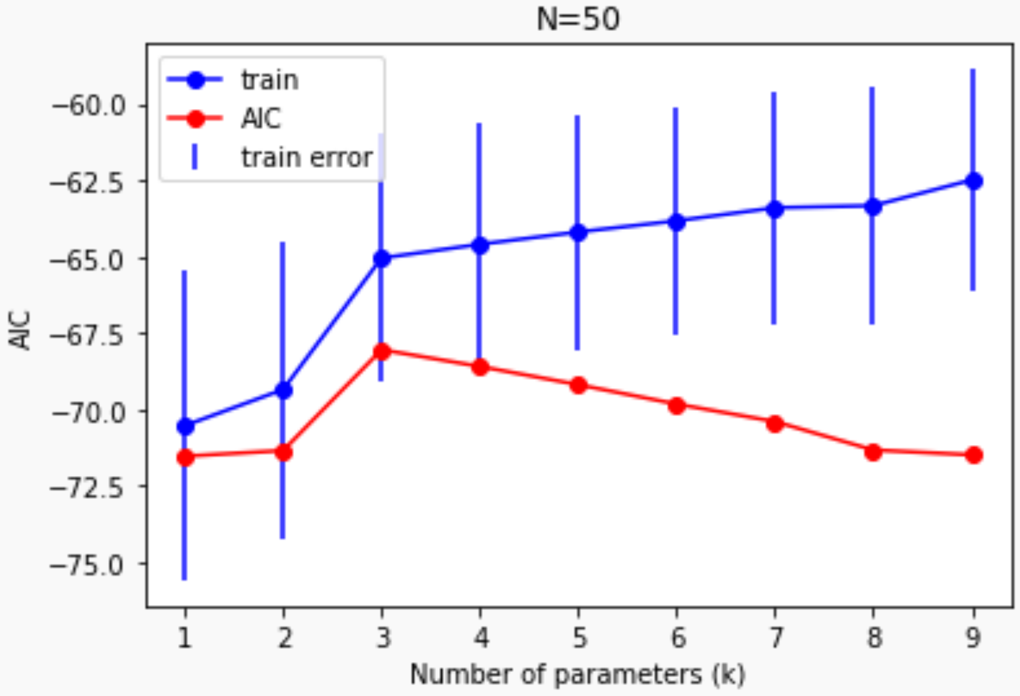
$$K_j = \int p(\mathbf{y}) \log q_j(\mathbf{y}|\boldsymbol{\theta}^{(j)}_{\text{MLE}}) d\mathbf{y}.$$

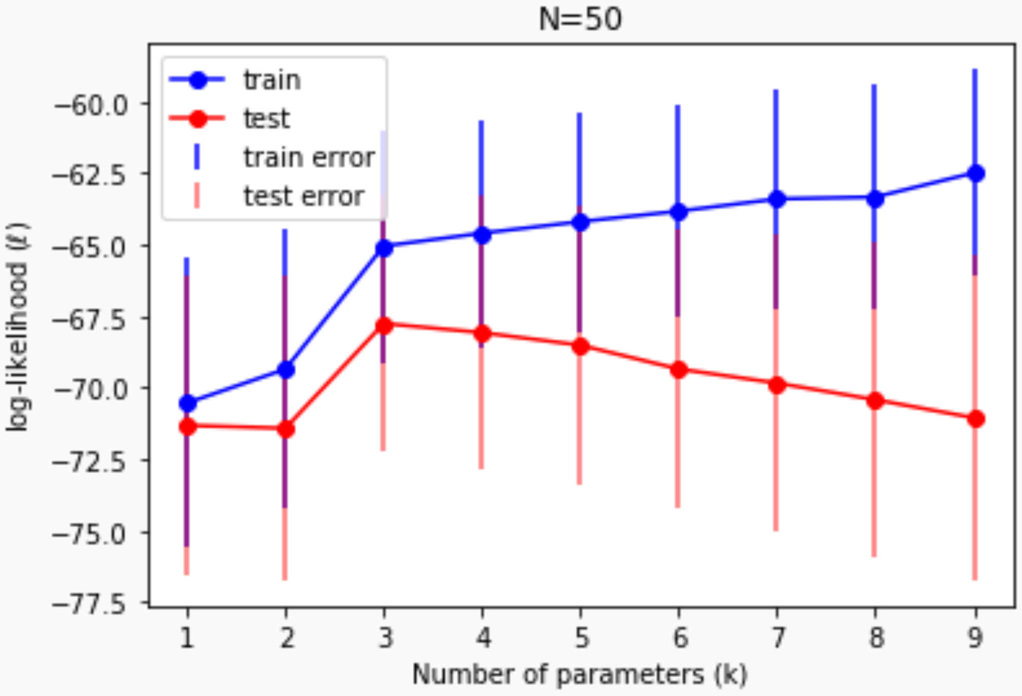
$$\bar{K}_j = \frac{1}{n}\sum_{i=1}^{n} \log q_j(y_i|\boldsymbol{\theta}^{(j)}_{\text{MLE}}) = \frac{\ell_j(\boldsymbol{\theta}^{(j)}_{\text{MLE}})}{n}$$

$$\begin{aligned} K_j &= \bar{K}_j - \frac{k_j}{n} \\ &= \frac{\ell_j(\boldsymbol{\theta}^{(j)}_{\text{MLE}})}{n} - \frac{k_j}{n}. \end{aligned}$$

$$\boxed{\begin{aligned} \text{AIC}(j) &= 2nK_j \\ &= 2\ell_j(\boldsymbol{\theta}^{(j)}_{\text{MLE}}) - 2k_j. \end{aligned}}$$

Numerical verification of AIC





Akaike Information Criterion (AIC): Proof

Asymptotic Expansion around true ideal MLE $\boldsymbol{\theta}_0$

$$K_j \simeq \int p(\mathbf{y})\left(\log q(\mathbf{y}|\boldsymbol{\theta}_0) + (\boldsymbol{\theta}_{\text{MLE}} - \boldsymbol{\theta}_0)^T s(\mathbf{y}|\boldsymbol{\theta}_0) + \frac{1}{2}(\boldsymbol{\theta}_{\text{MLE}} - \boldsymbol{\theta}_0)^T H(\mathbf{y}|\boldsymbol{\theta}_0)(\boldsymbol{\theta}_{\text{MLE}} - \boldsymbol{\theta}_0)\right) d\mathbf{y}$$
$$= K_0 + \frac{1}{2n} Z^T J(\mathbf{y}|\boldsymbol{\theta}_0) Z,$$

$$\bar{K}_j \simeq \frac{1}{n}\sum_{i=1}^{n}\left(\log q(y_i|\boldsymbol{\theta}_0) + (\boldsymbol{\theta}_{\text{MLE}} - \boldsymbol{\theta}_0)^T s(y_i|\boldsymbol{\theta}_0) + +\frac{1}{2}(\boldsymbol{\theta}_{\text{MLE}} - \boldsymbol{\theta}_0)^T H(y_i|\boldsymbol{\theta}_0)(\boldsymbol{\theta}_{\text{MLE}} - \boldsymbol{\theta}_0)\right)$$
$$= K_0 + A_n + \frac{Z^T S_n}{\sqrt{n}} - \frac{1}{2n} Z^T J_n Z^T,$$

Akaike Information Criterion (AIC): Proof

$$J(y|\boldsymbol{\theta}) = -\mathbb{E}_p\left[H(y|\boldsymbol{\theta})\right]$$

$$Z = \sqrt{n}\left(\boldsymbol{\theta}_{\mathrm{MLE}} - \boldsymbol{\theta}_0\right) \quad (\text{with } Z_i \text{ given by } \mathcal{N}(0, V_Z)),$$

$$S_n = \frac{1}{n}\sum_{i=1}^{n} s(y_i|\boldsymbol{\theta}_0)$$

$$A_n = \frac{1}{n}\sum_{i=1}^{n}\left(\log q(y_i|\boldsymbol{\theta}_0) - K_0\right)$$

$$\begin{aligned}\bar{K} - K &\simeq A_n + \frac{\sqrt{n}Z^T S_n}{n} \\ &= A_n + \frac{Z^T J Z}{n},\end{aligned}$$

$$\mathbb{E}_p\left[\bar{K} - K\right] = \mathbb{E}_p\left[A_n\right] + \mathbb{E}_p\left[\frac{Z^T J Z}{n}\right]$$

Akaike Information Criterion (AIC): Proof

$$\mathbb{E}_p\left[\bar{K} - K\right] = 0 + \text{trace}\left(\frac{J\,J^{-1}VJ^{-1}}{n}\right) = \frac{1}{n}\text{trace}\left(J^{-1}V\right).$$

$$K \simeq \bar{K} - \frac{1}{n}\text{trace}\left(J^{-1}V\right).$$

In the limit of a correct model: $\boldsymbol{\theta}_{\text{MLE}} = \boldsymbol{\theta}_0$, and thus, $J^{-1} = V$.

$$\boxed{K \simeq \bar{K} - \frac{k}{n}}$$

Review

- Maximum Likelihood Estimation (MLE)
  1. A powerful method to estimate the ideal fitting parameters of a model.
  2. Exponential distribution, a simple but useful example.
  3. Linear Regression Model as a special paradigm of MLE implementation.

- Model Selection & Information Criteria
  1. KL-divergence quantifies the “distance” between the fitting model and the “real” distribution.
  2. KL-divergence justifies the MLE and is used for model comparison.
  3. AIC: Estimates the number of model parameters and protects from overfitting.

Thank you

Office hours are:

Monday 6-7:30 (Marios)

Tuesday 6:30-8 (Trevor)