Parameter Estimation

- Value $\hat{\theta}$ is realization for a random variable
  \[ \theta^* = t(X_1, \ldots, X_n) \]  
  (2)
- The estimation equation (2) is also called as estimator
- The estimate is real, calculated value and the estimator is a model

Fisher Information

- In the following we consider only unbiased estimators. Who could the variance of the estimator to be studied?
- One way is to examine the absolute value of the second derivative of the likelihood (log likelihood) function.
- If the absolute value is large, the likelihood value is spiky at the maximum value, indicating small variance.
- On the other hand, if the absolute value is small, likelihood function is more flat and the estimate is less accurate

Parameter Estimation

- Random variable $X$ is defined by its distribution $p(x; \theta)$. On the other hand parameters $\theta$ define the distribution $p(x; \theta)$.
- We have set of observations (from measurements etc.) $x_1, \ldots, x_n$, which are realizations for random variables $X_1, \ldots, X_n$
- Estimation is forming values for $\theta$ i.e. the estimates $\hat{\theta}$, through estimator.
  \[ \hat{\theta} = t(x_1, \ldots, x_n) \]  
  (1)
  On the other words, mapping observations to parameter values

Estimators

- Characteristics of a good estimator:
  1. Unbiased: The estimator does not make a systematic error, i.e. $E(\theta^*) = 0$.
  2. Consistency: The accuracy of the estimator increases as size of the observation set increases.
  3. The variance of the estimator is small.
Fisher Information

- \( J(\theta) \) is formed by studying random variable \( V \)

\[
V = \frac{\partial}{\partial \theta} \ln f(X; \theta) = \frac{\frac{\partial f(X; \theta)}{f(X; \theta)}}{\frac{\partial f(X; \theta)}{\partial \theta}}
\] (6)

where \( X \sim f(X; \theta) \)

Fisher Information

- But the variance

\[
EV^2 = J(\theta) = E \left( \left( \frac{\partial \ln f(X; \theta)}{\partial \theta} \right)^2 \right)
\] (12)

has interesting properties

- We can show that \( J(\theta) \) gives the lower bound for the variance of an unbiased estimator.

- The proof is based on Cauchy - Schwarz inequality

\[
\left( \sum_i (x_i y_i) \right)^2 \leq \sum_i x_i^2 \sum_i y_i^2
\] (13)

Fisher Information

- Fisher information is defined as the expected value of the absolute value of the second derivate of the log likelihood function.

\[
l(\theta) = \ln(L(\theta))
\] (3)

\[
J(\theta) = -E \left[ \frac{\partial^2 l(\theta)}{\partial \theta^2} \right]
\] (4)

\[
= nE \left[ \left( \frac{\partial \ln p(X; \theta)}{\partial \theta} \right)^2 \right]
\] (5)

- Greater the \( J(\theta) \) is, more accurate the estimate is

Fisher Information

- The mean value of \( V \) is zero

\[
EV = \int \frac{\partial f(X; \theta)}{f(X; \theta)} f(X; \theta) dx
\] (7)

\[
= \int \frac{\partial f(X; \theta)}{\partial \theta} f(X; \theta) dx
\] (8)

\[
= \frac{\partial}{\partial \theta} \int f(X; \theta) dx
\] (9)

\[
= \frac{\partial}{\partial \theta} 1
\] (10)

\[
= 0
\] (11)
**Fisher Information**

- $E_\theta(VT)$ is

\[
E_\theta(VT) = \int \frac{\partial f(X; \theta)}{f(X; \theta)} T(X) f(X; \theta) dx \quad (16)
\]

\[
= \int \frac{\partial}{\partial \theta} f(X; \theta) T(X) dx \quad (17)
\]

\[
= \frac{\partial}{\partial \theta} \int f(X; \theta) T(X) dx \quad (18)
\]

\[
= \frac{\partial}{\partial \theta} E_\theta T \quad (19)
\]

\[
= \frac{\partial}{\partial \theta} \theta \quad (20)
\]

\[
= 1 \quad (21)
\]

**Cramer-Rao Inequality**

- The lower bound of unbiased estimator can be defined with $J(\theta)$ as:

\[
Var[\theta^*] = \frac{1}{J(\theta)} \quad (23)
\]

- The equation (23) is called **Cramer - Rao inequality**.

**Fisher Information**

- Define $T$ as estimator for $V$. By Cauchy-Swarz inequality we have

\[
(E_\theta[(V - E_\theta V)(T - E_\theta T)])^2 \leq E_\theta(V - E_\theta V)^2 E_\theta(T - E_\theta T)^2 \quad (14)
\]

- We know that $EV = 0$ and $var(V) = J(\theta)$. Applying these to above we have

\[
(E_\theta(VT))^2 \leq J(\theta) var(T) \quad (15)
\]

**Fisher Information**

- Because $E_\theta(VT) = 1$, $(E_\theta(VT))^2 \leq J(\theta) var(T)$ becomes

\[
VarT \leq \frac{1}{J(\theta)} \quad (22)
\]
Cramer-Rao Inequality: Example

- X is r.v.
  \[
  \frac{d^2\ell(p)^*}{dp^2} = -\frac{X}{p^2} - \frac{n - X}{(1 - p)^2}
  \] (27)

- Fisher information
  \[
  J(X) = -E\left[\frac{d^2\ell(p)^*}{dp^2}\right] = \frac{E[X]}{p^2} + \frac{n - E[X]}{(1 - p)^2}
  \] (28)

- X has binomial distribution, so \( E[X] = np \)
  \[
  J(X) = \frac{n}{p(1 - p)}
  \] (29)

- And Cramer-Rao inequality is
  \[
  Var[X^*] \geq \frac{1}{J(X)} = \frac{p(1 - p)}{n}
  \] (30)

Model Complexity

- Let’s examine the following example: We have observed some data (shown below) and we need to fit a model to it.

Kuva 1: Observed data

Cramer-Rao Inequality: Example

- Let \( X \sim Bin(n, p) \) and we have observed \( X = x \). Calculate \( J(\theta) \) and Cramer-Rao lower bound for variance

- Likelihood
  \[
  L(p) = \binom{n}{x} p^x (1 - p)^{n-x}
  \] (24)

- Log likelihood
  \[
  l(p) = \ln L(p) = \ln \left( \binom{n}{x} \right) + x \ln p + (n - x) \ln(1 - p)
  \] (25)

- Second derivative
  \[
  \frac{d^2\ell(p)}{dp^2} = -\frac{x}{p^2} - \frac{n - x}{(1 - p)^2}
  \] (26)

Model Complexity

- The data consists of observations, which consists of actual information \( a \) and noise \( \epsilon \)
  \[
  x = a + \epsilon
  \] (31)

- The purpose of the modeling is to tell apart these two components, the information and the noise.

- This is done by learning from the data, not by learning the data.
Model Complexity

- First, let’s fit linear model, i.e., $m = 1$

- It is clear that this model sucks

Model Complexity

- Then we use high model order.

Model Complexity

- This model order seems nice.

Model Complexity

- Used model class is $m$th order polynomial and we have $n$ observations

\[
y = a_0 x^0 + a_1 x^1 + a_2 x^2 + \ldots + a_m x^m = \sum_{i=0}^{m} a_i x^i \quad (32)
\]

- Model’s parameters are solved by pseudoinverse

\[
\hat{a} = (X^T X)^{-1} X^T y \quad (33)
\]

where $X$ is $n \times m$ matrix and $y$ is $n \times 1$ vector

- Does this model suck and if it does, what’s the level of the suckness?
Model Complexity

- A very important term in statistical modeling is the *complexity* of a model.
- Complexity can be considered as the model’s capability to generalize to a given data set.
- If the model has too low complexity, the model will have high systematic error.
- If the model has too high complexity, the model will fit to intrinsic noise in the data, instead of the actual phenomenon we are looking for.
- There is no actual value or a measure for the model complexity; only definitions 'good', 'too high', 'mediocre', etc.
- Although often some model properties, such as model order, are interpreted as the complexity.

Model Complexity

As a conclusion about the three previous slides, it seems to that:

- If the model order is too low, we have difficulties to learn from data
- If the model order is too high, we seem to learn the noise also

Model Complexity

- This one with high model order:

Model Complexity

- Let’s revisit the example: First, the linear model, $m = 1$

Model Complexity

- This model has too low complexity. It clearly cannot grasp the structure of the observed data. We could not use this model to e.g. predict the outputs of new inputs.
Model Complexity

- We cannot calculate the actual complexity but we can calculate the effective complexity.
- This value is meaningful only when inspecting model with large number of adjustable parameters such as multi-layer perceptron network (MLP).
- This value can be used to determine better complexity for the model during the training process.

Model Complexity

- Statistical methods for parameter selection have been created. These can be divided to:
  1. Classic hypotheses testing, $\chi^2$-testing etc.
  2. Two part information criteria
  3. Numerical information criteria
  4. Bayesian model selection

Model Complexity

- With this model order the complexity seems to be good.

- But there is still no way to absolutely determine this.

Model Complexity

- We have seen that model parameter selection affects (dictates) model performance.
- Therefore parameter selection requires some attention.
- Too often parameter selection is done by 'Stetson Harrison’ method, ie, different values are tested until suitable is found.
- This method provides inconsistency to model selection and is suboptimal.
Model Complexity

- Procedure is usually as follows:
  1. Model class is selected
  2. Parameter range is defined, i.e., limits to the number of parameters
  3. Information criterion score is calculated for the model with each number of parameters
  4. That number of parameters is selected, which has the best score

4. Select model order \( m \) by finding \( k \) for which information criterion score is the best
Akaike’s Information Criterion (AIC)

- AIC is based on log likelihood function, which is an estimator for the expected log likelihood.
- From the previous lectures, Greater the (expected) log likelihood, better the model.
- The core of the AIC is to estimate the mean expected value of log likelihood and use it to model selection.

From Relative Entropy to Maximum Likelihood

- From the previous lectures: By the law of large numbers
  \[ \frac{1}{N} \sum_{i=1}^{N} \log q_{e_t} \]  
  converges to expected value of log likelihood as \( N \to \infty \).
- Because random variable \( \log q(X) \) will have realization \( \log q(X = \omega_i) \) \( n_i \) times, the equation (35) can be rewritten to form
  \[ \frac{1}{N} \sum_{i=1}^{m} n_i \log q(\omega_i), \]
  where \( l(q) = \sum_{i=1}^{m} n_i \log q(\omega_i) \) is log likelihood for model \( q \).

Akaike’s Information Criterion (AIC)

- Also by Akaike
  \[ l'(\theta') = E[l(\hat{\theta})] - k/2 \]  
  (39)
- By combining equations (38) ja (39) we get
  \[ l'_n(k) = E[l(\hat{\theta})] - k \]  
  (40)
  ie, ML estimate is a biased estimator for mean expected log likelihood and the bias is the number of parameters \( k \).
- Therefore the unbiased estimator is
  \[ l(\hat{\theta}) - k, \]  
  (41)
  Which leads to the information criterion.
of 1.2, the difference is significant.

4. In some cases AIC is inconsistent.

---

**Akaike’s Information Criterion (AIC)**

- AIC is usually presented as

\[ AIC(k) = -2l(\hat{\theta}) + 2k, \]  

(42)

When using this from, that \( k \) is elected, for which the \( AIC(k) \) is lowest.

- Some issues when using AIC

  1. The maximum number of parameters \( m_{\text{max}} \) should not exceed \( 2\sqrt{n} \), where \( n \) is the number of observation. This is because larger \( m_{\text{max}} \) weakens assumption made in AIC definitions.

  2. There are cases, when AIC is monotonic, i.e., there is no solution. In most of these cases the culprit is poor selection of model class.

  3. If AIC score difference between two models is in magnitude

---

**Log likelihood:**

\[ l(a, \sigma^2) = -\frac{n}{2} \log 2\pi \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left( y - a_0 + \sum_{j=1}^{m} a_j x_{ji} \right)^2 \]  

(47)

---

**Akaike’s Information Criterion (AIC)**

- Example: AIC criterion for linear regression

\[ y = f(x; \theta) = a_0 + \sum_{i=1}^{m} a_i x_i + \epsilon \]  

(43)

\[ \epsilon \sim N(0, \sigma^2) \]  

(44)

- pdf:

\[ p(y | a, \sigma^2) = \frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{1}{2\sigma^2}(y - a_0 + \sum_{i=1}^{m} a_i x_i)^2} \]  

(45)

- Likelihood: \( n \) observations

\[ L(y | a, \sigma^2) = \prod_{i=1}^{n} p(y_i | a, \sigma^2) \]  

(46)
Akaike’s Information Criterion (AIC)

- AIC:
  \[ AIC(k) = -2l(\hat{\theta}) + 2k \]  
  \[ \text{(48)} \]

- The number of parameters in this model is \( m + 2 \) \((m + (a_0, \sigma^2))\).

- Parameters can be solved with eg. pseudoinverse and then the yare substituted to AIC criterion

  \[ AIC(k) = n(\ln(2\pi) + 1) + n\ln(\hat{\sigma}^2) + 2(m + 2) \]  
  \[ \text{(49)} \]