Aleatoric / Epistemic uncertainty

Aleatoric uncertainty is uncertainty arising from the inherent randomness of an event.
↳ Unexplainable Variance of $G(\cdot)$ due to non-deterministic behaviour. Cause for $Var(\hat\theta) > 0$.

Epistemic uncertainty is uncertainty due to lack of knowledge, which in principle can be diminished by increasing the amount of data or with an excessive number of experiments.

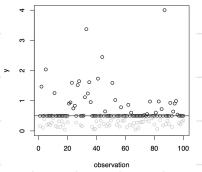
Likelihood principle

The likelihood principle states that for inference about parameter θ, all relevant information is contained in the likelihood function for the data at hand

→ Example

Assume $\log(Y) \sim \mathcal{N}(\mu, \sigma^2)$. Instead of $\min_\mu \sum_{i=1}^n (\log(y_i) - \mu)^2$ we solve

$$\max_\mu \ell(\mu, \sigma^2) = \max_\mu \sum_{i=1}^n I(y_i \geq c) \cdot \left(-\tfrac{1}{2}\log(\sigma^2) - \frac{(\log(y_i)-\mu)^2}{2\sigma^2}\right) + \sum_{i=1}^n I(y_i < c) \cdot \Phi\left(\frac{\log(c)-\mu}{\sigma}\right)$$



Kullback-Leibler Divergence

Estimator - Theory

$$KL(G, F(\cdot;\theta)) := \int \log\left(\frac{g(y)}{f(y;\theta)}\right) dG(y) = \int \log\left(\frac{g(y)}{f(y;\theta)}\right) \cdot g(y)\, d\mu(y)$$

$$= \underbrace{\int \log(g(y)) \cdot g(y)\, d\mu(y)}_{\text{Entropy}} - \int \log(f(y;\theta)) \cdot g(y) d\mu(y)$$

$$\underbrace{\frac{1}{n}\sum_{i=1}^n \log\left(\frac{g(y_i)}{f(y_i;\theta)}\right)}_{\text{empirical KL}} = \frac{1}{n}\sum_{i=1}^n \log(\underbrace{g(y_i)}_{\text{unknown}}) - \frac{1}{n}\sum_{i=1}^n \log(f(y_i;\theta)) \xrightarrow{n\to\infty} KL(G, F(\cdot;\theta))$$

To minimize empirical KL we need to maximize $\sum_{i=1}^n \log(f(y_i;\theta))$

The optimal parameter θ_0 is defined as $0 = \int \frac{\partial \log(f(y;\theta_0))}{\partial\theta} dG(y)$

and hence estimated through $\frac{\partial \ell}{\partial\theta}(\hat\theta) = 0$

Maximum Likelihood (MLE)

For a random sample $y_1, \dots, y_n$ the maximum likelihood estimate is defined as $\hat\theta := \arg\max_{\theta \in \Theta} \ell(\theta; y_1, \dots, y_n)$ which for fisher-regular distributions occurs when $s(\hat\theta; y_1, \dots, y_n) = 0$

Score function

$$s(\theta; y) := \frac{\partial \ell(\theta; y)}{\partial\theta} = \frac{\partial \log f(y;\theta)}{\partial\theta}$$

Under regularity we have $\mathbb{E}_\theta(s(\theta;y)) = \int \frac{\partial \log f(y;\theta)}{\partial\theta} \cdot f(y;\theta)\, dy$

$$= \int \frac{\partial f(y;\theta)}{\partial\theta} dy = \frac{\partial}{\partial\theta}\int f(y;\theta)\, d\theta = \frac{\partial 1}{\partial\theta} = 0.$$

Fisher Information

$$I(\theta) := Var_\theta[s(\theta;Y)] = \mathbb{E}_\theta[s(\theta;Y)^2] = \int \left(\frac{\partial \log f(y;\theta)}{\partial\theta}\right)^2 f(y;\theta)\, dy.$$

For certain assumptions: $I(\theta) = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(Y;\theta)\right]$

The observed Fisher Information is $J(\theta) = -\sum_{i=1}^n \frac{\partial^2}{\partial\theta^2}\log f(y_i;\theta)$.

! The Fisher information determines how quickly the observed score function converges to the shape of the true score function.

For the case of $\theta \in \mathbb{R}^n$

$$[I(\theta)]_{ij} := \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta_i}\log f(Y;\theta)\right)\left(\frac{\partial}{\partial\theta_j}\log f(Y;\theta)\right)\right]$$

For certain assumptions: $[I(\theta)]_{ij} = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(Y;\theta)\right]$

Cramér Rao Lower Bound (CRLB)

Given a statistical model $Y \sim F(\cdot;\theta)$ the CRLB provides a lower bound on the variance of an estimator $t(Y)$. If an unbiased estimator achieves the CRLB, the estimator must have a lower variance than any other unbiased estimator for all $\theta \in \Theta$.

Theorem

Let $Y = (Y_1, \dots, Y_n) \sim F(\cdot;\theta)$ Fisher regular dist

Let $t(Y)$ be a (biased) estimator of θ.

Then $Var(t(Y)) \geq \frac{\left(\frac{\partial}{\partial\theta}\mathbb{E}[t(Y)]\right)^2}{I(\theta)} := CRLB(\theta)$

If $t(\cdot)$ is unbiased, then we get $Var(t(Y)) \geq \frac{1}{I(\theta)}$

Properties of Estimates

Bias: $bias(\hat\theta, \theta_0) = \mathbb{E}[\hat\theta] - \theta_0$

MSE: $MSE(\hat\theta, \theta_0) := \mathbb{E}[(\hat\theta - \theta_0)^2] = Var(\hat\theta) + bias^2(\hat\theta, \theta_0)$

Consistency: $MSE(\hat\theta, \theta_0) \xrightarrow{n\to\infty} 0$

Efficiency: We call $\hat\theta_1$ more efficient than $\hat\theta_2$ if $MSE(\hat\theta_1, \theta_0) < MSE(\hat\theta_2, \theta_0)$

Sufficiency: $t(Y_1, \dots, Y_n)$ is called sufficient for θ $P(Y_1 = y_1, \dots, Y_n = y_n \mid t(Y_1, \dots, Y_n) = t_0;\ \theta)$ does not depend on θ.

Neyman - factorisation

$t(Y_1, \dots, Y_n)$ is called sufficient for θ iff

$$f(y_1, \dots, y_n;\theta) = h(y_1, \dots, y_n) \cdot g(t(y_1, \dots, y_n), \theta)$$

Minimal sufficient:

$t(y_1, \dots, y_n)$ is minimal sufficient for θ if

t is sufficient and for any other sufficient statistic $\tilde t(y_1, \dots, y_n)$ there exists a function m s.t. $t(Y_1, \dots, Y_n) = m(\tilde t(Y_1, \dots, Y_n))$.

Example

$Y_i \sim \mathcal{B}(\pi)$ and $t(y_1, \dots, y_n) = t_0$ and $n_0 = n \cdot t_0$.

$$P(Y_1 = y_1, \dots, Y_n = y_n \mid t(Y_1, \dots, Y_n) = t_0; \pi) = \begin{cases} \frac{1}{\binom{n}{n_0}} & \text{, for } \sum_{i=1}^n y_i = n_0 \\ 0 & \text{, otherwise} \end{cases}$$

Distribution is independent of π.
$\Rightarrow t(Y_1, \dots, Y_n) = \bar Y$ is sufficient.

Confidence interval

The interval $CI = [t_\ell(Y_1, \dots, Y_n), t_r(Y_1, \dots, Y_n)]$ is called confidence interval for θ with confidence level $1-\alpha$ if $\mathbb{P}_\theta(\theta \in [t_\ell(Y_1, \dots, Y_n), t_r(Y_1, \dots, Y_n)]) \geq 1-\alpha$ for all θ.

The value $(1-\alpha)$ is called the confidence level.

Confidence principle

If we were to repeat the calculation with new data of the same size multiple times, then in $\alpha \cdot 100\%$ of the cases the true parameter would not be within the CI.

✓: There is a $(1-\alpha)$ probability that the CI for θ with confidence level $(1-\alpha)$ calculated from a given future sample will cover the true value of θ.

X: For a CI for θ with confidence level $(1-\alpha)$ there is a $(1-\alpha)$ probability of covering the true value of θ. → $\mathbb{P}(\theta \in [-1, 1])$ is nonsense b.c. θ ain't a random variable.

68 - 95 - 99.7 - rule

For $X \sim \mathcal{N}(\mu, \sigma)$ we get $\mathbb{P}(X \in [\mu \pm \sigma]) \approx 68\%$

$\mathbb{P}(X \in [\mu \pm 2\sigma]) \approx 95\%$ $\mathbb{P}(X \in [\mu \pm 3\sigma]) \approx 99.7\%$

Pivotal statistic

A quantity $g(t(Y_1, \dots, Y_n); \theta)$ is called pivotal statistic if its distribution does not depend on θ. It's distribution is called a pivotal distribution.

Example

Assume $\hat\theta = t(Y_1, \dots, Y_n) \overset{a}{\sim} \mathcal{N}(\theta, Var(\hat\theta))$

We construct an approximate pivotal statistic with $g(t(Y_1, \dots, Y_n); \theta) = \frac{\hat\theta - \theta}{\sqrt{Var(\hat\theta)}} \overset{a}{\sim} \mathcal{N}(0,1)$

! $\mathcal{N}(0,1)$ is independent of θ

Confidence interval N / t - Distr.

assume $Y_i \sim \mathcal{N}(\mu, \sigma^2)$ i.i.d.

then $\hat\mu = \bar Y \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$

We obtain $CI = \left[\bar y \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt n}\right]$

assume $\frac{\bar Y - \mu}{\hat\sigma/\sqrt n} \sim t_{(n-1)}$

with $\hat\sigma^2 := S^2 = \frac{1}{n-1}\sum_{i=1}^n (y_i - \bar y)^2$

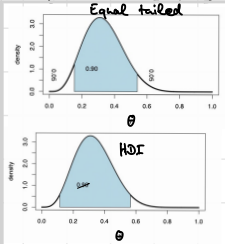
we obtain $CI = \left[\bar y \pm t_{(n-1),(1-\alpha/2)} \cdot \frac{S_n}{\sqrt n}\right]$

Example

$X_i \sim \mathcal{U}(0, \theta)$, i.i.d. $t(X) = \max\{X_1, \dots, X_n\}$

$$P(X_1 = x_1, \dots, X_n = x_n;\theta) = \underbrace{\frac{1}{\theta^n} \cdot I_{(-\infty,\theta)}(\max\{x_i \mid i=1,\dots,n\})}_{= g(t(x),\theta)} \cdot \underbrace{I_{(0,\infty)}(\min\{x_i \mid i=1,\dots,n\})}_{= h(x_1,\dots,x_n)}$$

So the sample maximum is a sufficient statistic for the population maximum.



Asymptotic Normality

MLE - Theory

Idea

The MLE is popular for multiple reasons, one of such being that MLE is asymptotically efficient: in the limit, the MLE achieves the CRLB. Recall that the MLE and other point estimator are themselves random variables. Therefore, a low-variance estimator $\hat\theta_N$ estimates the true parameter θ_0 more precisely.

Theorem Assuming a Fisher-regular distribution with parameter θ_0 from which an i.i.d. sample $Y_1, \dots, Y_n$ is drawn. ! So we assume $G(\cdot) \in \mathcal{F}$. Then the MLE is asymptotically normally distributed with

$$\hat\theta \xrightarrow{\mathcal{D}} \mathcal{N}(\theta_0, I(\theta_0)^{-1})$$

Maximum Likelihood in Misspecified Models

Idea

We derived the previous results under the assumptions that our model is specified correctly, i.e. $G(\cdot) \in \mathcal{F}$.

If we drop this assumption the identity $\mathbb{E}[s(\theta_0)] = \int \frac{\partial \ell(\theta_0)}{\partial\theta} dG(y)$ still holds, but the Fisher information $I(\theta_0) := -\int \frac{\partial^2 \ell(\theta)}{\partial\theta^2} dG(y)$ is no longer equal to the variance of the score $V(\theta_0) := \int s(\theta) \cdot s^T(\theta)\, dG(y) = Var(s(\theta_0))$.

This changes the asymptotic normality behaviour of the MLE.

Theorem Assuming Fisher-regular and i.i.d. sample $Y_1, \dots, Y_n$ drawn from $Y_i \sim G(\cdot)$.

The MLE is asymptotically normally distributed with $\hat\theta \overset{a}{\sim} \mathcal{N}(\theta_0, I^{-1}(\theta_0) \cdot V(\theta_0) \cdot I^{-1}(\theta_0))$

! Since $G(\cdot)$ is unknown, neither $I(\theta_0)$ nor $V(\theta_0)$ can be calculated analytically But empirical estimates are

$$\hat I(\theta_0) = \frac{1}{n}\sum_{i=1}^n -\frac{\partial^2 \log(f(y_i;\hat\theta))}{\partial\theta^2} \quad \text{and} \quad \hat V(\theta_0) = \frac{1}{n}\sum_{i=1}^n \left(\frac{\partial \log(f(y_i;\hat\theta))}{\partial\theta}\right)^2.$$

If $G \in \mathcal{F}$, then $V(\theta_0) = I(\theta_0)$.

Parameter Transformation

Let $\gamma = h(\theta)$ for some bijective transformation h with $\hat\gamma = h(\hat\theta)$. $\ell_\gamma(\gamma) := \ell_\theta(h^{-1}(\gamma))$.

$$I_\gamma(\gamma) = \mathbb{E}\left[-\frac{\partial^2 \ell_\gamma(\gamma)}{\partial\gamma^2}\right] = \frac{\partial\theta}{\partial\gamma} \cdot I_\theta(\theta) \cdot \frac{\partial\theta}{\partial\gamma}$$

It follows $\hat\gamma \overset{a}{\sim} \mathcal{N}(\gamma, \frac{\partial\gamma}{\partial\theta} I_\theta^{-1}(\theta) \frac{\partial\gamma}{\partial\theta})$

Test Theory

Statistical hypothesis test and significance level

A statistical hypothesis test assumes two states H_0 and H_a and two possible decisions "H_0" and "H_a". The decision rule is thereby constructed s.t. $P(\text{"}H_a\text{"} \mid H_0) \leq \alpha$ for a fixed (small) value of α.

The term α is called significance level.

! Rule of thumb: The statement for which we seek evidence is the one that determines H_a.

truth	decision: pred. positive "H_0"	decision: pred. negative "H_1"
positive H_0	correct decision TP	type I error (α error)
negative H_1	type II error (β error)	correct decision TN

p-value

The p-value is the probability of obtaining test results at least as extreme as the actually observed result, under the assumption that the null hypothesis is correct.

Misinterpretations

1. The p-value is not the probability that the null hypothesis is true, or the probability that the alternative hypothesis is false.
2. The p-value is not the probability that the observed effects were produced by random chance alone.
3. The p-value doesn't indicate the size or importance of the observed effect.

Connection between hypotesis test and p-value

The p-values give a measurement for the evidence against the hypothesis, whereas the result of a hypothesis test is binary in the sense that the outcome is "H_0" or "H_a". Still, the two concepts can be linked as follows: For a statistical hypothesis test with a significance level of α it holds that: p-value $\leq \alpha \Leftrightarrow$ "H_a".

Power of a Test

Idea Given a simple Hypothesis system $H_0: \theta = \theta_0$ and $H_a: \theta = \theta_a$ we can compute the type I-error for a given type I-error and for a given test. We are now interested in finding a test that minimizes the type I-error given a type I-error.

Power of a Test The power of a statistical hypothesis test is defined as $\mathbb{P}(\text{"}H_a\text{"} \mid H_a)$ and commonly denoted by $1-\beta$.

The power of a test quantifies the probability of correctly rejecting H_0 if H_a holds.

Likelihood Based Tests

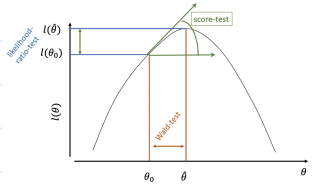
Wald test: $|\hat\theta - \theta_0| > c \Leftrightarrow$ "H_a" where $c = z_{1-\alpha/2} \cdot \sqrt{I_n^{-1}(\theta_0)}$

Likelihood-ratio-test: $2\cdot(\ell(\hat\theta) - \ell(\theta_0)) > c \Leftrightarrow$ "H_a" where $c = \chi^2_p(1-\alpha)$

Score-test: $|s(\theta_0; y_1,\dots,y_n)| > c \Leftrightarrow$ "H_a" where $c = z_{1-\alpha/2} \cdot \sqrt{I(\theta_0)}$

Note: In practice we use $I_n(\hat\theta)$ instead of $I_n(\theta_0)$.

(#) dimension of parameter



Chi-Squared Goodness-of-Fit test

Let $Y_i \in \{1,\dots,K\}$ with $\mathbb{P}(Y_i = k) = \pi_k$, where $\sum_{k=1}^{K} \pi_k = 1$. Assuming an i.i.d. setting with variables $Y_1,\dots,Y_n$ we obtain $M(n, (\pi_1,\dots,\pi_K))$.

define $n_k := \sum_{i=1}^{n} \mathbb{1}_{\{y_i = k\}}$ and $e_k := \mathbb{E}[n_k] = n \cdot \pi_k$ ← true prob.

We consider the hypothesis $H_0: \pi = \pi_0$! Note that the dimension of π is $K-1$ since $\pi_K = 1 - \sum_{k=1}^{K-1} \pi_k$. As test statistic we use

$\chi^2 = \sum_{k=1}^{K} \frac{(n_k - e_k)^2}{e_k} \sim \chi^2_{K-1}$ decision rule "H_a" $\Leftrightarrow \chi^2 \geq (\chi^2_{K-1-p})^{-1}(1-\alpha)$

Chi-squared test of independence

Let $Y_{i1} \in \{1,\dots,K_1\}$ and $Y_{i2} \in \{1,\dots,K_2\}$ and $Y_i = (Y_{i1}, Y_{i2})$. We assume that $\mathbb{P}(Y_{i1} = k, Y_{i2} = \ell) = \pi_{k\ell}$ and we consider the hypothesis

$H_0: \mathbb{P}(Y_{i1} = k, Y_{i2} = \ell) = \mathbb{P}(Y_{i1} = k) \cdot \mathbb{P}(Y_{i2} = \ell)$.

Y_{i1} \ Y_{i2}	1	2	…	K_2	
1	n_{11}	n_{12}		n_{1K_2}	$n_{1\cdot}$
2	n_{21}	n_{22}		n_{2K_2}	$n_{2\cdot}$
⋮					⋮
K_1	$n_{K_1 1}$	$n_{K_1 2}$		$n_{K_1 K_2}$	$n_{K_1\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	…	$n_{\cdot K_2}$	$n_{\cdot\cdot}$

and $\hat\pi_{k\cdot} = \frac{n_{k\cdot}}{n_{\cdot\cdot}}$ and $\hat\pi_{\cdot\ell} = \frac{n_{\cdot\ell}}{n_{\cdot\cdot}}$ and $e_{k\ell} = n_{\cdot\cdot} \cdot \hat\pi_{k\cdot} \cdot \hat\pi_{\cdot\ell} = \frac{n_{k\cdot} \cdot n_{\cdot\ell}}{n_{\cdot\cdot}}$.

This leads to the test-statistic

$\chi^2 = \sum_{k=1}^{K_1} \sum_{\ell=1}^{K_2} \frac{(n_{k\ell} - e_{k\ell})^2}{e_{k\ell}} \sim \chi^2_{(K_1-1)\cdot(K_2-1)}$

decision rule "H_a" $\Leftrightarrow \chi^2 \geq (\chi^2_{(K_1-1)\cdot(K_2-1)})^{-1}(1-\alpha)$

Gauss-Test and t-Test

We assume $Y_i \sim \mathcal{N}(\mu, \sigma^2)$ i.i.d. $i = 1,\dots,n$. We want to test $H_0: \mu \leq \mu_0$, $H_a: \mu > \mu_0$. decision rule "H_a" $\Leftrightarrow \bar Y > c$

Known variance: $c = \mu_0 + z_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}}$ with $\Phi(z_{1-\alpha}) = 1-\alpha$. Gauss-test.

Unknown variance: $c = \mu_0 + t_{(n-1),1-\alpha} \cdot \frac{\hat\sigma}{\sqrt{n}}$ with $\hat\sigma = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar Y)^2$ t-test.

We can use the setting to compare mean values of two groups. Let $Y_{1i} \sim \mathcal{N}(\mu_1, \sigma^2)$ i.i.d. $i = 1,\dots,n_1$ and $Y_{2i} \sim \mathcal{N}(\mu_2, \sigma^2)$ i.i.d. $i = 1,\dots,n_2$

We know $\bar Y_1 - \bar Y_2 \sim \mathcal{N}(\mu_1 - \mu_2, \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2})$. $H_0: \mu_1 - \mu_2 = \mu_0$ and $H_a: \mu_1 - \mu_2 \neq \mu_0$. For independent samples the rule is given by "H_a": $|\bar y_1 - \bar y_2| > c$

Known variance: $c = \mu_0 + z_{1-\alpha/2} \cdot \sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$.

Unknown variance: $c = \mu_0 + t_{(n_1+n_2-2),\,1-\frac{\alpha}{2}} \cdot \hat\sigma \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ with $\hat\sigma^2 = \frac{(n_1-1)\hat\sigma_1^2 + (n_2-1)\hat\sigma_2^2}{n_1+n_2-2}$.

Nonparametric statistics

Nonparametric statistics is based on either being distribution-free or having a specified distribution but with unspecified distribution's parameters.

Mann-Whitney U test

This nonparametric test can be used to test if for randomly selected values X and Y from two populations $\mathbb{P}(X > Y) = \mathbb{P}(Y > X)$. Let $X_1,\dots,X_n \overset{i.i.d.}{\sim} X$ and $Y_1,\dots,Y_m \overset{i.i.d.}{\sim} Y$ and both samples are independent of each other.

The Mann-Whitney U statistic is $U = \sum_{i=1}^{n} \sum_{j=1}^{m} S(X_i, Y_j)$

with $S(X,Y) = \begin{cases} 1 & ; \text{if } X > Y \\ \frac{1}{2} & ; \text{if } X = Y \\ 0 & ; \text{if } X < Y \end{cases}$.

Kolmogorov-Smirnov Test

This nonparametric test can be used to compare difference between the empirical distribution $F_n(y)$ and the hypothetical distribution $F(y;\theta)$.

As test statistic we use the Kolmogorov-Smirnov-Distance

$D_n := \sup_y |F_n(y) - F(y;\theta)|$

The decision rule is "H_a" $\Leftrightarrow D_n \geq KS_{1-\alpha}$ with KS being the Kolmogorov-Smirnov distr.

Testing and Confidence Interval

$[t_\ell(Y), t_r(Y)]$, where $\mathbb{P}_\theta[t_\ell(Y) \leq \theta \leq t_r(Y)] \geq 1-\alpha$ with $Y = (Y_1,\dots,Y_n)$

$\mathbb{P}[\text{"}H_a\text{"} \mid H_0] \leq \alpha$

- For a given confidence interval we can construct a hypothesis test with significance level α:
  $H_0: \theta = \theta_0$. decision rule "H_a" $\Leftrightarrow \theta_0 \notin [t_\ell(Y), t_r(Y)]$.
  Using the decision function $\varphi_\theta(Y) = \begin{cases} 0, & \text{if } \theta \in [t_\ell(Y), t_r(Y)] \\ 1, & \text{otherwise} \end{cases}$
  we obtain $\mathbb{P}(\text{"}H_a\text{"} \mid H_0) = 1 - \mathbb{P}(\text{"}H_0\text{"} \mid H_0)$
  $= 1 - \mathbb{P}(\varphi_\theta(Y) = 0 \mid \theta = \theta_0) = 1 - \underbrace{\mathbb{P}(t_\ell(Y) \leq \theta_0 \leq t_r(Y))}_{\geq 1-\alpha} \leq \alpha$.
- For a given hypothesis test we construct a CI with confidence level $1-\alpha$: Let $\varphi_\theta(Y) := \begin{cases} 0, & \text{if "}H_0\text{"} \\ 1, & \text{if "}H_a\text{"} \end{cases}$.
  Then $CI_{1-\alpha}(Y) = \{\theta \in \Theta \mid \varphi_\theta(Y) = 0\}$.
  $\mathbb{P}_\theta[\theta \in CI_{1-\alpha}(Y)] = 1 - \mathbb{P}_\theta[\theta \notin CI_{1-\alpha}(Y)]$
  $= 1 - \mathbb{P}_\theta[\varphi_\theta(Y) = 1] = 1 - \underbrace{\mathbb{P}[\text{"}H_a\text{"} \mid H_0]}_{\leq \alpha} \geq 1-\alpha$

Neyman-Pearson Lemma

Let $H_0: \theta = \theta_0$ be tested against $H_a: \theta = \theta_a$ with a statistical significance test using level α.

The most powerful test has the decision rule "H_a" $\Leftrightarrow \ell(\theta_0) - \ell(\theta_a) \leq c$ where c is determined s.t. $\mathbb{P}(\text{"}H_a\text{"} \mid H_0) \leq \alpha$. This is called the Neyman-Pearson test.

Example Power function

Let $Y_i \sim \mathcal{N}(\mu, \sigma^2)$ i.i.d. with σ^2 known and consider $H_0: \mu \leq \mu_0$. decision rule "H_a" $\Leftrightarrow \bar Y \geq \mu_0 + z_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}}$.

The power of the test is $\mathbb{P}(\text{"}H_a\text{"} \mid \mu) = \mathbb{P}(\bar Y \geq \mu_0 + z_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}} \mid \mu) = 1 - \Phi\left(z_{1-\alpha} + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}}\right)$.

Given a certain alternative Hypothesis $H_a: \mu = \mu_a$ and a fixed size α one can compute the sample size n that is needed to obtain a certain power $1-\beta$. → $\mu_0 - \mu = \mu_a$

Example Neyman-Pearson-Lemma

Let $Y_1,\dots,Y_n \sim \mathcal{N}(\mu, \sigma^2)$ i.i.d. $H_0: \mu = \mu_0$ and $H_a: \mu = \mu_a$. $\ell(\mu, \sigma^2; x_1,\dots,x_n) = \text{const} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2$

Likelihood ratio: $2\cdot(\ell(\mu_0) - \ell(\mu_a)) = \frac{1}{\sigma^2}\left(\sum_{i=1}^{n} (x_i - \mu_a)^2 - \sum_{i=1}^{n} (x_i - \mu_0)^2\right) \overset{!}{\leq} c$

$\sum_{i=1}^{n} (x_i - \mu_a)^2 - \sum_{i=1}^{n} (x_i - \mu_0)^2 \;\alpha\; \sum_{i=1}^{n} x_i$. Therefore we obtain "$H_a$" $\Leftrightarrow \ell(\mu_0) - \ell(\mu_a) < c \Leftrightarrow \frac{1}{n}\sum_{i=1}^{n} x_i < k$

with k being chosen s.t. $\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} x_i < k \mid \mu = \mu_0\right) \leq \alpha$. $\Rightarrow k = \mu_0 + z_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}}$

Multiple Testing

Motivation

More generalized: Assume we perform m tests and for each test we have a null hypothesis $H_{0j}: \theta_j = \theta_{0j}$, $j = 1,\dots,m$ and an alternative hypothesis H_{aj}, $j = 1,\dots,m$. We perform each test with significance level α. Then the overall type I-error is larger than α. Therefore we need to adjust the significance level s.t. the overall type I-error stays small.

Family-wise error rate (FWER)

probability of a at least one false rejection among the m tests. $\alpha_{FWER} := \mathbb{P}(\text{"}H_{a1}\text{"} \vee \text{"}H_{a2}\text{"} \vee \dots \vee \text{"}H_{am}\text{"} \mid H_{01} \wedge \dots \wedge H_{0m})$.

If we assume that the tests are independent, we get $\alpha_{FWER} = 1 - (1-\alpha)^m$

Bonferroni adjustment

If we set the significance level of each individual test to α/m, where m is the number of tests, then it follows $\alpha_{FWER} \leq 1 - (1 - \alpha/m)^m \leq 1 - (1 - m\frac{\alpha}{m}) = \alpha$. The adjusted significance level is defined by $\alpha_{adjust} = \alpha/m$.

! This adjustment is very conservative and not recommended for large m.

Nested and non-nested models

We look at two possible models which we denote as $\mathcal{F}_1$ and $\mathcal{F}_2$.

If one model is a subset of the other, than we call the models nested.

Assuming $\mathcal{F}_1$ to be the larger model this can be written as

$\mathcal{F}_1 = \{F(\cdot;\theta) \mid \theta \in \Theta_1\}$ and $\mathcal{F}_2 = \{F(\cdot;\theta) \mid \theta \in \Theta_2\}$ with $\Theta_2 \subset \Theta_1$

Alternatively, we can write non-nested models formally as

$\mathcal{F}_1 = \{F_1(\cdot;\theta) \mid \theta \in \Theta_1\}$ and $\mathcal{F}_2 = \{F_2(\cdot;\theta) \mid \theta \in \Theta_2\}$ with $\mathcal{F}_2 \nsubseteq \mathcal{F}_1$ and $\mathcal{F}_1 \nsubseteq \mathcal{F}_2$.

Model comparison idea

We prefer the model with the smaller Kullback-Leibler divergence, i.e. we choose model $\mathcal{F}_1$ over model $\mathcal{F}_2$ if

$KL(G(\cdot); F_1(\cdot;\hat\theta_1)) - KL(G(\cdot); F_2(\cdot;\hat\theta_2)) < 0 \Leftrightarrow \int \log\frac{f(y;\hat\theta_1)}{f(y;\hat\theta_2)}\, dG(y) > 0$

! The quantity $\sum_{i=1}^n \log\left(\frac{f_1(y_i;\hat\theta_1)}{f_2(y_i;\hat\theta_2)}\right)$ does not converge to the integral because we use the data twice: ones for $y_1,\dots,y_n$ and ones for estimating $\hat\theta_1, \hat\theta_2$.

This would lead to overfitting, b.c. more complex models are prefered.

Akaike Information Criterion (AIC)

AIC / BIC Theory

The AIC for model class $\mathcal{F} = \{F(\cdot;\theta) \mid \theta \in \Theta\}$ with $p := \dim(\Theta)$ is defined as $AIC(\mathcal{F}) := -2\cdot\sum_{i=1}^n \log(f(y_i;\hat\theta)) + 2p$.

The AIC is an estimate for the quantity $2\cdot\mathbb{E}\left[KL(G(\cdot); F(Y_1,\dots,Y_n;\hat\theta)\right] - 2\cdot\int \log(g(y))\, dG(y)$.

Therefore $AIC(\mathcal{F}_1) - AIC(\mathcal{F}_2)$ is an estimate for $-2\cdot\mathbb{E}\left[\int \log\frac{f(y;\hat\theta_1)}{f(y;\hat\theta_2)}\, dG(y)\right]$.

Hence, if $AIC(\mathcal{F}_1) < AIC(\mathcal{F}_2)$, we prefer $\mathcal{F}_1$ and otherwise $\mathcal{F}_2$.

Bayesian Information Criterion (BIC)

The idea is the same as with AIC but the BIC of a model $\mathcal{F}$ is defined as $BIC(\mathcal{F}) := -2\cdot\sum_{i=1}^n \log(f(y_i;\hat\theta)) + \log(n)\cdot p$

The BIC favors models with lower complexity, i.e. fewer parameters.

Corrected AIC (AICc)

In the case of small sample statistics the AIC needs to be modified and one should use the corrected AIC defined as $AICc := -2\cdot\ell(\hat\theta) + 2p\cdot\left(\frac{n}{n-p-1}\right)$.

Simulating Distr.

Known Distribution Function

We want to draw random samples from a known distribution $F(\cdot)$.

We need the inverse of $F(\cdot)$. We define $F^{-1}(u) := \inf\{y \mid F(y) \geq u\}$

To sample from $F(\cdot)$ we use the following property:

Let $Y = F^{-1}(U)$ where $U \sim \mathcal{U}(0,1)$ then $Y \sim F(\cdot)$.

Known Density Function (Rejection Sampling)

We want to draw random samples from an unknown distribution F by using its known density function f.

We use rejection sampling to do so: Assume we know a distribution F^*, its inverse and its density f^*. Assume further that $\exists a \in \mathbb{R}\ \forall y \in T_F:\ f(y) < a\cdot f^*(y)$ and that a is known.

↑ called umbrella distribution

Then the simulation procedure is a follows:

1. Draw Y^* from F^* using the method described before.
2. Draw U from $\mathcal{U}(0,1)$.
3. If $U \leq \frac{f(Y^*)}{a\cdot f^*(Y^*)}$ accept Y^*. Otherwise go back to 1. ! acceptence probability is $\frac{1}{a}$.

We can show that the accepted values Y^* follow the distribution F.

Importance Sampling

Same setup as in rejection sampling but we are interested in $\mathbb{E}_F[h(Y)]$ for some function $h(\cdot)$.

We use $\mathbb{E}_F[h(Y)] = \int_{\mathbb{R}} h(y) f(y)\, dy$

$= \int_{\mathbb{R}} h(y)\cdot\frac{f(y)}{f^*(y)} f^*(y)\, dy = \mathbb{E}_{F^*}\left[h(Y)\cdot\frac{f(Y)}{f^*(Y)}\right]$

Using now i.i.d. samples $Y_1^*,\dots,Y_N^* \sim F^*(\cdot)$

we obtain $\frac{1}{N}\cdot\sum_{i=1}^N h(y_i^*)\cdot\frac{f(y_i^*)}{f^*(y_i^*)} \xrightarrow{N\to\infty} \mathbb{E}_F[h(Y)]$

! The closer f is to f^* the smaller the variance of our approximation.

Metropolis (-Hasting) Algorithm

1. Select a starting value $y^{(0)}$ and set $y^{(t)} = y^{(0)}$
2. Based on $y^{(t)}$ propose a new value y^* from a proposal distribution $Y^* \sim H(y \mid y^{(t)})$ where $H(\cdot \mid y^{(t)})$ is a known distribution, with the same support (Träger) as F, from which we can sample. The corresponding density is $h(y \mid y^{(t)})$. In the Metropolis-Algorithm the proposal distribution is symmetric - for example $N(y^{(t)}, \sigma^2)$ with some fixed variance σ^2. In the Metropolis-Hasting-Algorithm the proposal distribution can be skewed. $H(\cdot \mid y^{(t)})$ can depend on $y^{(t)}$ but doesn't need to.
3. Compute the acceptence probability, which is defined as $\alpha(y^*, y^{(t)}) = \min\left\{1,\ \frac{f(y^*)}{f(y^{(t)})}\cdot\frac{h(y^{(t)} \mid y^*)}{h(y^* \mid y^{(t)})}\right\}$. If H is symmetric, then $\frac{h(y^{(t)} \mid y^*)}{h(y^* \mid y^{(t)})} = 1$.
4. Draw $U^* \sim \mathcal{U}(0,1)$ and define $y^{(t+1)} := \begin{cases} y^* & \text{, if } U^* \leq \alpha(y^*, y^{(t)}) \\ y^{(t)} & \text{, otherwise} \end{cases}$
5. Go back to step 2 until t is large enough.
6. Usually the first few values of $y^{(t)}$, i.e. $y^{(1)},\dots,y^{(1500)}$ are discarded (Burn-in)

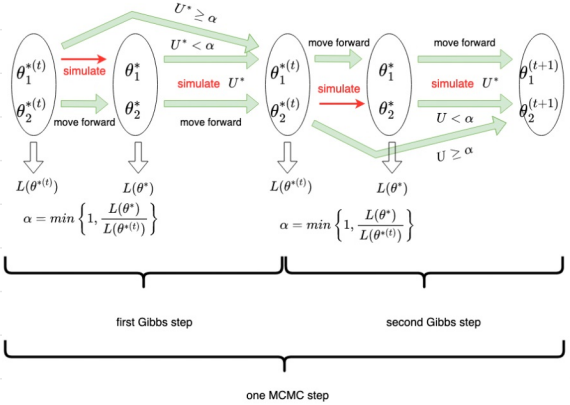
Result: The values $y^{(t)}$ after the Burn-in are distributed according to F, i.e. the stationary distribution is F.

Autocorrelation: The simulated values are not independent. To get a (quasi) i.i.d. sample Y_i^* we can choose only every v-th value as a sample (thinning).

Gibbs sampling

Note that $p_{\theta \mid y_1,\dots,y_n}(\theta \mid y_1,\dots,y_n) \propto \underbrace{\prod_{i=1}^n f(y_i;\theta)}_{:=L(\theta)}\cdot p(\theta;\gamma)$.

1. Let $\underline{\theta} = (\theta_1,\dots,\theta_p) \in \Theta$. Let $\underline{\theta}^* \in \Theta$ and set $(\theta_1^{*(t)},\dots,\theta_p^{*(t)}) = \underline{\theta}^{(t)} = \underline{\theta}_0$.
2. Set $\theta_j^{(t)} = \theta_j^{(t)}$ for $j = 1,\dots,p$ and execute the following steps:
   (i) Set $\underline{\theta}^* = \underline{\theta}^{(t)}$
   (ii) Replace θ_j^* with a sample from the proposal $q(\cdot;\theta_j^{(t)})$
   (iii) Sample $U^* \sim \mathcal{U}(0,1)$
   (iv) If $U^* \leq \alpha(\theta^*, \theta^{*(t)}) = \min\left\{1,\ \frac{L(\underline{\theta}^*)}{L(\underline{\theta}^{(t)})}\cdot\frac{q(\theta_j^{(t)} \mid \theta_j^*)}{q(\theta_j^* \mid \theta_j^{(t)})}\right\}$, then set $\theta_j^{*(t)} = \theta_j^*$. Otherwise leave $\theta_j^{*(t)}$ unchanged.
3. Set $\theta^{(t+1)} = \theta^{*(t)}$ and repeat step 2 until a stationary dist. is reached.



Bernstein-von Mises Theorem

For increasing sample size n and appropriately chosen prior we find $\theta \overset{a}{\sim} N(\hat\theta, I^{-1}(\hat\theta))$

AIC, BIC and Hypothesis Testing

To connect the ideas of model selection with hypothesis testing we assume two nested models $\mathcal{F}_0$ and $\mathcal{F}_1$ with $\mathcal{F}_0 \subset \mathcal{F}_1$. Assume $p_0 = |\Theta_0|$ and $p_1 = |\Theta_1|$ and $p = p_1 - p_0 > 0$.

Let $\hat\theta_0$ and $\hat\theta_1$ be the corresponding MLE's. It can be shown that if model $\mathcal{F}_0$ holds, then $2\cdot(\ell(\hat\theta_0) - \ell(\hat\theta_1)) \sim \chi^2_p$. Formulating the hypothesis $H_0: \mathcal{F} = \mathcal{F}_0$ and alternative $H_1: \mathcal{F} = \mathcal{F}_1$, we obtain the decision rule "H_1" $\Leftrightarrow 2\cdot(\ell(\hat\theta_0) - \ell(\hat\theta_1)) > \chi^2_{p,1-\alpha}$. On the other hand, the decision rule of the AIC can be written as "H_1" $\Leftrightarrow 2\cdot(\ell(\hat\theta_0) - \ell(\hat\theta_1)) > 2p$ and for the BIC we get "H_1" $\Leftrightarrow 2\cdot(\ell(\hat\theta_0) - \ell(\hat\theta_1)) > \log(n)p$

Bayes Inference

Prior and Posterior

Let $\mathcal{F} = \{f(\cdot;\theta) \mid \theta \in \Theta\}$ be our probability model for the i.i.d. data $y_1,\dots,y_n$.

For θ we formulate our (missing) knowledge as prior distribution $\theta \sim p(\cdot;\gamma)$ where parameter $\gamma \in \Gamma$ is called hyper-parameter. We define the prior-structure as the model class $\mathcal{P} = \{p(\cdot;\theta) \mid \gamma \in \Gamma\}$.

The posterior distribution $p_{\theta \mid y}(\theta;\ y_1,\dots,y_n) = \frac{\prod_{i=1}^n f(y_i;\theta)\, p(\theta;\gamma)}{f(y;\gamma)}$ with

$f(y;\gamma) = \int_\Theta \prod_{i=1}^n f(y_i;\theta)\cdot p(\theta;\gamma)\, d\theta$.

Conjugate prior distribution

For a given family of distributions $\mathcal{F}$, we call $\mathcal{P}$ the set of conjugate prior distributions, if it exists and if $p_{\theta \mid y}(\cdot) \in \mathcal{P}$. That is the posterior is from the same family of distributions as the prior distribution.

Example

Lets assume $Y \sim Poi(\lambda)$ and $\lambda \sim \Gamma(\alpha, \beta)$, i.e. $p(\lambda) \propto \lambda^{\alpha-1}\cdot\exp(-\lambda\beta)$.

Then $p_{\lambda \mid y_1,\dots,y_n}(\lambda \mid y_1,\dots,y_n) \underset{\text{Bayes Thm.}}{\propto} f(y_1,\dots,y_n \mid \lambda)\cdot p(\lambda) = \left(\prod_{i=1}^n \lambda^{y_i}\exp(-\lambda)\right)\cdot\lambda^{\alpha-1}\cdot\exp(-\lambda\beta) = \lambda^{\sum y_i + \alpha - 1}\cdot\exp(-\lambda(n+\beta))$

Consequently $\lambda \mid y_1,\dots,y_n \sim \Gamma(\sum_{i=1}^n y_i + \alpha,\ n+\beta)$. So given the set of Poisson distribution $\mathcal{F}$ with parameter λ, the set of gamma distributions $\mathcal{P}$ is a set of conjugate prior distributions.

Parameter	$\mathcal{F}$	$\mathcal{P}$
π	Binomial distribution	Beta distribution
λ	Poisson distribution	Gamma distribution
μ	Normal distribution	Normal distribution
λ	Exponential distribution	Gamma distribution

Table 10.1 Examples for conjugate distributions

Approximate Bayes Computation (ABC)

Concept

To simulate from the posterior $\theta|y_{obs}$ by using MCMC we needed to know/compute the Likelihood $f(y;\theta)$.

If the Likelihood is unknown/uncomputable, then we can use the ABC-method to at least approximatly simulate from the posterior.

Exact Algorithm

1. Generate θ^* from some prior dist. $p_\theta(\theta)$.
2. Generate y^* from the assumed dist. i.e. $f_Y(\cdot;\theta^*)$.
3. If $y^* = y$, with y being the available data, then accept θ^*. Otherwise go back to step 1.
4. Repeat 1-3. until a sufficient amount of simulated values is reached.

Approx. Algorithm

1. Generate θ^* from some prior dist. $p_\theta(\theta)$.
2. Generate y^* from the assumed dist. i.e. $f_Y(\cdot;\theta^*)$.
3. If $d(y^*,y) \le \varepsilon$, with y being the available data, then accept θ^*. Otherwise go back to step 1.
   ↳ $d(\cdot,\cdot)$ is a distance measure and ε needs to be chosen appropriately.
4. Repeat 1-3. until a sufficient amount of simulated values is reached.

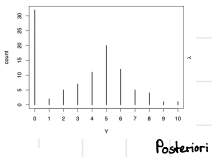
Example - Zero inflated data

Consider the same setup as above. We define $d(y^*,y) := \sum_{k=0}^{N} (\bar{x}_k^* - \bar{x}_k)^2$

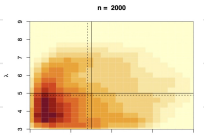
with $N = 10$, $\bar{x}_k := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(y_i = k)$ and $\bar{x}_k^* = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(y_i^* = k)$ and $\varepsilon = 0.05$.

Now we proceed with the Algorithm and simulate θ^*, y^*.

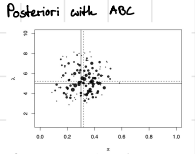
Acctual data $y_1,\dots,y_n$



Posteriori with MCMC



Posteriori with ABC



Point size $\propto d(y^*,y)$

Variational Bayes Reasoning

Concept

In this method we have the same setting and goal as in MCMC, but we trade the accuracy of MCMC (exact posterior) for speed/computation time. Instead of simulating exactly from the posterior we simulate from a more simple distribution $q(\cdot)$ that is approximatly the posterior $p(\theta|y)$. e.g. $q(\cdot) = \mathcal{N}(\hat{\theta}_p, I_p^{-1}(\hat{\theta}))$. (see Thm. below).

In general, let $q(\theta,\vartheta) \in \mathcal{Q} := \{q(\cdot;\vartheta) | \vartheta \in \Gamma\}$ be some easy-to-handle dist. and we are interested in $q(\cdot;\hat{\vartheta}) = \arg\min_{\vartheta\in\Gamma} \{KL(q(\cdot;\vartheta), p(\cdot|y))\}$.

This results in a less accurate estimate compared to MCMC but is less computationally expensive than MCMC

Solving the Optimization Problem

$$KL(q(\cdot;\vartheta), p(\cdot|y)) = \int \log\left(\frac{q(\theta;\vartheta)}{p(\theta|y)}\right) q(\theta;\vartheta)\, d\theta = \underbrace{-\int \log\left(\frac{f(y;\theta)\,p(\theta)}{q(\theta;\vartheta)}\right) q(\theta;\vartheta)\, d\theta}_{LB(q(\cdot;\vartheta))} + \overbrace{\log(f(y))}^{\text{unknown}}$$

To minimize $KL(q(\cdot;\vartheta), p(\cdot|y))$ we need to maximize $LB(q(\cdot;\vartheta))$.

$$\frac{\partial LB(q(\cdot;\vartheta))}{\partial \vartheta} = \int \log\left(\frac{f(y;\theta)\,p(\theta)}{q(\theta;\vartheta)}\right) \frac{\partial q(\theta;\vartheta)}{\partial \vartheta} d\theta - \int \log\left(\frac{f(y;\theta)\,p(\theta)}{q(\theta;\vartheta)}\right) \frac{\partial q(\theta;\vartheta)}{\partial \vartheta} d\theta - \underbrace{\int \frac{\partial \log(q(\theta;\vartheta))}{\partial \vartheta} \cdot q(\theta;\vartheta)\, d\theta}_{=0 \text{ (see score function)}}$$

$$= \mathbb{E}_{q(\cdot;\vartheta)}\left[\frac{\partial \log(q(\theta;\vartheta))}{\partial \vartheta} \cdot \log\left(\frac{f(y;\theta)\,p(\theta)}{q(\theta;\vartheta)}\right)\right] \overset{!}{\approx} 0$$

❗ Because of the unknown component we can't quantify how close our approximation actually is.

Approximating Solution of the Optimization Problem

We want to solve $\frac{\partial LB(q(\cdot;\vartheta))}{\partial \vartheta} = \mathbb{E}_{q(\cdot;\vartheta)}\left[\frac{\partial \log(q(\theta;\vartheta))}{\partial \vartheta} \cdot \log\left(\frac{f(y;\theta)\,p(\theta)}{q(\theta;\vartheta)}\right)\right] \overset{!}{\approx} 0$.

We can replace the expected value by simulated values, i.e. for given $\vartheta^{(t)}$ we draw $\theta^{*(k)} \sim q(\cdot;\vartheta^{(t)})$ for $k = 1,\dots,K$ and compute

$$\widehat{\frac{\partial LB(q(\cdot;\vartheta^{(t)}))}{\partial \vartheta}} = \frac{1}{K}\sum_{i=1}^{K} \frac{\partial \log(q(\theta^{*(k)};\vartheta^{(t)}))}{\partial \vartheta} \cdot \log\left(\frac{f(y;\theta^{*(k)})\,p(\theta^{*(k)})}{q(\theta^{*(k)};\vartheta^{(t)})}\right)$$

We can now use $\vartheta^{(t+1)} = \vartheta^{(t)} + r^{(t)} \cdot \widehat{\frac{\partial LB(q(\cdot;\vartheta^{(t)}))}{\partial \vartheta}}$ to solve the problem numerically.

↑ learning rate

Penalized version of Bernstein Thm.

For sufficiently large n we have $\theta|y \overset{\text{a}}{\sim} \mathcal{N}(\hat{\theta}_p, I_p^{-1}(\hat{\theta}_p))$

where $\hat{\theta}_p := \arg\max_{\theta\in\Theta} l_p(\theta) = \arg\max_{\theta\in\Theta} \sum_{i=1}^{n} \log(f(y_i;\theta)) + \log(p(\theta))$

where the priori p might depend on further hyperparameters and $I_p(\theta) := I(\theta) + \frac{\partial^2 p(\theta)}{\partial \theta^2}$.

Log-Linear Model (Multiple observation)

Multivariate Data

Let $Y^{(1)},\dots,Y^{(n)}$ be independent observations, i.e. $Y^{(i)} = (Y_1^{(i)},\dots,Y_p^{(i)})$.

$$\mathbb{P}(Y^{(i)} = y^{(i)},\dots,Y^{(n)} = y^{(n)}) \propto \prod_{i=1}^{n} \exp\left(\theta_0 + \sum_{j=1}^{p}\sum_{k_j=1}^{K_j-1} s_{i,k_j,\dots}(y_j^{(i)})\cdot\theta_{\cdot,k_j,\dots} + \sum_{j_1=2}^{p}\sum_{j_2=1}^{j_1-1}\sum_{k_{j_1}=1}^{K_{j_1}-1}\sum_{k_{j_2}=1}^{K_{j_2}-1} s_{i,k_{j_1},k_{j_2},\dots}(y_{j_1}^{(i)},y_{j_2}^{(i)})\cdot\theta_{\cdot,k_{j_1},k_{j_2},\dots} + \dots + \sum_{k_1=1}^{K_1-1}\dots\sum_{k_p=1}^{K_p-1} s_{i,k_1,k_2,\dots,k_p}(y_1^{(i)},\dots,y_p^{(i)})\cdot\theta_{k_1,k_2,\dots,k_p}\right)$$

$$= \exp\Bigg(\underbrace{n_{\dots,\dots}\cdot\theta_0}_{\text{zero variables}} + \underbrace{\sum_{j=1}^{p}\sum_{k_j=1}^{K_j-1} n_{k_j,\dots}\cdot\theta_{k_j,\dots}}_{\text{sums over one variable}} + \underbrace{\sum_{j_2=2}^{p}\sum_{j_1=1}^{j_2-1}\sum_{k_{j_1}=1}^{K_{j_1}-1}\sum_{k_{j_2}=1}^{K_{j_2}-1} n_{k_{j_1},k_{j_2},\dots}\cdot\theta_{\cdot,k_{j_1},k_{j_2},\dots}}_{\text{sums over two-wise pairs of variables}} + \dots + \underbrace{\sum_{k_1=1}^{K_1-1}\dots\sum_{k_p=1}^{K_p-1} n_{k_1,k_2,\dots,k_p}\cdot\theta_{k_1,k_2,\dots,k_p}}_{\text{sum over all p variables}}\Bigg)$$

We call the terms $\theta_{\dots}$ with more than one variable involved interaction terms, e.g. $\theta_{k_1,k_2,\dots}$ is the pairwise interaction between Y_1 and Y_2 for the outcome k_1,k_2.

❗ The interaction term is zero iff the involved variables are independent.

3-dimensional Log-Linear model and model hierarchy

We look at a Log-linear model for count-data with $p = 3$ and $K_1 = K_2 = K_3 = 2$.

i.e. a $2\times2\times2$ table. We use the notation 1 for Y_1, 2 for Y_2, 3 for Y_3.

The full/saturated model has $2\cdot2\cdot2 = 8$ parameters and is given below

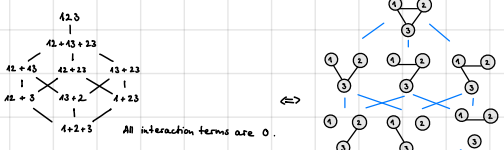
$$\mathbb{P}(Y^{(i)} = y^{(i)},\dots,Y^{(n)} = y^{(n)}) = \exp\left(\theta_0 + n_{1\dots}\theta_{1\dots} + n_{\cdot1\cdot}\theta_{\cdot1\cdot} + n_{\dots1}\theta_{\dots1} + n_{11\cdot}\theta_{11\cdot} + n_{1\cdot1}\theta_{1\cdot1} + n_{\cdot11}\theta_{\cdot11} + n_{111}\theta_{111}\right)$$

If we condition on the sample size, i.e. θ_0 and all the interaction terms are non-zero, then we can denote the model as $1*2*3$ or 123.

If we impose $\theta_{111} = 0$, then we can denote the model as $1*2 + 1*3 + 2*3$ or $12 + 13 + 23$.

If we further restrict $\theta_{11\cdot} = 0$, then we can denote the model as $1*3 + 2*3$ or $13 + 23$.

If we continue on setting different interaction terms to 0, then this leads to some model-hierarchy:



All interaction terms are 0. ⇔

Conditional Independence

Let $Y = (Y_1,\dots,Y_p)$ be a p-dimensional random variable with $f_{123\dots p}(y_1,\dots,y_p)$ as corresponding density or prob. function. We define $A,B,C \subseteq \{1,\dots,p\}$ s.t.

$A \cap B = A \cap C = B \cap C = \emptyset$ and $A \neq \emptyset$ and $B \neq \emptyset$.

We say A and B are conditionally independent given C, denoted as $(A \perp\!\!\!\perp B)|C$ iff $f_{(A\cup B)|C}(y_A, y_B|y_C) = f_{A|C}(y_A|y_C)\cdot f_{B|C}(y_B|y_C)$. This holds iff

$f_{A\cup B\cup C}(y_A,y_B,y_C) = h_{A\cup C}(y_A,y_C)\cdot h_{B\cup C}(y_B,y_C)$ for some function $h_{A\cup C}, h_{B\cup C}$.

Visualization of Conditional Independence

Let A,B,C be defined as before. We can visualize $(A \perp\!\!\!\perp B)|C$ as the graph 

.

The missing edge between A,B and the path that goes through C translates to $(A \perp\!\!\!\perp B)|C$.

If instead of sets of variables we want to visualize single variables, we construct the graph as follows: Let $G = (E,V)$ be a graph, i.e.

$E = \{1,\dots,p\}$ and $V \subseteq E\times E$. Let $i,j \in E$. $(i \perp\!\!\!\perp j)|E\setminus\{i,j\} \Leftrightarrow (i,j) \notin V$

Example

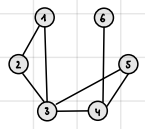
Let $p = 6$ and

$1 \perp\!\!\!\perp 4 | \{2,3,5,6\}$
$1 \perp\!\!\!\perp 5 | \{2,3,4,6\}$
$1 \perp\!\!\!\perp 6 | \{2,3,4,5\}$
$2 \perp\!\!\!\perp 4 | \{1,3,5,6\}$
$2 \perp\!\!\!\perp 5 | \{1,3,4,6\}$
$2 \perp\!\!\!\perp 6 | \{1,3,4,5\}$
$3 \perp\!\!\!\perp 6 | \{1,2,4,5\}$
$5 \perp\!\!\!\perp 6 | \{1,2,3,4\}$

From the graph we can see

$(\{1,2\} \perp\!\!\!\perp \{4,5,6\}) | \{3\}$

$(\{6\} \perp\!\!\!\perp \{1,2,3,5\}) | \{4\}$



Multivariate Normal Dist.

Let $Y \sim \mathcal{N}_p(\mu, \Sigma)$

$$f(y;\mu,\Sigma) = \frac{\exp\left(-\frac{1}{2}\cdot(y-\mu)^T\Sigma^{-1}(y-\mu)\right)}{\sqrt{(2\pi)^p\cdot\det(\Sigma)}}$$

Using $\Omega := \Sigma^{-1}$ we can rewrite f as

The inverse of the covariance matrix

$$f(y;\mu,\Sigma) \propto \exp\left(-\frac{1}{2}\sum_{j=1}^{p}\Omega_{jj}\cdot y_j^2 + \frac{1}{2}\sum_{j=1}^{p}\sum_{\ell\neq j}^{p}\Omega_{j\ell}\, y_j y_\ell\right)$$

If $\Omega_{j\ell} = 0$ and $C = \{1,\dots,p\}\setminus\{j,\ell\}$, then

$f(y;\mu,\Sigma) \propto h_{C\cup\{j\}}(y_C, y_j)\cdot h_{C\cup\{\ell\}}(y_C, y_\ell)$

Hence, $\Omega_{j\ell} = 0 \Leftrightarrow (Y_j \perp\!\!\!\perp Y_\ell)|Y_C$

Conditional Independence in Log-Linear Model

Let $Y = (Y_1, Y_2, Y_3)$ be a discrete random variable with $Y_i \in \{1,\dots,K_i\}$ $i = 1,2,3$.

$(Y_1 \perp\!\!\!\perp Y_2)|Y_3$ or $(1 \perp\!\!\!\perp 2)|3$ $\Leftrightarrow$ $\forall i \in \{1,\dots,K_3\}$: $\mathbb{P}(Y_1 = k_1|Y_2 = k_2, Y_3 = k_3) = \underbrace{\mathbb{P}(Y_1 = k_1|Y_3 = k_3)}_{=h_{13}(k_1,k_3)} \cdot \underbrace{\mathbb{P}(Y_2 = k_2|Y_3 = k_3)}_{=h_{23}(k_2,k_3)}$

$\Leftrightarrow$ $\forall i \in \{1,\dots,K_3\}$ $\forall k_1 \in \{1,\dots,K_1\}$: $\mathbb{P}(Y_1 = k_1|Y_2 = k_2, Y_3 = k_3) = \mathbb{P}(Y_1 = k_1|Y_3 = k_3)$

(given Y_3 we don't get additional information from Y_2)

In our log-linear model, this holds iff all parameters that involve components Y_1, Y_2 are 0.

i.e. $(Y_1 \perp\!\!\!\perp Y_2)|Y_3 \Leftrightarrow \forall i \in \{1,\dots,K_3\}$: $\theta_{k_1,k_2,i} = 0 \;\wedge\; \theta_{k_1,k_2,k_3} = 0$.

Using the model notation from before, this results in the model $13 + 23$.

Rare Events

Setup

Let $Y \sim B(n,\pi)$ and we want to estimate π based on the sample $y_1,\dots,y_n$.

But if $\forall i: y_i = 0$ or in general $n \cdot \pi \approx 0$ then our maximum likelihood estimation method fails and the usual confidence interval is meaningless.

Therefore we need another way to construct a confidence interval for rare events.

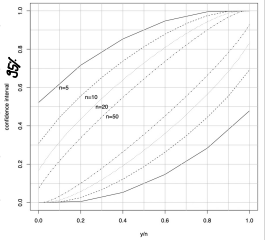
Clopper and Pearson Exact Confidence Interval

Let $Y \sim B(n,\pi)$. For $Y = y$ the exact confidence interval is (c_L, c_U) with

$c_L := \inf\{\pi \mid P(Y \le y;\pi) > \frac{\alpha}{2}\}$, $c_U := \sup\{\pi \mid P(Y \ge y;\pi) > \frac{\alpha}{2}\}$.

This can be computed by using the Beta-distribution:

$c_L = Be\left(\frac{\alpha}{2}, y, n-y+1\right)$, $c_U = Be\left(1-\frac{\alpha}{2}, y+1, n-y\right)$ (in R: qbeta())



Generalized Extreme Value Distribution

The GEV distribution has three parameters.

location paramater μ, scale parameter σ, shape parameter γ.

The distribution function ist given by

$$H(z) = \begin{cases} \exp(-\exp(-z)) & \text{, for } \gamma = 0 \\ \exp(-(1+\gamma z)^{-1/\gamma}) & \text{, for } \gamma \neq 0 \end{cases} \quad \text{for } z = \frac{x-\mu}{\sigma}.$$

1. Gumbel distribution ($\gamma = 0$): $F(x;\mu,\sigma,0) = \exp(-\exp(-\frac{x-\mu}{\sigma}))$
2. Fréchet distribution ($\gamma > 0$): $F(x;\mu,\sigma,\gamma) = \exp\left(-(1+\gamma\frac{x-\mu}{\sigma})^{-1/\gamma}\right) \cdot \mathbb{1}_{(0,\infty)}\left(1+\gamma\frac{x-\mu}{\sigma}\right)$
3. (Reversed) Weibull distr. ($\gamma < 0$): $F(x;\mu,\sigma,\gamma) = \begin{cases} \exp\left(-(1+\gamma\frac{x-\mu}{\sigma})^{-1/\gamma}\right) & \text{, if } 1+\gamma\frac{x-\mu}{\sigma} > 0 \\ 1 & \text{, otherwise} \end{cases}$

Extreme Data

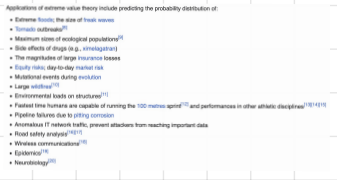
Setup

We are interested in the maximum of data. We therefore assume the classical setup $Y_i \sim F(\cdot)$ i.i.d. $i = 1,\dots,n$ and we are interested in the distr. of $Y_{max,n} = \max\{Y_1,\dots,Y_n\}$, i.e. $F_{Y_{max,n}}(y) = F_{Y_i}^n(y)$. In practice the distr. $F(\cdot)$ might be unknown. We can then use the Fisher-Tippett-Gnedenko-theorem to get an asymptotic result. Note: $F_{Y_{min,n}}(y) = 1 - (1 - F_{Y_i}(y))^n$

Fisher-Tippett-Gnedenko-Theorem

Suppose there exist two sequences of $(a_n)_{n\in\mathbb{N}} \subseteq \mathbb{R}^+$, $(b_n)_{n\in\mathbb{N}} \subseteq \mathbb{R}$, s.t. $F^n(a_n y + b_n) \xrightarrow{D}_{n\to\infty} H(y)$ with $H(\cdot)$ being a non-degenerate distribution, i.e. the support of H is not just a single point.

Then the limit distribution H belongs to one of the three distributions, that can be jointly written as the generalized extreme value (GEV) distribution.



Moving Average Process of order q

$$Y_t = Z_t + \theta_1 Z_{t-1} + \dots + \theta_q Z_{t-q} = \sum_{i=0}^{q} \theta_i B^i Z_{t-i}$$

with $Z_t \sim \mathcal{N}(0,\sigma^2)$ i.i.d. and $\theta_0 = 1$

Assume $\mu_Y(t) \equiv 0$ and $\gamma_Y(t,t') = \gamma_Y(|t-t'|) = \gamma_Y(h) = Cov(Y_t, Y_{t+h})$ with $h \in \{0,1,\dots,T\}$.

$\gamma_Y(h)$ is called autocovariance function.

The variance matrix is defined as $\Gamma = \begin{bmatrix} \gamma(0) & \gamma(1) & \dots & \gamma(T) \\ \gamma(1) & \gamma(0) & \ddots & \vdots \\ \gamma(T) & \dots & \ddots & \gamma(0) \end{bmatrix}$.

B.c. of $\gamma(0) = V(Y_t)$ we know that $\rho(h) = \frac{\gamma(h)}{\gamma(0)}$. We call $\rho(h)$ the autocorrelation function

The correlation matrix is defined as $R = \begin{bmatrix} \rho(0) & \rho(1) & \dots & \rho(T) \\ \rho(1) & \rho(0) & \ddots & \vdots \\ \rho(T) & \dots & \ddots & \rho(0) \end{bmatrix}$.

We define white noise as the time series $Y_t \sim \mathcal{N}(0,\sigma^2)$ with $\gamma_Y(h) = 0$ for $h \neq 0$

Missing Data

Setup

$Y = (Y_1,\dots,Y_p) = (Y_O, Y_M)$ $\quad O \in \{1,\dots,p\}$, $M = \{1,\dots,p\} \setminus O$.

Y_O denotes observed features in an observation

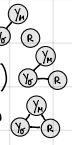
Y_M denotes the features in an observation that are NA.

Our data set consists of $Y^{(1)},\dots,Y^{(n)}$

MCAR: $P(R^{(i)} \mid Y^{(i)}) = P(R^{(i)})$

MAR: $P(R^{(i)} \mid Y^{(i)}) = P(R^{(i)} \mid Y_O^{(i)})$

MNAR: $P(R^{(i)} \mid Y^{(i)})$ depends on $Y_M^{(i)}$



Datenqualität

Let $\mu = \mathbb{E}[Y]$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{\sum R^{(i)} y^{(i)}}{\sum R^{(i)}}$

$$\mu - \bar{y} = \underbrace{Cor(R,Y)}_{\text{Qualität}} \cdot \underbrace{\sqrt{Var(Y)}}_{\text{Variabilität}} \cdot \underbrace{\frac{N-n}{n}}_{\text{Quantität}}$$

Auto Regressive Process AR(p)

$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \dots + \phi_p Y_{t-p} + Z_t$ with $Z_t \sim \mathcal{N}(0,\sigma^2)$ i.i.d.

We can rewrite this using the backshift operator B, i.e. $B^p Y_t = Y_{t-p}, p \in \mathbb{Z}$.

The AR(p) process can then be written as $(1 - \sum_{j=1}^{p} \phi_j B^j) Y_t = Z_t$

AR(1)

$Y_t = \phi Y_{t-1} + Z_t$. Then $Cov(Y_t, Y_{t-h}) = \phi Cov(Y_{t-1}, Y_{t-h}) = \phi\gamma(0)$

Hence $\rho(1) = \frac{\phi\gamma(0)}{\gamma(0)} = \phi$. In general $\rho(h) = \phi^h$. For $|\phi| < 1$ we get

$\Gamma = \begin{bmatrix} 1 & \phi & \phi^2 & \dots & \phi^T \\ \phi & 1 & & & \phi^{T-1} \\ \phi^2 & & \ddots & & \vdots \\ \vdots & & & & \phi \\ \phi^T & \phi^{T-1} & \dots & \phi & 1 \end{bmatrix}$. Note that $Y_t = \phi Y_{t-1} + Z_t = \phi^2 Y_{t-2} + \phi Z_{t-1} + Z_t = \dots = \sum_{j=0}^{\infty} \phi^j Z_{t-j}$

This results in $V[Y_t] = \sum_{j=0}^{\infty} (\phi^2)^j \cdot \sigma^2 = \frac{\sigma^2}{1-\phi^2}$

MA(1)

$Y_t = Z_t + \theta Z_{t-1}$. Then $\gamma(0) = \sigma^2(1+\theta^2)$

$\gamma(1) = \theta\sigma^2$, $\gamma(h) = 0$ for $h > 1$ and hence $\rho(h) = \begin{cases} 1 & , h = 0 \\ \frac{\theta}{1+\theta^2} & , h = 1 \\ 0 & , h > 1 \end{cases}$

ARMA(p,q)

$$Y_t = \underbrace{\phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \dots + \phi_p Y_{t-p}}_{AR(p)} + Z_t + \underbrace{\theta_1 Z_{t-1} + \dots + \theta_q Z_{t-q}}_{MA(q)}$$

We define $\Phi(B) := -\sum_{j=0}^{p} \phi_j B^j$ and $\Theta(B) := \sum_{i=0}^{q} \theta_i B^i$ with $\phi_0 = -1$, $\theta_0 = 1$

Then we can rewrite the ARMA(p,q) as $\Phi(B) Y_t = \Theta(B) Z_t$

! Representation is not unique with that notation. Can be solved if we find $\Phi^{-1}(B)$ s.t. $Y_t = \Phi^{-1}(B) \cdot \Theta(B) Z_t$.

Example

ARMA(1,1) process: $\Phi(B) = 1 - \phi B$, $\Theta(B) = 1 + \theta_1 B$

then $\Phi^{-1}(B) = \sum_{j=0}^{\infty} \phi^j B^j$ and $Y_t = \sum_{i=0}^{\infty} \Psi_i B^i Z_{t-i}$

with $\Psi_0 = 1$ and $\Psi_i = \phi^{i-1}(\phi + \theta)$

→ Hence, all AR components can be transfered into MA components.

Causal process

An ARMA(p,q) is causal if it can be rewritten to

$$Y_t = \sum_{j=0}^{\infty} \psi_j B^j Z_{t-j} = \Psi(B) Z_t$$

with $\sum_{j=0}^{\infty} |\psi_j| < \infty$

PACF for p

ACF for q

E.M-Algorithmus

1. Expectation step: Impute missing values based on $\theta_{(t)}$

   Compute $Q(\theta,\theta_{(t)}) := \sum_{i=1}^{n} \int l_i(\theta) f(y_{iM} \mid y_{iO}; \theta_{(t)})\, dy_{iM}$

   with $l_i(\theta) = \log(f(y_i;\theta))$ and $y_i = (y_{iO}, y_{iM})$

   ! Easy for exponential family.
2. Maximisation step: Calculate the new MLE using the imputed values.

   Maximise $Q(\theta,\theta_{(t)})$ w.r.t. θ, i.e. solve

   $s(\theta_{(t+1)};\theta_{(t)}) = 0$ with $s(\theta;\theta_{(t)}) = \frac{\partial Q(\theta,\theta_{(t)})}{\partial\theta}$
3. Return to step 1 until convergence