Uncertainty

Aleatoric / Epistemic uncertainty

Aleatoric uncertainty is uncertainty arising from the inherent randomness of an event. For some input x and output y the aleatoric uncertainty is the uncertainty originating from a stochastic / non-deterministic relationship between x and y that can be expressed as a probability model $Y \sim G(\cdot|x)$ with G being a random distribution based on x with $\mathrm{Var}(Y|x) > 0$.

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. With the probability model from above $Y \sim G(\cdot|x)$ we can say that epistemic uncertainty is due to the fact that we don't know $G(\cdot|x)$ and replace it by a numerical model.

Data generating process

For simplicity we ignore the input x and assume that our quantity Y is drawn from some unknown distr. $G(\cdot)$. i.e. $Y \sim G(\cdot)$ where $G(\cdot)$ is a distribution function that is unknown to us. We call $G(\cdot)$ the data generating process and assume that the drawn data $y_1, \dots, y_n$, which are realisations of $Y_i \sim G(\cdot)$, $i = 1, \dots, n$, is i.i.d.

Glevenko-Cantelli: $F_{emp}(y_1, \dots, y_n) \xrightarrow{n \to \infty} G(\cdot)$

Statistical model

We want to approximate $G(\cdot)$ through a parametric distribution function $F(\cdot\,; \theta)$, where $\theta = (\theta_1, \dots, \theta_p)^T$ is a parameter vector. We call the set of distribution functions $\mathcal{F} = \{F(\cdot\,; \theta) \mid \theta \in \Theta\}$ statistical model.

→ We assume that the data is generated by an element of set $\mathcal{F}$ i.e. $Y \sim F(\cdot\,; \theta_0)$ where θ_0 is called the "true" parameter.

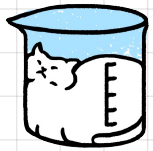
Note: In ML $\mathcal{F}$ is often called hypothesis space.

Estimator/Statistic

We aim to set parameter θ data driven to some value $\hat{\theta}$. i.e. for a function $t: \mathcal{D} \to \Theta$ we call $\hat{\theta} := t(y_1, \dots, y_n)$ an estimator and the function t a statistic

Note: In ML t is also called learner

Learning and Estimating



Kullback-Leibler-Divergence

Kullback-Leibler Divergence

Let $G(\cdot)$ and $F(\cdot\,; \theta)$ be two distribution functions and $g(y)$, $f(y;\theta)$ the corresponding density functions.

We define the Kullback-Leibler Divergence as

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

! KL divergence is not a metric or distance measure because it lacks symmetry. We take G as an anchor point.

For a set $\mathcal{F}$ we define the distance between G and $\mathcal{F}$ as $KL(G, \mathcal{F}) := \min_{F(\cdot;\theta) \in \mathcal{F}} KL(G, F(\cdot;\theta))$.

Solving $KL(G, \mathcal{F})$ - theoretically

If we assume $\mathcal{F}$ to be a set of Fisher-regular distributions, we can obtain the minimum through differentiation:

$$0 \overset{!}{=} \frac{\partial}{\partial\theta} \int \log\left(\frac{g(y)}{f(y;\theta)}\right) dG(y) \overset{(*)}{=} -\int \frac{\partial \log(f(y;\theta))}{\partial\theta}\, dG(y) = \mathbb{E}\left[\frac{\partial \log(f(Y;\theta))}{\partial\theta}(\theta_0)\right]$$

Expected value can't be calculated because G is unknown.

$$(*)\quad \frac{\partial}{\partial\theta} \log\left(\frac{g(y)}{f(y;\theta)}\right) = \frac{\partial}{\partial\theta}\left(\log(g(y)) - \log(f(y;\theta))\right) = -\frac{\partial \log(f(y;\theta))}{\partial\theta}$$

Positivity of KL

Note: $\log(y) \le y - 1$.

$$KL(G, F) = \int \log\left(\frac{g(y)}{f(y)}\right) \cdot g(y)\, d\mu(y)$$
$$= -\int \log\left(\frac{f(y)}{g(y)}\right) g(y)\, d\mu(y)$$
$$\ge \int \left(1 - \frac{f(y)}{g(y)}\right) g(y)\, d\mu(y)$$
$$= \int g(y) d\mu(y) - \int f(y)\, d\mu(y)$$
$$= 1 - 1 = 0$$

Fisher regular

A distribution $f(y;\theta)$ is Fisher-regular if the following properties hold:

(i) The support of Y is not dependent upon θ. i.e. $\{y \mid f(y;\theta) > 0\}$ does not depent on θ.

(ii) The possible parameter space Θ is open. i.e. if θ is univariate, Θ is an open interval (a, b).

(iii) The probability function $f(y;\theta)$ can be differentiated twice w.r.t. θ.

(iv) Integration with respect to some suitable measure and differentiation with respect to the parameter θ are exchangeable. i.e. $\frac{\partial}{\partial\theta} \int f(y;\theta)\, dy = \int \frac{\partial}{\partial\theta} f(y;\theta)\, dy$

Counterexample

$$f(y;\theta) = \begin{cases} 1/\theta & , y \in [0, \theta] \\ 0 & , \text{otherwise} \end{cases}$$

Support depends on θ.

Therefore f is not Fisher-regular.

Solving $KL(G, \mathcal{F})$ - empirically

Idea: Estimate $KL(G, \mathcal{F})$ with data and then minimize the estimate to find the F with the smallest estimated KL-divergence.

We know: $\frac{1}{n} \sum_{i=1}^{n} t(y_i) \xrightarrow{n \to \infty} \mathbb{E}[t(Y)]$.

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

($\log(g(y_i))$: unknown)

To minimize the estimate based on θ we need to maximize $\sum_{i=1}^{n} \log(f(y_i;\theta))$

Log-Likelihood

$$\ell(\theta) := \ell(\theta;\, y_1, \dots, y_n) = \sum_{i=1}^{n} \log(f(y_i;\theta))$$

Optimal parameter

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

The estimate for the optimal parameter is defined through $\frac{\partial \ell}{\partial\theta}(\hat{\theta}) = 0$

$\hat{\theta}$ is called maximum likelihood estimate. We will show that $F(\cdot;\hat{\theta}) \xrightarrow{n \to \infty} F(\cdot;\theta_0)$

Squared Loss and the Likelihood Principle

Squared Loss

Assume that Y is normal distributed and that the variance does not depend on x, then we get $\ell(\theta) = -\frac{1}{2}\sum_{i=1}^{n}(y_i - \mu(x_i;\theta))^2$. Maximizing $\ell(\theta)$ is equivalent to minimizing $\text{loss}(\theta) = \sum_{i=1}^{n}(y_i - \mu(x_i;\theta))^2$ $\longrightarrow$ Commonly used in ML, but is equivalent to assuming $Y \sim N(\theta)$

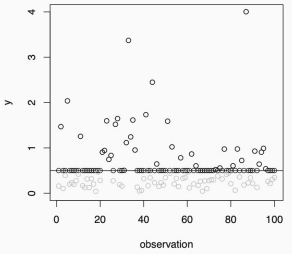
Likelihood principle

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 Y is log-normal distributed but values $Y < c$ can't be measured / observed (called censored). Let y_i be the measurement and $z_i = \log(y_i)$. Taking an L2-loss we would get $\min_{\mu}\sum_{i=1}^{n}(z_i - \mu)^2$. The result will be biased upwards. Taking the censoring into account we get

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



With L2-Loss: $\hat{\mu} = -0.42$, $\hat{\sigma}^2 = 0.22$

With likelihood-principal: $\hat{\mu} = -1.04$, $\hat{\sigma}^2 = 1.04$

$\mu_0 = -1$, $\sigma_0^2 = 1$

Properties of Estimates

Bias

The bias is defined as $\text{bias}(\hat{\theta},\theta_0) = \mathbb{E}[\hat{\theta}] - \theta_0$

An estimate $\hat{\theta}$ is called unbiased if $\text{bias}(\hat{\theta},\theta_0) = 0$

$\hat{\theta}$ is called asymptotically unbiased if $\text{bias}(\hat{\theta},\theta_0) \xrightarrow{n\to\infty} 0$

Mean Squared Error

For an estimate $\hat{\theta}$ the MSE is defined as $MSE(\hat{\theta},\theta_0) := \mathbb{E}[(\hat{\theta}-\theta_0)^2]$

$= Var(\hat{\theta}) + \text{bias}^2(\hat{\theta},\theta_0)$

Consistency

The estimate $\hat{\theta}$ is MSE consistent for θ_0 if for increasing data size n with data drawn from $G(\cdot)$ it holds $MSE(\hat{\theta},\theta_0) \xrightarrow{n\to\infty} 0$

Efficiency

Let $\hat{\theta}_1 = t(Y_1,\dots,Y_n)$ and $\hat{\theta}_2 = \tilde{t}(Y_1,\dots,Y_n)$ be estimators for θ_0. We call $\hat{\theta}_1$ more efficient than $\hat{\theta}_2$ if $MSE(\hat{\theta}_1,\theta_0) < MSE(\hat{\theta}_2,\theta_0)$

Example

Assume $Y_i \sim N(\mu,\sigma^2)$ i.i.d. and we propose the estimate $\hat{\mu} = t(\underline{y}) = \sum_{i=1}^{n}\omega_i \cdot y_i$ for some weights ω_i. For t to be unbiased we postulate $\sum\omega_i = 1$.

$$Var(t(\underline{y})) = \sum\omega_i^2\sigma^2 \overset{\text{set } \omega_i = \frac{1}{n}+d_i \text{ with } \sum d_i = 0}{=} \sum(\tfrac{1}{n}+d_i)^2\sigma^2 = \sum(\tfrac{1}{n}+d_1^2+\tfrac{2d_i}{=0})\sigma^2 = \frac{\sigma^2}{n} + \sigma^2\sum d_i^2 + \frac{2\sigma^2}{n}\underbrace{\sum d_i}_{=0} \geq \frac{\sigma^2}{n}$$

$\Rightarrow \omega_i = \frac{1}{n}$ gives the most efficient estimate out the possible weight-choices.

Sufficiency

A statistic $t(Y_1,\dots,Y_n)$ is called sufficient for θ if the conditional distribution $P(Y_1 = y_1,\dots,Y_n = y_n \mid t(Y_1,\dots,Y_n) = t_0;\theta)$ does not depend on θ.

Intuitiv: The unknown parameter θ interacts with the data $y_1,\dots,y_n$ only via the statistic $t(Y_1,\dots,Y_n) = t_0$

A bijective transformation of a sufficient statistic is also a sufficient statistic.

Neyman-factorisation

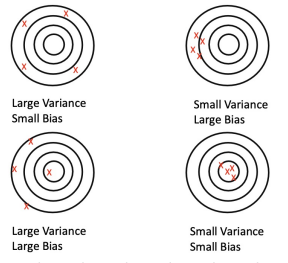
A statistic $t(Y_1,\dots,Y_n)$ is sufficient for θ iff the density decomposes to $f(y_1,\dots,y_n;\theta) = h(y_1,\dots,y_n)\cdot g(t(y_1,\dots,y_n);\theta)$ where h does not depend on θ and g depends on the data only through the statistic $t(y_1,\dots,y_n)$.

Minimal sufficient

The statistic $t(y_1,\dots,y_n)$ 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. $t(Y_1,\dots,Y_n) = m(\tilde{t}(Y_1,\dots,Y_n))$.

That is $\tilde{t}(y_1,\dots,y_n) = \tilde{t}(x_1,\dots,x_n) \Rightarrow t(y_1,\dots,y_n) = t(x_1,\dots,x_n)$

If two statistics are minimal sufficient, then there exists a one-to-one relationship between them.



Example

$Y_i \sim \mathcal{B}(\pi)$ and $t(Y_1,\dots,Y_n) = \bar{Y} \in \{0, \frac{1}{n}, \frac{2}{n}, \dots, 1\}$. For data $y_1,\dots,y_n$ we denote $t(y_1,\dots,y_n) = t_0$ and $n_0 = n\cdot t_0$.

$$P(Y_1 = y_1,\dots,Y_n = y_n \mid t(Y_1,\dots,Y_n) = t_0;\pi) = \frac{P(Y_1 = y_1,\dots,Y_n = y_n, \sum_{i=1}^{n} y_i = n_0;\pi)}{P(\sum_{i=1}^{n} y_i = n_0;\pi)}$$

$$= \begin{cases} \frac{\prod_{i=1}^{n}\pi^{y_i}\cdot(1-\pi)^{1-y_i}}{\binom{n}{n_0}\pi^{n_0}\cdot(1-\pi)^{n-n_0}} & \text{, for } \sum_{i=1}^{n} y_i = n_0 \\ 0 & \text{, otherwise} \end{cases}$$

$$= \begin{cases} \frac{1}{\binom{n}{n_0}} & \text{, for } \sum_{i=1}^{n} y_i = n_0 \\ 0 & \text{, otherwise} \end{cases}$$

Distribution is independent of π.

$\Rightarrow t(Y_1,\dots,Y_n) = \bar{Y}$ is sufficient.

Example

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

$$P(X_1 = x_1,\dots,X_n = x_n;\theta) = \prod_{i=1}^{n}\frac{1}{\theta}\cdot I_{(0,\theta)}(x_i)$$

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

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

Estimation and Uncertainty

Estimation variance

Due to $\hat{\theta} = t(y_1, \dots, y_n)$ being dependent on data from $G(\cdot)$, the outcome of t is subject to variance due to the aleatoric uncertainty. We call this the estimation variance.

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 $\hat{\theta} = t(Y_1, \dots, Y_n) \dot{\sim} N(\theta, Var(\hat{\theta}))$

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

! $N(0,1)$ is independent of θ

Confidence interval

To quantify the estimation variance of a univariate parameter θ we construct an interval with our statistic. 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 $P_\theta(\theta \in [t_\ell(Y_1, \dots, Y_n), t_r(Y_1, \dots, Y_n)]) \geq 1-\alpha$ for all θ.

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

Example

The pivotal statistic from below $g(\hat{\theta}, \theta) = \frac{\hat{\theta} - \theta}{\sqrt{Var(\hat{\theta})}} \sim N(0,1)$ can be used to construct the CI in the following way:

$$1-\alpha \doteq P\left(z_{\alpha/2} \leq \frac{\hat{\theta} - \theta}{\sqrt{Var(\hat{\theta})}} \leq z_{1-\alpha/2}\right) \Leftrightarrow 1-\alpha \doteq P\left(\hat{\theta} + z_{\alpha/2} \cdot \sqrt{Var(\hat{\theta})} \leq \theta \leq \hat{\theta} + z_{1-\alpha/2} \cdot \sqrt{Var(\hat{\theta})}\right)$$

Using $z_{\alpha/2} = -z_{1-\alpha/2}$ we obtain $CI = \left[\hat{\theta} - z_{1-\alpha/2} \cdot \sqrt{Var(\hat{\theta})},\ \hat{\theta} + z_{1-\alpha/2} \cdot \sqrt{Var(\hat{\theta})}\right]$

If we assume $Y_i \sim N(\mu, \sigma^2)$ i.i.d. then $\hat{\mu} = \bar{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

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

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 feature 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. $\longrightarrow$ $P(\theta \in [-1,1])$ is nonsense b.c. θ ain't a random variable.

t - Distribution

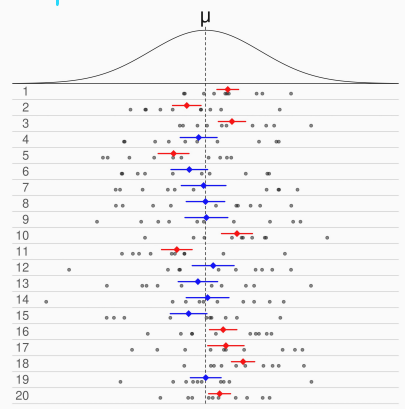
! In the CI above we assume σ^2 as given, but in real life σ^2 is unknown, hence we need to estimate it.

$\hat{\sigma}^2 := S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$. Note: $\frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$ but $\frac{\bar{Y} - \mu}{\hat{\sigma}/\sqrt{n}} \not\sim N(0,1)$

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

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

Example



20 samples with corresponding 50% CI for μ.

68 - 95 - 99.7 - rule

For $X \sim N(\mu, \sigma)$ we get

$P(X \in [\mu \pm \sigma]) \approx 68\%$

$P(X \in [\mu \pm 2\sigma]) \approx 95\%$

$P(X \in [\mu \pm 3\sigma]) \approx 99.7\%$

Credibility interval

Our knowledge of the parameter θ is given by the postirior distr. $p_\theta(\theta \mid y_1, \dots, y_n)$.

The credibility interval is defined as $P_\theta(\theta \in [t_\ell(y_1, \dots, y_n), t_r(y_1, \dots, y_n)] \mid y_1, \dots, y_n) = \int_{t_\ell(y_1, \dots, y_n)}^{t_r(y_1, \dots, y_n)} p_\theta(\theta \mid y_1, \dots, y_n)\, d\theta \geq 1-\alpha$

There are two natural ways to construct t_ℓ and t_r

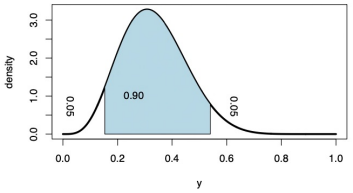
i) $\int_{-\infty}^{t_\ell} p_\theta(\theta \mid y_1, \dots, y_n)\, d\theta = \int_{t_r}^{\infty} p_\theta(\theta \mid y_1, \dots, y_n)\, d\theta = \frac{\alpha}{2}$

Problem: It occurs that $\theta_1 \notin [t_\ell, t_r]$ and $\theta_2 \in [t_\ell, t_r]$ but $p_\theta(\theta_1 \mid y) > p(\theta_2 \mid y)$

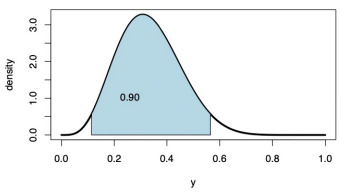
ii) Highest posterior density credibility interval:

$HDI(y_1, \dots, y_n) = \{\theta \mid p_\theta(\theta \mid y_1, \dots, y_n) \geq c\}$ with c chosen s.t. $\int_{HDI} p_\theta(\theta \mid y_1, \dots, y_n)\, d\theta = 1-\alpha$

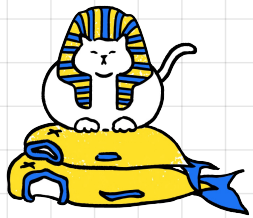
i)



ii)



Maximum Likelihood Estimation



Score equation

Score function

The first derivative of $l(\theta;y)$ w.r.t. θ is called the score function $s(\theta;y) := \frac{\partial l(\theta;y)}{\partial \theta} = \frac{\partial \log f(y;\theta)}{\partial \theta}$

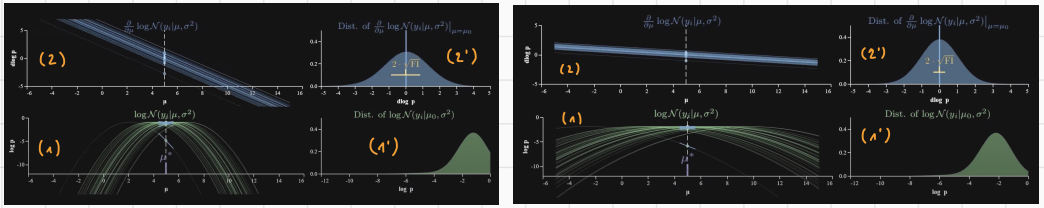
Under regularity we have $\mathbb{E}_\theta(s(\theta;y)) = \int \frac{\partial \log f(y;\theta)}{\partial \theta} \cdot f(y;\theta)\, dy$

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

Maximum Likelihood (MLE)

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

Intuition



(1) Log-likelihood function for samples from the data generating function (here $\mathcal{N}(y_i|\mu;\sigma^2)$) - left and right have different values of σ.

(2) slope/derivative of the log-likelihood functions in (1) - score functions

(2') Distribution of the score function value evaluated at different values of the parameter μ.

If the peak in (1) is sharper (left), it is 'easier' to tell what the true parameter μ^* is, than when the peak is broader (right)

This means that the same sample size results in different confidence about μ^* depending on the peakedness of the log-likelihood-functions.

The peakedness corresponds to the variance in the score-function-values evaluated at the true parameter μ^*. A sharper peak corresponds to higher variance and vice versa. We call this variance the Fisher information $I(\mu^*)$.

If the second derivative at μ^* exists, then its negative expected value is equal to the Fisher information.

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

Fisher Information

The Fisher Information is the variance of the score function, i.e. $I(\theta) := Var_\theta[s(\theta;Y)] = \mathbb{E}_\theta[s(\theta;Y)^2] = \int \left(\frac{\partial \log f(y;\theta)}{\partial \theta}\right)^2 f(y;\theta)\, dy$.

Under certain conditions and regularity it holds that $I(\theta) = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial \theta^2} \log f(Y;\theta)\right]$

! In this course we use the convention that the Fisher Information $I(\theta)$ refers not to only one random variable Y but to an random vector $Y_1, \dots, Y_n$ i.e. $I(\theta) := Var_\theta[s(\theta;Y_1,\dots,Y_n)]$. This also means, that the Fisher Information in this lecture depends on the sample size n.

The connection between those two definitions is, that if $Y_1, \dots, Y_n$ are i.i.d. then $Var_\theta[s(\theta;Y_1,\dots,Y_n)] = n \cdot Var_\theta[s(\theta;Y_i)]$

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

For the case of $\theta \in \mathbb{R}^N$ with $N \geq 2$ we transform the definition of the Fisher Information to a Fisher Information matrix:

We define $[I(\theta)]_{ij} := \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial \theta_i} \log f(Y;\theta)\right)\left(\frac{\partial}{\partial \theta_j} \log f(Y;\theta)\right)\right]$. Under certain conditions and regularity it holds that $[I(\theta)]_{ij} = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f(Y;\theta)\right]$

Example

Let $Y \sim B(n,\pi)$. Then $l(\pi) = y\log(\pi) + (n-y)\cdot\log(1-\pi)$ which leads to

$s(\pi;y) = \frac{\partial l}{\partial \pi}(\pi) = \frac{y}{\pi} - \frac{n-y}{1-\pi} = \frac{y-y\pi-n\pi+y\pi}{\pi(1-\pi)} = \frac{y}{\pi(1-\pi)} - \frac{n}{1-\pi}$

$\mathbb{E}[s(\pi,Y)] = \frac{n\cdot\pi}{\pi} - \frac{n-n\cdot\pi}{1-\pi} = 0$

$\frac{\partial^2 l}{\partial \pi^2}(\pi) = -\frac{y}{\pi^2} - \frac{n-y}{(1-\pi)^2}$

$I(\pi) = \mathbb{E}\left[-\frac{\partial^2 l}{\partial \pi^2}(\pi)\right] = \frac{n\cdot\pi}{\pi^2} + \frac{n-n\pi}{(1-\pi)^2} = \frac{n}{\pi} + \frac{n}{1-\pi} = \frac{n}{\pi(1-\pi)}$

$Var(s(\pi,Y)) = \frac{1}{\pi^2\cdot(1-\pi)^2} \cdot Var(Y) = \frac{n}{\pi(1-\pi)} = I(\theta)$

Cramér Rao Lower Bound (CRLB)

Idea

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 $Y = (Y_1, \dots, Y_n) \in \mathbb{R}^{\mathbb{N}}$ be a random vector with $Y \sim F(\cdot;\theta)$ where $F(\cdot;\theta)$ is a 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 $Var(t(Y)) \geq \frac{1}{I(\theta)}$

Asymptotic Normality

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

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

Proof of asymptotic normality

Using the mean-value theorem we get: $\exists \tilde{\theta}$ between $\hat{\theta}$ and θ_0 s.t. $\frac{\partial^2 \log f(\theta; Y_1,\dots,Y_n)}{\partial \theta^2}(\tilde{\theta}) = \frac{\frac{\partial \log f(\theta; Y_1,\dots,Y_n)}{\partial \theta}(\hat{\theta}) - \frac{\partial \log f(\theta; Y_1,\dots,Y_n)}{\partial \theta}(\theta_0)}{\hat{\theta} - \theta_0}$.

This is equivalent to $0 \overset{\text{by def.}}{=} s(\hat{\theta}; Y_1,\dots,Y_n) = s(\theta_0; Y_1,\dots,Y_n) + \frac{\partial^2 \log f(\theta; Y_1,\dots,Y_n)}{\partial \theta^2}(\tilde{\theta}) \cdot (\hat{\theta} - \theta_0)$

This results in $\hat{\theta} - \theta_0 = - \frac{s(\theta_0; Y_1,\dots,Y_n)}{\frac{\partial^2 \log f(\theta; Y_1,\dots,Y_n)}{\partial \theta^2}(\tilde{\theta})}$

! Here we used $G(\cdot) \in \mathcal{F}$ and therefore $\mathrm{Var}(s(\theta_0; Y_1,\dots,Y_n)) = I_n(\theta_0)$

$s(\theta; Y_1,\dots,Y_n) = \frac{\partial}{\partial \theta} \sum_{i=1}^{n} \log f(Y_i; \theta) = \sum_{i=1}^{n} \frac{\partial s_i(\theta; Y_i)}{\partial \theta} \Rightarrow s(\theta_0; Y_1,\dots,Y_n) = \sum_{i=1}^{n} \frac{\partial s_i}{\partial \theta}(\theta_0; Y_i) - n \cdot \overbrace{E[s(\theta_0, Y)]}^{=0} \xrightarrow{\mathcal{D}} \mathcal{N}(0, I(\theta_0))$ (by CLT)

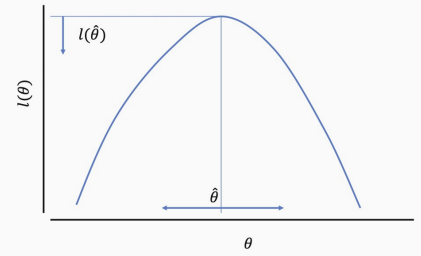
$\frac{1}{n} \cdot \frac{\partial^2 \log f(\theta; Y_1,\dots,Y_n)}{\partial \theta^2}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \log f(\theta; Y_i)}{\partial \theta^2} \xrightarrow{P} E\left[\frac{\partial^2 \log f(\theta; Y)}{\partial \theta^2}\right]$. In combination with $\hat{\theta} \xrightarrow{P} \theta_0$ and $\tilde{\theta} \in (\hat{\theta}, \theta_0)$ or $\tilde{\theta} \in (\theta_0, \hat{\theta})$ we get $\frac{\partial^2 \log f(\theta; Y_1,\dots,Y_n)}{\partial \theta^2}(\tilde{\theta}) \xrightarrow{P} n \cdot E\left[\frac{\partial^2 \log f(\theta_0; Y)}{\partial \theta^2}\right] = -I(\theta_0)$

Slutsky theorem: $s(\theta_0; Y_1,\dots,Y_n) \xrightarrow{\mathcal{D}} \mathcal{N}(0, I(\theta_0))$ and $\frac{\partial^2 \log f(\theta; Y_1,\dots,Y_n)}{\partial \theta^2}(\tilde{\theta}) \xrightarrow{P} -I(\theta_0)$ $\Rightarrow \hat{\theta} - \theta_0 = - \frac{s(\theta_0; Y_1,\dots,Y_n)}{\frac{\partial^2 \log f(\theta; Y_1,\dots,Y_n)}{\partial \theta^2}(\tilde{\theta})} \xrightarrow{\mathcal{D}} \mathcal{N}(0, I^{-1}(\theta_0))$

We conclude: $\hat{\theta} \xrightarrow{\mathcal{D}} \mathcal{N}(\theta_0, I^{-1}(\theta_0))$. So in the limit, MLE achieves the smallest possible variance.

Likelihood Ratio

Idea



Instead of looking at the distribution of $\hat{\theta} - \theta_0$ one can look at the distribution of $l(\hat{\theta}) - l(\theta_0)$.

Log likelihood ratio

$lr(\hat{\theta}; \theta) := l(\hat{\theta}) - l(\theta) = \log\left(\frac{L(\hat{\theta})}{L(\theta)}\right)$ Note: $lr(\hat{\theta}, \theta) \geq 0$

Theorem

For increasing sample size n the likelihood ratio for a Fisher regular distribution converges to a chi-squared distribution χ^2_p.

That is: $2 \cdot (l(\hat{\theta}) - l(\theta_0)) \overset{\mathcal{D}}{\approx} \chi^2_p$ with $p := \dim(\theta)$.

Maximum Likelihood and Parameter Transformation

Idea

Often, the parameter of a model is restricted, e.g. $X \sim B(n, \pi)$ with $\pi \in (0,1)$. Maximization with restricted parameters can be numerically clumsy and it is therefore often useful to transform the parameter to a unrestricted version.

Parameter Transformation

Generally, let θ be the parameter of a model and $l_\theta(\theta)$ the log likelihood.

Let $\gamma = h(\theta)$ for some bijective transformation h with $\hat{\gamma} = h(\hat{\theta})$.

The likelihood for γ is then defined as $l_\gamma(\gamma) := l_\theta(h^{-1}(\gamma))$.

It follows $\frac{\partial l_\gamma(\gamma)}{\partial \gamma} = \frac{\partial (l_\theta \circ h^{-1})}{\partial \gamma}(\gamma) = \frac{\partial l_\theta}{\partial \theta}(h^{-1}(\gamma)) \cdot \frac{\partial h^{-1}(\gamma)}{\partial \gamma}$ and $\frac{\partial l_\gamma}{\partial \gamma}(\hat{\gamma}) = \underbrace{\frac{\partial l_\theta}{\partial \theta}(\hat{\theta})}_{=0} \cdot \frac{\partial \theta}{\partial \gamma} = 0$

Hence, we obtain the MLE of γ by transforming the MLE of θ.

Furthermore, we see $\frac{\partial^2 l_\gamma(\gamma)}{\partial \gamma^2} = \frac{\partial}{\partial \gamma}\left(\frac{\partial l_\theta(\theta)}{\partial \theta} \cdot \frac{\partial \theta}{\partial \gamma}\right) = \frac{\partial \theta}{\partial \gamma} \cdot \frac{\partial^2 l_\theta(\theta)}{\partial \theta^2} \cdot \frac{\partial \theta}{\partial \gamma} + \underbrace{\frac{\partial l_\theta(\theta)}{\partial \theta}}_{E[\dots]=0} \cdot \frac{\partial^2 \theta}{\partial \gamma^2}$

$I_\gamma(\gamma) = E\left[-\frac{\partial^2 l_\gamma(\gamma)}{\partial \gamma^2}\right] = \frac{\partial \theta}{\partial \gamma} \cdot I_\theta(\theta) \cdot \frac{\partial \theta}{\partial \gamma}$. So the Fisher information for γ easily results from θ.

$\hat{\gamma} \overset{\mathcal{D}}{\approx} \mathcal{N}(\gamma, I_\gamma^{-1}(\gamma))$ or $\hat{\gamma} \overset{\mathcal{D}}{\approx} \mathcal{N}(\gamma, \frac{\partial \gamma}{\partial \theta} I_\theta^{-1}(\theta) \frac{\partial \gamma}{\partial \theta})$

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 $E[s(\theta_0)] = \int \frac{\partial l(\theta_0)}{\partial \theta} dG(y)$ still holds, but the Fisher information $I(\theta_0) := -\int \frac{\partial^2 l(\theta)}{\partial \theta^2} dG(y)$ is no longer equal to the variance of the score $V(\theta_0) := \int s(\theta) \cdot s^T(\theta) dG(y) = \mathrm{Var}(s(\theta_0))$. This changes the asymptotic normality behaviors of the MLE.

Theorem

Assuming a model of Fisher-regular distributions and an i.i.d. sample $Y_1,\dots,Y_n$ drawn from $Y_i \sim G(\cdot)$. The MLE is asymptotically normally distributed with $\hat{\theta} \overset{\mathcal{D}}{\approx} \mathcal{N}(\theta_0, I^{-1}(\theta_0) \cdot V(\theta_0) \cdot I^{-1}(\theta_0))$, where $\theta_0 := \underset{\theta \in \Theta}{\arg\min}\, KL(G(\cdot), F(\cdot; \theta))$ and $I(\theta_0)$ and $V(\theta_0)$ is defined as above.

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

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

Numerical Calculation of the MLE

Idea

We use a Newton-Raphson approach to compute the MLE by finding the root of the score-function.

Note that in first order approximation we have $0 = s(\hat{\theta}) \approx s(\theta_0) - I(\theta_0) \cdot (\hat{\theta} - \theta_0) \Rightarrow \hat{\theta} = \theta_0 + I^{-1}(\theta_0) \cdot s(\theta_0; y)$

The simple iteration scheme can be derived from the formula above:

(i) Initialize $\theta_{(0)}$ and $t = 0$.
(ii) Compute $\theta_{(t+1)} := \theta_{(t)} + I^{-1}(\theta_{(t)}) s(\theta_{(t)}; y)$
(iii) Repeat (ii) until $\|\theta_{(t+1)} - \theta_{(t)}\| \leq \varepsilon$ for some fixed error tolerance ε.
(iv) Set $\hat{\theta} = \theta_{(t+1)}$

Data Driven Decisions

Significance Tests and Statistical Hypothesis Tests

Hypothesis and p-value

In statistical tests we often make a statement about the parameters describing a population. We denote that statement as H_0 and call it the null hypothesis. If there is an alternative hypothesis, then we denote it as H_A.

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.

truth	decision	
	"H_0"	"H_1"
H_0	correct decision	type I error (α error)
H_1	type II error (β error)	correct decision

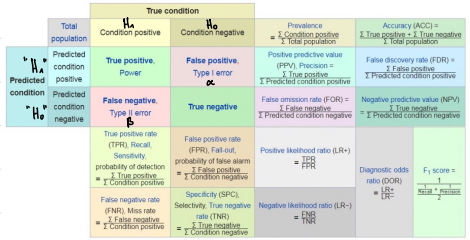
Suggested thresholds

Fisher argued that the smaller the p-value, the stronger the evidence against the null hypothesis H_0.

p-value $\leq 0.1 \Leftrightarrow$ weak evidence against H_0.

p-value $\leq 0.05 \Leftrightarrow$ increased evidence against H_0.

p-value $\leq 0.01 \Leftrightarrow$ strong evidence against H_0.



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.

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.

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".

Classical Statistical Parameter Test

Setup

We assume that $G(\cdot) \in \mathcal{F}(\cdot\,;\theta)$ where $\mathcal{F} \in \mathcal{F}$ is Fisher regular and $Y_1, \dots, Y_n \sim F(\cdot\,;\theta)$ i.i.d.

The decision problem equals $H_0: \theta \in \Theta_0$ and $H_A: \theta \in \Theta_A$ where $\Theta = \Theta_0 \cup \Theta_A$. We restrict the subsequent representation to cases where $\Theta_0 = \{\theta_0\}$ and $\Theta_A = \Theta \setminus \{\theta_0\}$ or $\Theta_0 = \{\theta \mid \theta \leq \theta_0\}$ and $\Theta_A = \Theta \setminus \Theta_0$

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 $H_0: \mu \leq \mu_0$, $H_A: \mu > \mu_0$. As test statistic we use $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$.

We formulate the decision rule "H_A" $\Leftrightarrow \bar{Y} > c$ where $c = \mu_0 + z_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}}$ with $\Phi(z_{1-\alpha}) = 1-\alpha$. This test is known as Gauss-test.

The Gauss-test requires that the variance σ^2 is known. If it's not, we can rely on the t-distribution and use

$c = \mu_0 + t_{(n-1),1-\alpha} \frac{\hat{\sigma}}{\sqrt{n}}$ with $\hat{\sigma} = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$. This test is known as Gauss-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 $\bar{Y}_1 - \bar{Y}_2 \sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}\right)$. $H_0: \mu_1 - \mu_2 = \mu_0$ and $H_A: \mu_1 - \mu_2 \neq \mu_0$. For independent samples the rule is given by "H_A": $|\bar{y}_1 - \bar{y}_2| > c$

with $c = \mu_0 + z_{1-\alpha/2} \cdot \sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$. If the variance is unknown, we replace it by an estimate which results in

$c = \mu_0 + t_{(n_1+n_2-2),\,1-\frac{\alpha}{2}} \cdot \hat{\sigma} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ with $\hat{\sigma}^2 = \frac{(n_1-1)\hat{\sigma}_1^2 + (n_2-1)\hat{\sigma}_2^2}{n_1+n_2-2}$.

Likelihood Based Tests

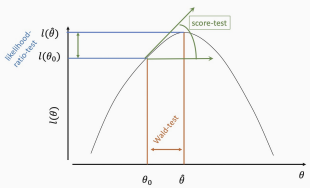
Wald test: $|\hat{\theta} - \theta_0| > c \Leftrightarrow$ "H_A" where $c = z_{1-\alpha/2} \cdot \sqrt{I_n^{-1}(\theta_0)}$

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 a multinomial distribution $M(n, (\pi_1, \dots, \pi_K))$. Let $Y = (Y_1, \dots, Y_n)$, then Y has the distribution $\mathbb{P}(Y = y;\pi) = \mathbb{P}((Y_1, \dots, Y_K) = (y_1, \dots, y_K);\pi) = \frac{n!}{y_1! \cdot \dots \cdot y_K!} \cdot \pi_1^{y_1} \cdot \dots \cdot \pi_K^{y_K}$. Let $\pi_0 = (\pi_{0_1}, \dots, \pi_{0_K})$ be the true probability value and define $n_k := \sum_{i=1}^{n} \mathbb{1}_{\{y_i = k\}}$ and $e_k := \mathbb{E}[n_k] = n \cdot \pi_{0_k}$.

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 the Chi-squared statistic $\chi^2 = \sum_{k=1}^{K} \frac{(n_k - e_k)^2}{e_k} \sim \chi^2_{K-1}$.

! More general: If $H_0: Y \sim F_0(\cdot)$, then $\chi^2 \sim \chi^2_{K-1-p}$ with p being the number of parameters that need to be estimated.

We formulate the decision rule "H_A" $\Leftrightarrow \chi^2 \geq \chi^2_{K-1-p}(1-\alpha)$

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

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

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 the 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_0" $\Leftrightarrow D_n \geq KS_{1-\alpha}$ with KS being the Kolmogorov-Smirnov distr.

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 II-error for a given type I-error and for a given test. We are now interested in finding a test that minimizes the type II-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.

Example

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

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

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 $1-\beta$.

Example

We assume $Y_i \sim \mathcal{N}(\mu,\sigma^2)$ i.i.d. $i=1,\dots,n$. We want to test $H_0: \mu \leq \mu_0$, $H_A: \mu > \mu_0$. As test statistic we use $Y_{med} = \text{median}(Y_1,\dots,Y_n)$

We can formulate a decision rule "H_A": $Y_{med} > c$ with c s.t. $\mathbb{P}(\text{"}H_A\text{"} \mid H_0) \leq \alpha$.

But the power of this test will be lower than the power of the gauss-test.

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

Let $X_1,\dots,X_n \sim \mathcal{N}(\mu,\sigma^2)$ 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: $2\cdot(\ell(\mu_0) - \ell(\mu_A)) = \frac{1}{\sigma^2}\left(\sum_{i=1}^{n}(x_i-\mu_A)^2 - \sum_{i=1}^{n}(x_i-\mu_0)^2\right) \overset{!}{\leq} c$

$\sum_{i=1}^{n}(x_i-\mu_A)^2 - \sum_{i=1}^{n}(x_i-\mu_0)^2 \propto \sum_{i=1}^{n} x_i$. 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. $\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} x_i < k \mid \mu=\mu_0\right) \leq \alpha$. $\Rightarrow k = \mu_0 + z_{1-\alpha}\cdot\frac{\sigma}{\sqrt{n}}$

Testing and Confidence Interval

Connection

We now have two concepts for drawing statistical inference:

- Confidence intervals $[t_l(Y), t_r(Y)]$, where $\mathbb{P}_\theta[t_l(Y) \leq \theta \leq t_r(Y)] \geq 1-\alpha$ with $Y=(Y_1,\dots,Y_n)$
- Hypothesis tests $\mathbb{P}[\text{"}H_A\text{"} \mid H_0] \leq \alpha$

Those two concepts turn out to be two sides of the same coin.

- For a given confidence interval we can construct a hypothesis test with significance level α:

  Let $H_0: \theta=\theta_0$. We then use the decision rule "H_0" $\Leftrightarrow \theta_0 \in [t_l(Y), t_r(Y)]$.

  Using the decision function $\varphi_{\theta_0}(Y) = \begin{cases} 0 & , \text{if } \theta \in [t_l(Y), t_r(Y)] \\ 1 & , \text{otherwise} \end{cases}$ we obtain

  $\mathbb{P}(\text{"}H_A\text{"} \mid H_0) = 1 - \mathbb{P}(\text{"}H_0\text{"} \mid H_0) = 1 - \mathbb{P}(\varphi_{\theta_0}(Y) = 0 \mid \theta=\theta_0) = 1 - \underbrace{\mathbb{P}(t_l(Y) \leq \theta_0 \leq t_r(Y))}_{\geq 1-\alpha} \leq \alpha$.
- For a given hypothesis test we construct a confidence interval with confidence level $1-\alpha$:

  Let $\varphi_{\theta_0}(Y) := \begin{cases} 0 & , \text{if "}H_0\text{"} \\ 1 & , \text{if "}H_A\text{"} \end{cases}$. Then $CI_{1-\alpha}(Y) = \{\theta \in \Theta \mid \varphi_{\theta_0}(Y) = 0\}$.

  $\mathbb{P}_\theta[\theta \in CI_{1-\alpha}(Y)] = 1 - \mathbb{P}_\theta[\theta \notin CI_{1-\alpha}(Y)] = 1 - \mathbb{P}_\theta[\varphi_{\theta_0}(Y) = 1] = 1 - \underbrace{\mathbb{P}[\text{"}H_A\text{"} \mid H_0]}_{\leq \alpha} \geq 1-\alpha$

Multiple Testing

Motivation

Suppose we consider the efficacy of a drug in terms of the reduction of any one of a number of disease symptoms. As more symptoms are considered, it becomes increasingly likely that the drug will appear to be an improvement over existing drugs in terms of at least one symptom.

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 so that the overall type I-error stays small.

Family-wise error rate (FWER)

The FWER is defined as $\alpha_{FWER} := \mathbb{P}(\text{"}H_{A1}\text{"} \vee \text{"}H_{A2}\text{"} \vee \dots \vee \text{"}H_{Am}\text{"} \mid H_{01} \wedge \dots \wedge H_{0m})$.

That is the probability of a at least one false rejection among the m tests.

If we assume that the tests are independent, we get $\alpha_{FWER} = 1-(1-\alpha)^m$ with α being the significance level for the single tests.

Bonferroni adjustment

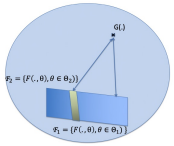
It can be shown that $\alpha_{FWER} \leq 1-(1-\alpha)^m$ (no independence assumption needed).

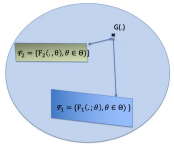
If we set the significance level of each individual test to $\frac{\alpha}{m}$, where m is the number of tests, then it follows $\alpha_{FWER} \leq 1-(1-\frac{\alpha}{m})^m \leq 1-(1-m\,\frac{\alpha}{m}) = \alpha$.

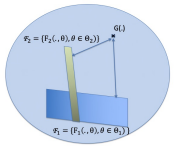
The adjusted significance level is defined by $\alpha_{adjust} = \frac{\alpha}{m}$.

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

Model Selection







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.

Training and Test Data

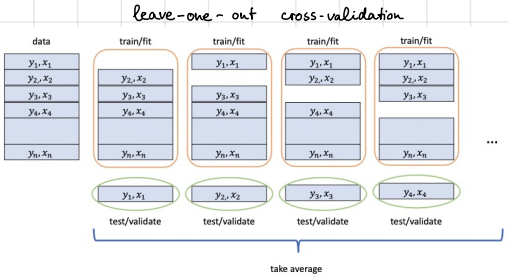
k-fold cross-validation

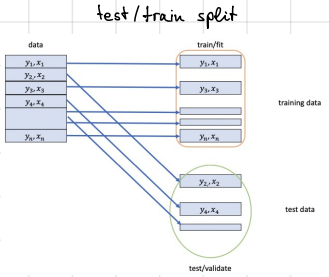
1. Divide the data $\{y_1,\dots,y_n\}$ in k disjoint sets of similar size and denote the resulting index sets as $N_1,\dots,N_k$.
2. We train the models on N_j seperately for all $j=1,\dots,k$. ,i.e. estimate $\hat{\theta}_1, \hat{\theta}_2$.
3. We calculate $\sum_{j=1}^{k} \sum_{i \in \{1,\dots,n\} \setminus N_j} \log\left(\frac{f(y_i;\hat{\theta}_1)}{f(y_i;\hat{\theta}_2)}\right)$

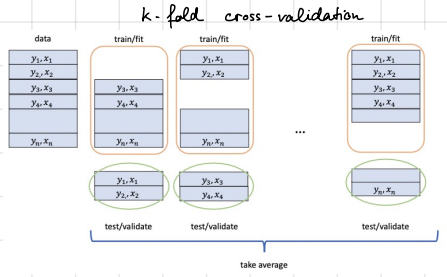
If the sum in (3.) is positive, we prefer $\mathcal{F}_1$, otherwise $\mathcal{F}_2$.

This method requires to refit the model k times.

If $k=n$, this method is called leave-one-out cross-validation. If $k=1$, this method is called







AIC and BIC

Akaike Information Criterion (AIC)

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. lower dimensional parameters.

Dimension of the Model

Setup

more on this in the script or see stepAIC-function in R.

$\frac{|\Theta|}{n}$ ← dimension of Θ

$\frac{|\Theta|}{n} \ll 1$: Classical statistics with asymptotic arguments

$\frac{|\Theta|}{n} \approx 1$: Either we have a high dimension - then we might wanna select the dimension data-driven - or we have a small sample - then we can use small sample corrections.

$\frac{|\Theta|}{n} \gg 1$: Beyond the framework of classical statistics but with applications in machine learning - labeled as deep learning

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 $AICc := -2 \cdot \ell(\hat{\theta}) + 2 \cdot p \cdot \left(\frac{n}{n-p-1}\right)$.

AIC, BIC and Hypothesis Testing

Setup

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_0" $\Leftrightarrow 2 \cdot (\ell(\hat{\theta}_0) - \ell(\hat{\theta}_1)) > \chi^2_{p,1-\alpha}$. On the other hand, the decision rule of the AIC can be written as "H_0" $\Leftrightarrow 2 \cdot (\ell(\hat{\theta}_0) - \ell(\hat{\theta}_1)) > 2p$ and for the BIC we get "H_0" $\Leftrightarrow 2 \cdot (\ell(\hat{\theta}_0) - \ell(\hat{\theta}_1)) > \log(n)p$

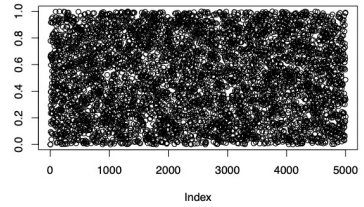
! Read L.Breiman (2001) - Statistical Modeling: The two cultures

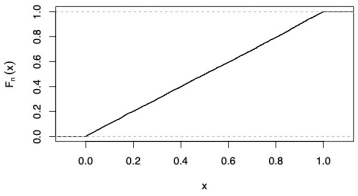
Simulation Based Inference

Simulating Uniformly Distributed Random Variables

Pseudo-random numbers

Numbers, which are produced by computers/algorithms are not random but deterministic. Therefore it is more suiteable to call them pseudo-random numbers. One way to generate such uniformly distributed random numbers on $[0,1]$ is by using $x_n := (a \cdot x_{n-1}) \bmod m$ with typically $a = 7^5$ and $m = 2^{31}-1$ with x_0 being called the seed. Then $u_n := x_n/m \in [0,1]$.





Usage Example

Let $g : \mathbb{R} \to \mathbb{R}_+$ be an integrable function with $\int_{\mathbb{R}} g(y)\,dy < \infty$.

Say we are interested in the value of $\int_{\mathbb{R}} g(y)\,dy$, then we can use $y = y(u) = \log\left(\frac{u}{1-u}\right) \iff u = u(y) = \frac{\exp(y)}{1+\exp(y)} \in [0,1]$ to compute

$$\int_{\mathbb{R}} g(y)\,dy = \int_0^1 \underbrace{g(y(u))\cdot\frac{1}{u(1-u)}}_{:=\tilde{g}(u)}\,du = \int_0^1 \tilde{g}(u)\,du = \mathbb{E}(\tilde{g}(U)) \quad \text{with } U \sim \mathcal{U}(0,1).$$

Drawing now $U_i \sim \mathcal{U}(0,1)$ $i = 1,\dots,n$ i.i.d. we can use the law of large numbers:

$$\bar{\tilde{g}} = \frac{1}{n}\cdot\sum_{i=1}^{n}\tilde{g}(U_i) \xrightarrow{n\to\infty} \mathbb{E}[\tilde{g}(U)] = \int_{\mathbb{R}} g(y)\,dy.$$

Simulating from a Distribution Function

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 $Y \sim F(\cdot)$.

Proof

We use $\mathbb{P}(U \leq u) = u$ for $U \sim \mathcal{U}(0,1)$.

$$\mathbb{P}(Y \leq y) = \mathbb{P}(F^{-1}(U) \leq y) = \mathbb{P}(U \leq F(y)) = F(y)$$

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. (f^*: called umbrella distribution)

Then the simulation procedure is as 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.

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

Proof

$$\mathbb{P}\left[Y^* \leq y \mid U \leq \frac{f(Y^*)}{a\cdot f^*(Y^*)}\right] = \frac{\mathbb{P}\left[Y^* \leq y, U \leq \frac{f(Y^*)}{a\cdot f^*(Y^*)}\right]}{\mathbb{P}\left[U \leq \frac{f(Y^*)}{a\cdot f^*(Y^*)}\right]} \overset{(*)}{=} \frac{\frac{1}{a}\cdot F(y)}{\frac{1}{a}} = F(y)$$

$$\overset{(*)}{\mathbb{P}}\left[Y^* \leq y, U \leq \frac{f(Y^*)}{a\cdot f^*(Y^*)}\right] = \int_{-\infty}^{y}\mathbb{P}\left[U \leq \frac{f(y^*)}{a\cdot f^*(y^*)}\right]\cdot f^*(y^*)\,dy^* = \int_{-\infty}^{y}\frac{f(y^*)}{a\cdot f^*(y^*)}\cdot f^*(y^*)\,dy^* = \frac{1}{a}\cdot F(y)$$

$$\mathbb{P}\left[U \leq \frac{f(Y^*)}{a\cdot f^*(Y^*)}\right] = \int_{-\infty}^{\infty}\mathbb{P}\left[U \leq \frac{f(y^*)}{a\cdot f^*(y^*)}\right]\cdot f^*(y^*)\,dy^* = \int_{-\infty}^{\infty}\frac{f(y^*)}{a\cdot f^*(y^*)}\cdot f^*(y^*)\,dy^* = \frac{1}{a}$$

Note that the last result also shows that the acceptence probability is $\frac{1}{a}$. So a large value of a increases the amount of samples needed to simulate F.

The number of samples until the sample is accepted follows a geometric distribution with mean $\frac{1}{a}$.

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.

Markov Chain Monte Carlo (MCMC)

For an intuition behind it read https://towardsdatascience.com/mcmc-intuition-for-everyone-5ae79fff22b1 or watch https://www.youtube.com/watch?v=OTO1DygELpY .

Motivation/Setup

We want to draw random samples from an unknown distribution F by using its density function f. But f is only known up to an unknown proportionality factor c, i.e. we know $\tilde{f}$ s.t. $f(y) = \frac{1}{c}\cdot\tilde{f}(y)$

That can be the case if the normalizing constant is unknown (or hard to compute) or when we want to sample from a posterior distribution $p(y|x) = \frac{p(x|y)\cdot p(y)}{p(x)}$ without knowing $p(x)$.

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 $\alpha(y^*, y^{(t)}) = \min\left\{1, \frac{\tilde{f}(y^*)}{\tilde{f}(y^{(t)})}\cdot\frac{h(y^{(t)}\mid y^*)}{h(y^*\mid y^{(t)})}\right\}$.
   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 10-th value as a sample (Thinning).

Simulating Multivariate Random Variables

Bayesian Inference

Bayesian Principle

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 $p_{\theta|y}(\theta; y_1,\dots,y_n) = \dfrac{\prod_{i=1}^n f(y_i;\theta)\, p(\theta;\gamma)}{f(y;\gamma)}$ 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 $p_{\theta|y}(\cdot) \in \mathcal{P}$. That is the posterior is from the same family of distributions as the prior distribution.

Bernstein-von Mises Theorem

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

Example

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

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

Consequently $\lambda|y_1,\dots,y_n \sim \Gamma(\sum_{i=1}^n y_i+\alpha,\ n+\beta)$. 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

Hyperparameters and Empirical Bayes

Concept of prior and flat- or improper prior

In general the hyperparameter $\gamma \in \Gamma$ should be set s.t. $p(\theta;\gamma)$ expresses the prior knowledge about θ (e.g. from previous analyses, expert knowledge, etc.). If there is no prior knowledge (or its not quantifiable), then one can choose a 'flat-prior', i.e. $p(\theta;\gamma) \propto$ const. ! Can result in an improper prior, i.e $\int_{\theta\in\Theta} p(\theta;\gamma)\, d\theta \neq 1$.

But even if we get an improper prior, the posterior can be a proper distribution.

Concept of empirical Bayes

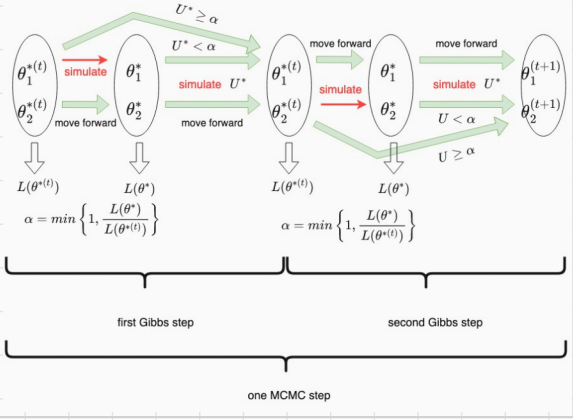
If we consider θ random and γ fixed, we can integrate out θ and derive $\tilde f(y;\gamma) = \int_{\theta\in\Theta} f(y;\theta)\cdot p(\theta;\gamma)\, d\theta$. So instead of working with $\mathcal{F}$ we work with $\tilde{\mathcal{F}} = \{\tilde f(\cdot;\gamma) \mid \gamma \in \Gamma\}$.

Inference based on MCMC

Gibbs sampling

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

1. Let $\underline\theta = (\theta_1,\dots,\theta_p) \in \Theta$. Let $\underline\theta^{(0)} \in \Theta$ and set $(\theta_1^{*(0)},\dots,\theta_p^{*(0)}) = \underline\theta^{(0)} = \underline\theta_0$.
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 $u^* \le \alpha(\theta^*,\theta^{*(t)}) = \min\left\{1,\ \frac{L(\theta^*)}{L(\theta^{*(t)})}\cdot\frac{q(\theta_j^{*(t)}\mid\theta_j^*)}{q(\theta_j^*\mid\theta_j^{*(t)})}\right\}$, 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.



Example - Zero inflated model

Let $Y \in \mathbb{N}_0$ and $\mathbb{P}(Y=k;\pi,\lambda) = \begin{cases}\pi + (1-\pi)\cdot\frac{\lambda^0}{0!}e^{-\lambda}, & \text{for } k=0\\ (1-\pi)\cdot\frac{\lambda^k}{k!}e^{-\lambda}, & \text{for } k\in\mathbb{N}\end{cases}$ with $\lambda>0$ and $\pi\in[0,1]$.

Given $y_1,\dots,y_n$, we want to model a two-parameter distribution.

As priors we assume $\pi \sim \mathcal{U}(0,1) \sim \text{Be}(1,1)$ and $\lambda \sim p(\lambda;\gamma) \sim \gamma e^{-\lambda\gamma}$. This results in the posterior $\lambda,\pi|y_1,\dots,y_n \propto \left(\prod_{i=1}^n \mathbb{P}(Y=y_i;\pi,\lambda)\right)\cdot\gamma e^{-\lambda\gamma}\cdot 1$ (prior for π)

! Problem: π,λ are constraint. In order to not control for that, we transform the parameters: $\eta = \text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) \Rightarrow \pi(\eta) = \frac{\exp(\eta)}{1+\exp(\eta)}$ and $\delta = \log(\lambda) \Leftrightarrow \lambda(\delta) = \exp(\delta)$.

As proposal distributions we now use $\delta^*\mid\delta^{*(t)} \sim \mathcal{N}(\delta^{*(t)},\sigma_\delta^2)$ and $\eta^*\mid\eta^{*(t)} \sim \mathcal{N}(\eta^{*(t)},\sigma_\eta^2)$ with $\sigma_\delta^2 = \sigma_\eta^2 = \frac{1}{2}$.

The new priors are $\eta \sim f_\eta(\eta) \overset{\text{Dichtetransformationssatz}}{=} f_\pi(\pi(\eta))\cdot\left|\frac{\partial\pi(\eta)}{\partial\eta}\right| = 1\cdot\pi(\eta)(1-\pi(\eta))$ and $\delta \sim f_\delta(\delta;\gamma) = p_\lambda(\lambda(\delta);\gamma)\cdot\left|\frac{\partial\lambda(\delta)}{\partial\delta}\right| = \gamma e^{-\lambda\gamma}\cdot e^\delta = \gamma e^{-\lambda\gamma}\cdot\lambda$

The acceptance probability is $\alpha(\delta^*,\eta^*,\delta^{*(t)},\eta^{*(t)}) = \min\left\{1,\ \frac{\prod_{i=1}^n \mathbb{P}(Y=y_i;\pi^*,\lambda^*)}{\prod_{i=1}^n \mathbb{P}(Y=y_i;\pi^{*(t)},\lambda^{*(t)})}\cdot\frac{f_\delta(\delta^*,\gamma)}{f_\delta(\delta^{*(t)},\gamma)}\cdot\frac{f_\eta(\eta^*)}{f_\eta(\eta^{*(t)})}\right\}$ with $\lambda^* = \exp(\delta^*)$ and $\pi^* = \frac{\exp(\eta^*)}{1+\exp(\eta^*)}$

For computational reasons we can rewrite $\frac{\prod_{i=1}^n \mathbb{P}(Y=y_i;\pi^*,\lambda^*)}{\prod_{i=1}^n \mathbb{P}(Y=y_i;\pi^{*(t)},\lambda^{*(t)})} = \exp\left(\sum_{i=1}^n \log\left(\mathbb{P}(Y=y_i;\pi^*,\lambda^*)\right) - \log\left(\mathbb{P}(Y=y_i;\pi^{*(t)},\lambda^{*(t)})\right)\right)$.

Using the above, we can perform the MCMC algorithm as explained in Gibbs sampling.

Approximate Bayes Computation (ABC)