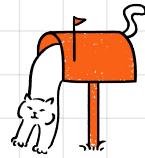




Uncertainty



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 $\text{Var}(Y|x) > 0$.

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.

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

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

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

! 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

$$KL(G, \mathcal{F}) := \min_{F(\cdot; \theta) \in \mathcal{F}} KL(G, F(\cdot; \theta)).$$

Data generating process

For simplicity we ignore the input x and assume that our quantity Y is drawn from some unknown dist. $G(\cdot)$. 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 realizations of $Y \sim G(\cdot)$, $i=1, \dots, n$, is i.i.d.

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

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 $\mathcal{F} = \{F(\cdot; \theta) | \theta \in \Theta\}$ 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 $t: \mathcal{D} \rightarrow \Theta$ we call $\hat{\theta} := t(y_1, \dots, y_n)$ an estimator and the function t a statistic.

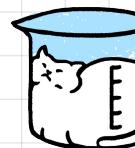
Note: In ML t is also called learner

Counterexample

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

Support depends on θ .

Therefore f is not Fisher-regular.



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

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

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

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

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

Positivity of KL

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

$$\begin{aligned} \text{KL}(G, \mathcal{F}) &= \int \log\left(\frac{g(y)}{f(y; \theta)}\right) g(y) d\mu(y) \\ &= - \int \log\left(\frac{f(y)}{g(y)}\right) g(y) d\mu(y) \\ &\geq - \int \left(1 - \frac{f(y)}{g(y)}\right) g(y) d\mu(y) \\ &= \int g(y) d\mu(y) - \int f(y) d\mu(y) \\ &= 1 - 1 = 0 \end{aligned}$$

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 $\frac{\partial \ell(\theta)}{\partial \theta} = 0$

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

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

Idea: Estimate $\text{KL}(G, \mathcal{F})$ with data and then minimize the estimate to find

the F with the smallest estimated KL-divergence.

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

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

unknown

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

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 loss(θ) = $\sum_{i=1}^n (y_i - \mu(x_i; \theta))^2$ → Commonly used in ML, but is equivalent to assuming $Y \sim N(\theta)$

Properties of Estimates

Bias

The bias is defined as $\text{bias}(\hat{\theta}, \theta_0) := E[\hat{\theta}] - \theta_0$. An estimate $\hat{\theta}$ is called unbiased if $\text{bias}(\hat{\theta}, \theta_0) = 0$. $\hat{\theta}$ is called asymptotically unbiased if $\text{bias}(\hat{\theta}, \theta_0) \xrightarrow{n \rightarrow \infty} 0$

Mean Squared Error

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

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 $\text{MSE}(\hat{\theta}_n, \theta_0) \xrightarrow{n \rightarrow \infty} 0$

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 $\text{MSE}(\hat{\theta}_1, \theta_0) < \text{MSE}(\hat{\theta}_2, \theta_0)$

Example

Assume $Y_i \sim N(\mu, \sigma^2)$ i.i.d. and we propose the estimate $\hat{\mu} = t(y) = \sum_{i=1}^n w_i \cdot y_i$ for some weights w_i . For t to be unbiased we postulate $\sum w_i = 1$. $\text{Var}(t(y)) = \sum w_i \sigma^2 = \sum (\frac{1}{n} + d_i)^2 \sigma^2 = \sum (\frac{1}{n} + d_i^2 + 2d_i) \sigma^2 = \frac{\sigma^2}{n} + \sigma^2 \sum d_i^2 + \frac{2\sigma^2}{n} \sum d_i \geq \frac{\sigma^2}{n}$ set $w_i = \frac{1}{n} + d_i$; with $\sum d_i = n$ 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 $P(Y_1 = y_1, \dots, Y_n = y_n | t(y_1, \dots, y_n) = t_0; \theta)$ does not depend on θ .

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

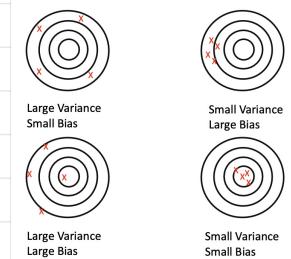
A bijective transformation of a sufficient statistic is also a sufficient statistic.

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

$Y_i \sim B(\pi)$ 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 $n_0 = n \cdot t_0$.

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

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

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

Distribution is independent of π . $\Rightarrow t(y_1, \dots, y_n) = \bar{Y}$ is sufficient.

Example

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

$$\begin{aligned} P(X_1 = x_1, \dots, X_n = x_n; \theta) &= \prod_{i=1}^n \frac{1}{\theta} \cdot I_{(0, \theta)}(x_i) \\ &= \underbrace{\frac{1}{\theta^n} \cdot I_{(-\infty, \theta)}(\max\{x_1, \dots, x_n\})}_{= t(X)} \underbrace{I_{(0, \infty)}(\min\{x_1, \dots, x_n\})}_{= h(x_1, \dots, x_n)} \end{aligned}$$

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.

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.

Pivotal statistic

A quantity $g(t(y_1, \dots, y_n); \theta)$ is called pivotal statistic if its distribution does not depend on θ . Its distribution is called a pivotal distribution.

Example

Assume $\hat{\theta} = t(y_1, \dots, y_n) \sim N(\theta, \text{Var}(\hat{\theta}))$

We construct an approximate pivotal statistic with $g(t(y_1, \dots, y_n); \theta) = \frac{\hat{\theta} - \theta}{\sqrt{\text{Var}(\hat{\theta})}} \sim N(0, 1)$
 $\sim N(0, 1)$ is independent of θ

Example

The pivotal statistic from below $g(\hat{\theta}, \theta) = \frac{\hat{\theta} - \theta}{\sqrt{\text{Var}(\hat{\theta})}} \sim N(0, 1)$

can be used to construct the CI in the following way:

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

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

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

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

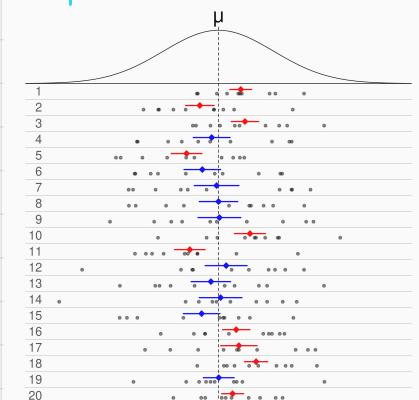
Confidence principle

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 (α) probability of covering the true value of θ . $\rightarrow P(\theta \in [-1, 1])$ is nonsense b.c. θ isn't a random variable.

Example



20 samples with corresponding 50% CI for μ .

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.

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

Instead we have $\frac{\bar{y} - \mu}{S/\sqrt{n}} \sim t_{(n-1)}$

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

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 posterior distr. $p_{\theta}(\theta | y_1, \dots, y_n)$.

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

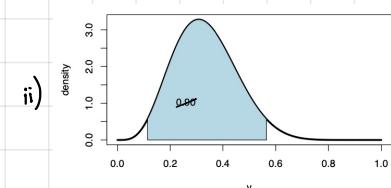
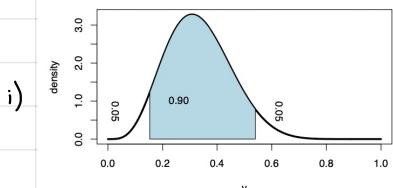
There are two natural ways to construct t_{ℓ} and t_r :

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

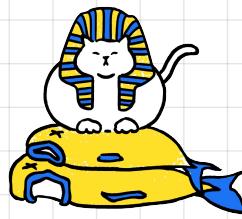
Problem: It occurs that $\theta_1 \in [t_{\ell}, t_r]$ and $\theta_2 \in [t_{\ell}, t_r]$ but $p_{\theta}(\theta_1 | y) > p_{\theta}(\theta_2 | y)$

ii) Highest posterior density credibility interval:

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



Maximum Likelihood Estimation



Score equation

Score function

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

Under regularity we have $E_\theta[s(\theta; y)] = \int \frac{\partial \log f(y; \theta)}{\partial \theta} \cdot f(y; \theta) dy = \int \frac{\partial f(y; \theta)}{\partial \theta} dy = \frac{\partial}{\partial \theta} \int f(y; \theta) dy = \frac{\partial}{\partial \theta} = 0$.

Fisher Information

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

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

! 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. $I(\theta) = \text{Var}_\theta[s(\theta; Y_1, \dots, Y_n)]$. 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 $\text{Var}_\theta[s(\theta; Y_1, \dots, Y_n)] = n \cdot \text{Var}_\theta[s(\theta; Y_1)]$

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

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

Example

Let $Y \sim \text{Bin}(n, \pi)$. Then $\ell(\pi) = y \log(\pi) + (n-y) \log(1-\pi)$ which leads to

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

$$E[s(\pi, y)] = \frac{n\pi}{\pi} - \frac{n-n\pi}{n-\pi} = 0$$

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

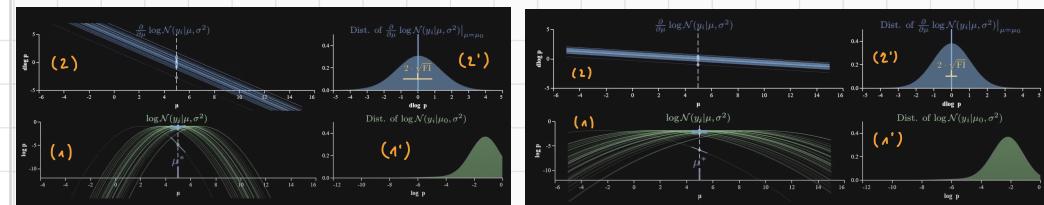
$$I(\pi) = E \left[-\frac{\partial^2 \ell(\pi)}{\partial \pi^2} \right] = \frac{n\pi}{\pi^2} + \frac{n-n\pi}{(1-\pi)^2} = \frac{n}{\pi} + \frac{n}{1-\pi} = \frac{n}{\pi(1-\pi)}$$

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

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 $s(\hat{\theta}; y_1, \dots, y_n) = 0$

Intuition



(1) log-likelihood function for samples from the data generating function (here $N(y_i | \mu, \sigma^2)$) - 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.

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 $\theta \in \Theta$.

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

Let $Y = (Y_1, \dots, Y_n) \in \mathbb{R}^N$ 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 $\text{Var}(t(Y)) \geq \frac{(\frac{\partial}{\partial \theta} E[t(Y)])^2}{I(\theta)} := \text{CRLB}(\theta)$

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

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 $G(\cdot) \in \mathbb{F}$.

Then the MLE is asymptotically normally distributed with $\hat{\theta} \xrightarrow{D} N(\theta_0, I(\theta_0)^{-1})$

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

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 $\text{Var}(s(\theta_0; Y_1, \dots, Y_n)) = I_{\theta_0}(\theta_0)$.

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

Beweis

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

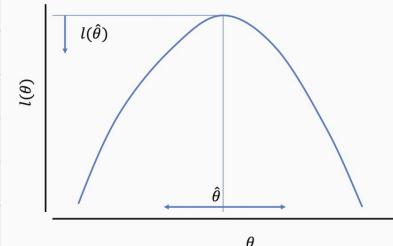
we get $\frac{\partial^2 \log f(\theta; Y_1, \dots, Y_n)}{\partial \theta^2}(\tilde{\theta}) \xrightarrow{P} n \cdot E\left[\frac{\partial^2 \log f(\theta; Y_i)}{\partial \theta^2}\right] = -I(\theta_0)$.

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

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

Likelihood Ratio

Idea



Log likelihood ratio

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

Theorem

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

That is: $2 \cdot (\text{lr}(\hat{\theta}) - \text{lr}(\theta_0)) \xrightarrow{D} \chi_p^2$ with $p = \dim(\theta)$.

Instead of looking at the distribution of $\hat{\theta} - \theta_0$,

one can look at the distribution of $\text{lr}(\hat{\theta}) - \text{lr}(\theta_0)$.

Maximum Likelihood and Parameter Transformation

Idea

Often, the parameter of a model is restricted, e.g. $X \sim \mathcal{B}(n, \pi)$ 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 $\ell_\theta(\theta)$ the log likelihood.

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

The likelihood for y is then defined as $\ell_y(y) := \ell_\theta(h^{-1}(y))$.

$$\text{It follows } \frac{\partial \ell_y(y)}{\partial y} = \frac{\partial \ell_\theta(h^{-1}(y))}{\partial y} = \frac{\partial \ell_\theta(h^{-1}(y))}{\partial \theta} \cdot \frac{\partial h^{-1}(y)}{\partial y} \quad \text{and} \quad \frac{\partial^2 \ell_y(y)}{\partial y^2} = \frac{\partial^2 \ell_\theta(h^{-1}(y))}{\partial \theta^2} \cdot \frac{\partial h^{-1}(y)}{\partial y} = 0$$

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

$$\text{Furthermore, we see } \frac{\partial^2 \ell_y(y)}{\partial y^2} = \frac{\partial}{\partial y} \left(\frac{\partial \ell_y(y)}{\partial y} \right) = \frac{\partial \theta}{\partial y} \frac{\partial \ell_\theta(\theta)}{\partial \theta} \frac{\partial \theta}{\partial y} + \underbrace{\frac{\partial \ell_\theta(\theta)}{\partial \theta} \frac{\partial \theta}{\partial y}}_{E[\cdot]=0}$$

$$I_y(y) = E\left[\frac{\partial^2 \ell_y(y)}{\partial y^2}\right] = \frac{\partial \theta}{\partial y} \cdot I_\theta(\theta) \cdot \frac{\partial \theta}{\partial y}. \quad \text{So the Fisher information for } y \text{ easily results from } \theta.$$

$$\hat{y} \xrightarrow{D} N(y, I_y^{-1}(y)) \quad \text{or} \quad \hat{y} \xrightarrow{D} N(y, \frac{\partial \theta}{\partial y} \cdot I_\theta(\theta) \cdot \frac{\partial \theta}{\partial y})$$

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 $E[s(\theta_0)] = \int \frac{\partial \ell(\theta_0)}{\partial \theta} dG(y)$ 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 $V(\theta_0) := \int s(\theta) \cdot s'(\theta) dG(y) = \text{Var}(s(\theta_0))$. This changes the asymptotic normality behaviors of the MLE.

Theorem

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

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

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 $\theta_{(t+1)} := \theta_{(t)} + I^{-1}(\theta_{(t)}) \cdot s(\theta_{(t)}; y)$

(iii) Repeat (ii) until $\|\theta_{(t+1)} - \theta_{(t)}\| < \epsilon$ for some fixed error tolerance ϵ .

(iv) Set $\hat{\theta} = \theta_{(t+1)}$.

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 $P(X>Y) = P(Y>X)$.

Let $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} X$ and $Y_1, \dots, Y_m \stackrel{i.i.d.}{\sim} Y$ 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 $S(X, Y) = \begin{cases} 1, & \text{if } X > Y \\ 0, & \text{if } X = Y \\ -1, & \text{if } X < Y \end{cases}$

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 " $\Leftrightarrow D_n \geq K_{S-\alpha}$ with K_S being the Kolmogorov - Smirnov dist.

Power of a Test

Idea

Given a simple Hypothesis system $H_0: \theta = \theta_0$ and $H_1: \theta = \theta_1$ 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 $P(H_1 | H_0)$ and commonly denoted by $1 - \beta$. The power of a test quantifies the probability of correctly rejecting H_0 if H_1 holds.

Example

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

The power of the test is now easily calculated through $P(H_1 | \mu) = P(\bar{Y} \geq \mu_0 + z_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}} | \mu) = 1 - \Phi(z_{1-\alpha} + \frac{\mu - \mu_0}{\sigma/\sqrt{n}})$.

Given a certain alternative Hypothesis $H_1: \mu = \mu_1$ 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 $Y_i \sim N(\mu, \sigma^2)$ i.i.d. $i=1, \dots, n$. We want to test $H_0: \mu \leq \mu_0$, $H_1: \mu > \mu_0$. As test statistic we use $Y_{\text{med}} = \text{median}(Y_1, \dots, Y_n)$. We can formulate a decision rule " H_0 " $\Leftrightarrow Y_{\text{med}} > c$ with c s.t. $P(H_0 | H_0) \leq \alpha$. 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 $H_1: \theta = \theta_1$ with a statistical significance test using level α . The most powerful test has the decision rule " H_0 " $\Leftrightarrow l(\theta_0) - l(\theta_1) \leq c$ where c is determined s.t. $P(H_0 | H_0) \leq \alpha$. This is called the Neyman - Pearson test.

Example

Let $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ i.i.d. $H_0: \mu = \mu_0$ and $H_1: \mu = \mu_1$. $l(\mu, \sigma^2; x_1, \dots, x_n) = \text{const} - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2$

Likelihood ratio: $2(l(\mu_0) - l(\mu_1)) = \frac{1}{\sigma^2} \left(\sum_{i=1}^n (x_i - \mu_0)^2 - \sum_{i=1}^n (x_i - \mu_1)^2 \right) \stackrel{!}{<} c$

$\sum_{i=1}^n (x_i - \mu_0)^2 - \sum_{i=1}^n (x_i - \mu_1)^2 \propto \sum_{i=1}^n x_i$. Therefore we obtain " H_0 " $\Leftrightarrow l(\mu_0) - l(\mu_1) < c \Leftrightarrow \frac{1}{n} \sum_{i=1}^n x_i < k$ with k being chosen s.t. $P\left(\frac{1}{n} \sum_{i=1}^n x_i < k \mid \mu = \mu_0\right) \leq \alpha \Rightarrow k = \mu_0 + z_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}}$

Testing and Confidence Interval

Connection

We now have two concepts for drawing statistical inference:

- Confidence intervals $[t_p(Y), t_r(Y)]$, where $P[t_p(Y) \leq \theta \leq t_r(Y)] \geq 1 - \alpha$ with $Y = (Y_1, \dots, Y_n)$
- Hypothesis tests $P(H_1 | H_0) \leq \alpha$

These 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 " $\Leftrightarrow \theta \notin [t_p(Y), t_r(Y)]$.

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

$$P(H_0 | H_0) = 1 - P(H_0 | H_0) = 1 - P(\varphi_\theta(Y) = 0 | \theta = \theta_0) = 1 - P(t_p(Y) \leq \theta_0 \leq t_r(Y)) \leq \alpha.$$

- For a given hypothesis test we construct a confidence interval with confidence level $1 - \alpha$:

Let $\varphi_\theta(Y) = \begin{cases} 0, & \text{if } H_0 \\ 1, & \text{if } H_1 \end{cases}$. Then $C_{1-\alpha}(Y) = \{\theta \in \Theta \mid \varphi_\theta(Y) = 0\}$.

$$P[\theta \in C_{1-\alpha}(Y)] = 1 - P[\theta \notin C_{1-\alpha}(Y)] = 1 - P[\varphi_\theta(Y) = 1] = 1 - P(H_1 | H_0) \geq 1 - \alpha$$

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 $H_{1j}, j=1, \dots, m$.

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 $\alpha_{FWER} = P(H_{11} \vee H_{12} \vee \dots \vee H_{1m} \mid H_{01} \wedge \dots \wedge H_{0m})$.

That is the probability of 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 α_m , where m is the number of tests, then it follows $\alpha_{FWER} \leq 1 - (1 - \alpha_m)^m \leq 1 - (1 - m \cdot \alpha_m) = \alpha$.

The adjusted significance level is defined by $\alpha_{\text{adjust}} = \alpha_m$.

! 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 .

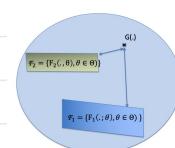
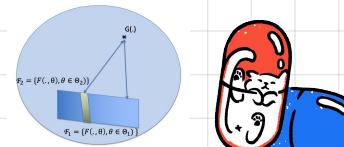
If one model is a subset of the other, than we call the models nested.

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

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

Alternatively, we can write non-nested models formally as

$$\mathcal{F}_1 = \{f_1(\cdot; \theta) \mid \theta \in \Theta_1\} \text{ and } \mathcal{F}_2 = \{f_2(\cdot; \theta) \mid \theta \in \Theta_2\} \text{ with } \mathcal{F}_2 \not\subseteq \mathcal{F}_1 \text{ and } \mathcal{F}_1 \not\subseteq \mathcal{F}_2.$$



Model comparison idea

We prefer the model with the smaller Kullback-Leibler divergence,

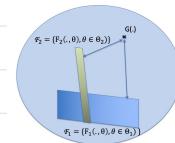
i.e. we choose model \mathcal{F}_1 over model \mathcal{F}_2 if

$$KL(G(\cdot), \mathcal{F}_1(\cdot, \hat{\theta}_1)) - KL(G(\cdot), \mathcal{F}_2(\cdot, \hat{\theta}_2)) < 0 \Leftrightarrow \int \log \frac{f_1(y; \hat{\theta}_1)}{f_2(y; \hat{\theta}_2)} dG(y) > 0$$

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 preferred.



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 .

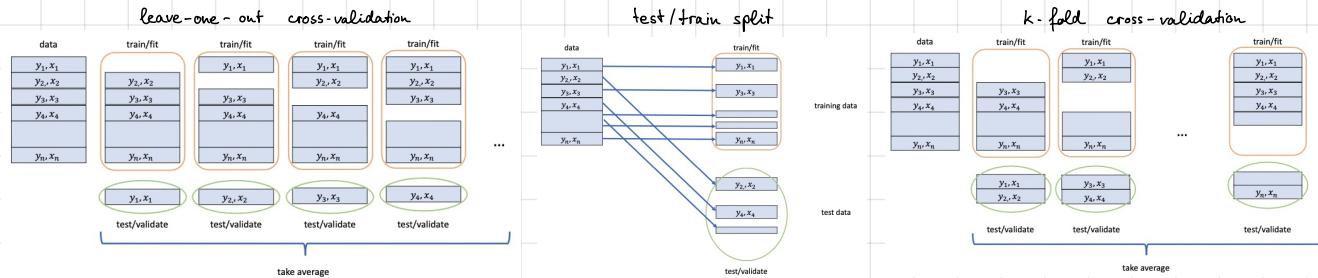
2. We train the models on N_j separately for all $j=1, \dots, k$, i.e. estimate $\hat{\theta}_1, \hat{\theta}_2$.

$$3. \text{ We calculate } \sum_{j=1}^k \sum_{i \in N_j \setminus N_j} \log \left(\frac{f(y_i; \hat{\theta}_j)}{f(y_i; \hat{\theta}_{j+1})} \right)$$

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 $p = \dim(\Theta)$

$$\text{is defined as } AIC(\mathcal{F}) := -2 \sum_{i=1}^n \log(f(y_i; \hat{\theta})) + 2p.$$

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

$$\text{Therefore } AIC(\mathcal{F}_1) - AIC(\mathcal{F}_2) \text{ is an estimate for } -2 \cdot E \left[\log \frac{f_1(y; \hat{\theta}_1)}{f_2(y; \hat{\theta}_2)} \right] dG(y).$$

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

$$BIC(\mathcal{F}) := -2 \sum_{i=1}^n \log(f(y_i; \hat{\theta})) + \log(n) \cdot p$$

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

Dimension of the Model

Setup

$\frac{1}{n} \text{ dimension of } \Theta \ll 1$: Classical statistics with asymptotic arguments

$\frac{1}{n} \text{ dimension of } \Theta \approx 1$: 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.

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

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

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 MLEs. It can be shown that if model \mathcal{F}_0 holds, then

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

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

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

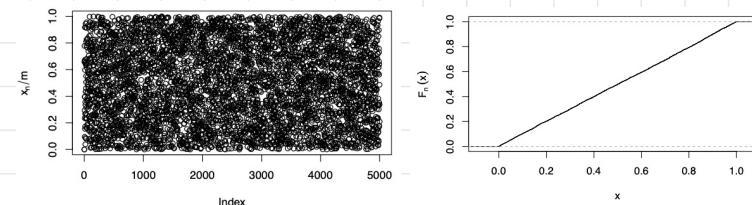
Simulation Based Inference



Simulating Uniformly Distributed Random Variables

Pseudo-random numbers

Numbers, which are produced by computers/algorithms are not random but deterministic. Therefore it is more suitable 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 $g: \mathbb{R} \rightarrow \mathbb{R}_+$ be an integrable function with $\int_{\mathbb{R}} g(y) dy < \infty$.

Say we are interested in the value of $\int g(y) dy$, then we can use $y = g(u) \Leftrightarrow u = u(y) = \frac{\exp(y)}{1 + \exp(y)} \in [0, 1]$ to compute

$$\int_{\mathbb{R}} g(y) dy = \int_{\mathbb{R}} g(g(u)) \cdot \frac{1}{u'(u)} du = \int_0^1 g(u) du = E[g(U)] \text{ with } U \sim U(0,1).$$

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

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

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 $F^{-1}(u) := \inf\{y | F(y) \geq u\}$

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

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

Proof

We use $P[U \leq u] = u$ for $U \sim U(0,1)$.

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

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 $\exists a \in \mathbb{R} \forall y \in T_F: f(y) \leq a \cdot f^*(y)$ and that a is known.
↑ called umbrella distribution

Then the simulation procedure is as follows:

1. Draw Y^* from F^* using the method described before.
2. Draw U from $U(0,1)$.
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

$$P[Y^* \leq y | U \leq \frac{f(Y^*)}{a \cdot f^*(Y^*)}] = \frac{P[Y^* \leq y, U \leq \frac{f(Y^*)}{a \cdot f^*(Y^*)}]}{P[U \leq \frac{f(Y^*)}{a \cdot f^*(Y^*)}]} \stackrel{(1)}{=} \frac{\frac{1}{a} \cdot F(y)}{\frac{1}{a}} = F(y)$$

$$(2) \quad P[Y^* \leq y, U \leq \frac{f(Y^*)}{a \cdot f^*(Y^*)}] = \int_{-\infty}^y P[U \leq \frac{f(y)}{a \cdot f^*(y)}] \cdot f^*(y) dy^* = \int_{-\infty}^y \frac{f(y)}{a \cdot f^*(y)} \cdot f^*(y) dy^* = \frac{1}{a} \cdot F(y)$$

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

! Note that the last result also shows that the acceptance 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 $E_F[h(Y)]$ for some function $h(\cdot)$.

$$\text{We use } E_F[h(Y)] = \int_{\mathbb{R}} h(y) f(y) dy = \int_{\mathbb{R}} h(y) \cdot \frac{f(y)}{f^*(y)} f^*(y) dy = E_{F^*}\left[h\left(\frac{f(y)}{f^*(y)}\right)\right]$$

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

$$\text{we obtain } \frac{N}{N} \sum_{i=1}^N h(Y_i^*) \frac{f(Y_i^*)}{f^*(Y_i^*)} \xrightarrow{N \rightarrow \infty} E_F[h(Y)]$$

! 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=OTQ1DygELpY>.

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 (-Hastings) 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 $Y^* \sim H(y|y^{(t)})$ where $H(\cdot | y^{(t)})$ is a known distribution, with the same support (Träger) as F , from which we can sample. The corresponding density is $h(y|y^{(t)})$. In the Metropolis-Hastings algorithm the proposal distribution is symmetric - for example $N(y^{(t)}, \sigma^2)$ with some fixed variance σ^2 . In the Metropolis-Hastings algorithm the proposal distribution can be skewed. $H(\cdot | y^{(t)})$ can depend on $y^{(t)}$ but doesn't need to.
3. Compute the acceptance probability, which is defined as

$$\alpha(y^*, y^{(t)}) = \min\{1, \frac{\tilde{f}(y^*) \cdot h(y^{(t)}|y^*)}{\tilde{f}(y^{(t)}) \cdot h(y^*|y^{(t)})}\}$$

If H is symmetric, then $\frac{h(y^{(t)}|y^*)}{h(y^*|y^{(t)})} = 1$.

$$4. \text{ Draw } U^* \sim U(0,1) \text{ and define } y^{(t+1)} = \begin{cases} y^*, & \text{if } U^* \leq \alpha(y^*, y^{(t)}) \\ y^{(t)}, & \text{otherwise} \end{cases}$$

5. Go back to step 2 until t is large enough.

(6) Usually the first few values of $y^{(t)}$, i.e. $y^{(1)}, \dots, y^{(500)}$ 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^* we can choose only every 10 -th value as a sample (thinning).

Simulating Multivariate Random Variables

Bayesian Inference



Bayesian Principle

Prior and Posterior

Let $\mathcal{F} = \{f(\cdot; \theta) | \theta \in \Theta\}$ 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 $\mathcal{P} = \{p(\cdot; \theta) | \theta \in \Theta\}$.

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

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. $p(\theta; \gamma) \propto \text{const.}$! Can result in an improper prior, i.e. $\int p(\theta; \gamma) d\theta \neq 1$. But even if we get an improper prior, the posterior can be a proper distribution.

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 $p_{\text{post}}(\cdot) \in \mathcal{P}$.

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 $\theta \xrightarrow{d} N(\hat{\theta}, I^{-1}(\hat{\theta}))$

Example

Let's assume $Y \sim \text{Poi}(2)$ and $\lambda \sim \Gamma(\alpha, \beta)$, i.e. $p(\lambda) \propto \lambda^{\alpha-1} \exp(-\beta\lambda)$.

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

Consequently $y_1, \dots, y_n \sim \Gamma\left(\frac{\sum y_i + n - 1}{\alpha + \beta}, \alpha + \beta\right)$. 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

Concept of empirical Bayes

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

Inference based on MCMC

Gibbs sampling

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

1. Let $\theta = (\theta_1, \dots, \theta_p) \in \Theta$. Let $\theta_j^{(t)} \in \Theta$ and set $(\theta_1^{(t)}, \dots, \theta_p^{(t)}) = \theta^{(t)}$.

2. Set $\theta_j^{(t+1)} = \theta_j^{(t)}$ for $j=1, \dots, p$ and execute the following steps:

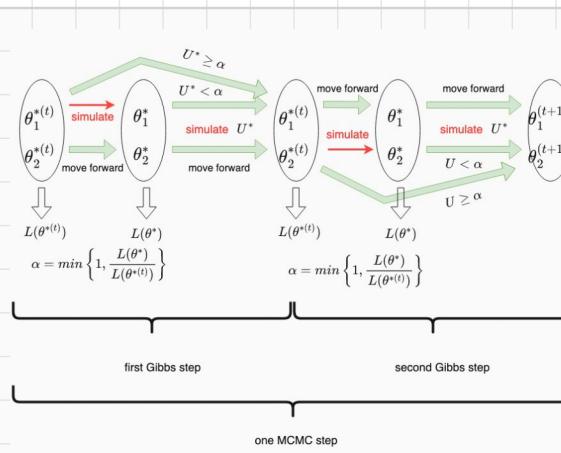
(i) Set $\theta_j^{(t)} = \theta_j^{(t)}$

(ii) Replace $\theta_j^{(t)}$ with a sample from the proposal $q(\cdot; \theta_j^{(t)})$

(iii) Sample $U^* \sim U(0, 1)$

(iv) If $U^* \leq \alpha(\theta_j^{(t)}, \theta_j^{(t)}) = \min\left\{1, \frac{L(\theta_j^{(t)})}{L(\theta_j^{(t)})} \cdot \frac{q(\theta_j^{(t)} | \theta_j^{(t)})}{q(\theta_j^{(t)} | \theta_j^{(t)})}\right\}$, then set $\theta_j^{(t+1)} = \theta_j^{(t)}$. Otherwise leave $\theta_j^{(t+1)}$ 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 $\mathbb{P}(Y=k; \pi, \lambda) = \begin{cases} \pi + (1-\pi) \frac{\lambda^k}{k!} e^{-\lambda}, & \text{for } k=0 \\ (1-\pi) \frac{\lambda^k}{k!} e^{-\lambda}, & \text{for } k \geq 1 \end{cases}$ 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 $\pi \sim U(0, 1) \sim Be(1, 1)$ and $\lambda \sim p(\lambda; \gamma) \sim \gamma e^{-\lambda^2}$. This results in the posterior $\pi, \lambda | y_1, \dots, y_n \propto \left(\prod_{i=1}^n \mathbb{P}(Y=y_i; \pi, \lambda)\right) \cdot \gamma e^{-\lambda^2} \cdot \pi$.

! 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 $\delta = \log(\lambda) \Rightarrow \lambda(\delta) = \exp(\delta)$.

As proposal distributions we now use $\delta^* | \delta^{(t)} \sim N(\delta^{(t)}, \sigma_\delta^2)$ and $\eta^* | \eta^{(t)} \sim N(\eta^{(t)}, \sigma_\eta^2)$ with $\sigma_\delta^2 = \sigma_\eta^2 = \frac{1}{2}$. The new priors are $\eta \sim f_{\pi}(\eta) = \frac{1}{\pi} \cdot (1-\pi)^{-1}$ and $\delta \sim f_\lambda(\delta) = \lambda \exp(\delta) \cdot \frac{\partial \lambda}{\partial \delta} = \lambda e^{-\delta^2} \cdot \exp(\delta) = \delta^2 e^{-\delta^2} \cdot \lambda$.

The acceptance probability is $\alpha(\delta^*, \eta^*, \delta^{(t)}, \eta^{(t)}) = \min\left\{1, \frac{\mathbb{P}(Y=y_1; \pi^*, \lambda^*)}{\mathbb{P}(Y=y_1; \pi^{(t)}, \lambda^{(t)})} \cdot \frac{f_\pi(\eta^*)}{f_\pi(\eta^{(t)})} \cdot \frac{f_\lambda(\delta^*)}{f_\lambda(\delta^{(t)})}\right\}$ with $\lambda^* = \exp(\delta^*)$ and $\pi^* = \frac{\exp(\eta^*)}{1+\exp(\eta^*)}$.

For computational reasons we can rewrite $\frac{\mathbb{P}(Y=y_1; \pi^*, \lambda^*)}{\mathbb{P}(Y=y_1; \pi^{(t)}, \lambda^{(t)})} = \exp\left(\sum_{i=2}^n \log(\mathbb{P}(Y=y_i; \pi^*, \lambda^*)) - \log(\mathbb{P}(Y=y_i; \pi^{(t)}, \lambda^{(t)}))\right)$.

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

Approximate Bayes Computation (ABC)

Concept

To simulate from the posterior $\pi(\theta|y_1, \dots, y_n)$ 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 approximately 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 $d(Y^*, y) \leq \epsilon$, 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

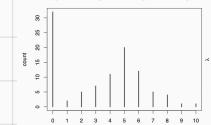
1. Generate θ^* from some prior dist. $p_\theta(\theta)$.
 2. Generate y^* from the assumed dist. i.e. $f_{Y^*}(\cdot; \theta^*)$.
 3. If $d(Y^*, y) \leq \epsilon$, with y being the available data, then accept θ^* . Otherwise go back to step 1.
 4. Repeat 1-3, until a sufficient amount of simulated values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

Example - Zero inflated data

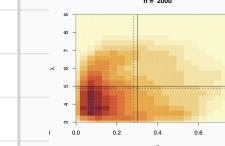
Consider the same setup as above. We define $d(Y^*, y) = \sum_{k=0}^N (\bar{x}_k^* - \bar{x}_k)^2$ with $N = 10$, $\bar{x}_k = \frac{1}{n} \sum_{i=1}^n \mathbb{1}(y_i = k)$ and $\bar{x}_k^* = \frac{1}{n} \sum_{i=1}^n \mathbb{1}(Y_i^* = k)$ and $\epsilon = 0.05$.

Now we proceed with the algorithm and simulate θ^*, Y^* .

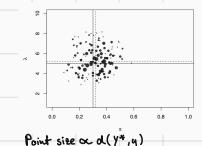
Actual data y_1, \dots, y_n



Posterior with MCMC
n = 2000



Posterior with ABC



Point site or $d(Y^*, y)$

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 approximately the posterior $p(\theta|y)$, e.g. $q(\cdot) = N(\hat{\theta}_p, I_p^{-1}(\hat{\theta}_p))$ (see Thm. below).

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

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

Solving the Optimization Problem

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

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

$$\begin{aligned} \frac{\partial LB(q(\cdot, y))}{\partial \theta} &= \int q(\theta, y) \frac{\partial \log(f(y|\theta))}{\partial \theta} d\theta - \left(\int q(\theta, y) \frac{\partial \log(f(y|\theta))}{\partial \theta} d\theta \right) \frac{\partial q(\theta, y)}{\partial \theta} d\theta \\ &= \mathbb{E}_{q(\cdot, y)} \left[\frac{\partial \log(f(y|\theta))}{\partial \theta} \log \left(\frac{f(y|\theta)p(\theta)}{q(\theta, y)} \right) \right] \stackrel{!}{=} 0 \quad (\text{see Score function}) \end{aligned}$$

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

Approximating Solution of the Optimization Problem

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

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

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

We can now use $y^{(t+1)} = y^{(t)} + \frac{1}{K} \sum_{k=1}^K \frac{\partial \widehat{LB}(q(\cdot, y^{(t)}))}{\partial \theta}$ to solve the problem numerically.

Penalized version of Bernstein Thm.

For sufficiently large n we have $\theta|y \approx N(\hat{\theta}_p, I_p^{-1}(\hat{\theta}_p))$,

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

where the prior p might depend on further hyperparameters and

$$I_p(\theta) = I(\theta) + \frac{\partial^2 p(\theta)}{\partial \theta^2}.$$

Multivariate Data

Log-Linear Model

Notation

Let $Y = (Y_1, \dots, Y_p)$ be a p -dimensional random vector. $Y_j \in \{1, \dots, K_j\}$ for $j=1, \dots, p$.

Multiple independent random vectors are denoted by $Y^{(1)}, \dots, Y^{(n)}$.

Let $k_j \in \{1, \dots, K_j-1\}$ for $j=1, \dots, p$.

$$s_{k_1 k_2 \dots k_p}(y) := \#\{y_1 = k_1, y_2 = k_2, \dots, y_p = k_p\} = \prod_{j=1}^p \#\{y_j = k_j\}$$

$$n_{k_1 k_2 \dots k_p} := \sum_{y \in \Omega} s_{k_1 k_2 \dots k_p}(y)$$

$$s_{k_1 \dots k_p}(y) := \#\{y_1 = k_1, \dots, y_p = k_p\} = \prod_{j=1}^p \#\{y_j = k_j\}$$

$$n_{k_1 \dots k_p} := \sum_{y \in \Omega} s_{k_1 \dots k_p}(y)$$

$$s_{k_1 \dots k_p}(y) := \#\{y_j = k_j\}$$

$$n_{k_1 \dots k_p} := \sum_{j=1}^p s_{k_1 \dots k_p}(y_j)$$

$$\text{Note: We exclude } k_j = K_j, \text{ b.c. of we have } \#\{y_j = K_j\} = 1 - \sum_{k_j=1}^{K_j-1} s_{k_1 \dots k_p}(y_j)$$

Log-Linear Model (General Form)

Let $Y = (Y_1, \dots, Y_p)$ be a discrete random-valued variable.

The distribution of Y follows a log-linear model if

$$P(Y=y) \propto \exp(\theta_0 + \sum_j s_j(y) \theta_j)$$

where $s_j(\cdot)$ are some linearly independent statistics from $y = (y_1, \dots, y_p)$ and $\theta = (\theta_0, \theta_1, \dots)$ is the parameter vector.

Log-Linear Model (Multiple observation)

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

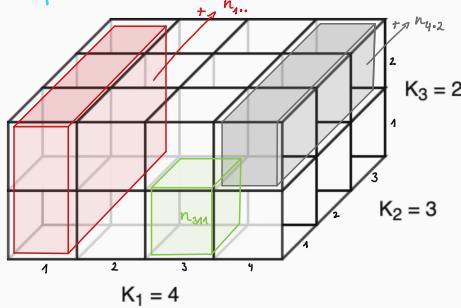
$$P(Y^{(1)}=y^{(1)}, \dots, Y^{(n)}=y^{(n)}) \propto \prod_{i=1}^n \exp\left(\theta_0 + \sum_{j=1}^p \sum_{k_j=1}^{K_j-1} s_{j k_1 \dots k_p}(y_j^{(i)}) \cdot \theta_{k_1 \dots k_p} + \dots + \sum_{j=2}^p \sum_{j=2}^{p-1} \sum_{k_j=1}^{K_j-1} \sum_{k_p=1}^{K_p-1} s_{j k_1 \dots k_p}(y_j^{(i)}) \cdot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

$$= \exp\left(\underbrace{\theta_0 + \dots + \theta_0}_{\text{Zero variables}} + \underbrace{\sum_{j=1}^p \sum_{k_j=1}^{K_j-1} n_{j k_1 \dots k_p} \cdot \theta_{k_1 \dots k_p}}_{\text{Sums over one variable}} + \underbrace{\sum_{j=2}^p \sum_{j=2}^{p-1} \sum_{k_j=1}^{K_j-1} \sum_{k_p=1}^{K_p-1} n_{j k_1 \dots k_p} \cdot \theta_{k_1 \dots k_p}}_{\text{Sums over two-wise pairs of variables}} + \dots + \underbrace{\sum_{k_1=1}^{K_1-1} \dots \sum_{k_p=1}^{K_p-1} n_{k_1 \dots k_p} \cdot \theta_{k_1 \dots k_p}}_{\text{Sums over all } p \text{ variables}}\right)$$

We call the terms θ_{\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.

Example



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 $2 \times 2 \times 2$ table. We use the notation 1 for Y_1 , 2 for Y_2 , 3 for Y_3 .

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

$$P(Y^{(1)}=y^{(1)}, \dots, Y^{(n)}=y^{(n)}) = \exp\left(\theta_0 + \theta_{111} + \theta_{112} + \theta_{121} + \theta_{122} + \theta_{211} + \theta_{212} + \theta_{221} + \theta_{222}\right)$$

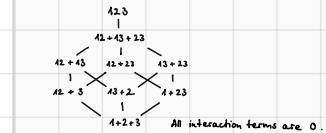
If we condition on the sample size, i.e. θ_0 and all the interaction terms are non-zero, then we can

denote the model as $1+2+3$ or 123 .

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

If we further restrict $\theta_{nn}=0$, 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:



Graphical Models

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 $A + B$ and $B + C$.

We say A and B are conditionally independent given C , denoted as $(A \perp\!\!\!\perp B) | C$

iff $f_{(A|B)}|C}(y_A, y_B | y_C) = f_{A|C}(y_A | y_C) \cdot f_{B|C}(y_B | y_C)$. This holds iff

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

Visualization of Conditional Independence

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

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

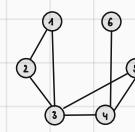
If instead of sets of variables we want to visualise single variables, we construct the graph as follows: Let $G = (E, V)$ be a graph, i.e. $E = \{1, \dots, p\}$ and $V \subseteq E \times E$.

Let $i, j \in E$. $(i \perp\!\!\!\perp j) | E \setminus \{i, j\} \Leftrightarrow (i, j) \notin V$

Example.

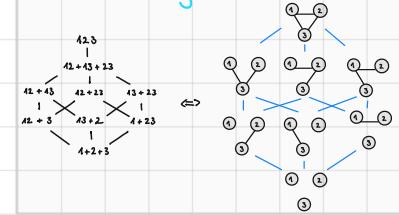
Let $p = 6$ and

$$\begin{aligned} 1 \perp\!\!\!\perp 4 & \{2, 3, 5, 6\} \\ 1 \perp\!\!\!\perp 5 & \{2, 3, 4, 6\} \\ 1 \perp\!\!\!\perp 6 & \{2, 3, 4, 5\} \\ 2 \perp\!\!\!\perp 4 & \{1, 3, 5, 6\} \\ 2 \perp\!\!\!\perp 5 & \{1, 3, 4, 6\} \\ 2 \perp\!\!\!\perp 6 & \{1, 3, 4, 5\} \\ 3 \perp\!\!\!\perp 6 & \{1, 2, 4, 5\} \\ 5 \perp\!\!\!\perp 6 & \{1, 2, 3, 4\} \end{aligned}$$



From the graph we can see
 $\{2, 1, 2\} \perp\!\!\!\perp \{4, 5, 6\} | \{3\}$
 $\{5, 6\} \perp\!\!\!\perp \{1, 2, 3, 4\} | \{4\}$

Model hierarchy



Rules of Conditional Independence

Let X, Y, Z, W be random variables.

Symmetry: $X \perp\!\!\!\perp Y \Rightarrow Y \perp\!\!\!\perp X$

Decomposition: $X \perp\!\!\!\perp Y, Z \Rightarrow X \perp\!\!\!\perp Y \wedge X \perp\!\!\!\perp Z$

Weak union: $X \perp\!\!\!\perp Y, Z \Rightarrow (X \perp\!\!\!\perp Y) | Z \wedge (X \perp\!\!\!\perp Z) | Y$

Contraction: $(X \perp\!\!\!\perp Y) | Z \wedge (Y \perp\!\!\!\perp Z) \Rightarrow X \perp\!\!\!\perp Y | Z$

Intersection: For strictly positive probability distributions it holds that

$$(X \perp\!\!\!\perp Y) | Z, W \wedge (X \perp\!\!\!\perp W) | Y \Rightarrow (X \perp\!\!\!\perp W, Y) | Z$$

Conditional Independence in Log-Linear Model

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

$$\begin{aligned} (Y_1 \perp\!\!\!\perp Y_2) | Y_3 \text{ or } (1 \perp\!\!\!\perp 2) | 3 & \Leftrightarrow \text{Var}_{\{1, 2, 3\}} \cup \{k_1 \in \{1, \dots, K_1\}, k_2 \in \{1, \dots, K_2\}, k_3 \in \{1, \dots, K_3\}}: P(Y_1 = k_1, Y_2 = k_2 | Y_3 = k_3) = \frac{P(Y_1 = k_1 | Y_3 = k_3) \cdot P(Y_2 = k_2 | Y_3 = k_3)}{= h_{123}(k_1, k_2)} \\ & \Leftrightarrow \text{Var}_{\{1, 2, 3\}} \cup \{k_1 \in \{1, \dots, K_1\}, k_2 \in \{1, \dots, K_2\}, k_3 \in \{1, \dots, K_3\}\}: P(Y_1 = k_1 | Y_2 = k_2, Y_3 = k_3) = P(Y_1 = k_1 | Y_3 = k_3) - \\ & \quad (\text{given } Y_3 \text{ we don't get additional information from } Y_2) \end{aligned}$$

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

$$\text{i.e. } (Y_1 \perp\!\!\!\perp Y_2) | Y_3 \Leftrightarrow \text{Var}_{\{1, 2, 3\}} \cup \{k_1 \in \{1, \dots, K_1\}, k_2 \in \{1, \dots, K_2\}, k_3 \in \{1, \dots, K_3\}\}: \theta_{k_1 k_2} = 0 \wedge \theta_{k_1 k_2 k_3} = 0.$$

Using the model notation from before, this results in the model $42 + 23$.

DAG's

Multivariate Normal Distribution

Multivariate Normal Dist.

Let $Y = (Y_1, \dots, Y_p)$ be a continuous random vector.

Y follows a multivariate normal distribution with mean

$$\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix}$$

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

The inverse of the covariance matrix $\Omega := \Sigma^{-1}$ is

called concentration matrix or precision matrix.

Using Ω we can rewrite f as

$$f(y; \mu, \Sigma) \propto \exp\left(-\frac{1}{2} \sum_{i=1}^p \Omega_{ii} y_i^2 + \frac{1}{2} \sum_{j=1}^p \sum_{i \neq j} \Omega_{ij} y_i y_j\right)$$

Conditional independence

Let $j, l \in \{1, \dots, p\}$ and $C = \{i \in \{1, \dots, p\} \setminus \{j, l\}\}$.

If $\Omega_{jl} = 0$, then $f(y; \mu, \Sigma) \propto h_{C \cup \{j\}}(y_C, y_l) \cdot h_{C \cup \{l\}}(y_C, y_j)$

Hence, we have the equivalence $\Omega_{jl} = 0 \Leftrightarrow (Y_j \perp\!\!\!\perp Y_l) | Y_C$.

Inference in Extreme Data

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 $y_i = 0$ or in general $n=0$ 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.

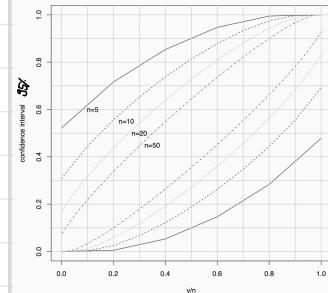
Clopper and Pearson Exact Confidence Interval

Let $Y \sim B(n, \pi)$. For $Y=y$ the exact confidence interval is (c_L, c_U) with $c_L := \inf \{ \pi \mid P(Y \leq y, \pi) > \frac{\alpha}{2} \}$, $c_U := \sup \{ \pi \mid P(Y \geq y, \pi) > \frac{\alpha}{2} \}$.

This can be computed by using the Beta-distribution:

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

Examples



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. $F_{Y_{\max, n}}(y) = F_y^n(y)$. In practice the distr.

$F(\cdot)$ might be unknown. We can then use the Fisher-Tippett-Gnedenko-theorem to get an asymptotic result.

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. $F^n(a_n y + b_n) \xrightarrow{n \rightarrow \infty} H(y)$ 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.

Generalized Extreme Value Distribution

The GEV distribution has three parameters.

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

The distribution function is given by

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

1. Gumbel distribution ($\gamma = 0$): $F(x; \mu, \sigma, 0) = \exp(-\exp(-\frac{x-\mu}{\sigma}))$

2. Fréchet distribution ($\gamma > 0$): $F(x; \mu, \sigma, \gamma) = \exp(-(1 + \gamma \frac{x-\mu}{\sigma})^{-\frac{1}{\gamma}}) \cdot \mathbf{1}_{(0, \infty)}(1 + \gamma \frac{x-\mu}{\sigma})$

3. (Reversed) Weibull distr. ($\gamma < 0$): $F(x; \mu, \sigma, \gamma) = \begin{cases} \exp(-(1 + \gamma \frac{x-\mu}{\sigma})^{-\frac{1}{\gamma}}) & \text{if } 1 + \gamma \frac{x-\mu}{\sigma} > 0 \\ 1 & \text{otherwise} \end{cases}$

Applications of extreme value theory include predicting the probability distribution of:

- Extreme floods; the size of freak waves
- Tornado outbreaks^[8]
- Maximum sizes of ecological populations^[9]
- Side effects of drugs (e.g., ximelagatran)
- The magnitudes of large insurance losses
- Equity risks; day-to-day market risk
- Mutational events during evolution
- Large wildfires^[10]
- Environmental loads on structures^[11]
- Fastest time humans are capable of running the 100 metres sprint^[12] and performances in other athletic disciplines^{[13][14][15]}
- Pipeline failures due to pitting corrosion
- Anomalous IT network traffic; prevent attackers from reaching important data
- Road safety analysis^{[16][17]}
- Wireless communications^[18]
- Epidemics^[19]
- Neurobiology^[20]