

---

# Statistics

## Collection of Formulas

---

# Contents

<b>1 Descriptive Statistics</b>	<b>3</b>	<b>5 Regression</b>	<b>17</b>
1.1 Summary Statistics . . . . .	3	5.1 Models . . . . .	17
1.1.1 Location . . . . .	3	5.1.1 Simple Linear Model . . . . .	17
1.1.2 Dispersion . . . . .	3	5.1.2 Multivariate Linear Model . . . . .	17
1.1.3 Concentration . . . . .	3	5.1.3 Bayesian Linear Model . . . . .	18
1.1.4 Shape . . . . .	4	5.1.4 Quantile Regression . . . . .	18
1.1.5 Dependence . . . . .	4	5.1.5 Flexible Regression . . . . .	18
1.2 Diagrams . . . . .	4	5.1.6 Generalized Regression . . . . .	19
1.2.1 Histogram . . . . .	5	5.1.7 Weighted Regression . . . . .	19
<b>2 Probability</b>	<b>6</b>	5.2 Goodness of Fit . . . . .	19
2.1 Combinatorics . . . . .	6	5.2.1 Coefficient of Determination . . . . .	19
2.2 Probability Theory . . . . .	6	<b>6 Survival Analysis</b>	<b>20</b>
2.3 Random Variables/Vectors . . . . .	6	6.1 Basics . . . . .	20
2.4 Probability Distributions . . . . .	7	6.2 Modelling . . . . .	20
2.4.1 Discrete Distributions . . . . .	7	6.3 Model Comparison . . . . .	21
2.4.2 Continuous Distributions . . . . .	8	<b>7 Spatial Statistics</b>	<b>22</b>
2.4.3 Exponential Family . . . . .	9	7.1 Markov Random Field . . . . .	22
2.5 Multivariate Distributions . . . . .	9	7.2 Geostatistics . . . . .	23
2.6 Limit Theorems . . . . .	10	7.3 Point Processes . . . . .	23
<b>3 Inference</b>	<b>11</b>	<b>8 Bayesian Statistics</b>	<b>24</b>
3.1 Method of Moments . . . . .	11	8.1 Basics . . . . .	24
3.2 Loss Functions . . . . .	11	8.2 Numerical Methods for the Posterior . . . . .	24
3.3 Maximum Likelihood (ML) . . . . .	12	<b>9 Sampling</b>	<b>26</b>
3.4 Consistency and Sufficiency . . . . .	13	<b>10 Model Selection</b>	<b>27</b>
<b>4 Statistical Hypothesis Testing</b>	<b>14</b>	<b>11 Dimensionality Reduction</b>	<b>28</b>
4.1 Significance and Confidence Intervals . . . . .	14	<b>12 Missing/Deficient Data</b>	<b>29</b>
4.2 Tests for One Sample . . . . .	14	<b>13 Experiment Design</b>	<b>30</b>
4.3 Tests for Goodness of Fit . . . . .	15		
4.4 Multiple Tests . . . . .	15		

# 1 Descriptive Statistics

## 1.1 Summary Statistics

### 1.1.1 Location

**Mode** Most frequent value of  $x_i$ . Two or more modes are possible (bimodal).

**Median**

$$\tilde{x}_{0.5} = \begin{cases} x_{((n+1)/2)} & \text{falls } n \text{ ungerade} \\ \frac{1}{2}(x_{(n/2)} + x_{(n/2+1)}) & \text{falls } n \text{ gerade} \end{cases}$$

**Quantile**

$$\tilde{x}_\alpha = \begin{cases} x_{(k)} & \text{falls } n\alpha \notin \mathbb{N} \\ \frac{1}{2}(x_{(n\alpha)} + x_{(n\alpha+1)}) & \text{falls } n\alpha \text{ ganzzahlig} \end{cases}$$

with

$$k = \min x \in \mathbb{N}, \quad x > n\alpha$$

**Minimum/Maximum**

$$x_{\min} = \min_{i \in \{1, \dots, N\}} (x_i) \quad x_{\max} = \max_{i \in \{1, \dots, N\}} (x_i)$$

### 1.1.2 Dispersion

**Range**

$$R = x_{(n)} - x_{(1)}$$

**Interquartile Range**

$$d_Q = \tilde{x}_{0.75} - \tilde{x}_{0.25}$$

**(Empirical) Variance**

$$s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2$$

Estimates the second centralized moment.

*Calculation Rules:*

$$\star \operatorname{Var}(aX + b) = a^2 \cdot \operatorname{Var}(X)$$

### 1.1.3 Concentration

**Gini Coefficient**

$$G = \frac{2 \sum_{i=1}^n i x_{(i)} - (n+1) \sum_{i=1}^n x_{(i)}}{n \sum_{i=1}^n x_{(i)}} = 1 - \frac{1}{n} \sum_{i=1}^n (v_{i-1} + v_i)$$

with

$$u_i = \frac{i}{n}, \quad v_i = \frac{\sum_{j=1}^i x_{(j)}}{\sum_{j=1}^i x_{(j)}} \quad (u_0 = 0, \quad v_0 = 0)$$

**Arithmetic Mean**

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Estimates the expectation  $\mu = E[X]$  (first moment).

*Calculation Rules:*

$$\star E(a + b \cdot X) = a + b \cdot E(X)$$

$$\star E(X \pm Y) = E(X) \pm E(Y)$$

**Geometric Mean**

$$\bar{x}_G = \sqrt[n]{\sum_{i=1}^n x_i}$$

For growth factors:  $\bar{x}_G = \sqrt[n]{\frac{B_n}{B_0}}$

**Harmonic Mean**

$$\bar{x}_H = \frac{\sum_{i=1}^n w_i}{\sum_{i=1}^n \frac{w_i}{x_i}}$$

$$\star \operatorname{Var}(X \pm Y) = \operatorname{Var}(X) + \operatorname{Var}(Y) + 2\operatorname{Cov}(X, Y)$$

**(Empirical) Standard Deviation**

$$s = \sqrt{s^2}$$

**Coefficient of Variation**

$$\nu = \frac{s}{\bar{x}}$$

**Average Absolute Deviation**

$$e = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|$$

Estimates the first absolute centralized moment.

These are also the values for the Lorenz curve.

$$\text{Range: } 0 \leq G \leq \frac{n-1}{n}$$

**Lorenz-Münzner Coefficient (normed  $G$ )**

$$G^+ = \frac{n}{n-1} G$$

$$\text{Range: } 0 \leq G^+ \leq 1$$

## 1.1.4 Shape

(Empirical) Skewness

$$\nu = \frac{n}{(n-1)(n-2)} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s} \right)^3$$

Estimates the third centralized moment, scaled with  $(\sigma^2)^{\frac{2}{3}}$

## 1.1.5 Dependence

*for two nominal variables*

$\chi^2$ -Statistic

$$\chi^2 = \sum_{i=1}^k \sum_{j=1}^l \frac{(n_{ij} - \frac{n_{i+}n_{+j}}{n})^2}{\frac{n_{i+}n_{+j}}{n}} = n \left( \sum_{i=1}^k \sum_{j=1}^l \frac{n_{ij}^2}{n_{i+}n_{+j}} - 1 \right)$$

Range:  $0 \leq \chi^2 \leq n(\min(k, l) - 1)$

Phi-Coefficient

$$\Phi = \sqrt{\frac{\chi^2}{n}}$$

Range:  $0 \leq \Phi \leq \sqrt{\min(k, l) - 1}$

Cramér's  $V$

$$V = \sqrt{\frac{\chi^2}{\min(k, l) - 1}}$$

Range:  $0 \leq V \leq 1$

Contingency Coefficient  $C$

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

Range:  $0 \leq C \leq \sqrt{\frac{\min(k, l) - 1}{\min(k, l)}}$

Corrected Contingency Coefficient  $C_{corr}$

$$C_{corr} = \sqrt{\frac{\min(k, l)}{\min(k, l) - 1}} \cdot \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

Range  $0 \leq C_{corr} \leq 1$

Odds-Ratio

$$OR = \frac{ad}{bc} = \frac{n_{ii}n_{jj}}{n_{ij}n_{ji}}$$

Range:  $0 \leq OR < \infty$

*for two ordinal variables*

Gamma (Goodman and Kruskal)

$$\gamma = \frac{K - D}{K + D}$$

$K = \sum_{i < m} \sum_{j < n} n_{ij}n_{mn}$  Number of concordant pairs

$D = \sum_{i < m} \sum_{j > n} n_{ij}n_{mn}$  Number of reversed pairs

Range:  $-1 \leq \gamma \leq 1$

## 1.2 Diagrams

(Empirical) Kurtosis

$$k = \left[ n(n+1) \cdot \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s} \right)^4 - 3(n-1) \right] \cdot \frac{n-1}{(n-2)(n-3)} + 3$$

Estimates the fourth centralized moment, scaled with  $(\sigma^2)^2$

Excess

$$\gamma = k - 3$$

Kendall's  $\tau_b$

$$\tau_b = \frac{K - D}{\sqrt{(K + D + T_X)(K + D + T_Y)}}$$

with

$T_X = \sum_{i=m} \sum_{j < n} n_{ij}n_{mn}$  Number of ties w.r.t.  $X$

$T_Y = \sum_{i < m} \sum_{j=n} n_{ij}n_{mn}$  Number of ties w.r.t.  $Y$

Range:  $-1 \leq \tau_b \leq 1$

Kendall's/Stuart's  $\tau_c$

$$\tau_c = \frac{2 \min(k, l)(K - D)}{n^2(\min(k, l) - 1)}$$

Range:  $-1 \leq \tau_c \leq 1$

Spearman's Rank Correlation Coefficient

$$\rho = \frac{n(n^2 - 1) - \frac{1}{2} \sum_{j=1}^J b_j(b_j^2 - 1) - \frac{1}{2} \sum_{k=1}^K c_k(c_k^2 - 1) - 6 \sum_{i=1}^n d_i^2}{\sqrt{n(n^2 - 1) - \sum_{j=1}^J b_j(b_j^2 - 1)} \sqrt{n(n^2 - 1) - \sum_{k=1}^K c_k(c_k^2 - 1)}}$$

or

$$\rho = \frac{srg_x rgy}{\sqrt{srg_x rgy srg_y rgy}}$$

Without ties:

$$\rho = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}$$

with

$d_i = R(x_i) - R(y_i)$  rank difference

Range:  $-1 \leq \rho \leq 1$

*for two metric variables*

Correlation Coefficient (Bravais-Pearson)

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{s_{xy}}{\sqrt{s_{xx}s_{yy}}}$$

with

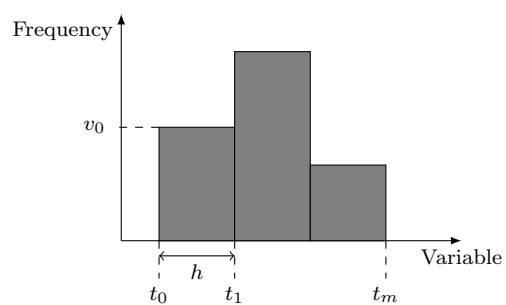
$$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})^2 \quad \text{or } s_{xy} = \frac{S_{xy}}{n}$$

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 \quad \text{or } s_{xx} = \frac{S_{xx}}{n}$$

$$S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 \quad \text{or } s_{yy} = \frac{S_{yy}}{n}$$

Range:  $-1 \leq r \leq 1$

## 1.2.1 Histogram



sample:  $X = \{x_1, x_2, \dots, x_n\}$

$k$ -th bin:  $B_k = [t_k, t_{k+1})$ ,  $k = \{0, 1, \dots, m-1\}$

Number of observations in the  $k$ -th bin:  $v_k$

bin width:  $h = t_{k+1} - t_k, \forall k$

### Scott's Rule

$$h^* \approx 3.5\sigma n^{-\frac{1}{3}}$$

For approximately normal distributed data (min. MSE)

## 2 Probability

### 2.1 Combinatorics

	without replacement	with replacement
Permutations	$n!$	$\frac{n!}{n_1! \cdots n_s!}$
Combinations:		
without order	$\binom{n}{m}$	$\binom{n+m-1}{m}$
with order	$\binom{n}{m} m!$	$n^m$

with:

$$n! = n \cdot (n-1) \cdot \dots \cdot 1$$

$$\binom{n}{m} = \frac{n!}{m!(n-m)!}$$

### 2.2 Probability Theory

**Laplace**

$$P(A) = \frac{|A|}{|\Omega|}$$

**Kolmogorov Axioms** mathematical definition of probability

- (1)  $0 \leq P(A) \leq 1 \quad \forall A \in \mathcal{A} = \sigma\text{-algebra}(\Omega)$
- (2)  $P(\Omega) = 1$
- (3)  $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$   
 $\forall A_i \in \mathcal{A}, i = 1, \dots, \infty$  with  $A_i \cap A_j = \emptyset$  for  $i \neq j$

Implications:

- $P(\bar{A}) = 1 - P(A)$
- $P(\emptyset) = 0$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $A \subseteq B \Rightarrow P(A) \leq P(B)$

**Probability (Mises)** frequentist definition of probability

$$P(A) = \lim_{n \rightarrow \infty} \frac{n_A(n)}{n}$$

with  $n$  repetitions of a random experiment and  $n_A(n)$  events  $A$

**Conditional Probability**

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{for } P(B) > 0$$

$$\Rightarrow P(A \cap B) = P(B|A)P(A) = P(A|B)P(B)$$

**Law of Total Probability**

$$P(B) = \sum_{i=1}^n P(B|A_i)P(A_i) \quad \text{for } \Omega = \bigcup_{i=1}^{\infty} A_i \text{ and } A_i \cap A_j = \emptyset$$

**Bayes' Theorem**

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad \text{for } P(A), P(B) > 0$$

**Stochastic Independence**

$$A, B \text{ independent} \Leftrightarrow P(A \cap B) = P(A) \cdot P(B)$$

$$X, Y \text{ independent} \Leftrightarrow f_{XY}(x, y) = f_X(x) \cdot f_Y(y) \quad \forall x, y$$

### 2.3 Random Variables/Vectors

**Random Variables**  $\in \mathbb{R}$

**Definition**

$$Y : \Omega \rightarrow \mathbb{R}$$

The subset of possible values for  $\mathbb{R}$  is called support.

Notation: Realisations of  $Y$  are depicted with lower case letters.

$Y = y$  means, that  $y$  is the realisation of  $Y$ .

**Discrete and Continuous Random Variables**

If the support is uncountably infinite, the random variable is called *continuous*, otherwise it is called *discrete*.

- **Density  $f(\cdot)$**  (positive, integrates out to 1):

$$\text{For continuous variables: } P(Y \in [a, b]) = \int_a^b f_Y(y) dy$$

For discrete variables the density (and other functions) can be depicted like the corresponding function for continuous variables, if the notation is extended as follows:

$$\int_{-\infty}^y f_Y(\tilde{y}) d\tilde{y} := \sum_{k: k \leq y} P(Y = k). \text{ This notation is used.}$$

$$\text{If } Y = g(X), \text{ then } f_Y(y) = \left| \frac{dg^{-1}(y)}{dy} \right| f_X(g^{-1}(y)).$$

- **Cumulative Distribution Function  $F(\cdot)$ :**

$$F_Y(y) = P(Y \leq y)$$

$$\text{with } \lim_{y \rightarrow -\infty} F_Y(y) = 0 \text{ and } \lim_{y \rightarrow \infty} F_Y(y) = 1$$

Relationship:

$$F_Y(y) = \int_{-\infty}^y f_Y(\tilde{y}) d\tilde{y}$$

**Moments**

- **Expectation (1. Moment):**  $\mu = E(Y) = \int y f_Y(y) dy$   
with  $E(a+bX) = a + bE(X)$  and  $E(X \pm Y) = E(X) \pm E(Y)$

- **Variance (2. centralized Moment):**

$$\sigma^2 = Var(Y) = E(\{Y - E(Y)\}^2) = \int (y - E(Y))^2 f_Y(y) dy$$

$$\text{with } Var(a + bX) = b^2 Var(X) \text{ and}$$

$$Var(X \pm Y) = Var(X) + Var(Y) \pm 2Cov(X, Y)$$

$$\text{Note: } E(\{Y - \mu\}^2) = E(Y^2) - \mu^2$$

Proof:

$$E(\{Y - \mu\}^2) = E(Y^2 - 2Y\mu + \mu^2) = E(Y^2) - 2\mu^2 + \mu^2 = E(Y^2) - \mu^2$$

- **kth Moment:**  $E(Y^k) = \int y^k f_Y(y) dy$ ,  
**kth centralized Moment:**  $E(\{Y - E(Y)\}^k)$

### Moment Generating Function

$$M_Y(t) = E_Y(e^{tY})$$

$$\text{with } \left. \frac{\partial^k M_Y(t)}{\partial t^k} \right|_{t=0} = E(Y^k)$$

$$\text{Cumulant Generating Function } K_Y(t) = \log M_Y(t)$$

A random variable is uniquely defined by its moment generating function and vice versa (as long as moments and cumulants are finite).

**Chebyshev's Inequality** with  $E(X) = \mu$  and  $\text{Var}(X) = \sigma^2$

$$P(|X - \mu| \geq c) \leq \frac{\sigma^2}{c^2}$$

## Random Vectors $\in \mathbb{R}^q$

### Definition

$$(Y_1, Y_2, \dots, Y_q)$$

with random variables  $Y_i$

### Density and Cumulative Distribution Function

$$F(y_1, \dots, y_q) = P(Y_1 \leq y_1, \dots, Y_q \leq y_q)$$

$$P(a_1 \leq Y_1 \leq b_1, \dots, a_q \leq Y_q \leq b_q) = \int_{a_1}^{b_1} \dots \int_{a_q}^{b_q} f(y_1, \dots, y_q) dy_1 \dots dy_q$$

### Marginal Density

$$f_{Y_1}(y_1) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(y_1, \dots, y_k) dy_2 \dots dy_k$$

## 2.4 Probability Distributions

### 2.4.1 Discrete Distributions

#### Discrete Uniform

$$Y \sim U(\{y_1, \dots, y_k\}), y \in \{y_1, \dots, y_k\}$$

$$P(Y = y_i) = \frac{1}{k}, i = 1, \dots, k$$

$$E(Y) = \frac{k+1}{2}, \text{Var}(Y) = \frac{k^2-1}{12}$$

#### Binomial successes in independent trials

with special case Bernoulli:  $Y \sim \text{Bin}(1, \pi)$

$$Y \sim \text{Bin}(n, \pi) \text{ with } n \in \mathbb{N}, \pi \in [0, 1], y \in \{0, \dots, n\}$$

$$P(Y = y|\lambda) = \binom{n}{y} \pi^y (1-\pi)^{n-y}$$

$$E(Y|\pi, n) = n\pi, \text{Var}(Y|\pi, n) = n\pi(1-\pi)$$

#### Poisson Counting model for rare events

### Conditional Density two-dimensional case

$$f_{Y_1|Y_2}(y_1|y_2) = \frac{f(y_1, y_2)}{f(y_2)} \text{ for } f(y_2) > 0$$

### Covariance and Correlation

$$\text{Cov}(Y_j, Y_k) = E(Y_j Y_k) - E(Y_j)E(Y_k)$$

$$\text{Cor}(Y_j, Y_k) = \frac{\text{Cov}(Y_j, Y_k)}{\sqrt{\text{Var}(Y_j)\text{Var}(Y_k)}}$$

### Iterated Expectation

$$E(Y) = E_X(E(Y|X))$$

Proof:

$$E(Y) = \int y f(y) dy = \int \int y f(y|x) dy f_X(x) dx = E_X(E(Y|X))$$

$$\text{Var}(Y) = E_X(\text{Var}(Y|X)) + \text{Var}_X(E(Y|X))$$

Proof:

$$\begin{aligned} \text{Var}(Y) &= \int (y - \mu_Y)^2 f(y) dy \\ &= \int (y - \mu_Y)^2 f(y|x) f(x) dy dx \\ &= \int (y - \mu_{Y|x} + \mu_{Y|x} - \mu_Y)^2 f(y|x) f(x) dy dx \\ &= \int (y - \mu_{Y|x})^2 f(y|x) f(x) dy dx + \\ &\quad \int (\mu_{Y|x} - \mu_Y)^2 f(y|x) f(x) dy dx + \\ &\quad 2 \int (y - \mu_{Y|x})(\mu_{Y|x} - \mu_Y) f(y|x) f(x) dy dx \\ &= \int \text{Var}(Y|x) f(x) dx + \int (\mu_{Y|x} - \mu_Y)^2 f(x) dx \\ &= E_X(\text{Var}(Y|X)) + \text{Var}_X(E(Y|X)) \end{aligned}$$

only one event at a time, no autocorrelation, mean number of events over time is constant and proportional to length of the considered time interval

$$Y \sim \text{Po}(\lambda) \text{ with } \lambda \in [0, +\infty], y \in \mathbb{N}_0$$

$$P(Y = y|\lambda) = \frac{\lambda^y \exp^{-\lambda}}{y!}$$

$$E(Y|\lambda) = \lambda, \text{Var}(Y|\lambda) = \lambda$$

The model tends to overestimate the variance (Overdispersion).

Approximation of the Binomial for small  $p$  and big  $n$

#### Geometric

$$Y \sim \text{Geom}(\pi) \text{ with } \pi \in [0, 1], y \in \mathbb{N}_0$$

$$P(Y = y|\pi) = \pi(1-\pi)^{y-1}$$

$$E(Y|\pi) = \frac{1}{\pi}, \text{Var}(Y|\pi) = \frac{1-\pi}{\pi^2}$$

### Negative Binomial

$$Y \sim \text{NegBin}(\alpha, \beta) \text{ with } \alpha, \beta \geq 0, y \in \mathbb{N}_0$$

## 2.4.2 Continuous Distributions

### Continuous Uniform

$$Y \sim U(a, b) \text{ with } \alpha, \beta \in \mathbb{R}, a \leq b, y \in [a, b]$$

$$p(y|a, b) = \frac{1}{b-a}$$

$$E(Y|a, b) = \frac{a+b}{2}, \text{Var}(Y|a, b) = \frac{(b-a)^2}{12}$$

### Exponential Time between Poisson events

$$Y \sim \text{Exp}(\lambda) \text{ with } \lambda > 0, y \geq 0$$

$$p(y|\lambda) = \lambda \exp(-\lambda y)$$

$$E(Y|\lambda) = \frac{1}{\lambda}, \text{Var}(Y|\lambda) = \frac{1}{\lambda^2}$$

### Univariate Normal symmetric with $\mu$ and $\sigma^2$

$$Y \sim N(\mu, \sigma^2) \text{ with } \mu \in \mathbb{R}, \sigma^2 > 0, y \in \mathbb{R}$$

$$p(y|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

$$E(Y|\mu, \sigma^2) = \mu, \text{Var}(Y|\mu, \sigma^2) = \sigma^2$$

### Approximation of

Binomial for  $np(1-p) \geq 9$  with  $N(np, np(1-p))$

Poisson for  $\lambda \geq 10$  with  $N(\lambda, \lambda)$

### Log-Normal

$$Y \sim \text{LogN}(\mu, \sigma^2) \text{ with } \mu \in \mathbb{R}, \sigma^2 > 0, y > 0$$

$$p(y|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}y} \exp\left(-\frac{(\log y - \mu)^2}{2\sigma^2}\right)$$

$$E(Y|\mu, \sigma^2) = \exp\left(\mu + \frac{\sigma^2}{2}\right),$$

$$\text{Var}(Y|\mu, \sigma^2) = \exp(2\mu + \sigma^2)(\exp(\sigma^2) - 1)$$

Relationship:  $\log(Y) \sim N(\mu, \sigma^2) \Rightarrow Y \sim \text{LogN}(\mu, \sigma^2)$

**non-standardized Student's t** statistical tests for  $\mu$  with unknown (estimated) variance and  $\nu$  degrees of freedom

$$Y \sim t_\nu(\mu, \sigma^2) \text{ with } \mu \in \mathbb{R}, \sigma^2, \nu > 0, y \in \mathbb{R}$$

$$p(y|\mu, \sigma^2, \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\Gamma(\sqrt{\nu\pi}\sigma)} \left(1 + \frac{(y-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}}$$

$$E(Y|\mu, \sigma^2, \nu) = \mu \text{ for } \nu > 1,$$

$$\text{Var}(Y|\mu, \sigma^2, \nu) = \sigma^2 \frac{\nu}{\nu-2} \text{ for } \nu > 2$$

Relationship:  $Y|\theta \sim N(\mu, \frac{\sigma^2}{\theta}), \theta \sim \text{Ga}(\frac{\nu}{2}, \frac{\nu}{2}) \Rightarrow Y \sim t_\nu(\mu, \sigma)$   
 $t_\nu(\mu, \sigma^2)$  has heavier tails than the normal distribution.  
 $t_\infty(\mu, \sigma^2)$  approaches  $N(\mu, \sigma^2)$ .

$$P(Y = y|\alpha, \beta) = \binom{\alpha+y-1}{\alpha-1} \left(\frac{\beta}{\beta-1}\right)^\alpha \left(\frac{1}{\beta+1}\right)^y$$

$$E(Y|\alpha, \beta) = \frac{\alpha}{\beta}, \text{Var}(Y|\alpha, \beta) = \frac{\alpha}{\beta^2}(\beta+1)$$

### Beta

$$Y \sim \text{Be}(a, b) \text{ with } a, b > 0, y \in [0, 1]$$

$$p(y|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} y^{a-1} (1-y)^{b-1}$$

$$E(Y|a, b) = \frac{a}{a+b},$$

$$\text{Var}(Y|a, b) = \frac{ab}{(a+b)^2(a+b+1)},$$

$$\text{mod}(Y|a, b) = \frac{a-1}{a+b-2} \text{ for } a, b > 1$$

### Gamma

$$Y \sim \text{Ga}(a, b) \text{ with } a, b > 0, y > 0$$

$$p(y|a, b) = \frac{b^a}{\Gamma(a)} y^{a-1} \exp(-by)$$

$$E(Y|a, b) = \frac{a}{b},$$

$$\text{Var}(Y|a, b) = \frac{a}{b^2},$$

$$\text{mod}(Y|a, b) = \frac{a-1}{b} \text{ for } a \geq 1$$

### Inverse-Gamma

$$Y \sim \text{IG}(a, b) \text{ with } a, b > 0, y > 0$$

$$p(y|a, b) = \frac{b^a}{\Gamma(a)} y^{-a-1} \exp\left(-\frac{b}{y}\right)$$

$$E(Y|a, b) = \frac{b}{a-1} \text{ for } a > 1,$$

$$\text{Var}(Y|a, b) = \frac{b^2}{(a-1)^2(a-2)} \text{ for } a \geq 2,$$

$$\text{mod}(Y|a, b) = \frac{b}{a+1}$$

Relationship:  $Y^{-1} \sim \text{Ga}(a, b) \Leftrightarrow Y \sim \text{IG}(a, b)$

**Chi-Squared** sum of squares for standard normal random variables with  $\nu$  degrees of freedom

$$Y \sim \chi^2(\nu) \text{ with } \nu > 0, y \in \mathbb{R}$$

$$p(y|\nu) = \frac{y^{\frac{\nu}{2}-1} e^{-\frac{y}{2}}}{2^{\frac{\nu}{2}} \Gamma(\frac{\nu}{2})}$$

$$E(Y|\nu) = \nu, \text{Var}(Y|\nu) = 2\nu$$

**Weibull** failure rate is proportional to a power of time

$$Y \sim \text{WB}(\lambda, k) \text{ with } \lambda > 0$$

$$p(y|\lambda, k) = \frac{k}{\lambda} \left(\frac{y}{\lambda}\right)^{k-1} e^{-(y/\lambda)^k}$$



## 2.4.3 Exponential Family

### Definition

The exponential family comprises all distributions, whose density can be written as follows:

$$f_Y(y, \theta) = e^{t^T(y)\theta - \kappa(\theta)} h(y)$$

with  $h(y) \geq 0$ ,  $t(y)$  vector of the canonical statistic, parameter vector  $\theta$  and  $\kappa(\theta)$  as the normalising constant.

### Normalising Constant

$$1 = \int \exp^{t^T(y)\theta} h(y) dy \exp^{-\kappa(\theta)}$$

$$\Leftrightarrow \kappa(\theta) = \log \int \exp^{t^T(y)\theta} h(y) dy$$

$\kappa(\theta)$  is the cumulant generating function, therefore e. g.

$$\frac{\partial \kappa(\theta)}{\partial \theta_1} = E(t_1(Y))$$

### Members

- **Poisson**
- **Geometric**
- **Exponential**
- **Normal**  $t(y) = \left(-\frac{y^2}{2}, y\right)^T$ ,  $\theta = \left(\frac{1}{\sigma^2}, \frac{\mu}{\sigma^2}\right)^T$ ,  $h(y) = \frac{1}{\sqrt{2\pi}}$ ,  $\kappa(\theta) = \frac{1}{2} \left(-\log \frac{1}{\sigma^2} + \frac{\mu^2}{\sigma^2}\right)$
- **Gamma**
- **Chi-Squared**
- **Beta**
- **Binomial**

## 2.5 Multivariate Distributions

**Multivariate Normal** symmetric with  $\mu_i$  and  $\Sigma$

$$Y \sim N(\mu, \Sigma) \text{ with } \mu \in \mathbb{R}^d, \Sigma \in \mathbb{R}^{d \times d} \text{ s.p.d., } y \in \mathbb{R}^d$$

$$p(y|\mu, \Sigma) = (2\pi)^{-\frac{d}{2}} \det(\Sigma)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(y - \mu)^T \Sigma^{-1}(y - \mu)\right)$$

$$E(Y|\mu, \Sigma) = \mu, \text{ Var}(Y|\mu, \Sigma) = \Sigma$$

### General Copulas

$$F(y_1, \dots, y_q) = C(F_1(y_1), \dots, F_q(y_q)) \text{ with } C: [0, 1]^q \rightarrow [0, 1]$$

with  $C$  monotonically increasing as a cdf on  $[0, 1]^q$

Modelled as follows:

1. marginal distributions  $F_j(y_j) = C(F_j(y_j), 1, \dots, 1)$
2. dependence structure  $\hat{u}_i = (\hat{u}_{i1}, \dots, \hat{u}_{iq}) \stackrel{iid}{\sim} C(\cdot)$  with  $\hat{u}_{ij} := \hat{F}_j(y_{ij})$ .

The copula density is  $c(u_{1:q}) = \frac{\partial^q C(u_{1:q})}{\partial u_1 \dots \partial u_q}$  and  $f(y_{1:q}) = c(F_1(y_1), \dots, F_q(y_q)) \prod_{j=1}^q f_j(y_j)$ .

### Tail Dependence

$$\text{upper: } \lambda_u := \lim_{u \rightarrow 1} P(Y_1 \geq F_1^{-1}(u) | Y_2 \geq F_2^{-1}(u))$$

$$= \lim_{u \rightarrow 1} \frac{1 - 2u + C(u, u)}{1 - u}$$

$$\text{lower: } \lambda_l := \lim_{u \rightarrow 0} P(Y_1 \leq F_1^{-1}(u) | Y_2 \leq F_2^{-1}(u))$$

$$= \lim_{u \rightarrow 0} \frac{C(u, u)}{u}$$

**Gaussian Copula** coefficients for pairwise dependences

$$c(u_{1:q}) = \frac{1}{|R|^{1/2}} \exp\left(-\frac{1}{2}u^T R^{-1}u\right)$$

For  $Y_{ij} \sim N(\mu_j, \sigma_j^2)$ :  $f(y_{ij}; \mu_j, \sigma_j^2) = \frac{1}{\sigma_j} \phi(Z_{ij})$  with  $Z_{ij}$  the standardized  $Y_{ij}$ . With  $u_{ij} = \Phi^{-1}(Z_{ij})$ ,  $R$  can be estimated.

$$\lambda_l = 0, \lambda_u = 0$$

**Archimedean Copulas** few parameters even in high dimensions

$$\psi(\cdot; \theta) : [0, 1] \rightarrow [0, \infty)$$

with the parametric generator function  $\psi(u, \theta)$  continuous, strictly decreasing, convex, and  $\psi(1, \theta) = 0 \forall \theta$

$$C(u_{1:q}; \theta) = \psi^{-1}(\psi(u_1; \theta) + \dots + \psi(u_q; \theta); \theta)$$

- **Clayton**  $\psi(t; \theta) = \frac{1}{\theta}(\theta^{-1} - 1)$ :  $\lambda_l = 2^{-1/\theta}$ ,  $\lambda_u = 0$
- **Frank**  $\psi(t; \theta) = -\log \frac{\exp(-\theta t) - 1}{\exp(-\theta) - 1}$ :  $\lambda_l = 0$ ,  $\lambda_u = 0$
- **Gumbel**  $\psi(t; \theta) = (-\log(t))^\theta$ :  $\lambda_l = 0$ ,  $\lambda_u = 2 - 2^{1/\theta}$

**Pair Copulas** flexible pairwise dependences

$$f_{123} = c_{12}c_{23}c_{23|1} \prod_{j=1}^3 f_j$$

### Generalized Extreme Value Distribution (GEV)

for block maxima  $M_n := \max(Y_{1:n})$ :

$$F_{M_n}(y) = P(M_n \leq y) = P(Y_{1:n} \leq y) = (F_Y(y))^n$$

$$\lim_{n \rightarrow \infty} f_{M_n}(y) = \begin{cases} 1, & \text{if } F(y) = 1 \\ 0, & \text{otherwise} \end{cases}$$

For  $\{a_n\}_{n=1}^\infty, \{b_n\}_{n=1}^\infty$  fixed sequences, the standardized maximum  $\frac{M_n - a_n}{b_n}$  converges to a GEV as  $n \rightarrow \infty$ .

$$G(x) = \begin{cases} \exp(-(1 + \gamma z)^{-1/\gamma}), & \text{for } \gamma \neq 0 \\ \exp(-\exp(-z)), & \text{for } \gamma = 0 \end{cases}$$

with location  $\mu$ , scale  $\sigma$ , and shape  $\gamma$  and  $z = \frac{x - \mu}{\sigma}$

- **Gumbel**  $\gamma = 0$
- **Weibull**  $\gamma > 0$
- **Frechet-Pareto**  $\gamma < 0$

## 2.6 Limit Theorems

### Law of Large Numbers

$$\lim_{n \rightarrow \infty} P(|\bar{X}_n - \mu| < c) = 1 \quad \forall c > 0$$

with  $X_i$  i.i.d.,  $E(X_i) = \mu$ , and  $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$

### Central Limit Theorem

$$Z_n \xrightarrow{d} N(0, \sigma^2)$$

with  $Z_n = \sum_{i=1}^n \frac{Y_i}{\sqrt{n}}$  and  $Y_i$  i.i.d. with expectation 0 and variance  $\sigma^2$

Proof:

For normal random variables  $Z \sim N(\mu, \sigma^2)$ :  $K_Z(t) = \mu t + \frac{1}{2} \sigma^2 t^2$ . The first two derivatives  $\left. \frac{\partial^k K_Z(t)}{\partial t^k} \right|_{t=0}$  are  $\mu$  and  $\sigma$ . All other moments are zero.

For  $Z_n = (Y_1 + Y_2 + \dots + Y_n)/\sqrt{n}$ :

$$\begin{aligned} M_{Z_n}(t) &= E\left(e^{t(Y_1+Y_2+\dots+Y_n)/\sqrt{n}}\right) \\ &= E\left(e^{tY_1/\sqrt{n}} \cdot e^{tY_2/\sqrt{n}} \cdot \dots \cdot e^{tY_n/\sqrt{n}}\right) \\ &= E\left(e^{tY_1/\sqrt{n}}\right) E\left(e^{tY_2/\sqrt{n}}\right) \dots E\left(e^{tY_n/\sqrt{n}}\right) \\ &= M_Y^n(t/\sqrt{n}) \end{aligned}$$

Analogously:  $K_{Z_n}(t) = nK_Y(t/\sqrt{n})$ .

$$\begin{aligned} \left. \frac{\partial K_{Z_n}(t)}{\partial t} \right|_{t=0} &= \frac{n}{\sqrt{n}} \left. \frac{\partial K_Y(t)}{\partial t} \right|_{t=0} = \sqrt{n} \mu \\ \left. \frac{\partial^2 K_{Z_n}(t)}{\partial t^2} \right|_{t=0} &= \frac{n}{n} \left. \frac{\partial^2 K_Y(t)}{\partial t^2} \right|_{t=0} = \sigma^2 \end{aligned}$$

Using the Taylor Expansion, we can write  $K_{Z_n}(t) = 0 + \sqrt{n} \mu t + \frac{1}{2} \sigma^2 t^2 + \dots$ , where the terms in  $\dots$  are tending towards 0 as  $n \rightarrow \infty$ .

Therefore:  $K_{Z_n}(t) \xrightarrow{n \rightarrow \infty} K_Z(t)$  with  $Z \sim N(\sqrt{n} \mu, \sigma^2)$ .

## 3 Inference

### 3.1 Method of Moments

The theoretical moments are estimated by their empirical counterparts:

$$E_{\hat{\theta}_{MM}}(Y^k) = m_k(y_1, \dots, y_n) = \frac{1}{n} \sum_{i=1}^n y_i^k$$

For the exponential family:  $\hat{\theta}_{MM} = \hat{\theta}_{ML}$

### 3.2 Loss Functions

**Loss**

$$\mathcal{L} : \mathcal{T} \times \Theta \rightarrow \mathbb{R}^+$$

with parameter space  $\Theta \subset \mathbb{R}$ ,  $t \in \mathcal{T}$  with  $t : \mathbb{R}^n \rightarrow \mathbb{R}$  a statistic, that estimates the parameter  $\theta$ ,  $\mathcal{L}(\theta, \theta) = 0$  holds

- **absolute loss (L1):**  $\mathcal{L}(t, \theta) = |t - \theta|$
- **quadratic loss (L2):**  $\mathcal{L}(t, \theta) = (t - \theta)^2$

As  $\theta$  is unknown, the loss is a theoretical quantity. It is also the realisation of a random variable as it depends on a sample.

**Risk**

$$\begin{aligned} R(t(\cdot), \theta) &= E_{\theta}(\mathcal{L}(t(Y_1, \dots, Y_n), \theta)) \\ &= \int_{-\infty}^{\infty} \mathcal{L}(t(Y_1, \dots, Y_n), \theta) \prod_{i=1}^n f(y_i; \theta) dy_i \end{aligned}$$

**Minimax Approach** minimizing the worst case

The risk still depends on the true parameter  $\theta$ .

Tentative estimation: Choose  $\theta$ , s. t. the risk is maximal and then  $t(\cdot)$ , so that the risk is minimized:

$$\hat{\theta}_{minimax} = \arg \min_{t(\cdot)} \left( \max_{\theta \in \Theta} R(t(\cdot); \theta) \right)$$

**Mean Squared Error (MSE)**

$$\begin{aligned} MSE(t(\cdot), \theta) &= E_{\theta}(\{t(Y) - \theta\}^2) \\ &= \text{Var}_{\theta}(t(Y_1, \dots, Y_n)) + \text{Bias}^2(t(\cdot); \theta) \end{aligned}$$

with  $\text{Bias}(t(\cdot); \theta) = E_{\theta}(t(Y_1, \dots, Y_n)) - \theta$

Proof:

Let  $\mathcal{L}(t, \theta) = (t - \theta)^2$

$$\begin{aligned} R(t(\cdot), \theta) &= E_{\theta}(\{t(Y) - \theta\}^2) \\ &= E_{\theta}(\{t(Y) - E_{\theta}(t(Y)) + E_{\theta}(t(Y)) - \theta\}^2) \\ &= E_{\theta}(\{t(Y) - E_{\theta}(t(Y))\}^2) + E_{\theta}(\{E_{\theta}(t(Y)) - \theta\}^2) \\ &\quad + 2E_{\theta}(\{t(Y) - E_{\theta}(t(Y))\}\{E_{\theta}(t(Y)) - \theta\}) \\ &= \text{Var}_{\theta}(t(Y_1, \dots, Y_n)) + \text{Bias}^2(t(\cdot); \theta) + 0 \end{aligned}$$

**Cramér-Rao Bound**

$$MSE(\hat{\theta}, \theta) \geq \text{Bias}^2(\hat{\theta}, \theta) + \frac{\left(1 + \frac{\partial \text{Bias}(\hat{\theta}, \theta)}{\partial \theta}\right)^2}{I(\theta)}$$

Proof:

For unbiased estimates:  $\theta = E_{\theta}(\hat{\theta}) = \int t(y)f(y; \theta)dy$

$$\begin{aligned} 1 &= \int t(y) \frac{\partial f(y; \theta)}{\partial \theta} dy \\ &= \int t(y) \frac{\partial \log f(y; \theta)}{\partial \theta} f(y; \theta) dy \\ &= \int t(y) s(y; \theta) f(y; \theta) dy \\ &= \int (t(y) - \theta) (s(y; \theta) - 0) f(y; \theta) dy && \text{1. Bartlett equation } E_{\theta}(s(\theta; y)) = 0 \\ &= \text{Cov}_{\theta}(t(Y); s(\theta; Y)) \\ &\geq \sqrt{\text{Var}_{\theta}(t(Y))} \sqrt{\text{Var}_{\theta}(s(\theta; Y))} && \text{Cauchy-Schwarz} \\ &= \sqrt{MSE(t(Y); \theta)} \sqrt{I(\theta)} \end{aligned}$$

**Kullback-Leibler Divergence** Comparing distributions

$$KL(\theta, t) = \int_{-\infty}^{\infty} \log \frac{f(\tilde{y}; \theta)}{f(\tilde{y}; t)} f(\tilde{y}; \theta) d\tilde{y}$$

The KL divergence is not a distance as it is not symmetric. It is 0 for  $t = \theta$  and  $\geq 0$  otherwise.

Proof:

Follows from  $\log(x) \leq x - 1 \forall x \geq 0$ , with equality for  $x = 1$ .

$R_{KL}(\theta, t(\cdot))$  is approximated by the MSE.

Proof:

$$\begin{aligned} R_{KL}(\theta, t(\cdot)) &= \int_{-\infty}^{\infty} \mathcal{L}_{KL}(t(Y_1, \dots, Y_n), \theta) \prod_{i=1}^n f(y_i; \theta) dy_i \\ &= \int \int \log \frac{f(\tilde{y}; \theta)}{f(\tilde{y}; t)} f(\tilde{y}; \theta) d\tilde{y} \prod_{i=1}^n f(y_i; \theta) dy_i \\ &= \int \int (\log f(\tilde{y}; \theta) - \log f(\tilde{y}; t)) f(\tilde{y}; \theta) d\tilde{y} \prod_{i=1}^n f(y_i; \theta) dy_i \\ &\approx - \int \underbrace{\left( \int \frac{\partial \log f(\tilde{y}; \theta)}{\partial \theta} f(\tilde{y}; \theta) d\tilde{y} \right)}_0 (t - \theta) \prod_{i=1}^n f(y_i; \theta) dy_i \\ &\quad + \frac{1}{2} \int \underbrace{\left( - \int \frac{\partial^2 \log f(\tilde{y}; \theta)}{\partial \theta^2} f(\tilde{y}; \theta) d\tilde{y} \right)}_{I(\theta)} (t - \theta)^2 \prod_{i=1}^n f(y_i; \theta) dy_i \end{aligned}$$

The last step is approximated by the Taylor Expansion:  
 $\log f(\tilde{y}, t) \approx \log f(\tilde{y}, \theta) + \frac{\partial \log f(\tilde{y}, \theta)}{\partial \theta} (t - \theta) + \frac{1}{2} \frac{\partial^2 \log f(\tilde{y}, \theta)}{\partial \theta^2} (t - \theta)^2$

### 3.3 Maximum Likelihood (ML)

#### Prerequisites

- $Y_i \sim F(y; \theta)$  i.i.d.
- $\theta \in \mathbb{R}^p$
- $f(\cdot; \theta)$  Fisher-regular:
  - $\{y : f(y; \theta) > 0\}$  independent of  $\theta$
  - Parameter space  $\Theta$  is open
  - $f(y; \theta)$  twice differentiable
  - $\int \frac{\partial}{\partial \theta} f(y; \theta) dy = \frac{\partial}{\partial \theta} \int f(y; \theta) dy$

#### Central Functions

- **Likelihood**  $L(\theta; y_1, \dots, y_n) = \prod_{i=1}^n f(y_i; \theta)$
- **log-Likelihood**  $l(\theta; y_1, \dots, y_n)$ :  
 $\log L(\theta; y_1, \dots, y_n) = \sum_{i=1}^n \log f(y_i; \theta)$
- **Score**  $s(\theta; y_1, \dots, y_n) = \frac{\partial l(\theta; y_1, \dots, y_n)}{\partial \theta}$
- **Fisher-Information**  $I(\theta) = -E_\theta \left( \frac{\partial s(\theta; Y)}{\partial \theta} \right)$
- **observed Fisher-Information**  $J(\theta) = -E_\theta \left( \frac{\partial s(\theta; y)}{\partial \theta} \right)$

#### Attributes of the Score-Function

first Bartlett-Equation:

$$E(s(\theta; Y)) = 0$$

Proof:

$$\begin{aligned} 1 &= \int f(y; \theta) dy \\ 0 &= \frac{\partial 1}{\partial \theta} = \int \frac{\partial f(y; \theta)}{\partial \theta} dy = \int \frac{\partial f(y; \theta) / \partial \theta}{f(y; \theta)} f(y; \theta) dy \\ &= \int \frac{\partial \log f(y; \theta)}{\partial \theta} f(y; \theta) dy = \int s(\theta; y) f(y; \theta) dy \end{aligned}$$

second Bartlett-Equation:

$$\text{Var}_\theta(s(Y; \theta)) = E_\theta \left( -\frac{\partial^2 \log f(Y; \theta)}{\partial \theta^2} \right) = I(\theta)$$

Proof:

$$\begin{aligned} 0 &= \frac{\partial 0}{\partial \theta} = \frac{\partial}{\partial \theta} \int \frac{\partial \log f(y; \theta)}{\partial \theta} f(y; \theta) dy \quad \text{see above} \\ &= \int \frac{\partial^2 \log f(y; \theta)}{\partial \theta^2} f(y; \theta) dy \\ &\quad + \int \frac{\partial \log f(y; \theta)}{\partial \theta} \frac{\partial f(y; \theta)}{\partial \theta} dy \\ &= E_\theta \left( \frac{\partial^2 \log f(Y; \theta)}{\partial \theta^2} \right) \\ &\quad + \int \frac{\partial \log f(y; \theta)}{\partial \theta} \frac{\partial \log f(y; \theta)}{\partial \theta} f(y; \theta) dy \\ &\Leftrightarrow E_\theta(s(\theta; Y)s(\theta; Y)) = E_\theta \left( -\frac{\partial^2 \log f(Y; \theta)}{\partial \theta^2} \right) \end{aligned}$$

Bartlett's second equation holds then as  $E(s(\theta; Y)) = 0$

#### ML-Estimate

$$\hat{\theta}_{ML} = \arg \max l(\theta; y_1, \dots, y_n)$$

for Fisher-regular distributions:  $\hat{\theta}_{ML}$  has asymptotically the smallest variance, given by the Cramér-Rao bound,

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

$$\hat{\theta} \stackrel{a}{\sim} N(\theta, I^{-1}(\theta))$$

If the true model is unknown, the distribution is

$\hat{\theta} \stackrel{a}{\sim} N(\theta, I^{-1}(\theta)V(\theta)I^{-1}(\theta))$  with  $V(\theta)$  variance of the score function.

The ML-estimate is invariant for a bijective and differentiable function  $g(\cdot)$ :  $\hat{\gamma} = g(\hat{\theta})$  if  $\gamma = g(\theta)$ .

Proof:

$$\gamma = g(\theta) \Leftrightarrow \theta = g^{-1}(\gamma)$$

For the log-likelihood of  $\gamma$  at the location  $\hat{\theta}$  holds:

$$\frac{\partial l(g^{-1}(\hat{\gamma}))}{\partial \gamma} = \frac{\partial g^{-1}(\gamma)}{\partial \gamma} \underbrace{\frac{\partial l(\hat{\theta})}{\partial \theta}}_{=0} = 0$$

Then, the Fisher information is  $\frac{\partial \theta}{\partial \gamma} I(\theta) \frac{\partial \theta}{\partial \gamma}$

Proof:

$$\begin{aligned} I_\gamma(\gamma) &= -E \left( \frac{\partial^2 l(g^{-1}(\hat{\gamma}))}{\partial \gamma^2} \right) = -E \left( \frac{\partial}{\partial \gamma} \left( \frac{\partial g^{-1}(\gamma)}{\partial \gamma} \frac{\partial l(\theta)}{\partial \theta} \right) \right) \\ &= -E \left( \underbrace{\frac{\partial^2 g^{-1}(\gamma)}{\partial \gamma^2} \frac{\partial l(\theta)}{\partial \theta}}_{\text{Expectation 0}} + \frac{\partial g^{-1}(\gamma)}{\partial \gamma} \frac{\partial^2 l(\theta)}{\partial \theta^2} \frac{\partial g^{-1}(\gamma)}{\partial \gamma} \right) \\ &= \frac{\partial g^{-1}(\gamma)}{\partial \gamma} I(\theta) \frac{\partial g^{-1}(\gamma)}{\partial \gamma} = \frac{\partial \theta}{\partial \gamma} I(\theta) \frac{\partial \theta}{\partial \gamma} \end{aligned}$$

Delta rule:  $\gamma \stackrel{a}{\sim} N(\hat{\gamma}, \frac{\partial \theta}{\partial \gamma} I^{-1}(\theta) \frac{\partial \theta}{\partial \gamma})$

**Numerical computation of the ML estimate** Fisher-Scoring as statistical version of the Newton-Raphson procedure

1. Initialize  $\theta_{(0)}$
2. Repeat:  $\theta_{(t+1)} := \theta_{(t)} + I^{-1}(\theta_{(t)})s(\theta_{(t)}; y)$
3. Stop if  $\|\theta_{(t+1)} - \theta_{(t)}\| < \tau$ ; return  $\hat{\theta}_{ML} = \theta_{(t+1)}$

Proof:

$$0 = s(\hat{\theta}_{ML}; y) \stackrel{\text{Taylor}}{\approx}_{\text{Series}} s(\theta; y) + \frac{\partial s(\theta; y)}{\partial \theta} (\hat{\theta}_{ML} - \theta) \Leftrightarrow$$

$$\hat{\theta}_{ML} \approx \theta - \left( \frac{\partial s(\theta; y)}{\partial \theta} \right)^{-1} s(\theta; y) \approx \theta - I^{-1}(\theta)s(\theta; y)$$

As  $\frac{\partial s(\theta; y)}{\partial \theta}$  is often complicated, its expectation  $I(\theta)$  is used.

The second part in 2 can be weighted with a step size  $\delta$  or  $\delta(t) \in (0, 1)$ , e.g. to ensure convergence.

If  $I(\theta)$  can't be analytically derived, simulation from  $f(y; \theta_{(t)})$  can be used. For the exponential family, step 2 then changes to  $\theta_{(t+1)} := \theta_{(t)} + \widehat{\text{Var}}_{\theta_{(t)}}(t(Y))^{-1} \widehat{E}_{\theta_{(t)}}(t(Y))$  as the ML estimate is the expectation.

### Log Likelihood Ratio

$$lr(\theta, \hat{\theta}) := l(\hat{\theta}) - l(\theta) = \log \frac{L(\hat{\theta})}{L(\theta)}$$

with  $2 \cdot lr(\theta, \hat{\theta}) \stackrel{a}{\sim} \chi_1^2$

Proof:

$$\begin{aligned} l(\theta) &\stackrel{\substack{\text{Taylor} \\ \text{Series}}}{\approx} l(\hat{\theta}) + \underbrace{\frac{\partial l(\hat{\theta})}{\partial \theta}}_{=0} (\theta - \hat{\theta}) + \frac{1}{2} \underbrace{\frac{\partial^2 l(\hat{\theta})}{\partial \theta^2}}_{\approx -I(\theta)} (\underbrace{\theta - \hat{\theta}}_{\approx I^{-1}(\theta)} s(\theta; Y))^2 \\ &\approx l(\hat{\theta}) - \frac{1}{2} \frac{s^2(\theta, Y)}{I(\theta)} \end{aligned}$$

$s(\theta, Y)$  is asymptotically normal.

If  $\theta \in \mathbb{R}^p$  the corresponding distribution is  $\chi_p^2$ .

### Relation to Kullback-Leibler divergence

$$\hat{\theta}_{ML} = \arg \min \text{KL}(g, f)$$

with  $f$  distributional model used and  $g$  true model

Proof:

$$\begin{aligned} KL(g, f) &= \int \log \frac{g(y)}{f(y)} g(y) dy \\ &= \int \log(g(y)) g(y) dy - \int \log(f(y)) g(y) dy \end{aligned}$$

To minimize that, the second component needs to be maximized. Its derivative is  $\int s(\theta; y) g(y) dy = E_g(s(\theta; Y)) = 0$

## 3.4 Consistency and Sufficiency

### Statistic

$$t : \mathbb{R}^n \rightarrow \mathbb{R}$$

$t(Y_1, \dots, Y_n)$  depends on sample size  $n$  and is a random variable

**(Weak) Consistency**  $\hat{\theta}$  gets closer to  $\theta$  as  $n$  grows

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

Proof:

$$P(|\hat{\theta} - E_{\theta}(\hat{\theta})| \geq \delta) \leq \frac{\text{Var}_{\theta}(\hat{\theta})}{\delta^2} \text{ using the inequality of Chebyshev and } MSE(t(\cdot), \theta) = \text{Var}_{\theta}(t(Y_1, \dots, Y_n)) + \text{Bias}^2(t(\cdot); \theta)$$

### Sufficiency

A statistic  $t(y_1, \dots, y_n)$  is sufficient for  $\theta$ , if the conditional distribution  $f(y_1, \dots, y_n | t_0 = t(y_1, \dots, y_n); \theta)$  is independent of  $\theta$ .

### Neyman criterion:

$$t(Y_1, \dots, Y_n) \text{ sufficient} \Leftrightarrow f(y; \theta) = h(y) g(t(y); \theta)$$

Proof:

“ $\Rightarrow$ ”:

$$f(y; \theta) = \underbrace{f(y | t=t(y); \theta)}_{h(y)} \underbrace{f_t(t(y); \theta)}_{g(t(y); \theta)}$$

“ $\Leftarrow$ ”:

$$f_t(t; \theta) = \int_{t=t(y)} f(y; \theta) dy = \int_{t=t(y)} h(y) g(t; \theta) dy$$

Therefore:

$$f(y | t=t(y); \theta) = \frac{f(y, t=t(y); \theta)}{f_t(t, \theta)} = \begin{cases} \frac{h(y) g(t; \theta)}{g(t; \theta)} & t = t(y) \\ 0 & \text{otherwise} \end{cases}$$

### Minimal Sufficiency:

$t(\cdot)$  is sufficient and  $\forall \tilde{t}(\cdot) \exists h(\cdot)$  s.t.  $t(y) = h(\tilde{t}(y))$

# 4 Statistical Hypothesis Testing

## 4.1 Significance and Confidence Intervals

### Significance Test

Assuming two states  $H_0$  and  $H_1$  and two corresponding decisions “ $H_0$ ” and “ $H_1$ ”, a decision rule (a threshold  $c \in \mathbb{R}$  for the test statistic  $T(X)$ ) is constructed s. t.:

$$p = P(\text{“}H_1\text{”} | H_0) \leq \alpha$$

	“ $H_0$ ”	“ $H_1$ ”
$H_0$	$1 - p$ (correct)	$p$ (type I error)
$H_1$	$\beta$ (type II error)	$1 - \beta$ (correct)

**Power** concerns the type II error

$$power = P(\text{“}H_1\text{”} | H_1) = 1 - \beta$$

**p-Value** measures the amount of evidence against  $H_0$

$$p \leq \alpha \Leftrightarrow \text{“}H_0\text{”}$$

The  $p$ -value is uniformly distributed on  $[0, 1]$  under  $H_0$ .

### Confidence Interval

$$[t_l(Y_{1:n}), t_r(Y_{1:n})] \text{ Confidence Interval}$$

$$\Leftrightarrow$$

$$P_\theta(t_l(Y_{1:n}) \leq \theta \leq t_r(Y_{1:n})) \geq 1 - \alpha \forall \theta$$

with  $1 - \alpha$  confidence level und  $\alpha$  significance level

### Corresponding Test

$$\theta_0 \notin [t_l(y_{1:n}), t_r(y_{1:n})] \Leftrightarrow \text{“}H_1\text{”}$$

**Specificity** or True Negative Rate (1 – empirical type I error)

$$TNR = \frac{\#TN}{\#N} = \frac{\#TN}{\#TN + \#FP}$$

**Sensitivity** or True Positive Rate, Recall (empirical power)

$$TPR = \frac{\#TP}{\#P} = \frac{\#TP}{\#TP + \#FN}$$

## 4.2 Tests for One Sample

**Normal Distribution**  $X_i \stackrel{iid}{\sim} N(\mu, \sigma^2)$

**Test for  $\mu$ , known  $\sigma^2$  (Simple Gauss-Test)**

$H_0: \mu = \mu_0$  vs.  $H_1: \mu \neq \mu_0$

$$T(X) = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \stackrel{H_0}{\sim} N(0, 1)$$

**Test for  $\mu$ , unknown  $\sigma^2$  (Simple t-Test)**

$H_0: \mu = \mu_0$  vs.  $H_1: \mu \neq \mu_0$

$$T(X) = \frac{\bar{X} - \mu_0}{\hat{\sigma}/\sqrt{n}} \stackrel{H_0}{\sim} t_{n-1}$$

with  $\hat{\sigma} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$

**Binomial Distribution**  $X_i \stackrel{iid}{\sim} Bin(1, p)$

**Approximate test for  $p$**

$H_0: p = p_0$  vs.  $H_1: p \neq p_0$

$$T(X) = \frac{\frac{X}{n} - p_0}{\sqrt{p_0(1-p_0)}/\sqrt{n}} \stackrel{H_0, appr.}{\sim} N(0, 1)$$

**ML Estimate**  $\hat{\theta} \stackrel{a}{\sim} N(\theta, I^{-1}(\theta))$

**Wald Test**

$H_0: \theta = \theta_0$  vs.  $H_1: \theta \neq \theta_0$

$$T(X) = |\hat{\theta} - \theta_0| \stackrel{H_0}{\sim} N(0, I^{-1}(\theta_0))$$

As  $\hat{\theta}$  converges to  $\theta_0$  under  $H_0$ , it can also be used to calculate the variance:  $I^{-1}(\hat{\theta})$ .

**Score Test**

$H_0: \theta = \theta_0$  vs.  $H_1: \theta \neq \theta_0$

$$T(X) = |s(\theta_0; y)| \stackrel{H_0}{\sim} N(0, I(\theta_0))$$

Advantage compared to the Wald Test:  $\hat{\theta}$  does not have to be calculated.

**Likelihood Ratio Test**

$H_0: \theta = \theta_0$  vs.  $H_1: \theta \neq \theta_0$

$$T(X) = 2(l(\hat{\theta}) - l(\theta_0)) \stackrel{H_0}{\sim} \chi_1^2$$

**Neyman-Pearson Test**

$H_0: \theta = \theta_0$  vs.  $H_1: \theta = \theta_1$

$$T(X) = l(\theta_0) - l(\theta_1)$$

For a given significance level  $\alpha$ , the Neyman Pearson Test is the most powerful test for comparing two estimates for  $\theta$ .

Proof:

Decision rule of the NP-Test:  $\varphi^* = \begin{cases} 1 & \text{if } \frac{f(y; \theta_0)}{f(y; \theta_1)} \leq e^c \\ 0 & \text{otherwise} \end{cases}$

Need to show:  $P(\varphi(Y)=1|\theta_1) \leq P(\varphi^*(Y)=1|\theta_1) \forall \varphi$

$$\begin{aligned} P(\varphi^*=1|\theta_1) - P(\varphi=1|\theta_1) &= \\ &= \int \{\varphi^*(y) - \varphi(y)\} f(y; \theta_1) dy \\ &\geq \frac{1}{e^c} \int_{\varphi^*=1} \{\varphi^*(y) - \varphi(y)\} f(y; \theta_0) dy \quad f(y; \theta_1) \geq \frac{f(y; \theta_0)}{e^c} \\ &+ \frac{1}{e^c} \int_{\varphi^*=0} \{\varphi^*(y) - \varphi(y)\} f(y; \theta_0) dy \quad f(y; \theta_1) \leq \frac{f(y; \theta_0)}{e^c} \\ &= \frac{1}{e^c} \int \{\varphi^*(y) - \varphi(y)\} f(y; \theta_0) dy = 0 \end{aligned}$$

As  $\alpha = \int \varphi^*(y) f(y; \theta_0) dy = \int \varphi(y) f(y; \theta_0) dy$

### 4.3 Tests for Goodness of Fit

**Discrete (Chi-Squared)**

$H_0: X_i \sim F_0$  vs.  $H_1: X_i \sim F \neq F_0$

$$T(X) = \sum_{k=1}^K \frac{(n_k - l_k)^2}{l_k} \stackrel{H_0}{\sim} \chi_{K-1-p}^2$$

with the following contingency table:

	1	2	...	K
observed	$n_1$	$n_2$	...	$n_K$
expected under $H_0$	$l_1$	$l_2$	...	$l_K$

$l_k > 5$  and  $l_k > n - 5$  for the  $\chi_{K-1-p}^2$ -distribution to hold,  $F_0$  needs to be known, but its  $p$  parameters can be estimated. The test can be applied to discretized continuous variables.

**Continuous (Kolmogorov-Smirnov Test)**

$H_0: X_i \sim F_0$  vs.  $H_1: X_i \sim F \neq F_0$

$$T(X) = \sup_x |F_n(x) - F_0(x; \theta)| \stackrel{H_0}{\sim} KS$$

with the distribution function  $F(x; \theta)$  and the empirical counterpart  $F_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{X_i \leq x\}}$

Proof:

$$\begin{aligned} P(\sup_x |F_n(x) - F(x; \theta)| \leq t) &= \\ &= P(\sup_y |F^{-1}(y; \theta) - x| \leq t) \quad \begin{matrix} x \in [0, 1], x = F^{-1}(y; \theta) \\ F(F^{-1}(y; \theta); \theta) = y \end{matrix} \\ &\stackrel{*}{=} P(\sup_y |\frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{U_i \leq y\}} - y| \leq t) \quad \text{with } U_i \sim U(0, 1) \\ *F_n(F^{-1}(y; \theta)) &= \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{X_i \leq F^{-1}(y; \theta)\}} = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{F(y; \theta) \leq y\}} \end{aligned}$$

For an estimated parameter the distribution of  $T(X)$  is not independent of  $F_0$ :  $T(X) \stackrel{H_0}{\sim} KS$  only holds asymptotically.

**Pivotal Statistic**

$g(Y; \theta)$  pivotal

$\Leftrightarrow$

distribution of  $g(Y; \theta)$  independent of  $\theta$

**Approximative Pivotal Statistic**

$H_0: X_i \sim F$  pivotal vs.  $H_1: X_i \sim F$  not pivotal

$$g(\hat{\theta}; \theta) = \frac{\hat{\theta} - \theta}{\sqrt{\text{Var}(\hat{\theta})}} \stackrel{H_0}{\sim} N(0, 1)$$

with  $\hat{\theta} = t(Y) \stackrel{H_0}{\sim} N(\theta, \text{Var}(\hat{\theta}))$

$$KI = \left[ \hat{\theta} - z_{1-\frac{\alpha}{2}} \sqrt{\text{Var}(\hat{\theta})}, \hat{\theta} + z_{1-\frac{\alpha}{2}} \sqrt{\text{Var}(\hat{\theta})} \right]$$

Proof:

$$1 - \alpha \approx P\left(z_{\frac{\alpha}{2}} \leq \frac{\hat{\theta} - \theta}{\sqrt{\text{Var}(\hat{\theta})}} \leq z_{1-\frac{\alpha}{2}}\right)$$

### 4.4 Multiple Tests

**Family-Wise Error Rate (FWER)** as  $p \sim U(0, 1)$

For  $m$  tests:

$$\alpha \leq P\left(\bigcup_{k=1}^m (p_k \leq \alpha) | H_{0k}, k = 1, \dots, m\right) \leq m\alpha$$

$$FWER := P(\exists k : "H_{1k}" | \forall k : H_{0k})$$

**Bonferroni Adjustment**

$$\alpha_B = \frac{\alpha}{m}$$

**Šidák Adjustment** only for independent tests

$$\alpha_S = 1 - (1 - \alpha)^{1/m}$$

Proof:

$$\begin{aligned}\alpha &\stackrel{!}{=} P(\cup_{k=1}^m (p_k \leq \alpha) | H_{0k}, k = 1, \dots, m) \\ &= 1 - (1 - \alpha)^{1/m}\end{aligned}$$

**Holm's Procedure** also takes power into account

Order the p-values:  $p_{(0)} \leq \dots \leq p_{(m)}$

Step  $x \in \{0, \dots, m\}$ : if  $p_{(x)} > \frac{\alpha}{m-x}$  reject  $H_{01}$  to  $H_{0x}$  and stop, else move on to step  $x + 1$ .

**False Discovery Rate (FDR)** balances type I and II errors, especially for  $n \ll m$  problems

$$FDR = E\left(\frac{\#\text{"H1"} | H_0}{\#\text{"H1"}}\right)$$

Order the p-values:  $p_{(1)} \leq \dots \leq p_{(m)}$ , choose  $\alpha \in (0, 1)$

$j$  is largest index s. t.  $p_{(j)} \leq \alpha j/m$ , reject all  $H_{0i}$  for  $i \leq j$

It can be shown that  $FDR \leq m_0 \alpha / m$ , with  $m_0 = \#H_0$



# 5 Regression

## 5.1 Models

### 5.1.1 Simple Linear Model

#### Theoretical Model

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

#### Empirical Model

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

#### Assumptions

- **Independent Observations**  $y_1, \dots, y_n$  are independent
- **Linearity of the Mean**  $E(Y|x) = \beta_0 + \beta_1 x$  or  $E(e|x) = 0$
- **Constant Variation**  $Var(Y|x) = \sigma^2$

For the normal linear model:

- **Normality**  $e|x \sim N(0, \sigma^2)$  ;  $Y|x \sim N(\hat{y}, \sigma^2)$

#### Attributes of the Regression Line

$$\begin{aligned}\hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_i = \bar{y} + \hat{\beta}_1 (x_i - \bar{x}) \\ \hat{e}_i &= y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \\ &= y_i - (\bar{y} + \hat{\beta}_1 (x_i - \bar{x})) \\ \sum_{i=1}^n \hat{e}_i &= \sum_{i=1}^n y_i - \sum_{i=1}^n \bar{y} - \hat{\beta}_1 \sum_{i=1}^n (x_i - \bar{x}) \\ &= n\bar{y} - n\bar{y} - \hat{\beta}_1 (n\bar{x} - n\bar{x}) = 0 \\ \bar{\hat{y}} &= \frac{1}{n} \sum_{i=1}^n \hat{y}_i = \frac{1}{n} (n\bar{y} + \hat{\beta}_1 (n\bar{x} - n\bar{x})) = \bar{y}\end{aligned}$$

### 5.1.2 Multivariate Linear Model

#### Theoretical Model

$$Y = X\beta + u$$

#### Empirical Model

$$Y = X\hat{\beta} + e$$

$$\hat{Y} = X\hat{\beta}$$

$$y = (y_1, \dots, y_n)^T, e = (e_1, \dots, e_n)^T, X = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix}$$

#### Assumptions

- **Independent Observations**  $y_1, \dots, y_n$  are independent
- **Linearity of the Mean**  $E(Y|x_{1:p}) = X\beta$  or  $E(e|x_{1:p}) = 0$
- **Constant Variation**  $Var(Y|x) = \sigma^2$

#### Estimates (OLS)

$$\hat{\beta}_1 = \frac{Cov(x, y)}{Var(x)} = \frac{S_{xy}}{S_{xx}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \cdot \sqrt{\frac{S_{yy}}{S_{xx}}} = r \sqrt{\frac{S_{yy}}{S_{xx}}}$$

Proof:

$$Cov(x, y) = Cov(x, \hat{\beta}_0 + \hat{\beta}_1 \bar{x}) = \hat{\beta}_1 Var(x) \iff \hat{\beta}_1 = \frac{Cov(x, y)}{Var(x)}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Proof:

$$E[y] = E[\hat{\beta}_0 + \hat{\beta}_1 x + \hat{e}] \iff \hat{\beta}_0 = E[y] - \hat{\beta}_1 E[x]$$

The estimates are the same as for the ML procedure.

#### Estimates (ML) $Y|x \sim N(\beta_0 + \beta_1 x, \sigma^2)$

$$\begin{aligned}\hat{\beta}_0 &= \frac{1}{n} \sum_{i=1}^n y_i - \frac{1}{n} \sum_{i=1}^n x_i \hat{\beta}_1 \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^n x_i (y_i - \hat{\beta}_0)}{\sum_{i=1}^n x_i^2} \\ \hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\beta}_0 - x_i \hat{\beta}_1)^2\end{aligned}$$

The  $\beta$ -estimates are the same as for the OLS procedure.

Proof:

$$l(\beta_0, \beta_1, \sigma^2) = \sum_{i=1}^n \left\{ -\frac{1}{2} \log \sigma^2 - \frac{1}{2} \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{\sigma^2} \right\}$$

For the normal linear model:

- **Normality**  $e_i|x_{1:p} \sim N(0, \sigma^2)$  ;  $Y|x \sim N(\hat{y}, \sigma^2)$

#### Estimates (ML) $Y|x_{1:p} \sim N(X\beta, \sigma^2)$

$$\begin{aligned}\hat{\beta} &= (X^T X)^{-1} X^T y \\ Var(\hat{\beta}) &= \sigma^2 (X^T X)^{-1} = I^{-1}(\beta)\end{aligned}$$

Proof:

$$l(\beta, \sigma^2) = -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} (y - X\beta)^T (y - X\beta)$$

The estimates are the same as for the OLS procedure.

$\hat{\beta}$  is the **Best Linear Unbiased Estimator**

Proof:

Unbiased because of the Gauß-Markov Theorem:  $E(\hat{\beta}) = (X^T X)^{-1} X^T E(Y|X) = (X^T X)^{-1} X^T X \beta = \beta$

$$\hat{\sigma}^2 = \frac{1}{n} (y - X\hat{\beta})^T (y - X\hat{\beta}); \quad \hat{\beta} \sim N(\beta, \sigma^2 (X^T X)^{-1})$$

The ML-estimate for  $\sigma^2$  is biased.

Proof:

$H := X(X^T X)^{-1} X^T$  hat matrix;  $HH = H = H^T$  (idempotent)

$$\begin{aligned} E((Y - X\hat{\beta})^T (Y - X\hat{\beta})) &= E((Y^T (I_n - H))^T ((I_n - H)Y)) \\ &= E(\text{tr}(Y^T (I_n - H)Y)) \\ &= E(\text{tr}((I_n - H)YY^T)) \\ &= \text{tr}((I_n - H)E(YY^T)) \\ &= \text{tr}((I_n - H)E(X\beta\beta^T X^T + \sigma I_n)) \\ &= \sigma^2 \text{tr}((I_n - H)) \\ &= \sigma^2 (n - p) \end{aligned}$$

$$s^2 = \frac{1}{n - p} (y - X\hat{\beta})^T (y - X\hat{\beta}); \quad \hat{\beta} \sim t_{n-p}(\beta, s^2 (X^T X)^{-1})$$

with  $s$  an unbiased estimator

### 5.1.3 Bayesian Linear Model

**Prior** flat prior

$$f_{\beta, \sigma^2}(\beta, \sigma^2) = \frac{1}{\sigma^2}$$

**Posterior**

Resulting posterior:

$$f_{\text{post}}(\beta, \sigma^2 | y) \propto (\sigma^2)^{-\frac{n}{2} + 1} e^{-\frac{1}{2\sigma^2} (y - X\beta)^T (y - X\beta)}$$

Note:  $f_{\text{post}}(\beta, \sigma^2 | y) = f(\beta | \sigma^2, y) f(\sigma^2 | y)$

$$\begin{aligned} \beta | \sigma^2, y &\sim N(\hat{\beta}, \sigma^2 (X^T X)^{-1}) \\ \sigma^2 | y &\sim \text{IG}\left(\frac{n - p}{2}, \frac{s^2 (n - p)}{2}\right) \\ \beta | y &\sim t_{n-p}(\hat{\beta}, s^2 (X^T X)^{-1}) \end{aligned}$$

The two distributions for  $\beta$  mirror the results for  $\hat{\beta}$  in the linear model.

### 5.1.4 Quantile Regression

**Prediction Interval** range of  $1 - \alpha$  fraction of the data

$$\text{Var}(\hat{Y} | x_{1:p}) = \text{Var}(X\hat{\beta}) + \sigma^2$$

Determined by estimation variance (usually captured by confidence intervals) plus residual variance.

**Quantile**

$$Q(\tau) = \inf\{y : F(y) \geq \tau\}$$

If  $F$  is invertible:  $Q(\tau) = F^{-1}(\tau)$ ,  $\tau \in (0, 1)$

**Model**

$$Q(\tau | x_{1:p}) = X\beta$$

For median regression:  $\hat{\beta} = \arg \min \sum_{i=1}^n |y_i - x_i^T \beta|$

In general:

$$\hat{Q}(\tau) = \arg \min_{\beta} \left( \sum_{i=1}^n \delta_{\tau}(y_i - x_i^T \beta) \right)$$

with check function  $\delta_{\tau}(y) = y(\tau - \mathbb{1}_{\{y < 0\}})$

Proof:

$$Q(\tau) = \arg \min_q E(\delta_{\tau}(Y - q))$$

$$= \arg \min_q \left\{ (\tau - 1) \int_{-\infty}^q (y - q) f(y) dy + \tau \int_q^{\infty} (y - q) f(y) dy \right\}$$

Differentiating w.r.t.  $q$  gives  $(\tau - 1) \int_{-\infty}^q f(y) dy - \tau \int_q^{\infty} f(y) dy = (1 - \tau)F(q) - \tau(1 - F(q)) = F(q) - \tau$

**Estimates**

The estimates for  $\beta$  can be computed with linear programming and are normally distributed with mean  $\beta$ .

### 5.1.5 Flexible Regression

**Assumptions**

- **Independent Observations**  $y_1, \dots, y_n$  are independent
- **Constant Variation**  $\text{Var}(Y|x) = \sigma^2$
- **Normality**  $e_i | x_{1:p} \sim N(0, \sigma^2)$ ;  $Y|x \sim N(\hat{y}, \sigma^2)$

**Knot Placement**

- equidistant
- based on quantiles (more structure where data is dense)
- all data points plus penalization

### Penalized Regression Splines

$$\|y - X\beta\|^2 + \lambda \int_{x_1}^{x_n} [f''(x)]^2 dx = \|y - X\beta\|^2 + \lambda \beta^T D \beta$$

$$l_p(\beta, \sigma^2, \lambda) = l(\beta, \sigma^2) - \frac{\lambda}{2\sigma^2} \beta^T D \beta$$

$$\hat{\beta} = (X^T X + \lambda D)^{-1} X^T y$$

### Difference Penalty

- first order:  $\beta^T D \beta = \sum_{j=1}^p (\beta_{j+1} - \beta_j)^2$
- second order:  $\beta^T D \beta = \sum_{j=1}^p (\beta_{j+1} - 2\beta_j + \beta_{j-1})^2$

**Choosing  $\lambda$**  Model complexity

## 5.1.6 Generalized Regression

### Assumptions

- **Independent Observations**  $y_1, \dots, y_n$  are independent
- **Linearity of the Mean**  $E(Y|x_{1:p}) = X\beta$  or  $E(e|x_{1:p}) = 0$
- **Exponential Family**  $Y|x \sim \exp\{t(y)\theta(x) - \kappa(\theta(x))\} h(y)$

### Link Function

Linear predictor  $\eta = X\beta$ ;  $\mu = \frac{\partial \kappa(\theta)}{\partial \theta} = E(t(Y); \theta)$

$$\mu = g^{-1}(\eta)$$

If  $\lambda = 0$ , *canonical link*:

$$\theta = \eta$$

## 5.1.7 Weighted Regression

**Different Precision** variance heterogeneity:  $e_i \sim N(0, \sigma_i^2)$

$$l(\beta, \sigma^2) = -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} (y - X\beta)^T W (y - X\beta)$$

with  $W = \text{diag}(\frac{1}{a_1}, \dots, \frac{1}{a_n})$  and  $a_i = \frac{\sigma_i^2}{\sigma^2}$

$$\hat{\beta}_{ML} = (X^T W X)^{-1} (X^T W y)$$

## 5.2 Goodness of Fit

### 5.2.1 Coefficient of Determination

$$R^2 = \frac{SS_{Explained}}{SS_{Total}} = 1 - \frac{SS_{Residual}}{SS_{Total}} = r^2$$

$$\dim(\lambda) = \text{tr} \left\{ (X^T X + \lambda D)^{-1} (X^T X) \right\}$$

$$AIC(\lambda) = \text{fit}(\lambda) + 2\dim(\lambda)$$

Numerically complex. Alternative: **Bayes**

$\beta \sim N(0, \sigma_\beta^2 D^-)$  with  $(D^-)^- = D$  (generalized inverse)

$$\log f(\beta, \sigma^2; \sigma_\beta^2 | y) \propto l(\beta, \sigma^2) - \frac{rk(D^-)}{2} \log(\sigma_\beta^2) - \frac{1}{2\sigma_\beta^2} \beta^T D^- \beta$$

As  $\lambda = \frac{1}{\sigma_\beta^2}$ , marginal posterior for  $\sigma_\beta^2$  can be derived. E.g. set  $\lambda$  to the posterior mode estimate.

- score function:  $s(\beta) = X^T (t(y) - E(t(Y); \eta))$

- estimate  $\hat{\beta} = X^T E(t(Y); \hat{\eta}) = X^T t(y)$

- Fisher matrix  $I(\beta) = X^T W X$   
with  $W$  diagonal and  $W_{ii} = \frac{\partial^2 \kappa(\eta_i)}{\partial \eta^2} = \text{Var}(t(Y_i), \eta_i)$

Examples:

- **Logistic**:  $\text{logit} P(Y_i=1|x_i) = \log \frac{P(Y_i=1|x_i)}{1-P(Y_i=1|x_i)} = \eta$   
 $\text{Var}(Y_i|x_i) = P(Y_i=1|x_i) \cdot (1-P(Y_i=1|x_i))$

- **Poisson**:  $\log E(Y_i|x_i) = \eta$   
 $\text{Var}(Y_i|x_i) = E(Y_i|x_i) = e^\eta$

$$\text{Var}(\hat{\beta}_{ML}) = \sigma^2 (X^T W X)^{-1}$$

### Different Group Representation

$$Y_i | x_{i,1:p}, z_i \sim N(x_{i,1:p} \beta_{z_i}, \sigma^2)$$

with  $z_i$  indicating group affiliation

Range:  $0 \leq R^2 \leq 1$

# 6 Survival Analysis

## 6.1 Basics

**Time-to-Event Data** outcome tuple  $(t_i, \delta_i)$  is observed with time  $0 < t < \infty$  and event/censoring indicator  $\delta_i$

typically:  $\delta_i = \begin{cases} 1, & i \text{ experienced event} \\ 0, & i \text{ is (right-)censored} \end{cases}$

**Censoring**  $T_i \sim F$  times to event &  $C_i \sim G$  censoring times

- **right-censoring:** observe  $t_i = \min(T_i, C_i)$
- **left-censoring:** observe  $t_i = \max(T_i, C_i)$
- **interval-censoring:** only known  $t_i \in [C_i^l, C_i^r]$

Censoring and event distributions need to be independent given explanatory variables for models to be unbiased

**Truncation** a form of sampling bias

- **left-truncation:** observations with  $t_i < a$  excluded
- **right-truncation:** observations with  $t_i > a$  excluded

**Description of the Distribution**

- **survival function**

$$S(t) := P(T > t) = 1 - F(t)$$

- **hazard rate**

$$h(t) := \lim_{\delta \rightarrow 0} \frac{P(T \in [t, t + \delta] | T \geq t)}{\delta} = \frac{f(t)}{S(t)}$$

- **cumulative hazard rate**

$$H(t) := \int_0^t h(u) du = -\log(S(t))$$

## 6.2 Modelling

**Kaplan-Meier** non-parametric estimate for  $S(t)$

For ordered event times  $t_{(1)}, \dots, t_{(f)}$ :

$$\hat{S}(t_{(f)}) = \prod_{i=1}^f \hat{P}(T > t_{(i)} | T \geq t_{(i)})$$

$\hat{P}(T > t_{(i)} | T \geq t_{(i)}) = 1 - \frac{d_i}{n_i}$ , with  $d_i$  number of events at  $t_{(i)}$  and  $n_i$  number of observations under risk at  $t_{(i)}$

- also called product-limit estimator
- formula closely related to path rule
- assumes piecewise constant survival (step function)

**Greenwoods' Formula** basis for confidence intervals

$$\widehat{Var}(\hat{S}(t)) = \hat{S}(t)^2 \sum_{t_{(k)} \leq t} \frac{d_k}{n_k(n_k - d_k)}$$

**Nelson-Aalen** non-parametric estimate for  $H(t)$

$$H(t) = \sum_{k: t_{(k)} \leq t} \hat{h}_k^d$$

with  $\hat{h}_k^d$  discrete time hazard rate for  $k$ -th interval  $[t_{(k-1)}, t_{(k)}]$

- results in piecewise constant estimate (step function)
- $\exp(-\hat{H}(t))$  gives Breslow estimate for  $S(t)$

**Accelerated Failure Time** parametric modelling of  $T$

$T \sim \text{WB}(\lambda, k)$ : leads to proportional hazards

$T \sim \text{Exp}(\lambda) = \text{WB}(\lambda, 1)$ : results in constant hazard rate

Location-Scale-Model (with  $\lambda = \exp(-\eta)$  and  $k = \frac{1}{\sigma}$ ):

$$\log(T) = x^T \beta + \sigma \epsilon \iff$$

$$T = \exp(\beta_0) \exp(\beta_1 x_1) \dots \exp(\beta_p x_p) \exp(\sigma \epsilon)$$

incrementing  $x_1$  means survival time stretches by  $\exp(\beta_1)$ , c. p.

**Diagnostics**

- $\ln(-\ln S(t)) = \ln(\lambda) + \sigma \ln(t)$ , use  $\hat{S}_{KM}$  for graphical check (for single categorical  $x$  only)

- Distribution assumption (for Weibull):  $\epsilon \sim \text{Gumbel}\left(\frac{y-\eta}{\sigma}\right)$

**ML estimation** under right-censoring

$$L(\theta) = \prod_{i=1}^n (f(t_i | \theta, x_i))^{\delta_i} (S(t_i | \theta, x_i))^{1-\delta_i}$$

standard likelihood theory applies

**Cox Proportional Hazards** semi-parametric

$$h(t, x) = h_0(t) \cdot \exp(x^T \beta)$$

with non-parametric baseline hazard rate  $h_0(t)$  independent of  $x$ , but can depend on (part of)  $x$  in stratified models

**Proportional Hazards** incrementing  $x_1$  means hazard rate multiplies by  $\exp(\beta_1)$ , ceteris paribus

Proof: 
$$\frac{h(t|x^{(a)})}{h(t|x^{(b)})} = \exp((x^{(a)} - x^{(b)})^T \beta)$$

**Diagnostics**

- **Proportional Hazard:** Schoenfeld residuals ( $E(r_i) = 0$ ) and corresponding test ( $H_0$ : proportional hazards)
- **Overall Fit:** Cox Snell residuals  $\sim \text{Exp}(1)$
- **Functional Form:** Martingale residuals ( $E(r_i) = 0$ )
- **Outliers:** Deviance residuals ( $E(r_i) = 0$ )
- **Leverage:** Jack-Knife residuals

**Partial Likelihood** consider risk sets  $R_i = R(t_{(i)})$  for ordered event times  $t_{(1)}, \dots, t_{(f)}$ :

$$\begin{aligned} PL(\beta) &= \prod_{i=1}^m P(t_i = t_{(i)} | \exists j \in R_i : t_j = t_{(i)}) \\ &= \prod_{i=1}^m \frac{\exp(x_{(i)}^T \beta)}{\sum_{j \in R_i} \exp(x_j^T \beta)} \end{aligned}$$

**Random Survival Forests** Machine Learning method

1. draw  $B$  Bootstrap samples from the data
2. grow binary decision tree for each sample finding splits with log rank tests out of  $p$  randomly drawn splitting candidates

3. stopping: minimum  $d_0 > 0$  deaths per node

4. calculate  $\hat{H}(t)$  via Nelson-Aalen and average for samples

5. get prediction error using Out-Of-Bag samples

## 6.3 Model Comparison

**General** Some measures (e. g. AIC) require equal likelihood estimation and are typically only valid within a model class.

Other methods (presented below) should be used on test data.

**C-Index** compare predictions if censor times allow it

$$C = \frac{\text{\#concordant pairs}}{\text{\#comparable pairs}}$$

with  $0 \leq C \leq 1$  and  $C = 0.5$  for random guessing

**Brier Score** equivalent to the MSE for predicted probabilities

$$\widehat{BS}(t, \hat{S}) = \frac{1}{n} \sum_{i=1}^n (\tilde{Y}_i(t) - \hat{S}(t|x_i))^2$$

with  $\tilde{Y}_i(t) = \mathbb{1}(t_i > t)$  and  $t_i = \min(T_i, C_i)$

the arithmetic mean above is often weighted using Inverse Probability of Censoring Weights

**Integrated Brier Score** as the Brier Score depends on  $t$

$$IBS(\tau) = \frac{1}{\tau} \int_0^\tau \widehat{BS}(u, \hat{S}) du$$

can also be shown graphically as the Brier Score over different time points in so-called prediction error curves (PEC)

# 7 Spatial Statistics

## 7.1 Markov Random Field

**Lattice Data** observations from a random process observed over spatial regions, supplemented by a neighborhood structure

**Random Field** synonymous to stochastic process

$$(X_1, \dots, X_n)$$

realisations of random variable  $X_i$  for locations  $D = \{1, \dots, n\}$

**Markov Properties** extend the temporal  $X_{t+1}|X_t \perp X_1, \dots, t-1$

- **pairwise:**  $X_i|X_{-i} \perp X_j$  for  $i \neq j$  and  $i \not\sim j$
- **local:**  $X_i|X_{-\{i, n \in \partial(i)\}} \perp X_{n \in \partial(i)}$  with  $\partial(i) := \{j : j \sim i\}$
- **global:**  $X_A|X_B \perp X_C$  for  $C$  separating  $A$  and  $B$

global implies local implies pairwise; equivalent in a GMRF

**Neighbourhood System** set of neighbourhoods

$$\partial = \{\partial(s) : s \in D\}$$

if  $s \notin \partial(s)$  (antireflexive) and  $v \in \partial(s) \Leftrightarrow s \in \partial(v)$  (symmetric)

**GMRF** Gauss Markov Random Field

$$f(x) = (2\pi)^{-\frac{n}{2}} |P|^{\frac{1}{2}} \exp\left(-\frac{1}{2}(x - \mu)^T P(x - \mu)\right)$$

with precision matrix  $P := \Sigma^{-1} > 0$  and  $p_{ij} \neq 0 \Leftrightarrow i \sim j$

$p_{ij}$  define neighbourhoods with  $X_i|X_{-i} \perp X_j \Leftrightarrow p_{ij}=0$  for  $i \neq j$

GMRFs fulfill the pairwise, local, and global Markov property

**canonical representation:**

$$X \sim N^k(m, P) := N(P^{-1}m, P^{-1})$$

$$\Leftrightarrow f(x) \propto |P|^{\frac{1}{2}} \exp\left(-\frac{1}{2}x^T P x + m^T x\right)$$

get  $m$  via completing the square

**improper GMRFs:**  $P$  s.p.d. with  $\text{rk}(P) = r < n$

$$f(x) = (2\pi)^{-\frac{r}{2}} \det^*(P)^{\frac{1}{2}} \exp\left(-\frac{1}{2}(x - \mu)^T P(x - \mu)\right)$$

with  $\det^*(P) = \prod_{j=1}^r \lambda_j$  with  $\lambda_1, \dots, \lambda_r$  non-zero eigenvalues

**intrinsic** GMRFs ( $\sum_{j=1}^n p_{ij} = 0 \ \forall i = 1, \dots, n$ ) with

irregular grids: rank decline is number of independent areas

**models:**

- **Conditional Autoregressive (CAR):**  
 $E(X_i|X_{-i}) = \mu_i + \sum_{j:i \sim j} c_{ij}(X_j - \mu_j)$ , resulting in a  
 GMRF with  $p_{ij} = \begin{cases} -c_{ij}/(\sigma_i \sigma_j), & i \neq j \\ 1/\sigma_i, & i = j \end{cases}$   
 Weights  $c_{ij}$  are symmetric and often correspond to distance
- **Intrinsic Autoregressive (ICAR)**  
 $p_{ij} = \begin{cases} -1/(\sigma_i \sigma_j), & i \sim j \\ |\partial(i)|/\sigma_i, & i = j \\ 0, & \text{else} \end{cases}$

- **Simultaneous Autoregressive (SAR)**

$$E(X_i|X_{-i}, Z) = Z_i \beta + \sum_{j:i \sim j} b_{ij} X_j \text{ with covariates } Z$$

**GMRF Regression** Bayesian prior equals likelihood penalty

$$Y = X\beta + Z\eta + \epsilon$$

with  $\epsilon_i \sim N(0, \sigma^2)$ ;  $\eta \sim N(0, (\tau Q)^{-1})$ ;  $\beta_k \sim N(0, \kappa_k)$

get hyperparameters: cross validation, empirical Bayes, or Bayes Gamma is non-informative, semi-conjugated prior on  $\tau$ ,  $\frac{1}{\sigma^2}$ , &  $\frac{1}{\kappa_k}$

**Drawing from a GMRF** Rue algorithm

1. perform Cholesky decomposition:  $P = LL'$
2. calculate  $w$  and  $u$  with  $Lw = m$  and  $L'u = w$
3. draw  $z \sim N(0, I_n)$  and calculate  $v$  with  $L'v = z$
4. calculate result  $y = u + v$

Proof:

$$u = P^{-1}m, \text{ as } m = Lw = LL'u = Pu$$

$$E(v) = E(L'^{-1}z) = L'^{-1}E(z) = 0$$

$$\text{Var}(v) = \text{Var}(L'^{-1}z) = (L'^{-1})^2 \text{Var}(z) = (L''L')^{-1}I = P^{-1}$$

Therefore  $y = u + v \sim N(mP^{-1}, P^{-1})$ .

**latent GMRF** solution if target is not Gaussian

hierarchical model with  $\tau$  as smoothing parameter:

$$Y_i|\eta_i \stackrel{iid}{\sim} F(\eta_i) \text{ e.g. Poisson for count data}$$

$$\eta \sim \text{GMRF}(\tau Q)$$

**Inference** with MCMC, with a penalized Likelihood, or with INLA (fully Bayesian using approximations)

**Ising Model**  $X_i \sim B(1, \pi_i)$ ,  $x_i \in \{-1, 1\}$ , e.g. black/white

Minimize:

$$H(f, x) = -c \sum_{i \in I} f_i x_i - \beta \sum_{i \sim j} f_i f_j - h \sum_{i \in I} f_i$$

- $-\sum_{i \in I} f_i x_i \propto$  Hamming distance
- $\sum_{i \sim j} f_i f_j$  penalty for differing neighbors
- $\sum_{i \in I} f_i$  to achieve a balance between values (optional)

**Bayesian View**

- $-\sum_{i \in I} f_i x_i = -l(f|x)$  for multiplicative noise ( $x_i = f_i \eta_i$ , with  $\eta_i$  flip indicator), the rest of the formula is the prior
- **Perfect Sampling:** Do MCMC from all possible start values. Combine paths with same status. Once all paths are combined, distribution is independent of starting value.

**Potts Model** as extension to categorical data

## 7.2 Geostatistics

**Spatially Continuous Random Field** random variables  $Y_S$

$$Y_S : (\mathbb{R}^d, \mathcal{B}^d, \mathbb{P}) \rightarrow (Z, \mathcal{F})$$

i. e. probability space  $\rightarrow$  status space (including  $\sigma$ -Algebra)

**Variogram** typically plot over distance  $|s - s'|$

$$\gamma(s, s') := \frac{1}{2} \text{Var}(Y(s) - Y(s'))$$

**Covariogram**  $c(s, s') := \text{Cov}(Y(s), Y(s'))$  (= covariance)

It holds:  $\gamma(s, s') = \frac{1}{2} \text{Var}(Y(s)) + \frac{1}{2} \text{Var}(Y(s')) - c(s, s')$

**Gauss Random Field**  $Y = \{Y(s), s \in D \subseteq \mathbb{R}^d\}$

$\forall s_1, \dots, s_n : (Y(s_1), \dots, Y(s_n))^T$  multivariate normal

uniquely defined by  $\mu(s)$  and  $c(s, s')$

**Stationarity** strong  $\Rightarrow$  weak  $\Rightarrow$  intrinsic

**intrinsic**  $E(Y(s)) = \text{const.}$  and  $\gamma(s, s') = \gamma(\Delta s)$

**weak**  $E(Y(s)) = \text{const.}$  and  $c(s, s') = c(\Delta s) \forall s, s' \in S$

**strong** all  $n$ -dimensional distributions translation invariant

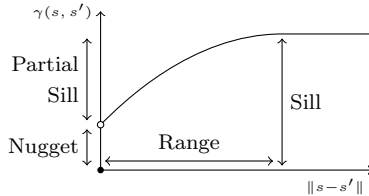
For stationary processes: **Correlogram**  $\rho(\Delta s) := \frac{c(\Delta s)}{\sigma^2}$ ,  
 $c(0) = \sigma^2 = \text{Var}(Y(s))$  nugget effect, and  $\gamma, c, \rho$  symmetric;  
 $\gamma(\Delta s) = c(0) - c(\Delta s) = \sigma^2(1 - \rho(\Delta s))$

**Isotropy** compare directional variograms

$$c(s, s') = c(\|s - s'\|) \Leftrightarrow Y \text{ isotrop}$$

- **geometric anisotropy** same sill, different range
- **zonal anisotropy** different sill

**Empirical Variogram** estimated via kernel functions and nearest neighbours (similar to gliding histograms)



**Parametric Models for Correlograms**

- **Exponential**  $\rho(h) = \exp(-(h/\phi))$
- **Gauss**  $\rho(h) = \exp(-(h/\phi)^2)$
- **Spherical**  $\rho(h) = \begin{cases} 1 - \frac{2}{3}(h/\phi) + \frac{1}{2}(h/\phi)^3, & 0 \leq h \leq \phi \\ 0, & h \geq \phi \end{cases}$
- **Matérn**  $\frac{1}{2\kappa - 1 \Gamma(\kappa)} (h/\phi)^\kappa K_\kappa(h/\phi)$  with  $K_\kappa$  Bessel function

## 7.3 Point Processes

Hi

**Simple Kriging** assumes known expectation and  $c(s, s')$

$$Y(s_i) = Z(s_i) + \epsilon(s_i)$$

with random field  $Z$  with  $\mu(s)$  and  $c(s, s') = \sigma^2 \rho(s, s')$

and  $\epsilon(s_i)$  i.i.d. with expectation 0 and variance  $\sigma_\epsilon^2$

minimizing MSPE  $\min \sigma_Y^2(s_0) = c(s_0, s_0) - c(s_0)' C_Y^{-1} c(s_0)$ :

$$\hat{Z}(s_0) = \mu(s_0) + c(s_0)' C_Y^{-1} (Y - \mu_Y)$$

**ordinary** constant unknown mean  $\mu(s) = \mu$

**universal** known beta in  $\mu(s) = x(s)' \beta$

**Gauss Model Kriging** same estimates as for simple kriging

Equivalent Approaches:

additive:  $Y(s) = \mu(s) + Z(s) + \epsilon(s)$ , with  $Z(s)$  as  $\text{GRF}(0, \tau^2)$

hierarchical:  $Y(s)|Z(s), \sigma^2 \sim N(Z(s), \sigma^2)$ ,  $Z(s)$  as  $\text{GRF}(\mu(s), \tau^2)$

**Inference:**

- ML: via profile-log-likelihood, but biased
- REML: via restricted (or marginal) likelihood, less biased
- fully Bayesian: MCMC but how to choose priors

**Kriging Procedure**

- |                                  |                              |
|----------------------------------|------------------------------|
| 1. variogram                     | 5. Kriging                   |
| 2. fit correlation function      | 6. residual variogram        |
| 3. test stationarity/isotropy    | 7. geoadditive regression    |
| 4. if so, transform for isotropy | 8. maybe generalized Kriging |

**Geoadditive Regression**

$$Y_i = x_i \beta + \sum_{j=1}^p f_j(z_{ij}) + f_{geo}(s_i) + \epsilon_i$$

Modelling  $f_{geo}(s) = z_{geo} \gamma_{geo}$  with  $\gamma_{geo} \sim N(0, (\tau_{geo} K_{geo})^{-1})$ :

- **lattice data:**  $f_{geo} = \phi_K$  for  $s_i = K$  and GMRF prior
- **geostatistical data:**  $f_{geo} = Z \gamma_{geo}$  using radial basis functions or 2D-splines

$$Y = X\beta + \sum_{j=1}^{p, geo} Z_j \gamma_j + \epsilon \quad \text{with } \gamma_j \sim N(0, (\tau_j K_j)^{-1})$$

with inference via REML, cross validation or fully Bayesian

**INLA** (Integrated Nested Laplace Approximation):

2 Laplace approximations for fully Bayesian procedure

AIC is impossible for spatial, but DIC works

# 8 Bayesian Statistics

## 8.1 Basics

### Bayes Theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad \text{für } P(A), P(B) > 0$$

or more general:

$$\begin{aligned} f_{\text{post}}(\theta|X) &= \frac{f(X|\theta) \cdot f_{\theta}(\theta)}{\int f(X|\tilde{\theta})f_{\theta}(\tilde{\theta})d\tilde{\theta}} \\ &= C \cdot f(X|\theta) \cdot f_{\theta}(\theta) \quad \text{choose } C \text{ s.t. } \int f_{\text{post}}(\theta|X) = 1 \\ &\propto f(X|\theta) \cdot f_{\theta}(\theta) \end{aligned}$$

### Point Estimates

$$\hat{\theta}_{\text{postmean}} = E_0(\vartheta|x) = \int_{\vartheta \in \Theta} \vartheta f_{\theta}(\vartheta|x)d\vartheta$$

$$\hat{\theta}_{\text{postmode}} = \underset{\vartheta}{\operatorname{argmax}} f_{\theta}(\vartheta, x)$$

$$\hat{\theta}_{\text{Bayesrisk}} = \underset{t(\cdot)}{\operatorname{argmin}} R_{\text{Bayes}}(t(\cdot))$$

with Bayes risk:  $R_{\text{Bayes}}(t(\cdot)) = \int_{\Theta} R(t(\cdot), \vartheta) f_{\theta}(\vartheta) d\vartheta$

$$\hat{\theta}_{\text{postBayesrisk}} = \underset{t(\cdot)}{\operatorname{argmin}} R_{\text{postBayes}}(t(\cdot)|y)$$

with posterior Bayes risk:

$$R_{\text{postBayes}}(t(\cdot)|y) = \int \mathcal{L}(t(y), \vartheta) f_{\text{post}}(\vartheta|y) d\vartheta = E_{\theta|y}(\mathcal{L}(t(y), \theta)|y)$$

For squared loss:  $\hat{\theta}_{\text{postBayesrisk}} = \hat{\theta}_{\text{postmean}}$

### Credibility Interval

$$P_{\theta}(\theta \in [t_l(y), t_r(y)] | y) = \int_{t_l(y)}^{t_r(y)} f_{\text{post}}(\vartheta|y) d\vartheta \stackrel{!}{=} 1 - \alpha$$

- symmetric:  $\int_{-\infty}^{t_l(y)} f_{\text{post}}(\vartheta|y) d\vartheta = \int_{t_r(y)}^{\infty} f_{\text{post}}(\vartheta|y) d\vartheta = \frac{\alpha}{2}$
- highest density:  $HDI = \{\theta : f_{\text{post}}(\theta|y) \geq c\}$ , choose  $c$  s.t.  $\int_{\vartheta \in HDI(y)} f_{\text{post}}(\vartheta|y) d\vartheta = 1 - \alpha$

**Bayes Factor** evidence contained in data for  $M_1$  vs.  $M_2$

$$\frac{P(M_1|y)}{P(M_0|y)} = \underbrace{\frac{f(y|M_1)}{f(y|M_0)}}_{\text{Bayes Factor}} \frac{P(M_1)}{P(M_0)}$$

with marginal likelihood  $f(y|M_i) = \int f(y|\vartheta) f_{\theta}(\vartheta|M_i) d\vartheta$

## Priors

**Flat (uninformative) Prior**

$f_{\theta}(\theta) = \text{const.}$  for  $\theta > 0$ , therefore:  $f(\theta|X) = C \cdot f(X|\theta)$

As  $\int f_{\theta}(\theta) = 1$  not possible like this, this is not a real density.

Changes for transformations of the parameter.

$$\text{Proof: For } \gamma = g(\theta): f_{\gamma}(\gamma) = f_{\theta}(g^{-1}(\gamma)) \left| \frac{\partial g^{-1}(\gamma)}{\partial \gamma} \right|$$

No prior is truly uninformative.

**Jeffrey's Prior** transformation-invariant

For Fisher-regular distributions:  $f(\theta) \propto \sqrt{I_{\theta}(\theta)}$

Proof:

For  $\gamma = g(\theta)$  and  $f_{\theta}(\theta) = \sqrt{I_{\theta}(\theta)}$ :

$$f_{\gamma}(\gamma) \propto f_{\theta}(g^{-1}(\gamma)) \left| \frac{\partial g^{-1}(\gamma)}{\partial \gamma} \right| \propto \sqrt{\frac{\partial g^{-1}(\gamma)}{\partial \gamma} I_{\theta}(g^{-1}(\gamma)) \frac{\partial g^{-1}(\gamma)}{\partial \gamma}} = \sqrt{I_{\gamma}(\gamma)}$$

Maximizes the information gained from the data (under appropriate regulatory conditions), i. e. maximizes

$$E(KL(f_{\theta}(\cdot), f_{\text{post}}(\cdot, x)))$$

### Empirical Bayes

Let the prior depend on a hyper-parameter:  $f_{\theta}(\theta, \gamma)$

Choose  $\gamma$  s.t.  $L(\gamma) = \int f(x; \gamma) = \int f(x; \vartheta) f_{\theta}(\vartheta, \gamma) d\vartheta$  is maximal.

Using the data to find the prior contradicts the Bayes approach of incorporating prior knowledge.

### Hierarchical Prior

$$x|\theta \sim f(x; \theta); \quad \theta|\gamma \sim f_{\theta}(\theta, \gamma); \quad \gamma \sim f_{\gamma}(\gamma)$$

### Conjugate Priors

If Prior and Posterior belong to the same family of distributions for a given likelihood function, they are called conjugate.

Examples:

Prior	Likelihood	Posterior
$\pi \sim \text{Be}(\alpha, \beta)$	$\text{Bin}(n, \pi)$	$\text{Be}(\alpha+k, \beta+n-k)$
$\mu \sim \text{N}(\gamma, \tau^2)$	$\text{N}(\mu, \sigma^2)$	$\text{N}(\cdot, \cdot) \xrightarrow{n \rightarrow \infty} \text{N}(\bar{y}, \frac{\sigma^2}{n})$
$\sigma^2 \sim \text{IG}(\alpha, \beta)$	$\text{N}(\mu, \sigma^2)$	$\text{IG}(\alpha + \frac{n}{2}, \beta + \frac{1}{2} \sum_{i=1}^n (y_i - \mu)^2)$
$\lambda \sim \text{Ga}(\alpha, \beta)$	$\text{Po}(\lambda)$	$\text{Ga}(\alpha+n\bar{y}, \beta+n)$

## 8.2 Numerical Methods for the Posterior

**Numerical Integration** here: trapezoid approximation

$$\begin{aligned} &\int_{\Theta} f(y|\vartheta) f_{\theta}(\vartheta) d\vartheta \approx \\ &\sum_{k=1}^K \frac{f(y; \theta_k) f_{\theta}(\theta_k) + f(y; \theta_{k-1}) f_{\theta}(\theta_{k-1})}{2} (\theta_k - \theta_{k-1}) \end{aligned}$$

only normalisation constant unknown, works well for one-dimensional integrals

### Laplace Approximation

$$\int_{\Theta} f(y|\vartheta) f_{\theta}(\vartheta) d\vartheta \approx f(y; \hat{\theta}_P) f_{\theta}(\hat{\theta}_P) (2\pi)^{p/2} \left| J_P(\hat{\theta}_P) \right|^{\frac{1}{2}}$$

with the one-dimensional  $J_P := -\frac{\partial^2 l_{(n)}(\theta, y)}{\partial \theta^2} - \frac{\partial^2 \log f_{\theta}(\theta)}{\partial \theta^2}$  Fisher information considering the prior,  $\hat{\theta}_P$  posterior mode estimate s.t.  $s_{P, \theta}(\hat{\theta}_P) = 0$



Proof:

For  $n$  independent samples:

$$f_{post}(\theta|y) = \frac{\prod_{i=1}^n f(y_i|\theta)f_{\theta}(\theta)}{\int \prod_{i=1}^n f(y_i|\theta)f_{\theta}(\theta)d\theta}$$

Denominator:  $\int e^{\{\sum_{i=1}^n \log f(y_i|\theta) + \log f_{\theta}(\theta)\}} d\theta =$

$$\int e^{\{l(\theta;y) + \log f_{\theta}(\theta)\}} d\theta \approx \int e^{(l_P(\hat{\theta}_P) - \frac{1}{2} J_P(\hat{\theta}_P)(\vartheta - \hat{\theta}_P)^2)} d\vartheta$$

Resembles the normal distribution, therefore the inverse of the normalisation constant can be calculated, which gives the inverse of the Laplace approximation in the univariate case.

Works well for large  $n$  and is numerically simple also for big  $p$ .

### Monte Carlo Approximations

The denominator can be written as  $E_{\theta}(f(y;\theta)) =$

$\int_{\Theta} f(y|\vartheta)f_{\theta}(\vartheta)d\vartheta$ , which can be estimated by the arithmetic mean for a sample of  $\theta_1, \dots, \theta_N$ , which needs to be drawn from the prior. The following methods to draw from non-standard distributions can be used for that.

#### • Inverse CDF

$F(X)$  known. Since  $F(x) = u$ ,  $F^{-1}(u) = x$ ,  $u \sim U(0, 1)$

1. Draw  $u \sim U(0, 1)$
2. Compute  $F^{-1}(u)$  to get a value  $x$

Proof:

$$P(x \leq y) = P(F^{-1}(u) \leq y) = P(u \leq F(y)) = F(y)$$

#### • Rejection Sampling

An umbrella distribution  $g(x)$  can be found s. t.

$$\frac{f(x)}{g(x)} \leq M \quad \forall x \text{ with } f(x) > 0 \text{ when } g(x) > 0$$

1. Draw candidate  $y \sim g(x)$
2. Acceptance probability  $\alpha$  for  $y$ :  $\alpha = \frac{f(y)}{Mg(y)}$
3. Draw  $u \sim U(0, 1)$  and accept if  $u \leq \alpha$ , else: step 1

Proof:

$$\begin{aligned} P\left(Y \leq x | U \leq \frac{f(Y)}{Mg(Y)}\right) &= \frac{P\left(Y \leq x, U \leq \frac{f(Y)}{Mg(Y)}\right)}{P\left(U \leq \frac{f(Y)}{Mg(Y)}\right)} \\ &= \frac{\int_{-\infty}^x \int_0^{\frac{f(y)}{Mg(y)}} du g(y) dy}{\int_{-\infty}^{\infty} \int_0^{\frac{f(y)}{Mg(y)}} du g(y) dy} = \frac{\int_{-\infty}^x \frac{f(y)}{g(y)} g(y) dy}{\int_{-\infty}^{\infty} \frac{f(y)}{g(y)} g(y) dy} \\ &= \frac{\int_{-\infty}^x f(y) dy}{\int_{-\infty}^{\infty} f(y) dy} = P(X \leq x) \end{aligned}$$

### • Importance Sampling

Directly estimate  $E_{\theta}(f(y;\theta))$ .

For sampling distribution  $g(x)$ ,

$$\frac{1}{N} \sum_{i=1}^N \frac{f(x)}{g(x)}$$

is a consistent estimator.

Proof:

$$E_g\left(\frac{1}{N} \sum_{i=1}^N \frac{f(x)}{g(x)}\right) = \int \frac{f(x)}{g(x)} g(x) dx = \int f(x) dx = f(x)$$

**Markov Chain Monte Carlo** sample from  $f_{post}(\theta|X)$

$f(y)$  unknown, however:

$$\frac{f_{post}(\theta|x)}{f_{post}(\tilde{\theta}|x)} = \frac{f(x|\theta)f_{\theta}(\theta)}{f(y)} \frac{f(y)}{f(x|\tilde{\theta})f_{\theta}(\tilde{\theta})} = \frac{f(x|\theta)f_{\theta}(\theta)}{f(x|\tilde{\theta})f_{\theta}(\tilde{\theta})}$$

**Metropolis-Hastings:** Draw Markov Chain  $\theta_1^*, \dots, \theta_n^*$ :

1. Draw candidate  $\theta^*$  from proposal distribution  $q(\theta|\theta_{(t)}^*)$
2. Accept  $\theta_{(t+1)}^* = \theta^*$  with probability

$$\alpha(\theta_{(t)}^*|\theta^*) = \min\left\{1, \frac{f_{post}(\theta^*|y) q(\theta_{(t)}^*|\theta^*)}{f_{post}(\theta_{(t)}^*|y) q(\theta^*|\theta_{(t)}^*)}\right\}$$

else choose  $\theta_{(t+1)}^* = \theta_{(t)}^*$

This sequence has a stationary distribution for  $n \rightarrow \infty$ .

Choice of  $q$ : trade-off between exploring  $\Theta$  and reaching a high  $\alpha$ .

Burn-in and thinning out give *i.i.d.* samples from  $f_{post}(\theta|X)$ .

**Gibbs Sampling:** For high dimensions  $\alpha$  is close to zero.

Sample from the marginal distributions separately:

$$\theta_{t+1,i}^* \sim f_{\theta_i|y,\theta_{\setminus i}}(\theta_i^*|y, \theta_{t^*,i}^*)$$

with  $\theta_{t^*,i}$  most recent estimates without  $\theta_i$

A Gibbs sampled sequence converges to  $f_{post}(\theta|X)$  as stationary.

Can also be used on its own, if marginal densities are known.

### Variational Bayes Principles

Approximate  $f_{post}(\theta|X)$  by  $q_{\theta} = \min_{q_{\theta} \in Q} KL(f_{post}(\cdot|X), q_{\theta}(\cdot))$

Restrict  $q_{\theta}$  to independence:  $q_{\theta}(\theta) = \prod_{k=1}^p q_k(\theta_k)$

Update each component iteratively. Works well for big  $p$ .

# 9 Sampling

## Bootstrap

1. Draw  $y_i^*$ :  $n$  samples with replacement from  $y$
2. Calculate the statistic of interest  $t(y_i^*)$
3. Repeat this  $B$  times
4. *Plug-in Principle*: Whenever the distribution function is involved in estimating a statistic, use the empirical version  $\hat{F}_n(y) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{(y_i \leq y)}$  instead.

In a **Parametric Bootstrap** the parameter is first estimated from the data and then Bootstrap samples are drawn from the resulting distribution.

## Bootstrap Probability

$$P(Y_i \in Y^*) = 1 - (1 - \frac{1}{n})^n \xrightarrow{n \rightarrow \infty} 1 - e^{-1} \approx 0.632$$

## Subsampling

- **replacement**  $m$ -out-of- $n$  bootstrap
- **non-replacement** subsampling directly from true  $F$

## Permutation Test for two variables

1. Calculate  $t(x, y)$ , e.g. differences in mean, correlation...
2. Draw samples  $x^*, y^*$  of size  $n$  from  $x$  and  $y$  without replacement ("shuffel")
3. Calculate  $t(x^*, y^*)$
4.  $p\text{-value} = \frac{1}{B} \sum_{b=1}^B \mathbb{1}_{\{t(x_b^*, y_b^*) \geq t(x, y)\}}$

For a **Bootstrap Test** do step 2 with replacement.

## Bootstrap in Regression

- **Residual based**: 1. Get Bootstrap sample  $e_i^*$  from fitted residuals  $\hat{e} = y - X\hat{\beta}$ , 2. Calculate new response  $y_i^* = x_i\hat{\beta} + e_i^*$ , 3. Calculate  $\hat{\beta}^*$
- **Model based** 1. Draw a sample from  $e_i \sim N(0, \hat{\sigma}_{ML}^2)$ , 2. Calculate new response  $y_i^* = x_i\hat{\beta} + e_i^*$ , 3. Calculate  $\hat{\beta}^*$
- **Pairwise** 1. Draw  $(y_i^*, x_i^*)$  from the original sample for  $i = 1, \dots, n$ , 2. Calculate  $\hat{\beta}^*$
- **Wild Set**  $\hat{e}_i^* = V_i^* \hat{e}_i$ , with  $V_i^*$  from the 2-point distribution  $P(V_i^* = \frac{\sqrt{5}+1}{2}) = \frac{\sqrt{5}-1}{2\sqrt{5}}$  and  $P(V_i^* = -\frac{\sqrt{5}-1}{2}) = \frac{\sqrt{5}+1}{2\sqrt{5}}$ , chosen as  $E(V_i^*) = 0$ ,  $Var(V_i^*) = 1$ ,  $E(V_i^{*3}) = 1$

## Consistency of a Bootstrap Estimator

$$\lim_{n \rightarrow \infty} P_n \left\{ \sup_t |G_n(t, F_n) - G_\infty(t, F)| > \epsilon \right\} = 0 \quad \forall \epsilon$$

with  $F_n(y) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{Y_i \leq y\}}$  empirical distribution function,  $G_n(t, F) = P(T_n \leq t)$  exact finite sample distribution of  $T_n = t(Y_1, \dots, Y_n)$ , and  $P_n$  joint probability of the sample

The bootstrap estimate is inconsistent for the maximum of a sample or if the  $\theta$  lies on the boundary of  $\Theta$ .

## Mallow's Metric

$$\rho_p(F, G) = \inf_{T_{XY}} \{E(|X - Y|)^p\}^{\frac{1}{p}}$$

for  $F, G$  in the set of distributions where  $\int_{-\infty}^{\infty} |t|^p dF(t) < \infty$ ;  $(X, Y) \sim T \in \mathcal{T}_{XY}$  with  $X \sim F$  and  $Y \sim G$

## Theorem of Beran and Ducharme

$G_n(\cdot, F_n)$  is consistent if  $\forall \epsilon > 0, F$  the following holds:

1.  $\lim_{n \rightarrow \infty} P_n(\rho(F_n, F) > \epsilon) = 0$
2.  $G_\infty(t, F)$  is a continuous function of  $t$
3.  $\forall t$  and sequences  $\{H_n\}$  s.t.  $\lim_{n \rightarrow \infty} \rho(H_n, F) = 0$  holds:  $G_n(t, H_n) \rightarrow G_\infty(t, F)$

# 10 Model Selection

## AIC (Akaike Information Criterion)

$$AIC = -2 \sum_{i=1}^n \log f(y_i; \hat{\theta}) + 2p$$

The AIC estimates  $2E_Y \{KL(g, f)\} - 2 \int \log(g(y))g(y)dy$ . The latter component is unknown, so the absolute value of the AIC is not informative. The AIC favours complex models.

For regressions:  $AIC = 2n \log(\hat{\sigma}^2) + 2(p+2)$

## The AIC as theoretical cross validation

The AIC minimizes  $E_{Y_{1:n}} \{E_Y [Y - \hat{\mu}]^2\}$  if we use the MSE instead of the Kullback-Leibler divergence. This can be estimated via cross validation.

## Bias Corrected AIC

$$AIC_{corr} = -2 \sum_{i=1}^n \log f(y_i; \hat{\theta}) + 2p \left( \frac{n}{n-p-1} \right)$$

should be preferred if  $\frac{n}{p} < 40$

## BIC (Bayesian Information Criterion)

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

approximately maximizes the posterior probability of a model and selects less complex models as the AIC

## DIC (Deviance Information Criterion) Bayesian AIC

$$DIC = D(y, \hat{\theta}_{postmean}) + 2p_D = \int D(y, \vartheta) f_{post}(\vartheta|y) d\vartheta + p_D$$

with deviance  $D(y; \theta) := -2l(\theta)$  the difference in likelihood compared to the full model and  $\Delta D(y; \theta, \hat{\theta}) = 2 \{l(\hat{\theta}) - l(\theta)\} \stackrel{a}{\sim} \chi_p^2$  the difference in deviance

$$p_D := E(\Delta D(y; \theta, \hat{\theta}_{postmean} | y)) = \int D(y, \vartheta) f_{post}(\vartheta|y) d\vartheta - D(y, \hat{\theta}_{postmean})$$

The integral can be approximated using MCMC.

**Model Averaging** Using probabilities as weights

$$P(M_k|y) := \frac{\exp(-\frac{1}{2}\Delta IC_k)}{\sum_{k'=1}^K \exp(-\frac{1}{2}\Delta IC_{k'})}$$

with  $\Delta IC_k = IC_k - \min(IC)$

For regressions:  $P(\text{covariate } x|y) = \sum_{k=1}^K \mathbb{1}_{\{x \text{ in } M_k\}} P(M_k|y)$

**Inference After Model Selection** neglect is a quiet scandal

$$Var(\hat{\theta}) = E_{model}(Var(\hat{\theta}|model)) + Var_{model}(E(\hat{\theta}|model))$$

$$= \sum_{k=1}^K \pi_k Var_k(\hat{\theta}) + \sum_{k=1}^K \pi_k (\theta_k - \bar{\theta})^2$$

The last component depends on the true parameter and will be biased if the estimates are used.

Solutions:

$$\bullet \widehat{Var}(\hat{\theta}) = \left[ \sum_{k=1}^K \pi_k \sqrt{\widehat{Var}_k(\hat{\theta}_k) + (\hat{\theta}_k - \bar{\hat{\theta}})^2} \right]^2$$

• Use the Variance of the full (saturated) model

• Use bootstrap for confidence intervals

**Lasso** least absolute shrinkage and selection operator

$$l_p(\theta, \lambda) = l(\theta) - \lambda \sum_{j=1}^p |\theta_j|$$

This penalized log likelihood can be solved with iterative quadratic programming using a Taylor expansion.

Using Bayesian view the penalty corresponds to a prior:

$$f_{\theta_j}(\theta_j) \propto \exp(-|\theta_j|) \text{ (Laplace prior)}$$

# 11 Dimensionality Reduction

## Covariance Matrix $\Sigma$

- symmetric,  $\in \mathbb{R}^{n \times n}$  therefore  $\frac{q(q+1)}{2}$  parameters
- positive definite, i.e.  $\forall a \in \mathbb{R}^q : a^T \Sigma a \geq 0$

## Marginal Independence

$$\Sigma_{jk} = 0 \Leftrightarrow Y_{ij} \text{ and } Y_{ik} \text{ are independent}$$

## Conditional Independence

$$\Omega = \Sigma_{jk}^{-1} = 0 \Leftrightarrow Y_{ij} \text{ and } Y_{ik} \text{ are independent given all other } Y \text{ with concentration matrix } \Omega$$

Proof:

$$f(y_{.j}, y_{.k} | y_{\overline{j,k}}) = \frac{f(y)}{f_{\overline{j,k}}} \propto f(y)^{N(\mu, \Sigma)} \exp \left\{ -\frac{1}{2} y^T \Sigma^{-1} y \right\}$$

## Graphical Models

visualize conditional dependences in a graph

## Principal Component Analysis (PCA)

1. Use singular value decomposition  $\Sigma = U \Lambda U^T$  with  $U$  matrix of orthonormal eigenvectors and  $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_q)$  matrix of sorted eigenvalues  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_q$  of  $\Sigma$
2. Prune smallest  $k = q - r$  eigenvalues in  $\tilde{\Lambda}$
3. Simplify model with spectral decomposition  $\tilde{Y} = \tilde{V} \tilde{\Lambda}^{1/2} \tilde{U}^T$  with  $\tilde{V}$ ,  $\tilde{U}$  first  $r$  eigenvectors of  $Y Y^T$  and  $Y^T Y$  respectively
4. explained variance  $\sum_{i=1}^r \lambda_i / \sum_{i=1}^q \lambda_i$

Proof:

Karhunen-Loève expansion:  $U \Lambda^{\frac{1}{2}} Z_{\cdot} \sim N(0, U \Lambda^{\frac{1}{2}} \Lambda^{\frac{1}{2}} U^T) = N(0, \Sigma)$  with  $Z_{\cdot} \sim N(0, \mathbb{1})$ , therefore  $\tilde{Y}_{\cdot} = \tilde{V} \tilde{\Lambda}^{\frac{1}{2}} \tilde{Z}_{\cdot}$

With spectral decomposition:  $Y = V \Lambda^{\frac{1}{2}} U^T$  (for column-centered  $Y$ )

## 12 Missing/Deficient Data

**Missing Completely at Random (MCAR)** independent

$$P(R_i|Y_i) = P(R_i)$$

with  $R_{ij} = \begin{cases} 0 & \text{if } Y_{ij} \text{ missing} \\ 1 & \text{otherwise} \end{cases}$  and  $R_i = (R_{i1}, \dots, R_{iq})$

A complete case analysis will lead to unbiased results.

**Missing at Random (MAR)** depends on observed variables

$$P(R_i|Y_i) = P(R_i|Y_{iO_i})$$

with  $O_i = \{j : R_{ij} = 1\}$  and  $M_i = \{j : R_{ij} = 0\}$

Complete case analysis  $P(Y|X, Z)$ :

- only response  $Y_i$  MAR: unbiased
- only covariate  $X_i$  MAR: biased  
Asymptotically unbiased with *inverse probability weighting*:
  1. Estimate  $\pi(y_i, z_i) = P(R_{X_i}=1|y_i, z_i)$
  2. Use weighted score function  $\hat{s}_w(\theta) = \sum_{i=1}^n \frac{R_{X_i}}{\pi} s_i(\theta)$
- both MAR: biased and  $\pi(y_i, z_i)$  can not be estimated due to missing  $Y_i$

**Missing Not at Random (MNAR)**

$$P(R_i|Y_i) \neq P(R_i|Y_{iO_i})$$

Can not be corrected to be unbiased.

**EM Algorithm** replace  $y_{iM}$  by  $E(Y_{iM}|y_{iO})$

Expectation Step:

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

Maximization Step:

$$\frac{\partial Q(\theta, \theta_{(t)})}{\partial \theta} = s(\theta, \theta_{(t)}) \stackrel{!}{=} 0$$

Iterate over E and M step until convergence

**Louis' Formula for Variance Estimates in EM Settings**

$$J_O(\theta) = \sum_{i=1}^n \{E(J_i(\theta)|y_{iO}) - E(s_i(\theta)s_i(\theta)|y_{iO}) + s_{iO}(\theta)s_{iO}(\theta)\}$$

**Multiple Imputation** EM but considers estimation variability

1. Create  $K$  complete datasets by simulating missing data  
 $\sim f_{post}(y_{iM}|y_{iO})$
2. Fit  $K$  models  $Y_i \sim f(y|\theta)$
3. Rubin's Rule:  $\hat{\theta}_{MI} = \frac{1}{K} \sum_{k=1}^K \hat{\theta}_{(k)}^*$ ;  
 $\widehat{\text{Var}}(\hat{\theta}_{MI}) = \hat{V} + (1 + \frac{1}{K})\bar{B}$  with  $\hat{V} = \frac{1}{K} \sum_{k=1}^K I^{-1}(\hat{\theta}_{(k)}^*)$  and  
 $\bar{B} = \frac{1}{K-1} \sum_{k=1}^K (\hat{\theta}_{(k)}^* - \hat{\theta}_{MI})(\hat{\theta}_{(k)}^* - \hat{\theta}_{MI})^T$

**Estimate Accuracy**

$$\hat{\mu}_g - \mu_g = \rho_{R_g} \times \sigma_g \times \sqrt{\frac{N-n}{n}}$$

with  $\rho_{R_g}$  data quality (correlation between  $R_j$  and  $g(Y_j)$ ),  $\sigma_g$  variability, and  $\sqrt{\frac{N-n}{n}}$  data quantity;  $g$  some known function

- MCAR:  $MSE(\hat{\mu}_g) = \frac{1}{N-1} \times \sigma_g^2 \times \frac{N-n}{n}$
- MNAR:  $MSE(\hat{\mu}_g) = E(\rho_{R_g}^2) \times \sigma_g^2 \times \frac{N-n}{n}$   
 $n_{eff} = \frac{\frac{n}{N}}{1 - \frac{n}{N} E(\rho_{R_g}^2)}$

**Measurement Error**

$$U = X - X_m \text{ with } E(U) = \mu_U \text{ and } \text{Var}(U) = \sigma_U^2$$

with  $\mu_U$  systematic error (bias/validity),  $\text{Var}(U)$  (reliability)

In Regression Settings:

- **error in Y:**  $Y_m = \beta_0 + \beta_1 X + \epsilon + U$  and  
 $E(Y_m|X) = \beta_0 + \mu_U + \beta_1 X$  leads to biased  $\hat{\beta}_0$
- **error in X:**  $Y = \beta_0 + \beta_1 X + \epsilon$  and  $X_m = X + U$  leads to biased  $\hat{\beta}_0$  and  $\hat{\beta}_1$ , the latter is attenuated by the inverse of reliability ratio  $rr = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_U^2} = \frac{\sigma_{X_m}^2 - \sigma_U^2}{\sigma_{X_m}^2}$   
Getting information about  $\sigma_U^2$ :

- **Validation Data** with both  $X$  and  $X_m$  observed
- **Replication Data** repeated measures of  $X_m$
- **Assumptions** e.g.  $\sigma_U^2 = 0$  (naive estimator)

# 13 Experiment Design

## Omitted Variables

Regression setting ignoring omitted Variables:

$$\int f_{Y|X,Z,U} f_{Z,U} dz du = \int \frac{f_{Y,X,Z,U}}{f_{X|Z,U}} dz du \neq f_{Y|X}$$

with  $Z$  observable and  $U$  unobservable quantities influencing  $Y$   
Solutions:

- *Randomization*: randomly assign  $X$  and then observe  $Y$
- *Balancing*: make  $X$  independent of  $Z$

## Analysis of Variances (ANOVA) of one categorical variable

Linear constraint:  $\sum_{k=1}^K n_k \beta_k = 0$  (usually controlled over  $\beta_K$ )

$$\hat{\beta}_{0,ML} = \hat{\mu}_{ML} = \sum_{k=1}^K \sum_{j=1}^{n_k} \frac{y_{kj}}{n} = \bar{y}_{..}$$

$$\hat{\beta}_{k,ML} = \sum_{j=1}^{n_k} \frac{y_{kj} - \bar{y}_{..}}{n_k} = \bar{y}_{k.} - \bar{y}_{..}$$

$$SS_{Total} = SS_{Explained} + SS_{Residual}$$

with

$$SS_{Total} = \sum_{i=1}^n (y_i - \bar{y}_{..})^2$$

$$SS_{Explained} = \sum_{i=1}^n (\hat{y}_i - \bar{y}_{..})^2$$

$$SS_{Residual} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n e_i^2 = S_{yy} - \hat{\beta}^2 S_{xx}$$

## F-Test

$$F = \frac{SS_{Explained}/(df_0 - df_X)}{SS_{Residual}/df_X} \sim \mathcal{F}_{df_0 - df_X, df_X}$$

with  $df_0 = n-1$  and  $df_X = n-K$

## Block Design account for block effects

$$Y_{kbj} = \mu + \beta_k + \alpha_k + \epsilon_{kbj}$$

Linear constraints:  $\sum_{k=1}^K n_k \beta_k = 0$  and  $\sum_{b=1}^B n_{.b} \alpha_b = 0$

For the F-Test:  $df_0 = df_Z = n-B$  and  $df_{X+Z} = n-K-B+1$

## Latin Squares Sudoku pattern for more variables

## Instrumental Variable

$$Y = \beta_0 + X\beta_X + \overbrace{U\beta_U}^{\tilde{\epsilon}} + \epsilon \Rightarrow \frac{\partial Y}{\partial X} = \beta_X + \frac{\partial \tilde{\epsilon}}{\partial X}$$

Construct instrumental variable  $Z$ :  $Cov(Z, \epsilon) = 0$ ,  $Cov(Z, X) \neq 0$ :

$$\frac{\partial Y}{\partial X} |_{(U=u)} = \frac{\partial Y / \partial Z}{\partial X / \partial Z}$$

i. e. fit two regressions  $Y|Z$  and  $X|Z$  and set  $\hat{\beta}_X = \frac{\hat{\beta}_{YZ}}{\hat{\beta}_{XZ}}$

## Propensity Score

$$\tau = E(Y(1)|D=1) - E(Y(0)|D=1)$$

with  $\tau$  average treatment effect on the treated,  $Y(1)$  response if treated,  $Y(0)$  analogous;  $D_i$  indicator if  $i$  is influenced by the treatment

$$E(Y(1)|D=1) - E(Y(0)|D=0) = \tau + \underbrace{E(Y(0)|D=1) - E(Y(0)|D=0)}_{\text{selection bias}}$$

If the selection bias is zero,  $D$  and  $X$  are unconfounded.

$$\hat{\tau} = \sum_{i=1}^n (Y_i(1) - Y_{j(i)}(0))$$

after matching treated individual  $i$  with individual  $j(i)$  from non-treatment group

