LIST OF FORMULAS FOR STK2100

(Version May 2026)

1 General issues

(a) Complexity/degrees of freedom

- (i) Assuming $\widehat{\boldsymbol{y}} = \boldsymbol{S}\boldsymbol{y}$, degrees of freedom is defined as $\text{df} = \text{trace}(\boldsymbol{S})$.
- (ii) Selection of df usually through cross-validation

(b) Loss functions

- (i) Quadratic loss: $L(y, \widehat{y}) = (y - \widehat{y})^2$. Typically used for regression. The optimal predictor based in input variable(s) $\boldsymbol{x}$ is then $\widehat{y} = E[Y|\boldsymbol{x}]$.
- (ii) 0-1 loss: $L(y, \widehat{y}) = I(y \neq \widehat{y})$, where $I(\cdot)$ is the indicator function. Typically used for classification. The optimal prediction based on covariates $\boldsymbol{x}$ is then $\widehat{Y} = \text{argmax}_k \Pr(Y = k|\boldsymbol{x})$.

(c) Model selection

- (i) for a model with parameters $\boldsymbol{\theta}$, log-likelihood $l(\boldsymbol{\theta})$ and maximum likelihood estimate $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ based on N observations, where $|\boldsymbol{\theta}|$ is the number of *free* parameters in the model:
  - i. $\text{AIC} = -2l(\widehat{\boldsymbol{\theta}}) + 2|\boldsymbol{\theta}|$
  - ii. $\text{BIC} = -2l(\widehat{\boldsymbol{\theta}}) + \log(N)|\boldsymbol{\theta}|$
- (ii) K-fold cross-validation:
  - i. Divide the N data points into K groups by randomization
  - ii. For $k = 1, ..., K$
    - A. Fit the model on all data *except* data from group k.
    - B. Predict $\widehat{y}_i$ for all data in group k based on the fitted model
    - C. Calculate $\widehat{\text{Err}}_i = L(y_i, \widehat{y}_i)$ for i in group k
  - iii. Calculate $\text{CV}_{(K)} = \frac{1}{N}\sum_{i=1}^{N} \widehat{\text{Err}}_i$
- (iii) Bootstrap:
  - i. Dram B bootstrap samples $(y_1^{*b}, \mathbf{x}_1^{*b}), \ldots, (y_N^{*b}, \mathbf{x}_N^{*b})$ from the data with replacement
  - ii. Fit a $\widehat{f}^{*b}$ to each bootstrap sample
  - iii. The leave-one-out bootstrap estimate of the prtediction error is

$$\widehat{\text{Err}}^{(1)} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|C^{-i}|}\sum_{b \in C^{-i}} L(y_i, \widehat{f}^{*b}(\mathbf{x}_i)),$$

where C^{-i} is the set of indices of the bootstrap samples

iv. The 0.632-estimator is fiven by

$$\widehat{\mathrm{Err}}^{(0.632)} = 0.368\ \overline{\mathrm{err}} + 0.632\ \widehat{\mathrm{Err}}^{(1)}$$

(d) Principal components: Let $\boldsymbol{S}$ be the sample covariance matrix of the covariate vectors $\mathbf{x}_1, \ldots, \mathbf{x}_N$

(i) First component : $\boldsymbol{z}_{i1} = \boldsymbol{a}_1^T \boldsymbol{x}_i$, $i = 1, \ldots, N$, with

$$\boldsymbol{a}_1 = \underset{\boldsymbol{a}}{\operatorname{argmax}}\, \boldsymbol{a}^T \boldsymbol{S} \boldsymbol{a} \quad \text{subject to} \quad \boldsymbol{a}^T \boldsymbol{a} = 1$$

(ii) Fo $k = 2, \ldots, p$: $\boldsymbol{z}_{ik} = \boldsymbol{a}_k^T \boldsymbol{x}_i$, $i = 1, \ldots, N$, with

$$\boldsymbol{a}_k = \underset{\boldsymbol{a}}{\operatorname{argmax}}\, \boldsymbol{a}^T \boldsymbol{S} \boldsymbol{a} \quad \text{subject to} \quad \boldsymbol{a}^T \boldsymbol{a} = 1,\ \boldsymbol{a}^T \boldsymbol{a}_1 = 0, \ldots,\ \boldsymbol{a}^T \boldsymbol{a}_{k-1} = 0$$

2 Linear regression

(a) Model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i\,;\ i = 1, 2, \ldots, N\,;$$

where the covariates x_{ij} have known values and the ϵ_i's are independent and $N(0, \sigma^2)$-distributed.

(b) Matrix form:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\mathbf{Y} = (Y_1, \ldots, Y_N)^T$ and $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^T$ are N- and $(p+1)$-dimenional vectors, respectively. Further, $\mathbf{X} = \{x_{ij}\}$ (with $x_{i1} = 1$) is an $N \times (p+1)$-dimensional matrix.

(c) The least squares estimator for $\boldsymbol{\beta}$ is $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$.

(d) Let $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \ldots, \widehat{\beta}_p)^T$. Then the $\widehat{\beta}_j$'s are normal distributed and unbiased, and

$$\mathrm{Var}(\widehat{\beta}_j) = \sigma^2 c_{jj} \quad \text{and} \quad \mathrm{Cov}(\widehat{\beta}_j, \widehat{\beta}_l) = \sigma^2 c_{jl}$$

where c_{jl} is element (j, l) in the $(p+1) \times (p+1)$ matrix $\mathbf{C} = (\mathbf{X}^T\mathbf{X})^{-1}$.

(e) Let $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \cdots + \widehat{\beta}_p x_{ik}$, and $\mathrm{RSS} = \sum_{i=1}^{N} (Y_i - \widehat{Y}_i)^2$. Then $\widehat{\sigma}^2 = \frac{\mathrm{RSS}}{n-(p+1)}$ is an unbiased estimator for σ^2, and $[N - (p+1)]\widehat{\sigma}^2/\sigma^2 \sim \chi^2_{N-(p+1)}$. Further, $\widehat{\sigma}^2$ and $\widehat{\boldsymbol{\beta}}$ are independent.

(f) Let $\mathrm{SE}(\widehat{\beta}_j)^2$ be the variance estimator for $\widehat{\beta}_j$ that we get by replacing σ^2 with $\widehat{\sigma}^2$ in the formulae for $\mathrm{Var}(\widehat{\beta}_j)$ in point (d). Then $(\widehat{\beta}_j - \beta_j)/\mathrm{SE}(\widehat{\beta}_j) \sim t_{N-(p+1)}$.

(g) We can test the hypothesis

$$H_0 : \beta_{i_1} = \beta_{i_2} = \cdots = \beta_{i_q} = 0$$

by using the test statistic

$$\mathrm{F} = \frac{(\mathrm{RSS}_0 - \mathrm{RSS}_1)/q}{\mathrm{RSS}_1/(N-p-1)} \overset{H_0}{\sim} F_{q,N-p-1},$$

where $\mathrm{RSS}_0 = \sum_{i=1}^N (Y - \tilde{Y}_i)^2$ with $\tilde{Y}_i$ the prediction of Y_i *under* H_0, while $\mathrm{RSS}_1 = \sum_{i=1}^N (Y - \widehat{Y}_i)^2$ with $\widehat{Y}_i$ the prediction of Y_i from the full model.

(h) Ridge/Lasso regression: Minimize with respect to $\boldsymbol{\beta}$

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p |\beta_j|^q$$

with $q = 2$ for Ridge, $q = 1$ for Lasso.

(i) Best subset selection: Select a subset of at most k covariates, for some k, by

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2, \quad \text{subject to} \quad \sum_{j=1}^p I(\beta_j \neq 0) \leq k$$

Usually some greedy algorithm (e.g. forward/backward selection) is applied.

3 General regression methods

(a) General setting: Assume

$$Y_i = f(\boldsymbol{x}_i) + \varepsilon_i$$

where $f(\cdot)$ is some general function while $\varepsilon_i, i = 1, ..., N$ are noise terms with $\mathrm{E}(\varepsilon) = 0$ and $\mathrm{Var}(\varepsilon) = \sigma^2$.

(b) K-nearest neighbour regression:

$$\widehat{f}(\boldsymbol{x}_0) = \frac{1}{K} \sum_{\boldsymbol{x}_i \in \mathcal{N}_k(\boldsymbol{x}_0)} y_i$$

where $\mathcal{N}_k(\boldsymbol{x}_0) \subset \{\boldsymbol{x}_1, ..., \boldsymbol{x}_N\}$ contains the K nearest points to $\boldsymbol{x}_0$ in the training set.

(c) Additive models: $f(\boldsymbol{x}) = \sum_{j=1}^p f_j(x_j)$, with basis expansions $f_j(x) = \sum_{m=1}^{M_j} \beta_{jm} h_{jm}(x)$, for instance

(i) Cubic spline with knots $\xi_1, \ldots, \xi_K$: Piecewise polyomial with basis functions

$$\begin{aligned} h_1(x) =& 1, \quad h_2(x) = x, \quad h_3(x) = x^2, \quad h_4(x) = x^3, \\ h_{3+k}(x) =& (x - \xi_k)_+^3, \quad k = 1, ..., K. \end{aligned}$$

(*ii*) Natural cubic splines: cubic spline with linearity constraint in the two end intervals

(*iii*) Smoothing spline, defined by

$$\underset{f_1,\ldots,f_p}{\operatorname{argmin}} \sum_{i=1}^{N} \left(y_i - \alpha - \sum_{j=1}^{p} f_j(x_{ij}) \right)^2 + \sum_{j=1}^{p} \lambda_j \int f_j''(t_j)^2 dt_j$$

(*d*) Kernel methods/Local polynomial regression with kernel function $K_\lambda(x_0, x_i) = D\left(\frac{|x_i - x_0|}{\lambda}\right)$:

$$f(x_0) = \alpha(x_0) + \beta(x_0)x_0,$$

where $(\alpha(x_0), \beta(x_0))$ is estimated by

$$(\widehat{\alpha}(x_0), \widehat{\beta}(x_0)) = \underset{\alpha(x_0),\beta_j(x_0)}{\operatorname{argmin}} \sum_{i=1}^{N} K_\lambda(x_0, x_i)[y_i - \alpha(x_0) - \sum_{j=1}^{d} \beta_j(x_0)x_i^j]^2$$

(*e*) Neural network with one hidden layer:

$$f(\boldsymbol{x}_i) = g\left(\beta_0 + \sum_{m=1}^{M} \beta_m \sigma(\sum_{j=1}^{p} \alpha_{jm} x_{ij})\right)$$

where σ and g are some chosen (nonlinear) activation and output functions, respectively.

(*f*) Tree-based methods: $f(\boldsymbol{x}) = \sum_{m=1}^{M} c_m I(\boldsymbol{x} \in R_m)$ where the covariate space has been divided into the regions $R_1, \ldots, R_M$ through sequential splitting based on one variable at a time.

(*i*) Boosted trees:

$$f(\mathbf{x}) = \sum_{m=1}^{M} b_m(\mathbf{x}; \gamma_m),$$

where $b_m(\mathbf{x}; \gamma_m)$ are regression trees, sequentially fitted to $r_{im} = y_i - \sum_{l=1}^{m-1} b_l(\mathbf{x}; \gamma_l)$, $i = 1, \ldots, N$, keeping $b_1, \ldots, b_{m-1}$ fixed.

(*ii*) Random forrest:

$$\widehat{f}_{\text{avg}}(\boldsymbol{x}) = \frac{1}{B} \sum_{b=1}^{B} \widehat{f}^b(\boldsymbol{x}),$$

where the $\widehat{f}^b$s are regression trees fitted to different bootstrap samples from the data, where each split is based on a randomly drawn subset of size m of the covariates

4 Some methods for classification

(a) Logistic regression for binary responses:

$$\Pr(Y=1|\boldsymbol{x}) = \frac{\exp(\beta_0 + \sum_{j=1}^p \beta_j x_j)}{1+\exp(\beta_0 + \sum_{j=1}^p \beta_j x_j)} \Leftrightarrow \log \frac{\Pr(Y=1|\boldsymbol{x})}{\Pr(Y=0|\boldsymbol{x})} = \beta_0 + \sum_{j=1}^p \beta_j x_j.$$

Can be combined with Ridge, Lasso and subset selection procedures.

(b) Several classes:

$$\log \frac{\Pr(Y=k|\boldsymbol{x})}{\Pr(Y=K|\boldsymbol{x})} = \beta_{k0} + \sum_{j=1}^p \beta_{kj} x_j.$$

(c) K-nearest neighbour classification:

$$\Pr(Y=k|\boldsymbol{X}=\boldsymbol{x}_0) \approx \frac{1}{K} \sum_{\boldsymbol{x}_i \in \mathcal{N}_0(\boldsymbol{x}_0)} I(y_i = k).$$

(d) Additive logistic regression:

$$\log \frac{\Pr(Y=1|\boldsymbol{x})}{\Pr(Y=0|\boldsymbol{x})} = \beta_0 + \sum_{j=1}^p f_j(x_j),$$

with basis functions f_j defined as for regression.

(e) Linear discriminant analysis (LDA): Let $\pi_k = \Pr(Y=k)$ and assume $[\mathbf{X}|Y=k] \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$, $k=1,\ldots,K$, with $p(\mathbf{x}|Y=k) = f_k(\mathbf{x})$, and use Bayes theorem to find:

$$\Pr(Y=k|X=x) = \frac{\pi_k f_k(\boldsymbol{x})}{\sum_{l=1}^K \pi_l f_l(\boldsymbol{x})}.$$

(f) Quadratic discriminant analysis (QDA): Same as LDA, but with $[\mathbf{X}|Y=k] \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$.

5 Unsupervised learning

(a) Clustering: divide the data into the clusters $C_1, ..., C_K$, where

(i) $C_1 \cup C_2 \cup \cdots \cup C_K = \{1,\ldots,N\}$ (all data points belong to a cluster)

(ii) $C_k \cap C_{k'} = \emptyset$ for $k \neq k'$ (each data point belongs to exactly one cluster)

(b) The total point scatter T for the dissimilarity measure $d_{ii'} \equiv d(x_i, x_{i'})$ is given by

$$\begin{aligned} T =& \frac{1}{2} \sum_{i=1}^N \sum_{i'=1}^N d_{ii'} \\ =& \frac{1}{2} \sum_{k=1}^K \sum_{i:C(i)=k} \left(\sum_{\{i':C(i')=k\}} d_{ii'} + \sum_{\{i':C(i')\neq k\}} d_{ii'} \right) \\ =& W(C) + B(C), \end{aligned}$$

where $W(C)$ and $B(C)$ are the within and between cluster point scatter, repectively.

(c) The clusters are typically found by minimising $W(C)$.

(d) K-means clustering: divide the data into K clusters, using the squared Euclidean distance as dissimilarity measure, i.e $d_{ii'} = ||\mathbf{x}_i - \mathbf{x}_{i'}||^2$. Algorithm: iterate between the two steps

(i) identify the centroids:

$$\mathbf{m}_1^*, \ldots, \mathbf{m}_K^* = \underset{\mathbf{m}_1,\ldots,\mathbf{m}_K}{\operatorname{argmin}} \sum_{k=1}^{K} N_k \sum_{\{i:C(i)=k\}} ||\mathbf{x}_i - \mathbf{m}_k||^2$$

(ii) allocate observations to closest centroid to minimise $W(C)$ w.r.t. C:

$$C(i) = \underset{k}{\operatorname{argmin}} ||\mathbf{x}_i - \mathbf{m}_k||^2$$

(e) K-medoids clustering: works with more general dissimilarity measures. The algorithm is the same as for K-means clustering, except that the centroids $\mathbf{m}_k$ are chosen as the observation x_{i^*} in each cluster k that minimises the total distance to all the other points in the cluster, i.e.

$$i^* = \underset{\{i:C(i)=k\}}{\operatorname{argmin}} \sum_{\{i':C(i')=k\}} d(\mathbf{x}_i, \mathbf{x}_{i'})$$

6 The maximum likelihood method

Assume that the random variables $X_1, X_2, \ldots, X_N$ have the joint probability density function/point mass function $f(x_1, \ldots, x_N; \boldsymbol{\theta})$, which is given by $f(x_1, \ldots, x_N; \boldsymbol{\theta}) = \prod_{i=1}^{N} f_i(x_i; \boldsymbol{\theta})$ if $X_1, X_2, \ldots, X_N$ are independent.

(a) Given the observed values $X_i = x_i$, $i = 1, \ldots, N$, the likelihood function is given by $L(\boldsymbol{\theta}) = f(x_1, \ldots, x_N; \boldsymbol{\theta})$ and the log-likelihood function by $\ell(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta})$.

(b) The maximum likelihood estimate $\widehat{\boldsymbol{\theta}}$ is given by

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, L(\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \ell(\boldsymbol{\theta}), \quad \ell(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta}).$$

(c) $\widehat{\boldsymbol{\theta}}$ is typically found as the solution of $\frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta}) = \mathbf{0}$.

(d) When there is no analytical solution, one may use the Newton-Raphson algorithm:

$$\boldsymbol{\theta}^{(s+1)} = \boldsymbol{\theta}^{(s)} - \left[\frac{\partial^2 \ell(\boldsymbol{\theta}^{(s)})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right]^{-1} \frac{\partial \ell(\boldsymbol{\theta}^{(s)})}{\partial \boldsymbol{\theta}}$$

(e) Under certain regularity conditions, $\widehat{\boldsymbol{\theta}} \approx N(\boldsymbol{\theta}, J(\widehat{\boldsymbol{\theta}})^{-1})$ with

$$J(\widehat{\boldsymbol{\theta}}) = -\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \ell(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}$$

7 Expectation, variance, standard deviation, covariance and correlation

(a) The expectation of a random variable X is given by

$$\mathrm{E}(X) = \sum_j x_j p(x_j) \qquad \text{(discrete)}$$

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x f(x) dx \qquad \text{(continuous)}$$

(b) For a real function g

$$\mathrm{E}[g(X)] = \sum_j g(x_j) p(x_j) \qquad \text{(discrete)}$$

$$\mathrm{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f(x) dx \qquad \text{(continuous)}$$

(c) $\mathrm{E}(a + bX) = a + b\mathrm{E}(X)$

(d) For a real function $g(X, Y)$ of two random variables X and Y

$$\mathrm{E}\big[g(X,Y)\big] = \sum_i \sum_j g(x_i, y_j) p(x_i, y_j) \qquad \text{(discrete)}$$

$$\mathrm{E}\big[g(X,Y)\big] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f(x,y) dy dx \qquad \text{(continuous)}$$

(e) If X and Y are independent $\mathrm{E}\big[g(X)h(Y)\big] = \mathrm{E}\big[g(X)\big] \cdot \mathrm{E}\big[h(Y)\big]$, and in particular $\mathrm{E}(XY) = \mathrm{E}(X) \cdot \mathrm{E}(Y)$

(f) $\mathrm{E}\left(a + \sum_{i=1}^{n} b_i X_i\right) = a + \sum_{i=1}^{n} b_i \mathrm{E}(X_i)$

(g) Conditional expectation:

$$\mathrm{E}(Y|X = x_i) = \sum_j y_j p_{Y|X}(y_j|x_i) \qquad \text{(discrete)}$$

$$\mathrm{E}(Y|X = x) = \int_{-\infty}^{\infty} y f_{Y|X}(y|x) dy \qquad \text{(continuous)}$$

(h) The variance and standard deviation of a random variable X are defined as

$$\mathrm{Var}(X) = \mathrm{E}[(X - \mu)^2]$$
$$\mathrm{sd}(X) = \sqrt{\mathrm{Var}(X)}$$

(i) $\mathrm{Var}(X) = \mathrm{E}(X^2) - \big(\mathrm{E}(X)\big)^2$

(j) $\mathrm{Var}(a + bX) = b^2\,\mathrm{Var}(X)$

(k)

$$\mathrm{Var}\left(a + \sum_{i=1}^{n} b_i X_i\right) = \sum_{i=1}^{n} {b_i}^2\,\mathrm{Var}(X_i) + \sum_{i=1}^{n}\sum_{j\neq i} b_i b_j \mathrm{Cov}(X_i, X_j)$$

(l) Let X og Y be random variables with $\mu_X = \mathrm{E}(X)$, $\sigma_X^2 = \mathrm{Var}(X)$, $\mu_Y = \mathrm{E}(Y)$ og $\sigma_Y^2 = \mathrm{Var}(Y)$. Then the covariance and correlation of X and Y are defined as

$$\mathrm{Cov}(X, Y) = \mathrm{E}\big[(X - \mu_X)(Y - \mu_Y)\big]$$
$$\rho = \mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

(m) $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$

(n) $\mathrm{Cov}(X, Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)$

(o) X, Y independent $\Rightarrow \mathrm{Cov}(X, Y) = 0$

(p)

$$\mathrm{Cov}\left(a + \sum_{i=1}^{n} b_i X_i\,,\; c + \sum_{j=1}^{m} d_j Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} b_i d_j \mathrm{Cov}(X_i, Y_j)$$

f) $-1 \leq \mathrm{Corr}(X, Y) \leq 1$ og $\mathrm{Corr}(X, Y) = \pm 1$ if and only if there exist numbers a, b so that $Y = a + bX$ (except potentially in an area with probability 0)

(q) If $X_1, \ldots, X_p$ are random variables, we say that $\mathbf{X} = (X_1, ..., X_p)^T$ is a random vector

(r) The expectation vector and covariance matrix of $\mathbf{X}$ are defined by by

$$\mathrm{E}\left(\mathbf{Z}\right) = \begin{pmatrix} \mathrm{E}\left(Z_1\right) \\ \mathrm{E}\left(Z_2\right) \\ \vdots \\ \mathrm{E}\left(Z_p\right) \end{pmatrix}, \quad \mathrm{Var}\left(\mathbf{Z}\right) = \begin{pmatrix} \mathrm{Var}\left(Z_1\right) & \mathrm{Cov}\left(Z_1, Z_2\right) & \cdots & \mathrm{Cov}\left(Z_1, Z_p\right) \\ \mathrm{Cov}\left(Z_2, Z_1\right) & \mathrm{Var}\left(Z_2\right) & \cdots & \mathrm{Cov}\left(Z_2, Z_p\right) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}\left(Z_p, Z_1\right) & \mathrm{Cov}\left(Z_p, Z_2\right) & \cdots & \mathrm{Var}\left(Z_p\right) \end{pmatrix}$$

(s) For $\mathbf{A}$ and $\mathbf{b}$ constant

$$\begin{aligned} \mathrm{E}\left(\mathbf{AZ} + \mathbf{b}\right) =& \mathbf{A}\mathrm{E}\left(\mathbf{Z}\right) + \mathbf{b}, \\ \mathrm{Var}\left(\mathbf{AZ} + \mathbf{b}\right) =& \mathbf{A}\mathrm{Var}\left(\mathbf{Z}\right)\mathbf{A}^T \end{aligned}$$

8 Some probability distributions

(a) <u>Binomial distribution</u>:

Point mass function: $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \qquad k = 0, 1, \ldots, n$

Expectation: $\mathrm{E}(X) = np$

Variance: $\mathrm{Var}(X) = np(1-p)$

Approximation: $Z = \dfrac{X - np}{\sqrt{np(1-p)}}$ is approximately normal distributed when np og $n(1-p)$ are both sufficiently large (at least 10)

Summation rule: $X \sim$ binomial (n, p), $Y \sim$binomial (m, p) og X, Y independent $\Rightarrow X + Y \sim$ binomial $(n + m, p)$

(b) <u>Multinomial distribution</u>:

Point mass function: $P(N_1 = n_1, \ldots, N_r = n_r) = \frac{n!}{n_1! \cdots n_r!} p_1^{n_1} \cdots p_r^{n_r}$,
where $\sum_{i=1}^{r} p_i = 1$ and $\sum_{i=1}^{r} n_i = n$

Marginal distributions: $N_i \sim$ binomial (n, p_i), $i = 1, \ldots, n$

(c) <u>Normal distribution</u>:

Density: $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} \qquad -\infty < x < \infty$

Expectation: $\mathrm{E}(X) = \mu$

Variance: $\mathrm{Var}(X) = \sigma^2$ Transformation: $X \sim N(\mu, \sigma^2) \Rightarrow a + bX \sim N(a + b\mu, b^2\sigma^2)$
$X \sim N(\mu, \sigma^2) \Rightarrow Z = \frac{X-\mu}{\sigma} \sim N(0, 1)$

Summation rule: $X \sim N(\mu_X, \sigma_X^2)$, $Y \sim N(\mu_Y, \sigma_Y^2)$, X, Y independent
$\Rightarrow X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$

(d) <u>Chi square (χ^2 distribution)</u>:

Density: $f(u) = \frac{1}{2^{n/2}\Gamma(n/2)} u^{(n/2)-1} e^{-u/2} \qquad u > 0$
n is the number of degrees of freedom

Expectation: $\mathrm{E}(U) = n$

Variance: $\mathrm{Var}(U) = 2n$

Summation rule: $U \sim \chi_n^2$, $V \sim \chi_m^2$, U and V independent
$\Rightarrow U + V \sim \chi_{n+m}^2$

Result: $Z \sim N(0, 1) \Rightarrow Z^2 \sim \chi_1^2$

(e) <u>Student's t distribution</u>:

Density: $f(t) = \frac{\Gamma[(n+1)/2]}{\sqrt{n\pi}\Gamma(n/2)}(1+\frac{t^2}{n})^{-(n+1)/2} \qquad -\infty < t < \infty$

n is the number of degrees of freedom

Expectation: $\mathrm{E}(T) = 0 \qquad (n \geq 2)$

Variance: $\mathrm{Var}(T) = n/(n-2) \qquad (n \geq 3)$

Result: $Z \sim N(0,1), \quad U \sim \chi^2_n, \quad Z, U$ independent $\Rightarrow Z/\sqrt{U/n} \sim t_n$

(f) <u>Multivariate normal distribution</u>:

Let $\mathbf{X} = (X_1, \ldots, X_d)^T$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)^T$, $\boldsymbol{\Sigma} = \{\Sigma_{ij}\}_{i,j=1,\ldots,d}$

Density:

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

Suppose $\mathbf{X}$ is partitioned as

$$\mathbf{X} = \begin{pmatrix}\mathbf{X}_1\\ \mathbf{X}_2\end{pmatrix} \text{ with } \boldsymbol{\mu} = \begin{pmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{pmatrix} \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{pmatrix}$$

Marginal distributions: $\mathbf{X}_k \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_{kk}), \quad k = 1, 2$ and
$X_i \sim N(\mu_i, \Sigma_{ii}), \quad i = 1, \ldots, d$

Expectation vector: $\mathrm{E}(\mathbf{X}) = \boldsymbol{\mu}$

Covariance matrix: $\mathrm{Var}(\mathbf{X}) = \boldsymbol{\Sigma}$

Correlations: $\mathrm{Corr}(X_i, X_j) = \frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii}\Sigma_{jj}}}$

Conditional distributions:
Then $\mathbf{X}_1|\mathbf{X}_2 = \mathbf{x}_2 \sim N(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})$

Independence: $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent if and only if $\boldsymbol{\Sigma}_{12}$ $(= \boldsymbol{\Sigma}_{21}^T) = \mathbf{0}$

If $\mathbf{X} \sim N(\boldsymbol{\mu}, \sigma^2\mathbf{I})$, then $X_1, \ldots, X_d$ are independent and $X_i \sim N(\mu_i, \sigma^2)$

Summation rule: If $\mathbf{U} = \mathbf{A}\mathbf{X} + \mathbf{b}$, then $\mathbf{U} \sim N(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)$