

UNIVERSITY OF OSLO

Faculty of mathematics and natural sciences

Exam in: STK2100 — Machine Learning and Statistical
Methods for Prediction and Classification

Day of examination: Friday June 12th, 2026

Examination hours: 09.00–13.00

This problem set consists of 19 pages.

Appendices: Formulas in STK2100

Permitted aids: Approved calculator

Please make sure that your copy of the problem set is complete before you attempt to answer anything.

Problem 1

In this problem, we will consider the **Concrete** data, that concerns the compressive strength of concrete based on attributes of the concrete mixture. The data set treated in this problem contains the variables:

- **cement**: Quantity of cement used in the concrete mixture in kg/m^3
- **slag**: Amount of blast furnace slag in kg/m^3
- **ash**: Proportion of fly ash in kg/m^3
- **water**: Volume of water added to the mixture in kg/m^3
- **superplastic**: Dosage of superplasticizer in kg/m^3
- **fineagg**: Amount of fine aggregate in kg/m^3
- **age.gr**: Age of the concrete specimen at the time of testing, grouped as Group 1: $\text{Age} \leq 28$ days, Group 2: $28 \text{ days} < \text{Age} \leq 56$ days, Group 3: $\text{Age} > 56$ days
- **strength**: Compression strength of concrete specimen in MPa

The response variable throughout this problem is **strength**. Further, **age.gr** is a categorical covariate, while **cement**, **slag**, **ash**, **superplastic** and **fineagg** are numerical covariates. All models in this problem are fitted to a randomly selected training set of 687 observations, whereas the remaining 343 observations are assigned to the test set.

(Continued on page 2.)

First, a linear regression model with all the covariates is fitted to the data. We call this model **Model 1**. Output from R is shown below.

```

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 28.503928 10.351264 2.754 0.00605 **
cement      0.095092 0.006287 15.125 < 2e-16 ***
slag        0.080491 0.007175 11.219 < 2e-16 ***
ash         0.042917 0.010886 3.942 8.91e-05 ***
water       -0.153019 0.026157 -5.850 7.65e-09 ***
superplastic 0.469698 0.105079 4.470 9.17e-06 ***
fineagg     -0.009843 0.007213 -1.365 0.17284
age.gr2     16.892463 1.309944 12.896 < 2e-16 ***
age.gr3     18.522281 1.040646 17.799 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.695 on 678 degrees of freedom
Multiple R-squared: 0.666, Adjusted R-squared: 0.6621
F-statistic: 169 on 8 and 678 DF, p-value: < 2.2e-16

```

a

- Based on the output above, which covariates seem to be important, and which do not seem to be important?
- What is the interpretation of the effect of **cement**?
- What is the interpretation of the effect of **age.gr**?

Next, a model without the covariate **fineagg** is fitted. We denote this model **Model 2**. Fitted coefficients and corresponding P-values from the Wald tests and the AIC and BIC for Models 1 and 2 are shown below.

Covariate	Model 1	Model 2
Intercept	28.5 (6.05e-3)	16.0 (9.14e-4)
cement	0.0951 (< 2e-16)	0.101 (< 2e-16)
slag	0.0805 (< 2e-16)	0.0863 (< 2e-16)
ash	0.0429 (8.91e-5)	0.05108 (2.88e-8)
water	-0.153 (7.65e-9)	-0.138 (9.86e-9)
superplastic	0.470 (9.17e-6)	0.428 (2.39e-5)
fineagg	-0.00984 (0.173)	-
age.gr2	16.9 (< 2e-16)	16.8 (< 2e-16)
age.gr3	18.5 (< 2e-16)	18.6 (< 2e-16)
AIC	5084	5084
BIC	5134	5129

b

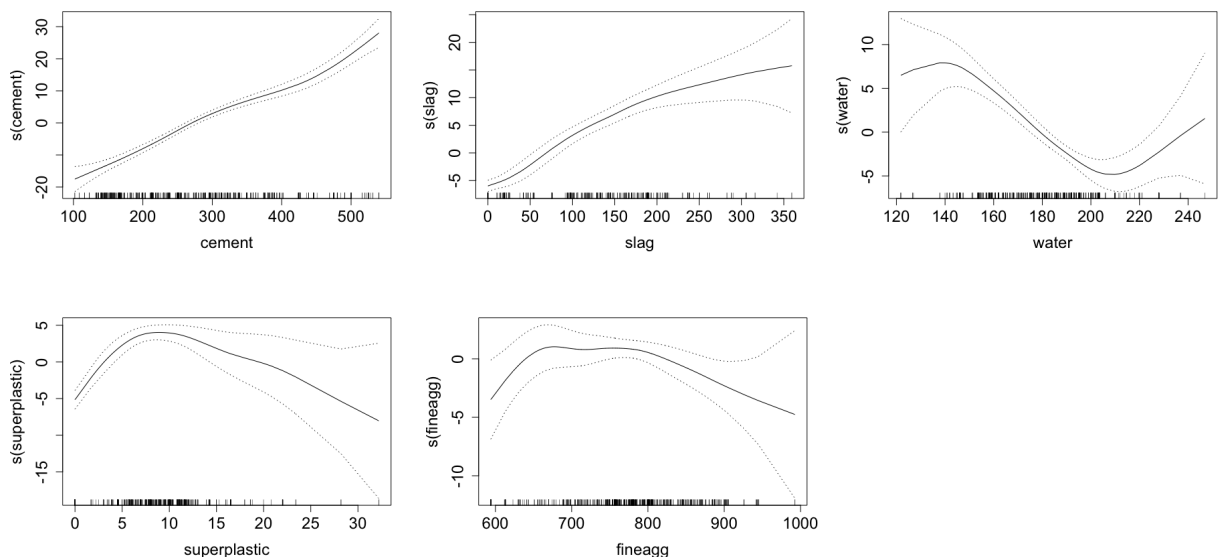
- How are the AIC and BIC computed, respectively? Use this to explain why the BIC is consistently larger than the AIC.
- Given the above results, which of the two models would you choose? Justify your answer.

(Continued on page 3.)

An additive linear model is now fitted to the data, where the effect of each of the covariates `cement`, `slag`, `water`, `superplastic` and `fineagg` is modelled with a smoothing spline. This model is denoted **Model 3**. Corresponding output from R is shown below, including plots of the effects modelled with a spline. Further, the AIC and BIC for the fitted additive model are 4981 and 5027, respectively. Finally, the results of a hypothesis test of Model 3 against a corresponding model without the covariate `fineagg` is also shown below.

Anova for Nonparametric Effects

Npar	Df	Npar	F	Pr(F)
(Intercept)				
s(cement)	3	4.7366	0.0028106	**
s(slag)	3	3.2307	0.0219939	*
ash				
s(water)	3	10.2514	1.324e-06	***
s(superplastic)	3	19.7391	2.890e-12	***
s(fineagg)	3	6.0695	0.0004447	***
age.gr				



Analysis of Deviance Table

Model 1: `strength ~ s(cement) + s(slag) + ash + s(water) + s(superplastic) + age.gr`

Model 2: `strength ~ s(cement) + s(slag) + ash + s(water) + s(superplastic) + s(fineagg) + age.gr`

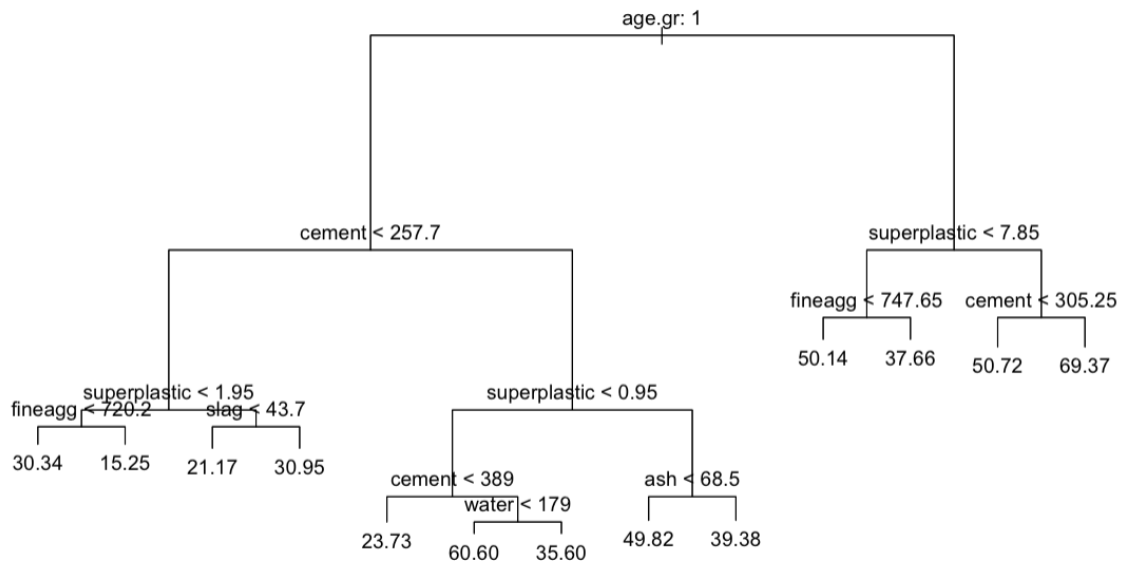
Resid.	Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	667	54152			
2	663	52713	4.0001	1438.7	0.001182 **

(Continued on page 4.)

c

- Interpret the effects of the covariates **water** and **fineagg** based on the plots.
- Assess whether **fineagg** is an important covariate in Model 3, and comment on potential differences in that respect compared to the linear model, considered in **a**. Why do you think the conclusion might be different for Model 3?
- Is there reason to believe that there are non-linear effects of the covariates on the compression strength of concrete? Justify your answer.

A regression tree is fitted to the **Concrete** data, and is called **Model 4**. The resulting tree is shown below.



Further, the table below shows predictions of the concrete strength in MPa from each of the 4 models fitted in this problem, for concrete specimens with the following attributes: 395kg/m³ cement, 30kg/m³ slag, 60 kg/m³ ash, 150 kg/m³ water and 710 kg/m³ fineagg, as well as

- Case 1: Age 20 days and 0.7 kg/m³ superplastic
- Case 2: Age 20 days and 8 kg/m³ superplastic
- Case 3: Age 70 days and 0.7 kg/m³ superplastic
- Case 4: Age 70 days and 8 kg/m³ superplastic.

The table also shows the mean squared errors (MSEs) $\frac{1}{343} \sum_{i \in \text{test set}} (y - \hat{y}_i)^2$ when using each of the 4 models to predict on the test set.

Setting	Model 1	Model 2	Model 3	Model 4
Case 1	41.4		40.3	
Case 2	44.9	44.1	48.2	49.8
Case 3	59.9	59.6	59.0	
Case 4		62.7	66.8	
MSE	94.4	94.2	76.8	104.9

(Continued on page 5.)

d

- What are the most important covariates in the tree model, i.e Model 4? Comment on the differences from Models 1 to 3 with respect to covariate importance.
- Fill out the missing numbers in the table of predictions.
- What do the above results indicate concerning the potential presence of interaction effects?

Finally, a boosted tree model is fitted to the data, and denoted **Model 5**. The best model was obtained with trees of depth 1. The MSE on the test set for this model is 72.6.

e

- Explain the main differences between a regression tree, such as Model 4, and a boosted tree model, such as Model 5, in particular with respect to how they are fitted to data.
- What do the results from fitting a boosted tree model indicate concerning the presence of interaction effects?
- How would you order the five considered models according to performance? Justify your answer.

Problem 2

In this problem, we will consider the **Credit card** data, that are records of a bank's approval or not of credit card applicants. The data set contains the variables:

- **GENDER**: Gender, with levels F=female, M=male, O=other
- **Car_Owner**: Indicator of whether the applicant owns a car, encoded as Y=yes, N=no
- **Propert_Owner**: Indicator of whether the applicant owns property, encoded as Y=yes, N=no
- **CHILDREN**: Number of children of the applicant
- **Annual_income**: Annual income of the applicant in USD
- **Type_Income**: Income type of the applicant, with levels Commercial associate, Pensioner, State servant and Working
- **EDUCATION**: Education level of the applicant, with levels Higher education, Incomplete higher, Lower secondary and Secondary / secondary special
- **Marital_status**: Marital status of the applicant, with levels Civil marriage, Married, Separated, Single / not married and Widow
- **Housing_type**: Housing type of the applicant, with levels Co-op apartment, House / apartment, Municipal apartment, Office apartment, Rented apartment and With parents

(Continued on page 6.)

- **Birthday_count**: Number of days since the birth of the applicant, counted backwards (-1000 means 1000 days ago)
- **Employed_days**: Number of days since the applicant was last employed, where a positive value means that the applicant is unemployed and a negative value means that the applicant is employed
- **Work_Phone**: Indicator of whether the applicant has a work phone, encoded as 0=no and 1=yes
- **Family_Members**: Number of family members of the applicant
- **credit.ind**: Indicator of whether the applicant was approved for a credit card or not, encoded as 0=no and 1=yes

The response in this problem is `credit.ind`, and is binary. Further, the covariates `GENDER`, `Car_Owner`, `Propert_Owner`, `Type_Income`, `EDUCATION`, `Marital_status`, `Housing_type` and `Work_Phone` are categorical, whereas `CHILDREN`, `Annual_income`, `Birthday_count`, `Employed_days` and `Family_Members` are numerical. Moreover, the data set is randomly divided into a training set of 220 observations and a test set of 110 observations. The ratio of 1s among the y s is 50% in both the training and the test set.

We will in this problem fit logistic regression models

$$\Pr(Y = 1|\mathbf{x}) = \frac{\exp(\beta_0 + \sum_{j=1}^{25} \beta_j x_j)}{1 + \exp(\beta_0 + \sum_{j=1}^{25} \beta_j x_j)},$$

both without and with regularisation, i.e. the model parameters β are estimated by minimising the loss function

$$D(\hat{\beta}) = -\sum_{i=1}^{220} \left(y_i \mathbf{x}_i^T \beta - \log \left(1 + e^{\mathbf{x}_i^T \beta} \right) \right) + \lambda \sum_{j=1}^{25} |\beta_j|^q. \quad (1)$$

for a given value of λ . In (1), the first term is the negative log-likelihood function of the logistic regression model, $\lambda \geq 0$ is the regularisation parameter, controlling the amount of regularisation and q is set to obtain the desired form of regularisation.

R output from fitting the logistic regression model without regularisation ($\lambda = 0$) to the `Credit card` data is shown below

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept)	20.31330	5602.62627	0.004	0.9971
GENDERF	-18.14371	2796.02184	-0.006	0.9948
GENDERM	-17.55567	2796.02184	-0.006	0.9950
Car_OwnerY	-0.22142	0.39240	-0.564	0.5726
Propert_OwnerY	-0.16061	0.35499	-0.452	0.6510
CHILDREN	-0.61540	0.61554	-1.000	0.3174
Annual_income	0.10280	0.18221	0.564	0.5726
Type_IncomePensioner	3.73633	3.41386	1.094	0.2738

(Continued on page 7.)

Type_IncomeState servant	-0.65439	0.91141	-0.718	0.4728
Type_IncomeWorking	0.02428	0.41114	0.059	0.9529
EDUCATIONIncomplete higher	-0.61438	1.22149	-0.503	0.6150
EDUCATIONLower secondary	51.86596	3183.12516	0.016	0.9870
EDUCATIONSecondary / secondary special	-0.78931	0.40877	-1.931	0.0535
Marital_statusMarried	17.43546	1516.33551	0.011	0.9908
Marital_statusSeparated	20.57897	1516.33610	0.014	0.9892
Marital_statusSingle / not married	18.62725	1516.33570	0.012	0.9902
Marital_statusWidow	17.07706	1516.33605	0.011	0.9910
Housing_typeHouse / apartment	-20.09509	4612.20219	-0.004	0.9965
Housing_typeMunicipal apartment	-17.55549	4612.20229	-0.004	0.9970
Housing_typeOffice apartment	-19.82101	4612.20242	-0.004	0.9966
Housing_typeRented apartment	-20.69746	4612.20234	-0.004	0.9964
Housing_typeWith parents	-37.71578	4884.24408	-0.008	0.9938
Birthday_count	-0.37837	0.22954	-1.648	0.0993
Employed_days	-1.42205	1.31626	-1.080	0.2800
Work_Phone	-0.27063	0.41040	-0.659	0.5096
Family_Members	0.91636	0.71633	1.279	0.2008

a

- What is the number of degrees of freedom of the logistic regression model above? How would regularisation affect the number of degrees of freedom?
- Explain why it is sensible to consider regularisation of the logistic regression model for the **Credit card** data.
- What value should q be set to in order to obtain ridge and Lasso type regularisation, respectively?
- What are the main differences between solutions obtained with ridge and Lasso type regularisation?

b

The regularisation parameter λ is typically chosen with K -fold cross-validation. Sketch how cross-validation can be used for this purpose (a detailed algorithm is not necessary).

In addition to the logistic regression model without regularisation (**Log. reg.**), one model with ridge (**ridge**) and one model with Lasso (**Lasso**) regularisation were fitted to the **Credit card** data. The corresponding models were then used to predict $\Pr(Y = 1|\mathbf{x})$ on the test set, resulting in predictions $\hat{p}_i$. These were then used to compute the area under the curve (AUC). The results are shown in the table below.

	Log. reg.	ridge	Lasso
AUC	0.514	0.562	0.529

c

- Explain how one obtains a ROC curve based on predictions $\hat{p}_i$.
- How is the AUC defined, and how can that be used to assess the performance of a binary classification model?

(Continued on page 8.)

- Based on the above results, how would you rank the 3 models? Justify your answer.

Problem 3

In this problem, we will consider local regression models.

$$Y = f(\mathbf{x}) + \varepsilon, \quad E(\varepsilon) = 0, \quad \text{with} \quad f(\mathbf{x}) = \mathbf{b}^T \boldsymbol{\theta}_{\mathbf{x}}, \quad (2)$$

$\mathbf{b}$ being the chosen basis expansion for the model, that depends on $\mathbf{x}$ and hence is known, whereas $\boldsymbol{\theta}_{\mathbf{x}}$ are unknown parameters. Here, $f(\mathbf{x})$ is allowed to adapt locally by re-fitting $f(\mathbf{x})$ at each new point $\mathbf{x}$ of interest, putting more weight on the observations that are closest to $\mathbf{x}$ than on the ones that are far away from $\mathbf{x}$. More specifically, the model is fitted to data $(y_i, \mathbf{x}_i)$, $i = 1, \dots, N$ by minimising the loss function

$$R(\boldsymbol{\theta}_{\mathbf{x}}) = \sum_{i=1}^N w(\mathbf{x}, \mathbf{x}_i) (y_i - \mathbf{b}_i^T \boldsymbol{\theta}_{\mathbf{x}})^2 \quad (3)$$

with respect to $\boldsymbol{\theta}_{\mathbf{x}}$, where $w(\mathbf{x}, \mathbf{x}_i)$ is the weight put on observation $\mathbf{x}_i$, and is given, which is a form of weighted least squares.

Now, let

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} \mathbf{b}_1^T \\ \vdots \\ \mathbf{b}_N^T \end{pmatrix} \quad \text{and} \quad \mathbf{W}_{\mathbf{x}} = \begin{pmatrix} w(\mathbf{x}, \mathbf{x}_1) & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & w(\mathbf{x}, \mathbf{x}_N) \end{pmatrix}.$$

a

- Show that the loss function $R(\boldsymbol{\theta}_{\mathbf{x}})$ from (3) can be written on the form

$$R(\boldsymbol{\theta}_{\mathbf{x}}) = (\mathbf{y} - \mathbf{B}\boldsymbol{\theta}_{\mathbf{x}})^T \mathbf{W}_{\mathbf{x}} (\mathbf{y} - \mathbf{B}\boldsymbol{\theta}_{\mathbf{x}}).$$

- Show that the estimate $\hat{\boldsymbol{\theta}}_{\mathbf{x}}$ of $\boldsymbol{\theta}_{\mathbf{x}}$ is then

$$\hat{\boldsymbol{\theta}}_{\mathbf{x}} = (\mathbf{B}^T \mathbf{W}_{\mathbf{x}} \mathbf{B})^{-1} \mathbf{B}^T \mathbf{W}_{\mathbf{x}} \mathbf{y}.$$

(Hint: Recall that if $\mathbf{a}$ and $\mathbf{b}$ are $p \times 1$ vectors and $\mathbf{A}$ a $p \times p$ matrix, then $\frac{\partial \mathbf{a}^T \mathbf{b}}{\partial \mathbf{b}} = \mathbf{a}$ and $\frac{\partial \mathbf{b}^T \mathbf{A} \mathbf{b}}{\partial \mathbf{b}} = (\mathbf{A} + \mathbf{A}^T) \mathbf{b}$).

Now, we will assume that there is a single covariate x . Then, the local linear regression model

$$\hat{f}(x) = \mathbf{b}^T \hat{\boldsymbol{\theta}}_x = \hat{\theta}_{x,0} + \hat{\theta}_{x,1} x$$

is obtained with the basis expansion $\mathbf{b} = (1, x)^T$ and weights $w(x, x_i) = K_{\lambda}(x, x_i)$, where $K_{\lambda}(x, x_i)$ is a kernel, with $K_{\lambda}(x, x_i) = D\left(\frac{|x - x_i|}{\lambda}\right)$, for

some $D(t) = \begin{cases} 1, & t \leq 1 \\ 0, & t > 1 \end{cases}$ and $\lambda > 0$.

(Continued on page 9.)

For other choices of $\mathbf{b}$ and $w(x, x_i)$, one can for instance obtain the kernel smoother

$$\hat{f}(x) = \frac{\sum_{i=1}^N K_\lambda(x, x_i) y_i}{\sum_{i=1}^N K_\lambda(x, x_i)}, \quad (4)$$

or k -nearest neighbour

$$\hat{f}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i = \frac{\sum_{i=1}^N I(x_i \in N_k(x)) y_i}{\sum_{i=1}^N I(x_i \in N_k(x))} \quad (5)$$

where $N_k(x)$ are the k nearest neighbours of x and $I(\cdot)$ is the indicator function $I(A) = \begin{cases} 1, & A \text{ is true} \\ 0, & A \text{ is not true} \end{cases}$.

b

- Use the results from **a** to show that the kernel smoother (4) is obtained with $\mathbf{b} = 1$ (so that $\mathbf{B} = (1, \dots, 1)^T$) and $w(x, x_i) = K_\lambda(x, x_i)$.
- What do $\mathbf{b}$ and $w(x, x_i)$ have to be to obtain k -nearest neighbour (5)?

Assume now that $\mathbf{x}$ has higher dimension, i.e. that we have more than one covariate. With appropriate $\mathbf{b}$ and $w(\mathbf{x}, \mathbf{x}_i)$, the local regression model (2) can be made into a local linear regression model, a kernel smoother or a k -nearest neighbour model with several covariates.

c

Explain why such local models tend to work poorly in higher dimensions.

END

APPENDIX: LIST OF FORMULAS FOR STK2100

General issues

(a) Complexity/degrees of freedom

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

(b) Loss functions

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

(c) Model selection

- (i) for a model with parameters $\boldsymbol{\theta}$, log-likelihood $l(\boldsymbol{\theta})$ and maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ based on N observations, where $|\boldsymbol{\theta}|$ is the number of *free* parameters in the model:

- i. $\text{AIC} = -2l(\hat{\boldsymbol{\theta}}) + 2|\boldsymbol{\theta}|$
- ii. $\text{BIC} = -2l(\hat{\boldsymbol{\theta}}) + \log(N)|\boldsymbol{\theta}|$

- (ii) K -fold cross-validation:

- i. Divide the N data points into K groups by randomization
- ii. For $k = 1, \dots, K$
 - A. Fit the model on all data *except* data from group k .
 - B. Predict $\hat{y}_i$ for all data in group k based on the fitted model
 - C. Calculate $\widehat{\text{Err}}_i = L(y_i, \hat{y}_i)$ for i in group k
- iii. Calculate $\text{CV}_{(K)} = \frac{1}{N} \sum_{i=1}^N \widehat{\text{Err}}_i$

- (iii) Bootstrap:

- i. Draw B bootstrap samples $(y_1^{*b}, \mathbf{x}_1^{*b}), \dots, (y_N^{*b}, \mathbf{x}_N^{*b})$ from the data with replacement
- ii. Fit a $\hat{f}^{*b}$ to each bootstrap sample
- iii. The leave-one-out bootstrap estimate of the prediction error is

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

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

- iv. The 0.632-estimator is given by

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

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

(Continued on page 11.)

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

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

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

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

Linear regression

(a) Model:

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

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

(b) Matrix form:

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

where $\mathbf{Y} = (Y_1, \dots, Y_N)^T$ and $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)^T$ are N - and $(p+1)$ -dimensional 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 $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$.

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

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

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

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

(g) We can test the hypothesis

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

by using the test statistic

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

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

(Continued on page 12.)

(h) Ridge/Lasso regression: Minimize with respect to β

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

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

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

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

General regression methods

(a) General setting: Assume

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

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

(b) K -nearest neighbour regression:

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

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

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

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

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

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

(iii) Smoothing spline, defined by

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

(Continued on page 13.)

- (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

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

- (e) Neural network with one hidden layer:

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

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

- (f) Tree-based methods: $f(\mathbf{x}) = \sum_{m=1}^M c_m I(\mathbf{x} \in R_m)$ where the covariate space has been divided into the regions $R_1, \dots, R_M$ 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 $r_{im} = y_i - \sum_{l=1}^{m-1} b_l(\mathbf{x}; \gamma_l)$, $i = 1, \dots, N$, keeping $b_1, \dots, b_{m-1}$ fixed.

- (ii) Random forest:

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

where the $\hat{f}^b$ 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

Some methods for classification

- (a) Logistic regression for binary responses:

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

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

- (b) Several classes:

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

(Continued on page 14.)

(c) K -nearest neighbour classification:

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

(d) Additive logistic regression:

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

with basis functions f_j defined as for regression.

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

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

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

Unsupervised learning

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

- (i) $C_1 \cup C_2 \cup \dots \cup C_K = \{1, \dots, N\}$ (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, respectively.

(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

(Continued on page 15.)

(i) identify the centroids:

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

(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.

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

The maximum likelihood method

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

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

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

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

(c) $\hat{\boldsymbol{\theta}}$ 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, $\hat{\boldsymbol{\theta}} \approx N(\boldsymbol{\theta}, J(\hat{\boldsymbol{\theta}})^{-1})$ with

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

Expectation, variance, standard deviation, covariance and correlation

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

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

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

(Continued on page 16.)

(b) For a real function g

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

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

(c) $E(a + bX) = a + bE(X)$

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

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

$$E[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y)f(x, y)dydx \quad (\text{continuous})$$

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

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

(g) Conditional expectation:

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

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

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

$$\begin{aligned} \text{Var}(X) &= E[(X - \mu)^2] \\ \text{sd}(X) &= \sqrt{\text{Var}(X)} \end{aligned}$$

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

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

(k)

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

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

$$\begin{aligned} \text{Cov}(X, Y) &= E[(X - \mu_X)(Y - \mu_Y)] \\ \rho = \text{Corr}(X, Y) &= \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} \end{aligned}$$

(Continued on page 17.)

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

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

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

(p)

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

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

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

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

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

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

$$E(\mathbf{AZ} + \mathbf{b}) = \mathbf{A}E(\mathbf{Z}) + \mathbf{b},$$

$$\text{Var}(\mathbf{AZ} + \mathbf{b}) = \mathbf{A}\text{Var}(\mathbf{Z})\mathbf{A}^T$$

Some probability distributions

(a) Binomial distribution:

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

Expectation: $E(X) = np$

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

Approximation: $Z = \frac{X - np}{\sqrt{np(1-p)}}$ is approximately normal distributed

when np and $n(1-p)$ are both sufficiently large (at least 10)

Summation rule: $X \sim \text{binomial}(n, p), Y \sim \text{binomial}(m, p)$
and X, Y independent $\Rightarrow X + Y \sim \text{binomial}(n + m, p)$

(b) Multinomial distribution:

Point mass function: $P(N_1 = n_1, \dots, N_r = n_r) = \frac{n!}{n_1! \cdots n_r!} p_1^{n_1} \cdots p_r^{n_r}$,

where $\sum_{i=1}^r p_i = 1$ and $\sum_{i=1}^r n_i = n$

Marginal distributions: $N_i \sim \text{binomial}(n, p_i), i = 1, \dots, r$

(Continued on page 18.)

(c) Normal distribution:

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

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

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

$$\begin{aligned} \text{Summation rule: } X &\sim N(\mu_X, \sigma_X^2), \quad Y \sim N(\mu_Y, \sigma_Y^2), \quad X, Y \text{ independent} \\ &\Rightarrow X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2) \end{aligned}$$

(d) Chi square (χ^2 distribution):

$$\text{Density: } f(u) = \frac{1}{2^{n/2}\Gamma(n/2)} u^{(n/2)-1} e^{-u/2} \quad u > 0$$

n is the number of degrees of freedom

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

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

$$\begin{aligned} \text{Summation rule: } U &\sim \chi_n^2, \quad V \sim \chi_m^2, \quad U \text{ and } V \text{ independent} \\ &\Rightarrow U + V \sim \chi_{n+m}^2 \end{aligned}$$

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

(e) Student's t distribution:

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

n is the number of degrees of freedom

$$\text{Expectation: } E(T) = 0 \quad (n \geq 2)$$

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

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

(f) Multivariate normal distribution:

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

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

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

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

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

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

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

Conditional distributions:

(Continued on page 19.)

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 $X_1, \dots, X_d$ are independent and $X_i \sim N(\mu_i, \sigma^2)$

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