A. Giovanidis 2022

08. Classification – pt.B
Bayesian methods

Network Data Analysis - NDA'22
Anastasios Giovanidis

Sorbonne-LIP6

November 15, 2022

Bibliography

A. Giovanidis 2022

B.1 Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. "An introduction to statistical learning: with applications in R". Springer Texts in Statistics. ISBN 978-1-4614-7137-0
Chapter 2, Chapter 4
DOI 10.1007/978-1-4614-7138-7

Bayes Classifier

A. Giovanidis 2022

Optimal Classifier: (If all misclassifications are equally important) Assign each observation to the most likely class, given its predictor values:

$$\max_{1 \leq j \leq M} \; Pr\left(Y = j \mid X = x_o\right)$$

- We consider *conditional probabilities* given the observed x_o.

☞ In a two-class problem

$$Pr\left(Y = 1 \mid X = x_o\right) + Pr\left(Y = 2 \mid X = x_o\right) = 1:$$

$$\begin{array}{ll} \text{Class } 1, & \text{if } Pr\left(Y = 1 \mid X = x_o\right) > 0.5 \\ \text{Class } 2, & \text{if } Pr\left(Y = 2 \mid X = x_o\right) > 0.5 \end{array}$$

☞ Decision boundary $Pr\left(Y = 1 \mid X = x_o\right) = Pr\left(Y = 2 \mid X = x_o\right)$

Drawback...

A. Giovanidis 2022

There is one problem however: For real data we do not know the conditional distribution $P(Y|X)$,

(unless we have generated data ourselves, in which case we know the joint distribution $P(X, Y)$).

Bayes classifier serves as an unreachable golden standard!

If we do not know exactly $P(Y|X)$ we can try to estimate it.

Naive Bayes

A. Giovanidis 2022

☞ The Naive Bayes classifier:

- Assumes that the K features are independent.
- Uses a simple MAP or ML estimator

$$\begin{aligned} P(Y \mid \mathcal{D}_n) &\propto P(\mathcal{D}_n \mid Y)P(Y) \quad \textbf{[MAP]} \\ P(Y \mid \mathcal{D}_n) &\propto P(\mathcal{D}_n \mid Y) \qquad\quad \textbf{[ML]} \end{aligned}$$

where Y is the class label.
We choose MAP or ML, depending on the prior information over the class distribution Y.

Naive Bayes with discrete features

A. Giovanidis 2022

✐ Let us classify texts (e.g. books, sentences) in one of two classes:

1. History
2. Science

Bag-of-words

A. Giovanidis 2022

To do so, we will use some features from the available data (texts).

These are a certain bag-of-words: {'king', 'food', 'equals', 'proof'}

	Bag-Of-Words				Label	
	1:'king'	2:'food'	3:'equals'	4:'proof'	History	Science
Text 1	No	Yes	Yes	Yes	No	Yes
Text 2	No	No	Yes	No	No	Yes
Text 3	Yes	Yes	No	Yes	Yes	No
...	...	...	...	...	...	...
Text n	Yes	No	Yes	Yes	No	Yes

Forming the variables

A. Giovanidis 2022

✐ If X contains K binary state features, with $X_{t,k} \in \{0,1\}$, then

$$X_t \quad = \quad (X_{t,1}, \ldots, X_{t,K}), \quad t = 1, \ldots, n.$$

$X_{t,k}$ says whether feature k appears or not in the t-th data sample of $\mathcal{D}_n$.

Also, Y is the label of each text. Then, let

$$Y_t = \begin{cases} 0 & \text{if 'History'} \\ 1 & \text{if 'Science'} \end{cases}$$

Discrete Estimators

A. Giovanidis 2022

- Mean estimators

$$p_{1} = p_{Sc} = P(Y=1) = \frac{1}{n}\sum_{t=1}^{n} Y_t,$$

$$p_{0} = p_{Hi} = P(Y=0) = \frac{1}{n}\sum_{t=1}^{n}(1-Y_t)$$

- ML estimators

$$p_{1,k} = p_{Sc,k} = P(X_k = 1 \mid Y = 1) = \frac{\sum_{t=1}^{n} Y_t \cdot X_{t,k}}{\sum_{t=1}^{n} Y_t}$$

$$p_{0,k} = p_{Sc,k} = P(X_k = 1 \mid Y = 0) = \frac{\sum_{t=1}^{n}(1-Y_t) \cdot X_{t,k}}{\sum_{t=1}^{n}(1-Y_t)}$$

Test likelihood

A. Giovanidis 2022

☞ How does Naive Bayes work in 2 classes ('History'-'Science'), for a **Test sample** (X_o, y_o)?

- We make use of the estimated likelihood!
- Suppose the distribution for each feature k per class j is $\mathrm{Bernoulli}(p_{j,k})$ and independent of other features.

For each feature the test-data likelihood is:

$$L(p_{j,k}; X_{o,k}) = p_{j,k}^{X_{o,k}}(1 - p_{j,k})^{1-X_{o,k}}, \qquad \textit{for class } j \in \{0, 1\}$$

Posteriors

A. Giovanidis 2022

- Then for ML posteriors (with feature independence):

$$P(Y=0 \mid X_o) \propto P(X_o \mid Y=0) = \prod_{k=1}^{K} p_{0,k}^{X_{o,k}} (1-p_{0,k})^{1-X_{o,k}},$$

$$P(Y=1 \mid X_o) \propto P(X_o \mid Y=1) = \prod_{k=1}^{K} p_{1,k}^{X_{o,k}} (1-p_{1,k})^{1-X_{o,k}}$$

- For MAP posteriors we need also the Prior distribution over classes, i.e. $p_0 = P(Y=0)$ and $p_1 = P(Y=1)$,

$$P(Y=j \mid \mathcal{D}_n) = P(\mathcal{D}_n \mid Y=j) \cdot P(Y=j), \;\; j \in \{0,1\}.$$

Example

A. Giovanidis 2022

Calculate the Naive Bayes classification for the following example:

	Bag-Of-Words				Label	
	1:'king'	2:'food'	3:'equals'	4:'proof'	History	Science
Text 1	No	Yes	Yes	Yes	No	Yes
Text 2	No	No	Yes	No	No	Yes
Text 3	Yes	Yes	No	Yes	Yes	No
Text 4	Yes	No	Yes	Yes	No	Yes
Test	No	No	Yes	Yes	??	??

Naive Bayes with continuous features

A. Giovanidis 2022

✐ Suppose that X contains K continuous state features.

- Suppose the distribution for each feature k per class j is Gaussian $\mathcal{N}(\mu_{j,k}, \sigma^2_{j,k})$.
- Prior distribution over classes, is assumed uniform, i.e. $P(Y = 0) = P(Y = 1) = 0.5$, non-uniform arbitrary, or estimated from dataset (as before).

Continuous estimates

A. Giovanidis 2022

- ML estimates for mean.

$$n_1 = \sum_{t \in \mathcal{D}_n} \mathbf{1}(Y_t = 1), \qquad \overline{X}_{1,k} = \frac{1}{n_1} \sum_{t \in \mathcal{D}_n, Y_t = 1} X_{t,k},$$

$$n_0 = \sum_{t \in \mathcal{D}_n} \mathbf{1}(Y_t = 0), \qquad \overline{X}_{0,k} = \frac{1}{n_0} \sum_{t \in \mathcal{D}_n, Y_t = 0} X_{t,k}$$

- ML estimates for variance.

$$\overline{S}^2_{1,k} = \frac{1}{n_1 - 1} \sum_{t \in \mathcal{D}_n, Y_t = 1} (X_{t,k} - \overline{X}_{1,k})^2,$$

$$\overline{S}^2_{0,k} = \frac{1}{n_0 - 1} \sum_{t \in \mathcal{D}_n, Y_t = 0} (X_{t,k} - \overline{X}_{0,k})^2.$$

ML and MAP estimators

A. Giovanidis 2022

Given a Test sample (X_o, y_o), the estimated class is the one which maximizes the Likelihood (ML) estimator, i.e. the maximum between

$$P(Y=0 \mid \mathcal{D}_n) \propto \prod_{k=1}^{K} \frac{1}{(2\pi \overline{S}_{0,k}^2)^{1/2}} \exp\left(-\frac{(X_{o,k} - \overline{X}_{0,k})^2}{2\overline{S}_{0,k}^2}\right) \quad \textit{for} \quad \textit{Class } 0$$

$$P(Y=1 \mid \mathcal{D}_n) \propto \prod_{k=1}^{K} \frac{1}{(2\pi \overline{S}_{1,k}^2)^{1/2}} \exp\left(-\frac{(X_{o,k} - \overline{X}_{1,k})^2}{2\overline{S}_{1,k}^2}\right) \quad \textit{for} \quad \textit{Class } 1$$

and similarly as in the discrete case for MAP estimators

$$P(Y=j \mid \mathcal{D}_n) \quad = \quad P(\mathcal{D}_n \mid Y=j) \cdot P(Y=j), \quad j \in \{0,1\}\,.$$

Linear Discriminant Analysis (LDA)

A. Giovanidis 2022

For classification of two or multiple classes, we often use LDA:

- Instead of modelling $Pr\,(Y = j|X = x)$ directly as in Logistic Regression, it does this indirectly by modelling $Pr\,(X = x|Y = j)$ (Likelihood again!).
- Makes use of the Bayes' Theorem and the Bayes classifier.
- Assumes the distribution of X's is approximately Gaussian.

☞ The class boundaries are linear, as in Logistic Regression.

> Applies to **continuous** feature variables $X_n = (X_{n,1}, \ldots, X_{n,K})$

Bayes' Theorem in Classification

A. Giovanidis 2022

We want to calculate the conditional probability for each class

$$
\begin{aligned}
Pr\left(Y=j|X=x\right) &\overset{Bayes'}{=} \frac{Pr\left(X=x|Y=j\right)Pr\left(Y=j\right)}{Pr\left(X=x\right)} \\
&\overset{Total}{=} \frac{Pr\left(X=x|Y=j\right)Pr\left(Y=j\right)}{\sum_{m=1}^{M} Pr\left(X=x|Y=m\right)Pr\left(Y=m\right)} \\
&= \frac{f_j(x)\cdot\pi_j}{\sum_{m=1}^{M} f_m(x)\cdot\pi_m}
\end{aligned}
$$

☞ We need the conditional probability of X given the class, and the frequency of each class.

Bayes' Theorem in Classification

A. Giovanidis 2022

We want to calculate the conditional probability for each class

$$
\begin{aligned}
Pr\left(Y=j|X=x\right) &\overset{Bayes'}{=} \frac{Pr\left(X=x|Y=j\right)Pr\left(Y=j\right)}{Pr\left(X=x\right)} \\
&\overset{Total}{=} \frac{Pr\left(X=x|Y=j\right)Pr\left(Y=j\right)}{\sum_{m=1}^{M} Pr\left(X=x|Y=m\right)Pr\left(Y=m\right)} \\
&= \frac{f_j(x)\cdot\pi_j}{\sum_{m=1}^{M} f_m(x)\cdot\pi_m}
\end{aligned}
$$

☞ We need the conditional probability of X given the class, and the frequency of each class.

☞ Given these, we can choose for $X = x_o$, the class with $\max_{1\leq j\leq M} Pr\left(Y=j|X=x_o\right)$ (Bayes classifier).

LDA for 1 predictor $K = 1$

A. Giovanidis 2022

We can **assume** that $f_j(x)$ is normal or Gaussian.

- For $K = 1$ feature:

$$f_j(x) = \frac{1}{\sigma_j\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma_j^2}(x-\mu_j)^2\right),$$

  μ_j and σ_j^2 are the mean and variance for the j-th class.

- Let us further assume that $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_M^2 = \sigma^2$, hence there is a shared variance among all classes.
- The π_j's are also called prior probabilities.

Q: Is the gaussian assumption reasonable?

LDA ($K = 1$)

A. Giovanidis 2022

Plugging in (1), we get:

$$Pr\left(Y = j|X = x\right) \quad = \quad \frac{\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu_j)^2\right) \cdot \pi_j}{\sum_{m=1}^{M} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu_m)^2\right) \cdot \pi_m}$$

Unknowns:

- prior probabilities π_m,
- means μ_m, $m = 1, \ldots, M$, and
- common variance σ.

LDA ($K = 1$) classification

A. Giovanidis 2022

Let us take the log in the above expression.

We assign for $X = x$, the class m^* such that

$$
\begin{aligned}
m^* &= \arg \max_{1 \leq m \leq M} Pr\left(Y = m | X = x\right) \\
&= \arg \max_{1 \leq m \leq M} \log Pr\left(Y = m | X = x\right) \\
&= \arg \max_{1 \leq m \leq M} \left\{ x \cdot \frac{\mu_m}{\sigma^2} - \frac{\mu_m^2}{2\sigma^2} + \log(\pi_m) \right\} \\
&= \arg \max_{1 \leq m \leq M} \{ x \cdot c_{1,m} + c_{0,m} \} \quad (\textit{linear}!)
\end{aligned}
$$

Estimating the decision function

A. Giovanidis 2022

For each class m we have the linear discriminant function of x:

$$\delta_m(x) = x \cdot \frac{\mu_m}{\sigma^2} - \frac{\mu_m^2}{2\sigma^2} + \log(\pi_m),$$

and to calculate it from the dataset D_n we use the estimates:

$$\hat{\mu}_m = \frac{1}{n_m} \sum_{t:y_t=m} x_t,$$

$$\hat{\sigma}^2 = \frac{1}{n-M} \sum_{m=1}^{M} \sum_{t:y_t=m} (x_t - \hat{\mu}_m)^2,$$

$$\hat{\pi}_m = \frac{n_m}{n}.$$

2-class example

A. Giovanidis 2022

For $M = 2$ classes, suppose $\pi_1 = \pi_2$ additionally (uniform prior).
Then the discriminant functions become:

$$\begin{aligned}\delta_1(x) &= x \cdot \frac{\mu_1}{\sigma^2} - \frac{\mu_1^2}{2\sigma^2} + \log(\pi_1)\\ \delta_2(x) &= x \cdot \frac{\mu_2}{\sigma^2} - \frac{\mu_2^2}{2\sigma^2} + \log(\pi_2)\end{aligned}$$

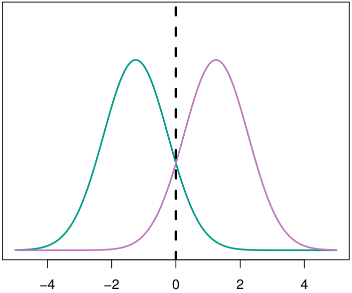
so that x is assigned class 1, if $\delta_1(x) > \delta_2(x)$ or,

$$2x(\mu_1 - \mu_2) > \mu_1^2 - \mu_2^2$$

The decision boundary are the points x, s.t.

$$x = \frac{\mu_1 + \mu_2}{2}.$$

A. Giovanidis 2022



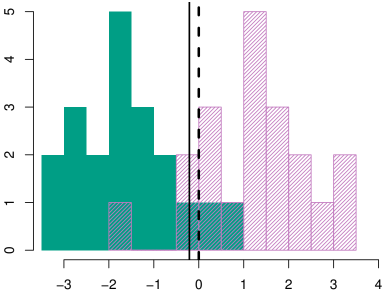


Figure: Two normal density functions and decision boundary. [1]

[1] Source [B.1]

LDA for $K > 1$ dimensions

A. Giovanidis 2022

How does the LDA perform, when the predictors X have more than 1 dimension? say $X = (X_1, \dots, X_K)$.

☞ Assume a multivariate Gaussian distribution instead of a 1-dimensional $X \sim \mathcal{N}(\mu, \mathbf{\Sigma})$.

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

- mean $\mu = (\mu_1, \dots, \mu_K)^T$,
- common covariance matrix $\mathbf{\Sigma}$.

Bivariate Gaussian distribution

A. Giovanidis 2022

Two random variables X and Z are said to have a **bivariate Gaussian distribution** with parameters

$$\mu_X,\ \sigma_X^2,\ \mu_Z,\ \sigma_Z^2,\ \rho,$$

if their joint PDF is given by

$$f_{XZ}(x,z) = \frac{1}{2\pi\sigma_X\sigma_Z\sqrt{1-\rho^2}} \cdot$$

$$\cdot \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 + \left(\frac{z-\mu_Z}{\sigma_Z}\right)^2 - 2\rho\frac{(x-\mu_X)(z-\mu_Z)}{\sigma_X\sigma_Z}\right]\right\},$$

where $\rho \in (-1,1)$ the correlation coefficient $\rho = \frac{Cov(X,Z)}{\sqrt{Var(X)Var(Z)}} = \frac{\sigma_{XZ}}{\sigma_X\sigma_Z}$.

Matrix form (general)

A. Giovanidis 2022

For any number of features $K > 1$

$$f_{\mathbf{X}}(X_1 = x_1, \ldots, X_K = x_K) \quad = \quad \frac{\exp\left(-\frac{1}{2}(\mathbf{x} - \mu)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \mu)\right)}{\sqrt{(2\pi)^K |\boldsymbol{\Sigma}|}},$$

- The covariance matrix $\boldsymbol{\Sigma}$ should be positive definite.
- The mean vector $\mu = (\mu_1, \ldots, \mu_K)^T$.
- In the case $K = 2$ with $(X_1, X_2) = (X, Z)$

$$\mu = (\mu_X, \mu_Z)^T, \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Z \\ \rho\sigma_X\sigma_Z & \sigma_Z^2 \end{pmatrix}$$

Bivariate properties

A. Giovanidis 2022

Property 1
Suppose X and Z follow a bivariate Gaussian distribution.
Then, given $X = x$, the variable Z is Gaussian distributed, with

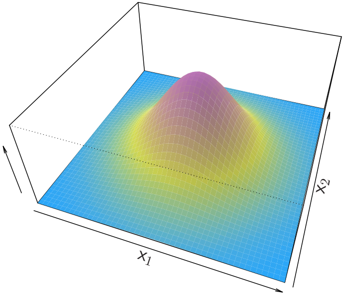
$$\begin{aligned} \mathbb{E}[Z \mid X = x] &= \mu_Z + \rho\sigma_Z \frac{x - \mu_X}{\sigma_X}, \\ Var(Z \mid X = x) &= (1 - \rho^2)\sigma_Z^2. \end{aligned}$$

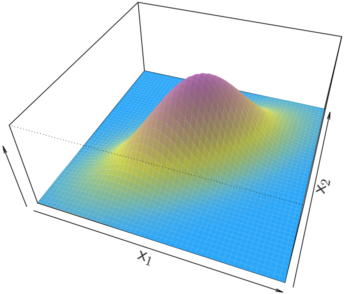
Property 2
Suppose X and Z follow a bivariate Gaussian distribution.
Then, if $X,\ Z$ are uncorrelated $\rho = 0$ they are independent.

$$\mu = (\mu_X, \mu_Z)^T, \qquad \mathbf{\Sigma} = \begin{pmatrix} \sigma_X^2 & 0 \\ 0 & \sigma_Z^2 \end{pmatrix}$$

A. Giovanidis 2022

Example bivariate





In the following we will make use of the expressions

$$\begin{aligned} |\mathbf{\Sigma}| &= (1-\rho^2)\sigma_X^2\sigma_Z^2, \\ \mathbf{\Sigma}^{-1} &= \frac{1}{|\mathbf{\Sigma}|}\begin{pmatrix} \sigma_Z^2 & -\rho\sigma_X\sigma_Z \\ -\rho\sigma_X\sigma_Z & \sigma_X^2 \end{pmatrix} \end{aligned}$$

Linear discriminant function (general)

A. Giovanidis 2022

Linear Discriminant Function for K features:

$$\delta_j(\mathbf{x}) \quad = \quad \mathbf{x}^T \mathbf{\Sigma}^{-1}\mu_j - \frac{1}{2}\mu_j^T \mathbf{\Sigma}^{-1}\mu_j + \log(\pi_j)$$

Q: Is it linear? Check for $K = 2$.

A. Giovanidis 2022

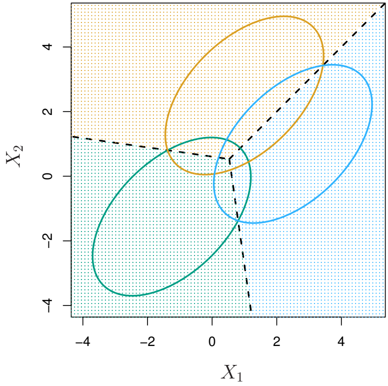


Figure: Classification for $M = 3$ classes and $K = 2$ dimensions. [2]

[2] Source [B.1]

Quadratic Discriminant Analysis (QDA)

A. Giovanidis 2022

LDA assumed for each class a different vector for the feature-means μ_j and same covariance matrix $\mathbf{\Sigma}$.

☞ QDA assumes different covariance matrix per class. That is, an observation from the j-th class is of the form $X \sim \mathcal{N}(\mu_j, \mathbf{\Sigma}_j)$.

Quadratic Discriminant Function:

$$\begin{aligned} \delta_j(\mathbf{x}) &= -\frac{1}{2}\mathbf{x}^T \mathbf{\Sigma}_j^{-1}\mathbf{x} + \mathbf{x}^T \mathbf{\Sigma}_j^{-1}\mu_j - \frac{1}{2}\mu_j^T \mathbf{\Sigma}_j^{-1}\mu_j - \\ &\quad -\frac{1}{2}\log|\mathbf{\Sigma}_j| + \log(\pi_j) \end{aligned}$$

QDA is more flexible than LDA: Bias vs Variance tradeoff !

QDA examples

A. Giovanidis 2022

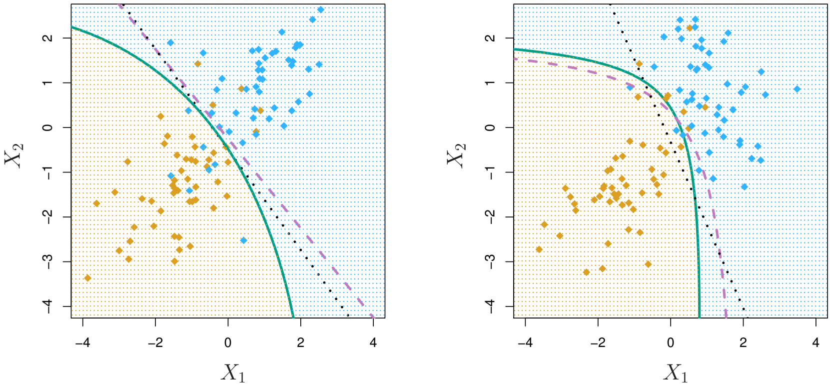


Figure: (left:) Truth common $\mathbf{\Sigma}$, (right:) Truth different $\mathbf{\Sigma}_1$, $\mathbf{\Sigma}_2$. [3]

[3] Source [B.1]

Method comparison: linear

A. Giovanidis 2022

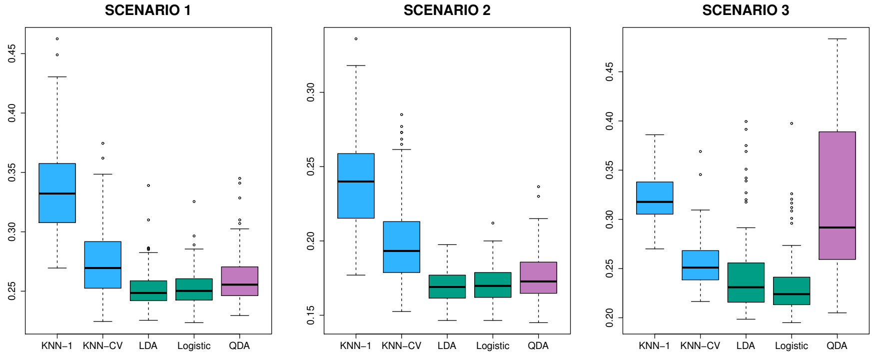


Figure: (1) uncorr., $\mathcal{N}$, $\mu_1 \neq \mu_2$, (2) corr., $\mathcal{N}$, (3) uncorr., t-distr.[4]

[4] Source [B.1]

Method comparison: non-linear

A. Giovanidis 2022

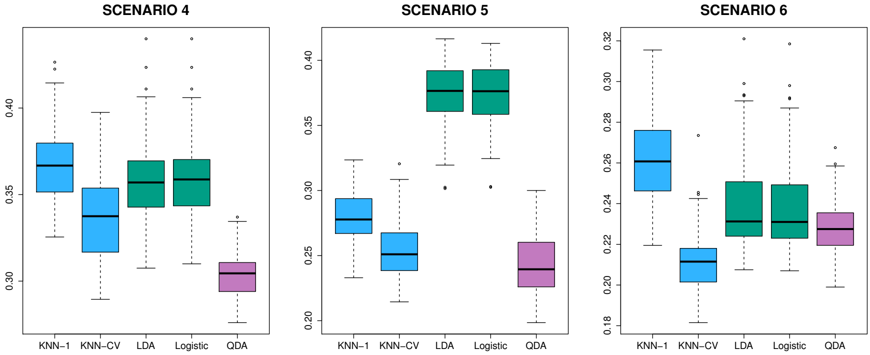


Figure: (4) corr. $\mathcal{N}$, $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$, (5) X_1^2, X_2^2, X_1X_2 (6) more-NL. [5]

[5] Source [B.1]

A. Giovanidis 2022

END