Impact of OVL Variation on AUC Bias Estimated by Non-parametric Methods

Carina Silva[1,3]
Antónia Turkman[2,3]
Lisete Sousa[2,3]

[1]Escola Superior de Tecnologia da Saúde de Lisboa -IPL
[2]Faculdade de Ciências da Universidade de Lisboa
[3]Centro de Estatística e Aplicações, Universidade de Lisboa

ESCOLA SUPERIOR DE TECNOLOGIA DA SAÚDE DE LISBOA
INSTITUTO POLITÉCNICO DE LISBOA

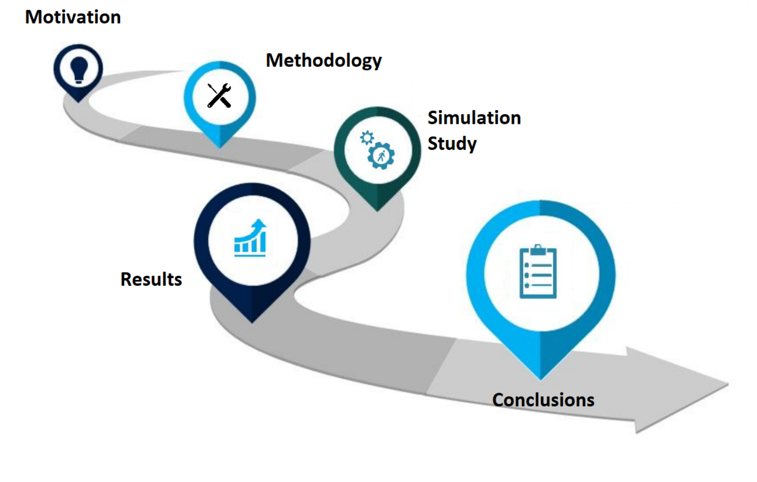
CEAUL
Centro de Estatística e Aplicações
Universidade de Lisboa

FCT
Fundação para a Ciência e a Tecnologia

This work is partially financed by national funds through FCT Fundação para a Ciência e a Tecnologia under the project UIDB/00006/2020

Roadmap

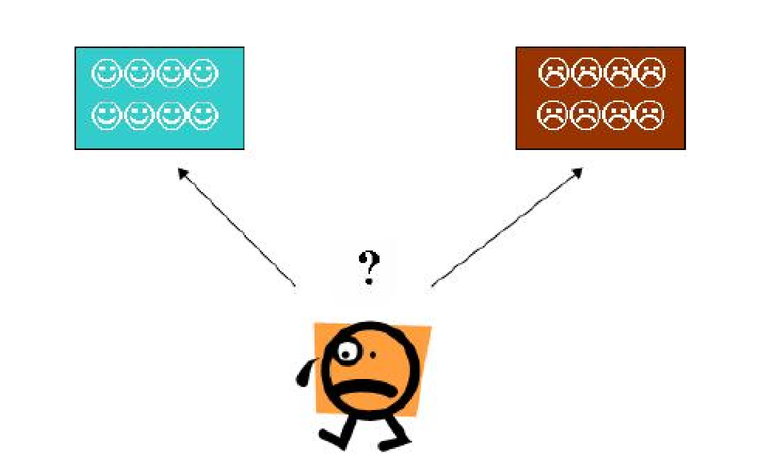
Roadmap



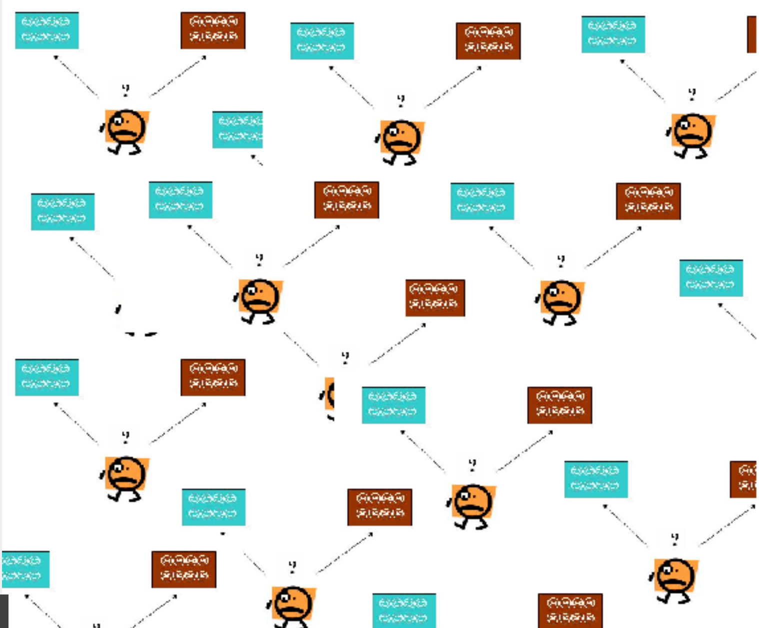
Motivation
Methodology
Simulation
Study
Results
Conclusions

Motivation

Motivation



Motivation



Arrow plot

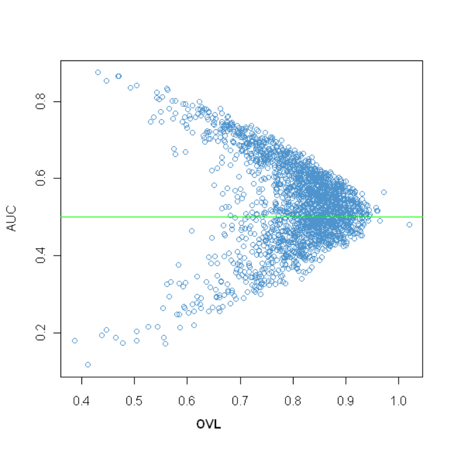
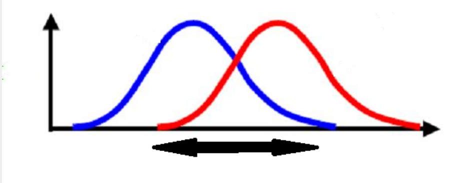
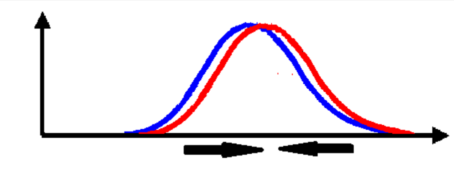


Figure: Silva-Fortes *et al.* (2012)

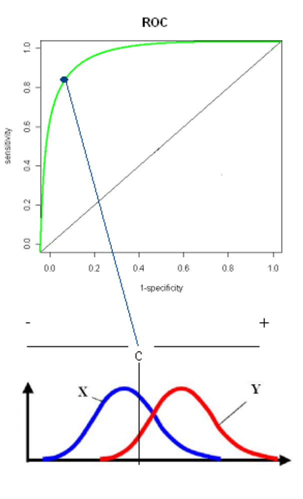
ROC (Receiver Operating Characteristic)

- Evaluates the accuracy of a binary classification system.



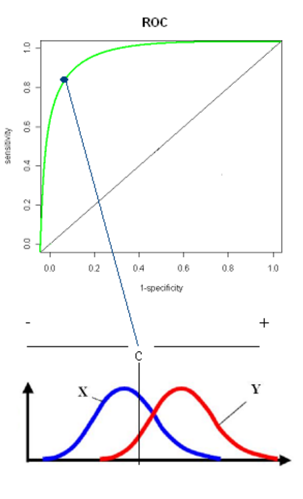


ROC curve

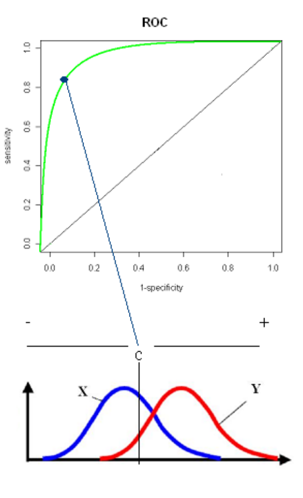


- The ROC curve results from the relationship between the proportion of true positives (sensitivity) and proportion of false positives (1-specificity) obtained for each cut-off point of the variable of decision.

ROC curve



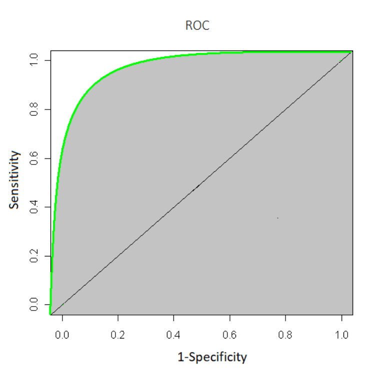
- The ROC curve results from the relationship between the proportion of true positives (sensitivity) and proportion of false positives (1-specificity) obtained for each cut-off point of the variable of decision.
- These proportions depend from the classification rule.



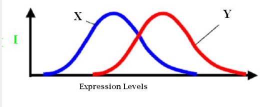
- The ROC curve results from the relationship between the proportion of true positives (sensitivity) and proportion of false positives (1-specificity) obtained for each cut-off point of the variable of decision.
- These proportions depend from the classification rule.
- Traditionally high values of the decision variable, correspond to the presence of the artifact of interest.

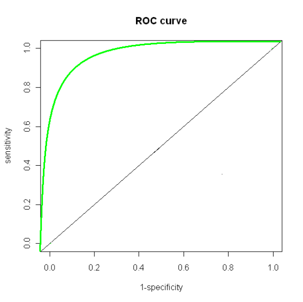
AUC - Area Under Curve

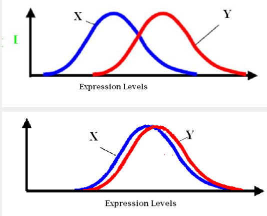
- Index of global accuracy evaluation.
- AUC $\in$ [0.5,1].

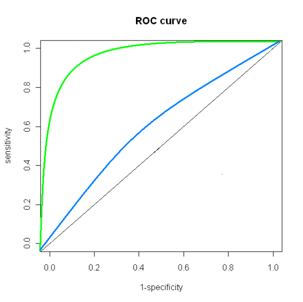


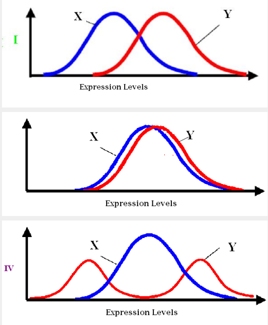
ROC curve in the Arrow plot



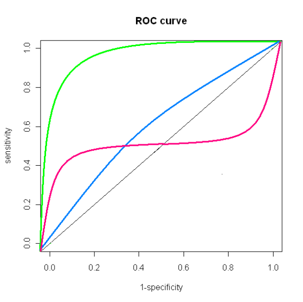



X
Y
I
Expression Levels
X
Y
Expression Levels


ROC curve
sensitivity
1-specificity
0.0
0.2
0.4
0.6
0.8
1.0

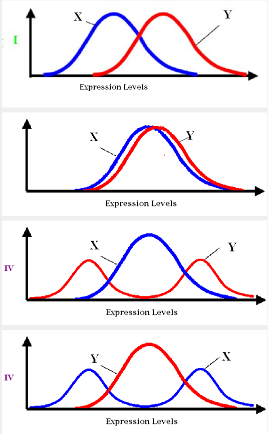


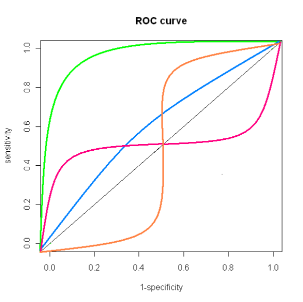
I
X
Y
Expression Levels
X
Y
Expression Levels
IV
X
Y
Expression Levels



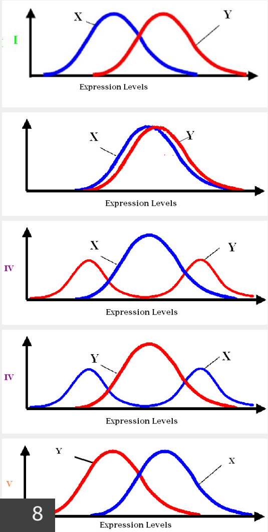
ROC curve
sensitivity
1-specificity
0.0
0.2
0.4
0.6
0.8
1.0

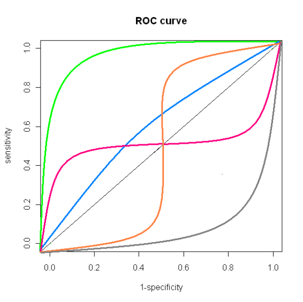
ROC curve in the Arrow plot



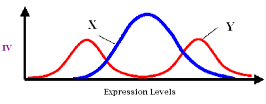


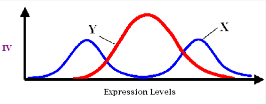
ROC curve in the Arrow plot





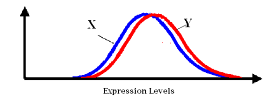
Special Genes



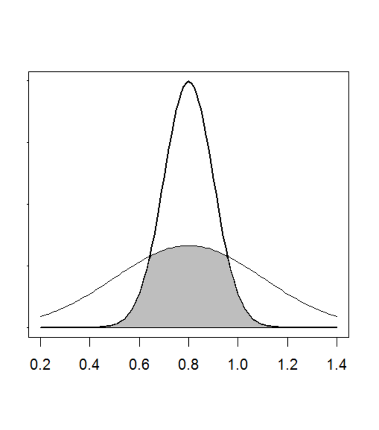


AUC$\approx$ 0.5.

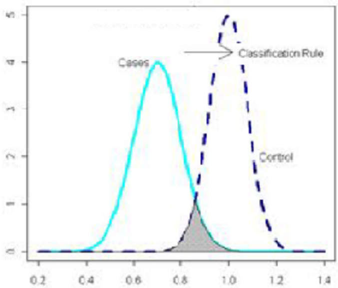
Genes non DE

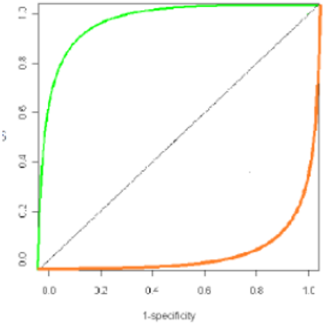


- OVL is the common area shared by the two densities, Weitzman (1970).
- $OVL(X, Y) = \int_{-\infty}^{+\infty} min[f_X(c), g_Y(c)]dc$
- OVL $\in [0, 1]$



OVL *vs.* AUC



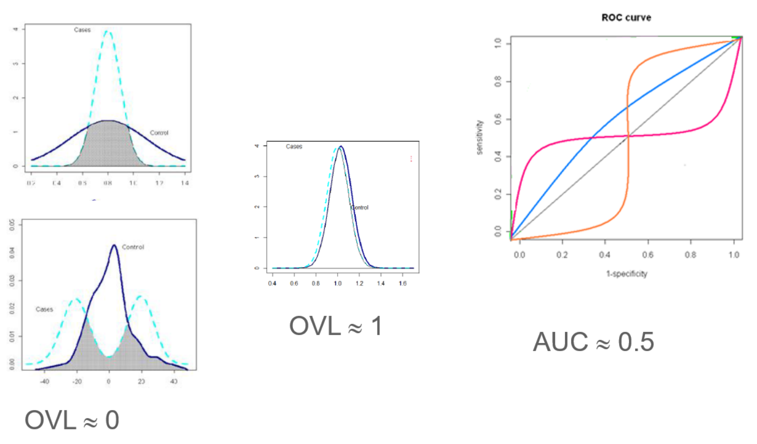


OVL ⬇

AUC ⬆

AUC < 0.5 (1-AUC⬆)

OVL *vs.* AUC



Cases
Control
OVL ≈ 0
Cases
Control
OVL ≈ 1
ROC curve
sensitivity
1-specificity
AUC ≈ 0.5

Arrow Plot

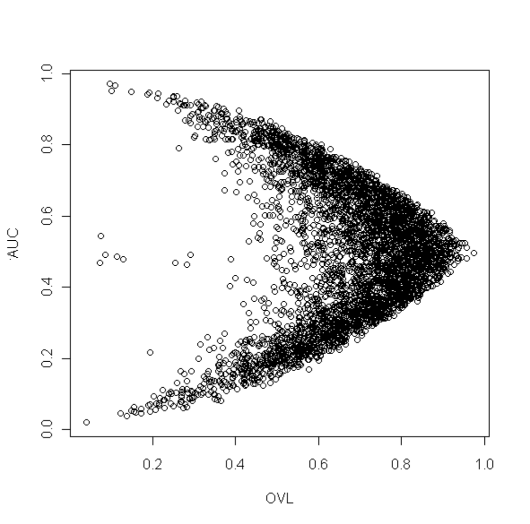


Figure: Silva-Fortes *et al.* (2012)

Arrow Plot

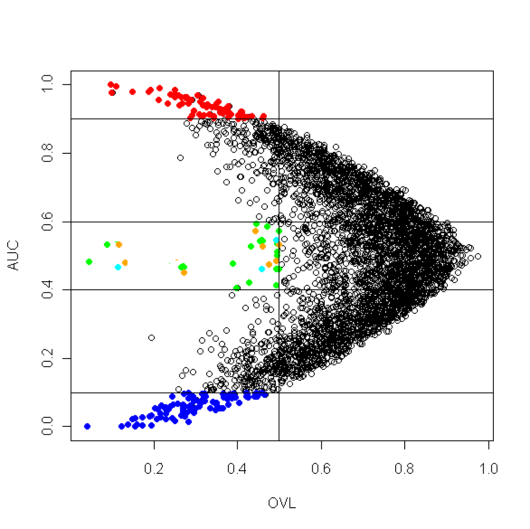
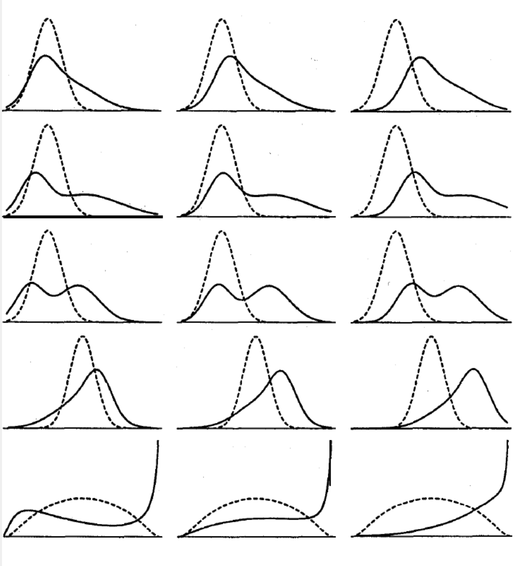


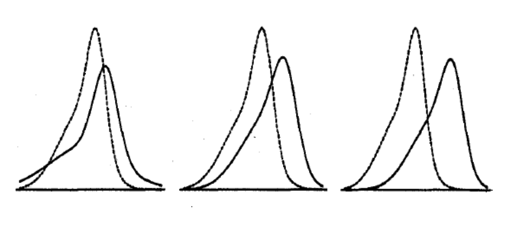
Figure: Silva-Fortes *et al.* (2012)

Methodology

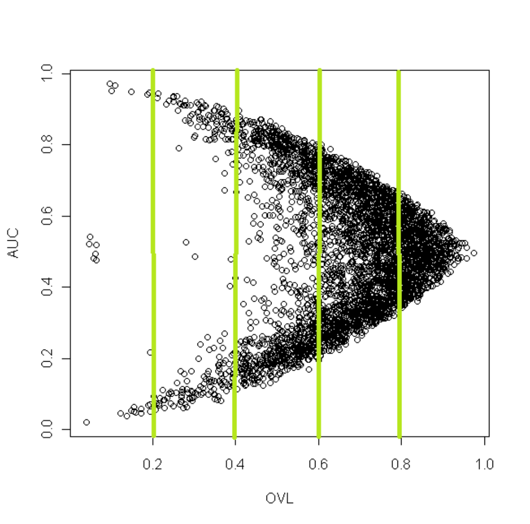
Simulation Study



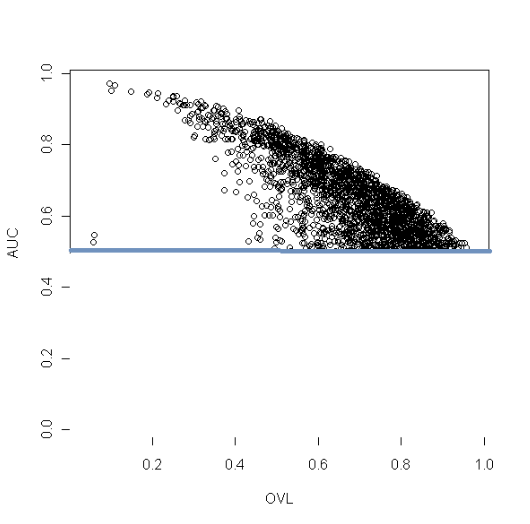
Simulation Study



Simulation Study



Simulation Study



Simulation Study

Scenario	Control	Experimental	OVL	AUC
Bi-normal	$N(0,1)$	$N(0,0.1)$	0.2	0.5
μ **fixed**	$N(0,1)$	$N(0,0.4)$	0.4	0.5
	$N(0,1)$	$N(0,2.4)$	0.6	0.5
	$N(0,1)$	$N(0,1.5)$	0.8	0.5
Bi-normal	$N(0,1)$	$N(2.55,1)$	0.2	0.96
σ **fixed**	$N(0,1)$	$N(1.65,1)$	0.4	0.89
	$N(0,1)$	$N(1.04,1)$	0.6	0.77
	$N(0,1)$	$N(0.5,1)$	0.8	0.64
Bi-Lognormal	$LN(0,1)$	$LN(0,0.1)$	0.2	0.5
	$LN(0,1)$	$LN(1.65,1)$	0.4	0.87
	$LN(0,1)$	$LN(1.04,1)$	0.6	0.77
	$LN(0,1)$	$LN(0.5,1)$	0.8	0.64
Bi-Exponential	$Exp(1)$	$Exp(0.05)$	0.2	0.95
	$Exp(1)$	$Exp(0.15)$	0.4	0.87
	$Exp(1)$	$Exp(0.32)$	0.6	0.76
	$Exp(1)$	$Exp(0.58)$	0.8	0.63

- Consider f_X e g_Y the probability density functions (PDF) associated to the controls and experimental condition respectively;

- Consider f_X e g_Y the probability density functions (PDF) associated to the controls and experimental condition respectively;
- F_X e G_Y their related distribution functions;

- Consider f_X e g_Y the probability density functions (PDF) associated to the controls and experimental condition respectively;
- F_X e G_Y their related distribution functions;
- Each cut-off point t defines a binary classification rule and $\mathrm{F}_X(t) > G_Y(t)$.

- For each scenario, samples of equal dimensions were simulated in the two conditions for $n = 15, 30, 50, 100$.

- **Empirical**

The empirical distribution functions F_X e F_Y are given respectively by:

$$\widehat{F}_X(t) = \frac{1}{n_0} \sum_{j=1}^{n_0} I[X_j \leq t], \tag{1}$$

$$\widehat{F}_Y(t) = \frac{1}{n_1} \sum_{k=1}^{n_1} I[Y_k \leq t], \tag{2}$$

where I is the indicator function.
The empirical estimator of the AUC corresponds to the Mann-Whitney statistic (McNeil e Hanley, 1984):

$$\widehat{\mathrm{AUC}} = \frac{1}{n_0 n_1} \sum_{j=1}^{n_0} \sum_{k=1}^{n_1} \left(I[X_j < Y_k] + \frac{1}{2} I[X_j = Y_k] \right). \tag{3}$$

- **Kernel**

$\widetilde{f}_X$ and $\widetilde{f}_Y$ represent the kernel estimators of f_X e f_Y:

$$\widetilde{f}_X(t) = \frac{1}{n_0 h_0} \sum_{i=1}^{n_0} K\left(\frac{t - X_i}{h_0}\right), \tag{4}$$

$$\widetilde{f}_Y(t) = \frac{1}{n_1 h_1} \sum_{j=1}^{n_1} K\left(\frac{t - Y_j}{h_1}\right). \tag{5}$$

Lloyd (1997) showed that when a Gaussian kernel is considered, the AUC is estimated as:

$$\widetilde{\mathrm{AUC}} = \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} \Phi\left(\frac{Y_j - X_i}{\sqrt{h_0^2 + h_1^2}}\right). \tag{6}$$

- **Estimation methods for the bandwith** h
- [$BW = nrd0$] Silverman (1992) [12] considers the following expression as an optimal bandwidth when the kernel is Gaussian:

$$h = \left(\frac{4}{3}\right)^{\frac{1}{5}} \times \min\left(s, \frac{R}{1.34}\right) n^{-\frac{1}{5}}, \tag{7}$$

where R is the interquartile range and s the empirical standard deviation.
- [$BW = nrd$] Scott (1992) considers the following expression when a Gaussian kernel is used.:

$$h = 1.06 \times sn^{-\frac{1}{5}}, \tag{8}$$

- [$BW = SJ$] Sheather e Jones (1991) proposed the *plug-in* method *solve-the-equation* for the optimal bandwidth.

Parametric *Bootstrap* estimates

- It was considered B=1000 bootstrap replicates in each scenario.

$$\widehat{\text{AUC}}_B = \frac{1}{1000} \sum_{i=1}^{1000} \widehat{\text{AUC}}_i^*, \tag{9}$$

where $\widehat{\text{AUC}}_i^*$ is the AUC estimate (empirical or kernel) in each *bootstrap* replicate.

Bootstrap estimator of the standard error (SE) of the AUC

$$\widehat{se}_B(\widehat{\text{AUC}}) = \sqrt{\frac{1}{999}\sum_{i=1}^{1000}\left(\widehat{\text{AUC}}_i^* - \widehat{\text{AUC}}_B\right)^2}. \tag{10}$$

Bootstrap estimator of the bias of the bootstrap AUC

The *bootstrap* estimate of the bias of $\widehat{\text{AUC}}$ is given by:

$$\widehat{\text{viés}}_B(\widehat{\text{AUC}}) = \widehat{\text{AUC}}_B - \text{AUC}, \tag{11}$$

where AUC corresponds to the exact value.

Bootstrap estimator of the root mean squared error (RMSE) of the AUC

$$\widehat{rmse}_B(\widehat{\mathrm{AUC}}) = \sqrt{\frac{1}{1000}\sum_{i=1}^{1000}(\widehat{\mathrm{AUC}}_i^* - \mathrm{AUC})^2}. \tag{12}$$

Scheme of the simulation procedure

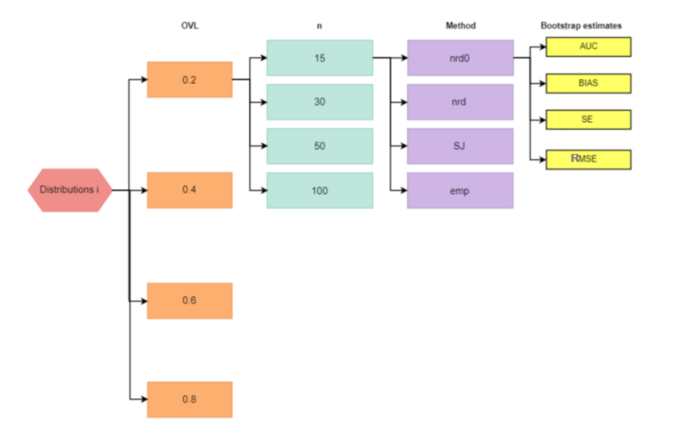


Figure: Silva *et al.* (2020)

Results

Boostrap AUC

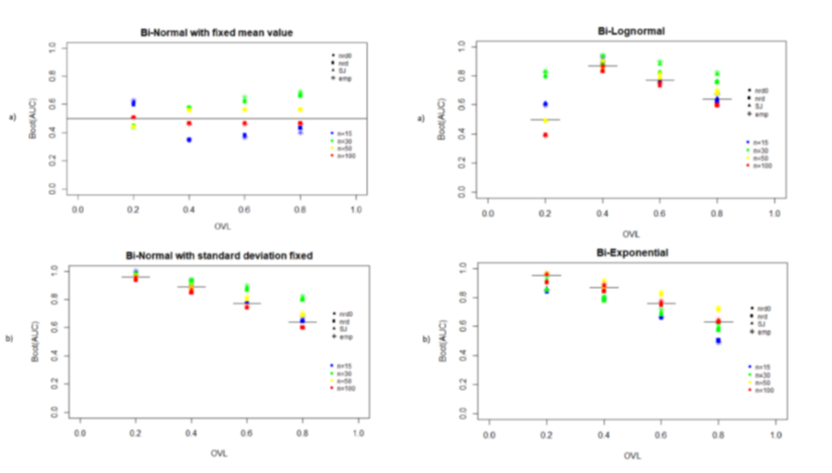


Figure: Silva *et al.* (2020)

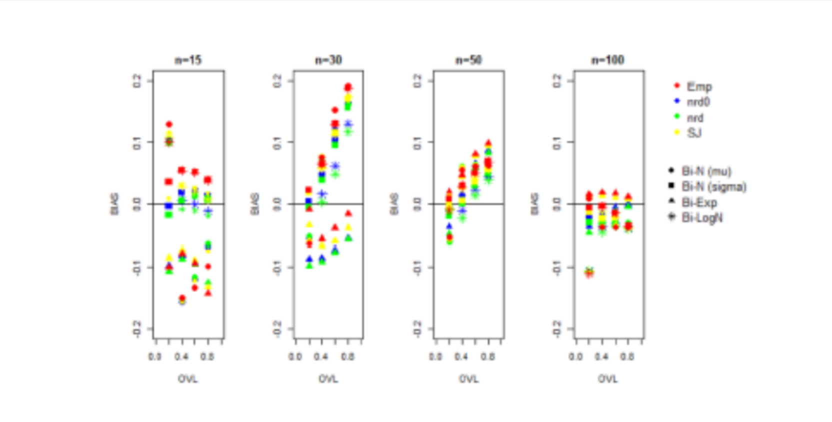
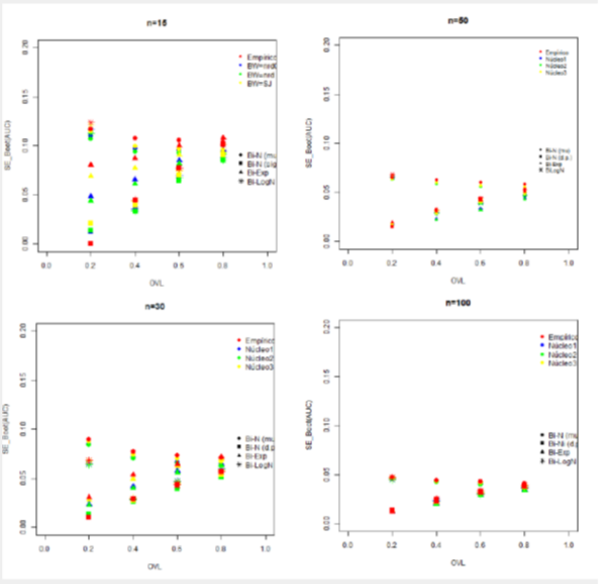


Figure: Silva *et al.* (2020)

SE of AUC


n=15
n=50
n=30
n=100
SE_Boot(AUC)
OVL
Empirico
Núcleo1
Núcleo2
Núcleo3
Bi-Exp
Bi-LogN

Conclusions

- Non-parametric methods for estimating AUC showed similar behaviors both in terms of bias and precision;

Conclusions

- Non-parametric methods for estimating AUC showed similar behaviors both in terms of bias and precision;
- When n increases the bias decreases and the precision increases.

Considering high values of OVL (≥ 0.6)

- Greater variability in the results obtained for bias.

Considering high values of OVL (≥ 0.6)

- Greater variability in the results obtained for bias.
- Less variability in results obtained for accuracy.

Considering high values of OVL (≥ 0.6)

- Greater variability in the results obtained for bias.
- Less variability in results obtained for accuracy.
- For small *n* dimensions there is a tendency to overestimate AUC when Bi-Normal and Bi-Lognormal distributions are considered and an underestimation for Exponential distributions.

Considering low values of OVL (≤ 0.4)

- Tendency for AUC be underestimate.

Considering low values of OVL (≤ 0.4)

- Tendency for AUC be underestimate.
- Less precision when considering Bi-Normal distributions when mean value is fixed.

- Precision is higher for small values of OVL, however this is not true when AUC values are around 0.5 and are obtained from distributions with the same mean value, leading to not proper ROC curves.

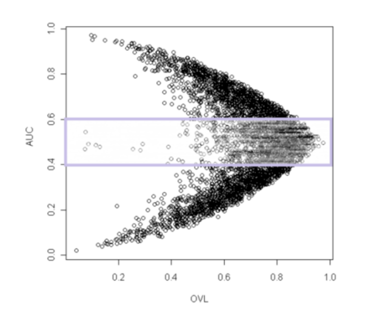


Figure: Silva *et al.* (2020)

Future Work

- Simulations considering discrete distributions.

References

- Lloyd, C. J. (1997). The use of smoothed ROC curves to summarize and compare diagnostic systems. Journal of *American Statistical Association*, 93:1356–1364.
- McNeil, B. J. e Hanley, J. A. (1984). Statistical Approaches to the Analysis of ROC curves. *Medical Decision Making*, 4(2):136–149.
- Silva-Fortes, C., Turkman, M., Sousa, L., 2012. Arrow plot: a new graphical tool for selecting up and down regulated genes and genes differentially expressed on sample subgroups. *BMC Bioinformatics* 13 (1), 147-–148.
- Silverman, B.W. (1998). *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.
- Weitzman, M.L., 1970. *Measures of overlap of income distributions of white and Negro families in the United States*. Vol. 22, US Bureau of the Census.
- Scott, D. (1979). On optimal and data-based histograms. *Biometrika*. 66 (3): 605-–610.
- Sheather, S. J. e Jones, M. C. (1991). A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation. Journal of the Royal Statistical Society: Series B (Methodological), 53(3): 683–690.
- Wald, A. (1947). *Sequential Analysis*. John Wiley Sons, New York.
- Silva C., Turkman M.A.A., Sousa L. (2020). Impact of OVL Variation on AUC Bias Estimated by Non-parametric Methods. In: Gervasi O. et al. (eds) Computational Science and Its Applications – ICCSA 2020. ICCSA 2020. Lecture Notes in Computer Science, vol 12251. Springer, Cham.


Thank you