Online supplementary material to: Optimal classification scores based on multivariate marker transformations

Pablo Martínez-Camblor[1]*, Sonia Pérez-Fernández[2]. Susana Díaz-Coto[2]

[1] Department of Biomedical Data Science, Geisel School of Medicine at Dartmouth, NH, USA

[2] Department of Statistics, Oviedo University, Asturies, Spain

As supplementary material of this paper we provide the `R` code used for computing plots and models reported herein. Main provided function, `optimalT`, incorporates a general k-fold cross-validation procedure for controlling the potential overfitting. `R` packages `nsROC` (developed by Pérez-Fernández et al. [2]) and `ks` (developed by Duong [1]) are required. The used dataset is freely available at `http://archive.ics.uci.edu/ml/datasets/QSAR+fish+toxicity#`. Results of additional simulations are provided in Tables S1, S2 and S3.

R code: description of function `optimalT`

```
optimalT <- function(X, D, H.method = c("Hbcv","Hscv","Hpi","Hns","Hlscv",
                    "Hbcv.diag","Hscv.diag","Hpi.diag","Hlscv.diag"), K = 1,
                    add.densityContour = TRUE, removeNA = FALSE, X1.lim =
                    NULL, X2.lim = NULL, levels.method = c("fpr","pretty"),
                    figures = c("A","B","C"), new.window = TRUE, seed = 623)

## Input parameters:

  # X: marker data (n x 2 matrix)
  # D: response (vector of length n)
  # H.method: method for computing the bandwidth among those proposed by
```

*Pablo Martínez-Camblor. 7 Lebanon Street, Suite 309, Hinman Box 7261, Hanover, NH 03755, USA. E-mail: Pablo.Martinez.Camblor@Dartmouth.edu

```
  Duong (2007) ("Hbcv" by default, which is a biased cross validation estimate)
  # K: number of folds considered for cross-validation. If K = 1 (default),
  optimal transformation is estimated
  # add.densityContour: TRUE if density contours of the bivariate density
  estimates for both populations should be shown over the optimal
  transformation contour plot
  # removeNA: TRUE if the region displayed should be adjusted removing
  NA-values in the optimal transformation (f+g estimate is zero)
  # X1.lim, X2.lim: limits for the region displayed (vectors of length 2)
  # levels.method: "fpr" if the optimal transformation contour levels
  displayed correspond to the sequence 0:0.1:1 of false-positive rates
  for the transformation; "pretty" if the default by filled.contour should
  be considered
  # figures: vector containing "A", "B" and/or "C" indicating which plots
  should be displayed:
   # "A": contour plot for the bivariate kernel density estimate for
positive (red) and negative (blue) populations
   # "B": contour plot for the optimal score estimate
   # "C": ROC curve estimate
  # new.window: TRUE if new windows should be opened for each figure displayed
  # seed: seed used for grouping in K-fold CV

## Output parameters:
  # X: marker data
  # D: response
  # tX: optimal transformation estimate (score) for X
  # auc: Area Under the ROC Curve for the score
  # x.grid, y.grid: grid used for each component of the bivariate marker
  # z.grid: matrix containing the values of the score over the grid
  x.grid * y.grid
  # tX.CV: if K>1, optimal transformation estimate (score) for X resulting
  from the k-fold cross validation performed
```

```
# auc.CV: Area Under the ROC Curve for the score tX.CV
```

Simulation study: additional tables

Table S1: Means for the integrate absolute error (Integ. absolute error) between the real ROC curve, $\mathcal{R}_T(\cdot)$, and its estimation, $\hat{\mathcal{R}}_{\hat{T}_N}(\cdot)$ ($\int_0^1 |\hat{\mathcal{R}}_{\hat{T}_N}(t) - \mathcal{R}_T(t)|dt$) and for the AUC from 2,000 Monte Carlo simulations for the six considered models **without using any cross-validation** procedure. Considered bandwidths were smooth cross-validation (SCV), plug-in (PI), normal scale (NS) and biased cross-validation (BCV). RL stands for model based on standard binary logistic regression.

			AUC					Integ. absolute error				
n	m	$\mathcal{A}$	**SCV**	**PI**	**NS**	**BCV**	**RL**	**SCV**	**PI**	**NS**	**BCV**	**RL**
Model 0												
400	400	0.50	0.637	0.647	0.648	0.637	**0.526**	0.138	0.148	0.149	0.139	**0.030**
	600	0.50	0.628	0.637	0.638	0.628	**0.523**	0.129	0.138	0.139	0.129	**0.026**
Model I												
400	400	0.70	**0.729**	0.733	0.733	**0.729**	0.526	**0.031**	0.034	0.034	**0.031**	0.174
	600	0.70	0.725	0.729	0.729	**0.724**	0.522	0.027	0.030	0.030	**0.026**	0.178
400	400	0.80	0.818	0.820	0.820	**0.815**	0.526	0.020	0.022	0.022	**0.019**	0.276
	600	0.80	0.816	0.818	0.818	**0.812**	0.522	0.018	0.019	0.019	**0.016**	0.280
Model II												
400	400	0.75	**0.768**	0.771	0.771	**0.768**	0.526	**0.025**	0.027	0.027	**0.025**	0.223
	600	0.75	0.764	0.766	0.766	**0.763**	0.523	0.022	0.023	0.023	**0.022**	0.227
400	400	0.80	0.814	0.816	0.816	**0.805**	0.526	0.022	0.023	0.023	**0.018**	0.272
	600	0.80	0.811	0.812	0.812	**0.800**	0.522	0.019	0.020	0.020	**0.017**	0.277
Model III												
400	400	0.70	0.727	0.731	0.731	0.726	**0.691**	0.033	0.036	0.036	0.032	**0.022**
	600	0.70	0.725	0.729	0.729	0.724	**0.693**	0.031	0.034	0.034	0.030	**0.020**
400	400	0.80	0.815	0.818	0.818	0.815	**0.797**	0.023	0.024	0.024	0.022	**0.019**
	600	0.80	0.816	0.818	0.818	0.815	**0.798**	0.022	0.023	0.023	0.021	**0.017**
Model IV												
400	400	0.75	0.776	0.780	0.780	**0.775**	0.722	0.031	0.034	0.034	**0.030**	0.034
	600	0.75	0.773	0.776	0.777	**0.772**	0.724	0.028	0.030	0.030	**0.027**	0.031
400	400	0.85	0.865	0.867	0.867	0.864	**0.843**	0.022	0.024	0.024	0.022	**0.018**
	600	0.85	0.865	0.867	0.867	0.864	**0.845**	0.021	0.023	0.023	0.021	**0.017**
Model V												
400	400	0.75	0.770	0.775	0.761	**0.752**	0.641	0.025	0.030	0.020	**0.017**	0.108
	600	0.75	0.766	0.771	0.756	**0.748**	0.640	0.022	0.026	0.017	**0.016**	0.109
400	400	0.85	0.866	0.869	0.861	**0.856**	0.815	0.023	0.025	0.019	**0.017**	0.036
	600	0.85	0.865	0.868	0.859	**0.855**	0.815	0.021	0.023	0.018	**0.016**	0.035
Model VI												
400	400	0.80	0.816	0.819	0.816	0.812	**0.789**	0.022	0.023	0.022	**0.020**	**0.020**
	600	0.80	0.815	0.818	0.815	0.812	**0.791**	0.020	0.022	0.020	**0.019**	**0.019**
400	400	0.85	0.861	0.863	0.861	0.858	**0.843**	0.018	0.020	0.018	0.017	**0.016**
	600	0.85	0.861	0.863	0.861	0.858	**0.845**	0.017	0.018	0.017	0.016	**0.015**

Table S2: Means for the integrate absolute error (Integ. absolute error) between the real ROC curve, $\mathcal{R}_T(\cdot)$, and its estimation, $\hat{\mathcal{R}}_{\hat{T}_N}(\cdot)$ ($\int_0^1 |\hat{\mathcal{R}}_{\hat{T}_N}(t) - \mathcal{R}_T(t)|dt$) and for the AUC from 2,000 Monte Carlo simulations for the six considered models **by using 2-fold cross-validation** procedure. **Small sample sizes, n and m for positive and negative groups respectively, were considered.** $\mathcal{A}$ is the real AUC. Considered bandwidths were smooth cross-validation (SCV), plug-in (PI), normal scale (NS) and biased cross-validation (BCV). RL stands for model based on standard binary logistic regression.

			AUC					Integ. absolute error				
n	m	$\mathcal{A}$	**SCV**	**PI**	**NS**	**BCV**	**RL**	**SCV**	**PI**	**NS**	**BCV**	**RL**
Model 0												
100	100	0.50	0.496	0.497	0.497	0.497	0.484	0.051	0.052	0.053	0.052	0.050
	200	0.50	0.496	0.495	0.495	0.495	0.478	0.046	0.046	0.046	0.046	0.046
Model I												
100	100	0.70	0.656	0.651	0.654	0.632	0.481	0.061	0.064	0.063	0.077	0.216
	200	0.70	0.662	0.658	0.660	0.642	0.484	0.053	0.055	0.054	0.067	0.216
100	100	0.80	0.774	0.771	0.773	0.739	0.478	0.044	0.046	0.045	0.067	0.320
	200	0.80	0.777	0.774	0.776	0.747	0.481	0.037	0.039	0.038	0.059	0.319
Model II												
100	100	0.75	0.713	0.709	0.713	0.716	0.482	0.053	0.056	0.054	0.052	0.264
	200	0.75	0.727	0.722	0.724	0.728	0.483	0.043	0.046	0.044	0.042	0.265
100	100	0.80	0.772	0.768	0.771	0.732	0.480	0.043	0.046	0.044	0.072	0.315
	200	0.80	0.777	0.774	0.776	0.736	0.486	0.040	0.042	0.040	0.067	0.311
Model III												
100	100	0.70	0.650	0.645	0.647	0.651	0.664	0.064	0.068	0.067	0.064	0.053
	200	0.70	0.656	0.652	0.654	0.658	0.672	0.057	0.060	0.058	0.056	0.043
100	100	0.80	0.767	0.762	0.765	0.769	0.779	0.051	0.054	0.052	0.049	0.042
	200	0.80	0.772	0.768	0.770	0.774	0.785	0.043	0.046	0.044	0.042	0.034
Model IV												
100	100	0.75	0.712	0.707	0.709	0.715	0.703	0.055	0.058	0.056	0.053	0.057
	200	0.75	0.721	0.716	0.718	0.722	0.707	0.046	0.049	0.048	0.046	0.051
100	100	0.85	0.824	0.821	0.822	0.827	0.831	0.043	0.045	0.045	0.042	0.038
	200	0.85	0.830	0.827	0.828	0.832	0.834	0.037	0.039	0.038	0.036	0.033
Model V												
100	100	0.75	0.677	0.674	0.674	0.671	0.601	0.076	0.079	0.079	0.081	0.145
	200	0.75	0.678	0.677	0.676	0.672	0.610	0.076	0.076	0.077	0.080	0.138
100	100	0.85	0.807	0.804	0.806	0.809	0.799	0.052	0.054	0.052	0.050	0.055
	200	0.85	0.808	0.805	0.808	0.809	0.802	0.049	0.052	0.049	0.047	0.051
Model VI												
100	100	0.80	0.762	0.759	0.761	0.766	0.770	0.051	0.054	0.051	0.048	0.044
	200	0.80	0.772	0.769	0.771	0.775	0.779	0.042	0.045	0.043	0.041	0.036
100	100	0.85	0.817	0.814	0.816	0.821	0.828	0.044	0.046	0.044	0.041	0.037
	200	0.85	0.827	0.825	0.827	0.830	0.835	0.036	0.038	0.036	0.034	0.030

Table S3: Means for the integrate absolute error (Integ. absolute error) between the real ROC curve, $\mathcal{R}_T(\cdot)$, and its estimation, $\hat{\mathcal{R}}_{\hat{T}_N}(\cdot)$ ($\int_0^1 |\hat{\mathcal{R}}_{\hat{T}_N}(t) - \mathcal{R}_T(t)|dt$) and for the AUC from 2,000 Monte Carlo simulations for the six considered models **without using any cross-validation** procedure. **Small sample sizes, n and m for positive and negative groups respectively, were considered.** $\mathcal{A}$ is the real AUC. Considered bandwidths were smooth cross-validation (SCV), plug-in (PI), normal scale (NS) and biased cross-validation (BCV). RL stands for model based on standard binary logistic regression.

			AUC					Integ. absolute error				
n	m	$\mathcal{A}$	**SCV**	**PI**	**NS**	**BCV**	**RL**	**SCV**	**PI**	**NS**	**BCV**	**RL**
Model 0												
100	100	0.50	0.696	0.720	0.719	0.704	0.551	0.201	0.225	0.224	0.209	0.060
	200	0.50	0.678	0.699	0.699	0.685	0.540	0.180	0.202	0.201	0.187	0.048
Model I												
100	100	0.70	0.763	0.777	0.775	0.767	0.550	0.068	0.081	0.079	0.072	0.147
	200	0.70	0.750	0.762	0.761	0.752	0.541	0.054	0.064	0.063	0.055	0.159
100	100	0.80	0.836	0.845	0.843	0.835	0.550	0.043	0.049	0.048	0.042	0.248
	200	0.80	0.829	0.836	0.835	0.825	0.540	0.035	0.039	0.039	0.032	0.260
Model II												
100	100	0.75	0.795	0.806	0.804	0.796	0.551	0.055	0.063	0.062	0.056	0.196
	200	0.75	0.784	0.792	0.791	0.785	0.543	0.045	0.050	0.049	0.045	0.205
100	100	0.80	0.831	0.839	0.838	0.817	0.550	0.042	0.048	0.047	0.034	0.245
	200	0.80	0.825	0.831	0.830	0.806	0.542	0.037	0.040	0.040	0.030	0.255
Model III												
100	100	0.70	0.759	0.774	0.772	0.761	0.695	0.068	0.081	0.079	0.070	0.040
	200	0.70	0.749	0.761	0.760	0.750	0.693	0.057	0.067	0.066	0.057	0.034
100	100	0.80	0.833	0.841	0.840	0.833	0.799	0.046	0.052	0.050	0.046	0.036
	200	0.80	0.828	0.835	0.834	0.828	0.798	0.038	0.042	0.042	0.038	0.030
Model IV												
100	100	0.75	0.801	0.813	0.812	0.803	0.725	0.060	0.070	0.069	0.061	0.045
	200	0.75	0.793	0.802	0.801	0.793	0.724	0.050	0.057	0.057	0.051	0.042
100	100	0.85	0.879	0.885	0.885	0.879	0.845	0.041	0.046	0.045	0.042	0.033
	200	0.85	0.875	0.880	0.879	0.875	0.844	0.036	0.039	0.039	0.036	0.030
Model V												
100	100	0.75	0.794	0.808	0.792	0.779	0.647	0.055	0.066	0.052	0.043	0.099
	200	0.75	0.784	0.796	0.779	0.767	0.642	0.043	0.052	0.039	0.032	0.106
100	100	0.85	0.882	0.890	0.881	0.873	0.817	0.044	0.050	0.042	0.038	0.043
	200	0.85	0.876	0.883	0.874	0.867	0.813	0.038	0.042	0.035	0.031	0.043
Model VI												
100	100	0.80	0.831	0.841	0.836	0.828	0.788	0.041	0.047	0.043	0.039	0.035
	200	0.80	0.825	0.833	0.829	0.823	0.791	0.036	0.040	0.038	0.034	0.031
100	100	0.85	0.871	0.879	0.875	0.869	0.843	0.034	0.038	0.035	0.032	0.030
	200	0.85	0.868	0.874	0.871	0.866	0.845	0.031	0.033	0.032	0.030	0.027

Acknowledgment

This work is supported by the Grants MTM2015-63971-P and MTM2014-55966-P from the Spanish Ministerio of Economía y Competitividad; FC-GRUPIN-IDI/2018/000132 and Severo Ochoa Grant BP16118 (this one for S. Pérez-Fernández) from the Asturias Goverment (Gobierno del Principado de Asturias).

References

[1] Duong, T. (2019). *ks: Kernel Smoothing*. R package version 1.11.5.

[2] Pérez-Fernández, S., P. Martínez-Camblor, P. Filzmoser, and N. Corral (2018). nsROC: An R package for Non-Standard ROC Curve Analysis. *The R Journal 10*(2), 55–77.