Mathematical Biosciences and Engineering

MBE, 20 (3): 5352–5378.
DOI: 10.3934/mbe.2023248
Received: 04 August 2021
Revised: 09 December 2021
Accepted: 18 January 2022
Published: 11 January 2023

http://www.aimspress.com/journal/MBE

Research article

Identification of influential observations in high-dimensional survival data through robust penalized Cox regression based on trimming

Hongwei Sun[1,2,*], Qian Gao[2], Guiming Zhu[1], Chunlei Han[1], Haosen Yan[1] and Tong Wang[2,*]

[1] Department of Health Statistics, School of Public Health and Management, Binzhou Medical University, Yantai City, Shandong 264003, China
[2] Department of Health Statistics, School of Public Health, Shanxi Medical University, Taiyuan City, Shanxi 030001, China

* **Correspondence:** Email: hwsun2000@163.com, tongwang@sxmu.edu.cn;
Tel: +86-535-6913408; +86-351-4135397.

Supplementary

S1. Performance measures

The evaluation criteria were divided into three categories. The first category concerns the variable selection accuracy. The second one concerns outliers identification. And the third one concerns prediction.

1 Indicators which evaluates accuracy of variables selection

(1) Model size: the number of non-zero coefficients in the estimated model.
(2) Positive select rate (*PSR*) and false discovery rate (*FDR*):

$$PSR = \frac{TP}{TP+FN},$$

$$FDR = \begin{cases} \frac{FP}{TP+FP}, & TP + FP > 0 \\ 0, & TP + FP = 0 \end{cases},$$

where true positive *TP* is the number of coefficients that are non-zero in the true model and were estimated as non-zero. In the true model, false positive *FP* represents the zero coefficients that were estimated as non-zero. False negative *FN* represents the number of non-zero coefficients that were estimated as zero. *PSR* represents the proportion of *TP* in non-zero coefficients in the actual model. Additionally, *FDR* represents the ratio of *FP* in non-zero estimated coefficients.

(3) The geometric mean of *PSR* and (1-*FDR*) (*GM*): We calculated the geometric mean of *PSR* and (1-*FDR*) to evaluate the selection performance of the methods comprehensively.

2 Indicators which evaluates the accuracy of outlier detection.

(1) *Num*: The number of outliers detected by a method.

(2) Sensitivity (*Sn*) and false positive rate (*FPR*):

$$Sn = \frac{TP^*}{TP^*+FN^*},$$

$$FPR = \frac{FP^*}{FP^*+TN^*},$$

where true positive TP^* represents the number of actual outliers that were also detected as outliers. False positive FP^* represents the number of normal individuals that were detected as outliers. False negative FN^* represents the number of actual outliers that were misclassified as normal individuals. True negative TN^* represents the number of normal individuals that were also identified as normal ones.

Sn represents the proportion of actual outliers that were correctly identified. *FPR* represents the proportion of normal individuals that were wrongly categorized as outliers

3 Indicators evaluates the prediction accuracy.

Log likelihood is used to evaluate the prediction of the model.

S2. Simulation setting

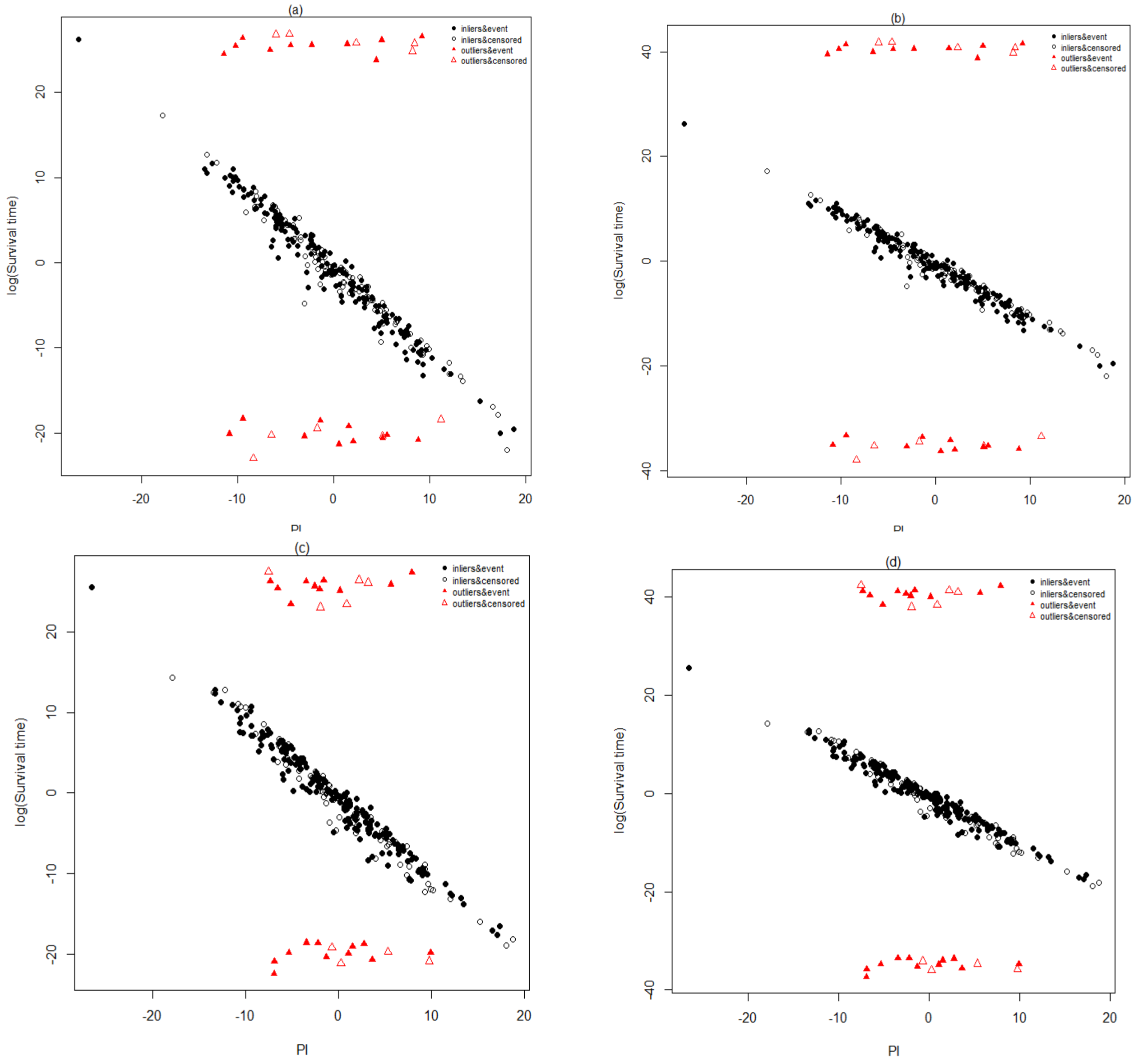

Figure S1. four outlier settings of scenario 3 (scatter plot of the logarithm of survival time and prognosis index PI; a, simulation 3 (a); b, simulation 3 (b); c, simulation 3 (c); d, simulation 3 (d); Black solid dots: normal points with outcomes; black hollow dots: censored normal points; red solid triangles: outliers with outcomes; red hollow triangles: censored outliers.)

S3. Results of simulation study

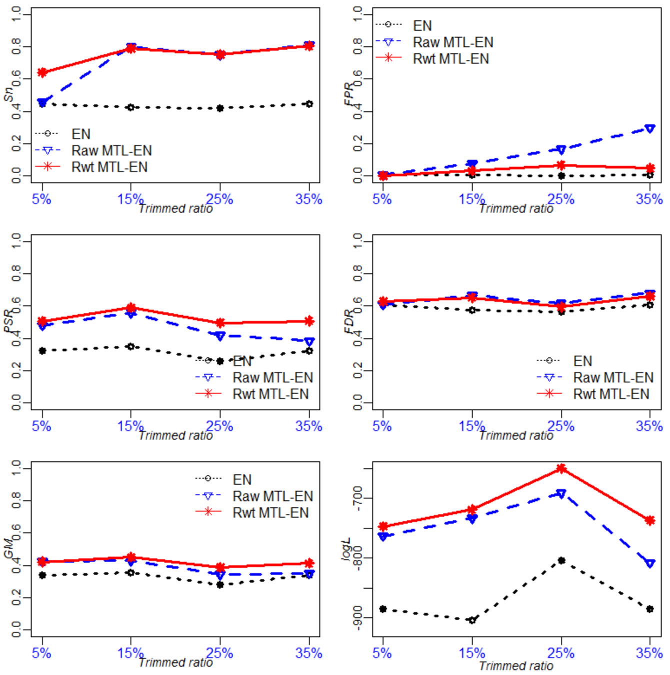

Figure S2. Comparison of results between EN and MTPL-EN under different trimmed ratios (n=300, p=1,000).

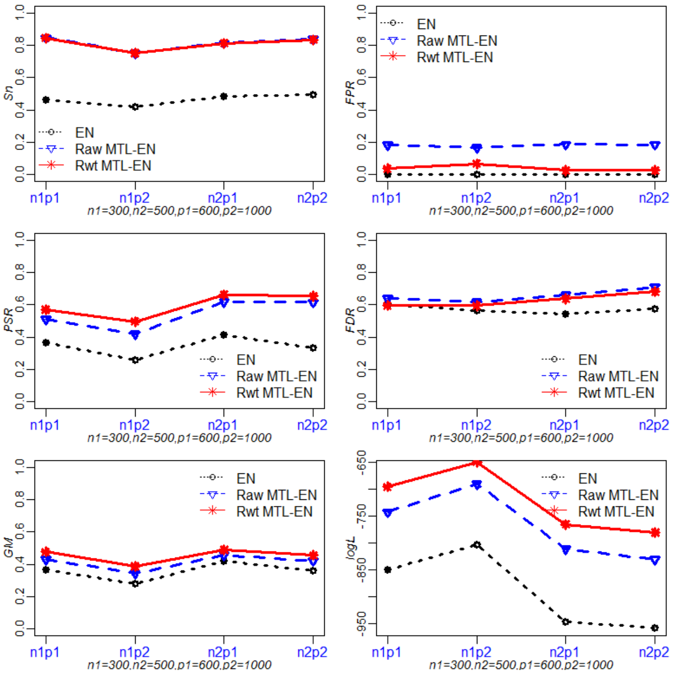

Figure S3. Comparison of results between EN and MPTL-EN under different samples and dimensions (Censored rate = 35%) (Since the value of the likelihood function is affected by sample size, the logarithmic likelihood function at n = 500 is multiplied by 0.6 to facilitate comparison with n = 300.)

S4. Results of real data analysis

Table S1. Genes and their coefficients identified by EN for glioma dataset.

Gene	Coefficient	Gene	Coefficient	Gene	Coefficient
ADAMDEC1#*	0.086	GPR17#*	-0.002	PLXNB3*	-0.020
AGBL3#	0.012	HIST1H2AC	0.026	POLL#*	-0.020
ANKRD55	-0.012	HIST1H2AD	0.016	POLR2J4	0.039
ARHGAP24#	-0.029	HIST1H2BF	0.008	POU4F1	0.015
ARL6IP1#	0.009	HIST1H2BH	0.005	PROM2#	-0.035
BCR#	-0.002	HIST1H2BK	0.002	PSTPIP1#*	-0.062
BMP5	0.002	HIST2H2BE	0.011	PTTG2*	0.011
C16orf62	-0.012	HOTAIR*	0.006	RBM45#	-0.003
C21orf45#	0.054	HOXA3#*	0.031	RDM1	0.007
C9orf40#	0.021	HOXC13#*	0.062	RGR	-0.019
CCDC34#	0.045	HOXC8#*	0.048	SC4MOL	-0.015
CENPH	0.001	HSD17B14	-0.012	SCYL2#	0.040
CEP68#	-0.054	HYOU1	-0.017	SELL#*	-0.013
CHST15*	-0.018	KAT2A#*	-0.016	SH3RF3	-0.001
CKS2#	0.059	KIAA0141#	-0.067	SMARCA1	0.002
CTBP2*	-0.017	KIAA1199	-0.025	SPATA17#	0.031
DGCR6L	-0.028	KIAA1462#	-0.034	SPATA9	-0.023
DIO2#*	-0.032	KIF18A	0.041	SYAP1	0.004
DISP2	-0.007	KL	-0.002	TBC1D17	-0.008
DKFZp434L192	0.021	LOC100289600#	-0.007	TBC1D5#	-0.081
DNAJB1	0.028	LOC285548	0.028	TOX	-0.051
FAM108C1	-0.028	LRRIQ1	0.039	TPCN1	-0.009
FAM13B	0.004	MN1	-0.018	TRIM73	0.010
FARP2	-0.074	MSH2*	0.017	TTC15	-0.009
FASN*	-0.013	OSBP2	-0.008	TYSND1	-0.019
FOXA1*	0.001	PAX3*	0.042	USP34#	-0.006
FOXO3	-0.028	PDGFC*	-0.005	WWOX	-0.018
FOXRED2	-0.004	PIGS	-0.022	WWP2*	-0.030
GPLD1	-0.017	PLK4*	0.018	ZNF367	0.008

*: Genes reported in literature. #: Genes selected that were coincident with those by Rwt MTPL-EN

Table S2. Fifty six Genes identified by Rwt MTPL-EN and their coefficients.

Gene	Coefficient	Gene	Coefficient	Gene	Coefficient
ABHD4	-0.024	CEP68#	-0.016	PPARA*	-0.008
ABLIM3	-0.008	CKS2#	0.098	PROM2#	-0.031
ADAMDEC1*#	0.095	DIO2*#	-0.041	PSMC3IP*	0.002
AGBL3#	0.032	GPR17*#	-0.042	PSTPIP1*#	-0.167
ANKAR	-0.008	HOXA3*#	0.084	PTPRE*	-0.050
ANKRD32	0.003	HOXC13*#	0.020	PUSL1	0.006
APITD1	0.032	HOXC8*#	0.019	RBM45#	-0.013
ARHGAP24#	-0.042	HSPB11*	0.041	RRP7B*	-0.003
ARL6*	0.059	ISL2	0.004	SCYL2#	0.030
ARL6IP1#	0.005	KAT2A*#	-0.005	SELL*#	-0.009
ATOH8*	-0.079	KIAA0141#	-0.072	SHE	-0.002
BCR#	-0.039	KIAA1462#	-0.095	SLC24A3	-0.021
C21orf45#	0.031	KTI12	0.030	SNRPA1	0.014
C21orf49	-0.056	LOC100288798	-0.007	SPATA17#	0.037
C9orf30	0.046	LOC100289600#	-0.015	TBC1D5#	-0.146
C9orf40#	0.009	LOC283788	-0.012	TMPO*	0.004
CCDC137	0.018	NAV1	-0.037	UBE2T	0.009
CCDC34#	0.091	PDCD5*	0.022	USP34#	-0.075
CCS	-0.008	POLL*#	-0.051		

*: Genes reported in literature. #: Genes selected that were coincident with those by EN

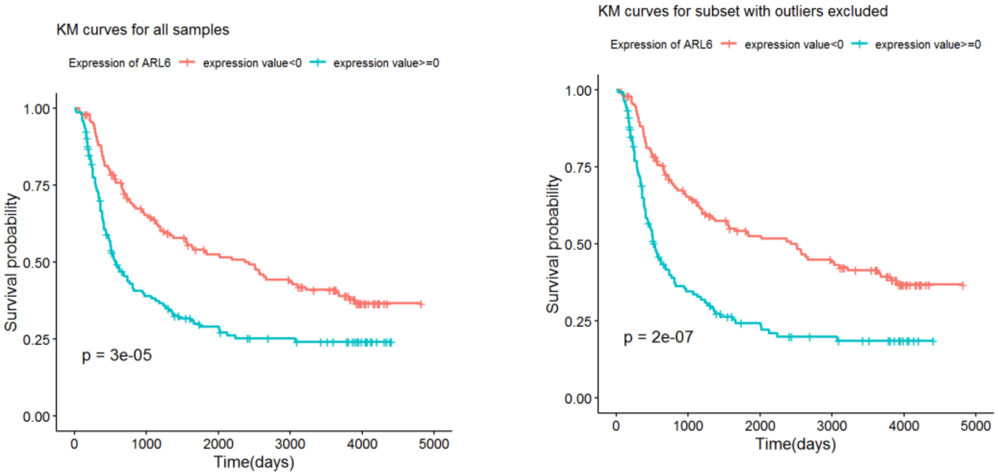

Figure S4. Kaplan-Meier Curves of high and low expression of ARL6 for all samples and subset with outliers excluded.