

Editorial

Research Article

Jiehuan Sun, Jose D. Herazo-Maya, Jane-Ling Wang, Naftali Kaminski, and Hongyu Zhao*

LCox: A tool for selecting genes related to survival outcomes using longitudinal gene expression data – supplementary materials

DOI ..., Received ...; accepted ...

PACS: ...

Communicated by: ...

Dedicated to ...

1 Additional Simulation Studies

In this section, we conduct additional simulations to study the effects of length of follow-up time and number of visits (or time points) per patient on the performance of LCox.

1. To study the effect of length of follow-up time, we adopt the same scenarios as in the paper, but the time points are now drawn from $\text{Uniform}(0, l)$, where l is the maximum length of follow-up time. Here, we choose $l = 0.6$ and $l = 0.8$ ($l = 1$ in the main text of the paper). From Table S1, we can see that LCox has better or similar performance to other methods for different lengths of follow-up time. When comparing the performance of LCox in scenarios with $l = 0.8$ to that in scenarios with $l = 0.6$, we can see that the performance diminishes as the length of follow-up time decreases if the survival outcome depends on the genes on the whole history (Scenarios 1 and 4 in particular). For example, in Scenario 4, the hazard depends on

Jiehuan Sun, Hongyu Zhao, Department of Biostatistics, Yale School of Public Health, 60 College Street, New Haven, CT 06510, e-mail: hongyu.zhao@yale.edu

Jose D. Herazo-Maya, Naftali Kaminski, Internal Medicine: Pulmonary, Critical Care and Sleep Medicine, Yale School of Medicine, 300 Cedar Street, New Haven, CT 06519

Jane-Ling Wang, Department of Statistics, University of California, Davis, One Shields Avenue, Davis, CA 95616

informative biomarker through its values in the history ($t \in (0, 1)$). If we only observe data on $t \in (0, 0.6)$, it is difficult to capture the effect of the biomarker on the survival outcome and hence the performance of LCox (and other methods as well) diminishes in this scenario.

Therefore, if the maximum length of follow-up time is not long enough to capture the effects of certain genes on the survival outcome, it is very likely that LCox cannot successfully detect such genes, neither can other methods. In these situations, however, LCox still has larger power than other methods at most of the time.

2. To study the effect of number of visits per patient, we adopt the same scenarios as in the paper, but the number of visits per patient is changed to be 2 (and 3). From Table S2, when number of visits is 3 for all subjects, the results are similar to that in the main text of the paper, where the number of visits is randomly selected from $\{2, 3, 4, 5\}$ for each patient. This suggests that LCox performs well in a sparse setting where the number of visits is only 3 per patient. However, in the most difficult scenario where the number of visits per patient is 2, the performance of LCox diminishes and the type one error rates seem to be inflated, which possibly is due to the instability of the estimation in the FPCA step. It is extremely difficult to estimate any function of complex forms when there are only 2 time points per patient. The best we can do is probably to estimate well the intercept and slope for each patient. Therefore, it makes sense that FPCA (and hence LCox) does not perform well when number of visits is 2. In this case, we recommend to use simple methods such as linear mixed-effects models to estimate the intercepts and slopes, which can then be used to test the association of the gene and the survival outcome.

To summarize, FPCA (and hence LCox) requires a reasonable number of time points per patient (i.e. not too sparse) to estimate the parameters well. Based on our simulation study, it seems that LCox performs well when the average number of visits per patient is larger than or equal to 3.

Table S1: Comparisons of performance of LCox, JM, Cox-base, and Cox-avg using 200 simulated datasets for each of the six scenarios with different lengths of follow-up time. The number in each of the cells indicates the power of each method with type one error rate in the parenthesis.

Scenarios	LCox		JM		Cox-base		Cox-avg	
	len=0.6	len=0.8	len=0.6	len=0.8	len=0.6	len=0.8	len=0.6	len=0.8
1	0.66 (0.07)	0.82 (0.04)	0.73 (0.07)	0.9 (0.02)	0.1 (0.06)	0.06 (0.04)	0.52 (0.04)	0.64 (0.04)
2	0.86 (0.08)	0.85 (0.05)	0.62 (0.12)	0.46 (0.04)	0.83 (0.07)	0.86 (0.04)	0.89 (0.05)	0.9 (0.06)
3	0.66 (0.08)	0.71 (0.06)	0.57 (0.24)	0.67 (0.25)	0.62 (0.04)	0.65 (0.06)	0.8 (0.06)	0.81 (0.06)
4	0.68 (0.04)	0.86 (0.07)	0.79 (0.09)	0.86 (0.02)	0.05 (0.04)	0.05 (0.06)	0.44 (0.02)	0.56 (0.08)
5	0.86 (0.02)	0.94 (0.04)	0.84 (0.05)	0.9 (0.04)	0.04 (0.04)	0.06 (0.04)	0.58 (0.03)	0.62 (0.08)
6	0.83 (0.08)	0.82 (0.06)	0.66 (0.18)	0.36 (0.21)	0.06 (0.07)	0.06 (0.06)	0.57 (0.07)	0.08 (0.04)

Table S2: Comparisons of performance of LCox, JM, Cox-base, and Cox-avg using 200 simulated datasets for each of the six scenarios with different numbers of visits per patient. The number in each of the cells indicates the power of each method with type one error rate in the parenthesis.

Scenarios	LCox		JM		Cox-base		Cox-avg	
	visits=2	visits=3	visits=2	visits=3	visits=2	visits=3	visits=2	visits=3
1	0.42 (0.11)	0.89 (0.1)	0.7 (0.07)	0.84 (0.02)	0.1 (0.04)	0.12 (0.06)	0.56 (0.06)	0.82 (0.06)
2	0.62 (0.1)	0.82 (0.06)	0.56 (0.04)	0.27 (0.03)	0.82 (0.04)	0.84 (0.04)	0.8 (0.04)	0.74 (0.06)
3	0.56 (0.14)	0.68 (0.02)	0.65 (0.27)	0.63 (0.2)	0.64 (0.06)	0.68 (0.05)	0.76 (0.06)	0.82 (0.06)
4	0.57 (0.1)	0.94 (0.06)	0.86 (0.06)	0.82 (0.01)	0.04 (0.06)	0.06 (0.07)	0.48 (0.06)	0.7 (0.04)
5	0.5 (0.08)	0.92 (0.04)	0.77 (0.07)	0.9 (0.04)	0.05 (0.06)	0.09 (0.05)	0.34 (0.12)	0.59 (0.08)
6	0.3 (0.12)	0.8 (0.07)	0.13 (0.09)	0.29 (0.11)	0.06 (0.04)	0.04 (0.04)	0.06 (0.03)	0.03 (0.04)

Table S3: The estimates, standard errors, 95% confidence interval, Z statistics, and P values of each functional scores from the top 10 most significant genes in the IPF dataset.

Gene	Beta	SE	Z	P value	95% CI
ACTB	3.15	1.6	1.97	0.0491	(0.0128,6.28)
ACTB	-17.5	4.41	-3.97	7.17E-05	(-26.1,-8.86)
MAGEA2B	-2.79	1.97	-1.42	0.157	(-6.65,1.07)
MAGEA2B	-16.1	4.74	-3.39	0.000687	(-25.4,-6.8)
MAGEA2B	-48.4	17	-2.84	0.0045	(-81.7,-15)
CRYGB	-4.24	1.67	-2.53	0.0113	(-7.53,-0.962)
CRYGB	-4.14	2.92	-1.42	0.156	(-9.86,1.58)
CRYGB	26.1	7.01	3.72	0.000201	(12.3,39.8)
CRYGB	-62.5	17.6	-3.55	0.000386	(-97,-28)
STOX2	3.56	2.08	1.71	0.0877	(-0.525,7.64)
STOX2	-5.29	2.18	-2.42	0.0153	(-9.56,-1.01)
STOX2	-13.4	3.5	-3.84	0.000124	(-20.3,-6.57)
STOX2	2.87	3.82	0.751	0.453	(-4.62,10.4)
LOC146429	-3.28	1.61	-2.04	0.0415	(-6.43,-0.127)
LOC146429	-12.6	3.28	-3.84	0.000124	(-19,-6.16)
LOC146429	-4.46	3.91	-1.14	0.253	(-12.1,3.19)
LOC146429	-57.4	20.2	-2.84	0.00448	(-97,-17.8)
RGS7BP	4.01	2.22	1.8	0.0713	(-0.347,8.37)
RGS7BP	-7.81	2.96	-2.64	0.00827	(-13.6,-2.01)
RGS7BP	12.9	23.3	0.554	0.579	(-32.8,58.7)
C9orf43	6.57	2.09	3.14	0.00169	(2.47,10.7)
C9orf43	-5.69	3.27	-1.74	0.0824	(-12.1,0.729)
C9orf43	-0.377	3.4	-0.111	0.912	(-7.05,6.3)
C9orf43	-25	10.2	-2.45	0.0145	(-45.1,-4.97)
ZP2	-3.53	1.17	-3.02	0.00256	(-5.82,-1.23)
ZP2	-10.8	3.26	-3.32	0.000911	(-17.2,-4.42)
ZP2	2.8	7.52	0.372	0.71	(-11.9,17.5)
APOH	3.51	1.45	2.43	0.0152	(0.676,6.35)
APOH	3.06	1.94	1.57	0.116	(-0.754,6.87)
APOH	10.5	2.94	3.57	0.000353	(4.74,16.3)
LSM6	-0.511	2.28	-0.224	0.823	(-4.98,3.96)
LSM6	-77.6	21	-3.7	0.000216	(-119,-36.5)
LSM6	-186	51.8	-3.59	0.000332	(-288,-84.4)