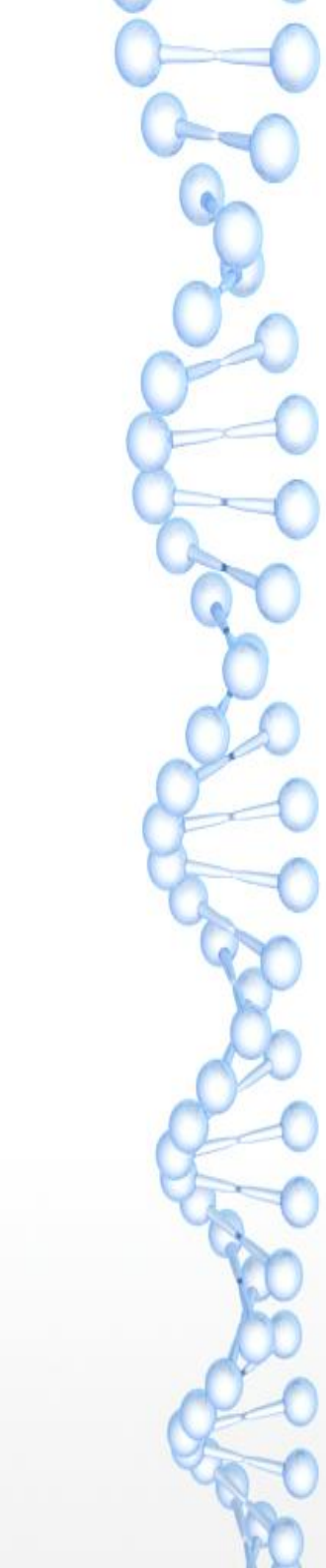
Generalized procedure for selecting methylation CpGs associated with cancer

Marija Stanojevic
Knowledge Discovery and Data Mining Class
Temple University
15th April 2018

Background and goal



- 1.7 million cases in 2017 in US
- The most common cancers: breast, lung and bronchus, prostate, colorectal
- Per year: 0.02% of population dies; $156 billion spent
- DNA methylation depends on genetics and environment; changes slowly
- **Can we find methylation positions that can diagnose/predict cancer?**
- Using cancers: breast, colorectal, lung and kidney with around 1000 samples per each.
- 2/3 samples – cancerous tissue; rest normal
- Around 23500 features (CpGs) per sample

Questions for this paper

1. Does methodology [1] (Lasso and Lasso Cox) find significant CpGs for all cancers? How well does it model time to death?
2. Is there an overlap in selected features between different common cancers?
3. Is it possible to use only Lasso without feature processing for feature selection and get good results?
4. Spatial dependency in feature selection
5. Is deviation or mean difference score better for pre-processing step?

[1] Wei, Jin-Huan, et al. "A CpG-methylation-based assay to predict survival in clear cell renal cell carcinoma." Nature communications 6 (2015): 8699.

Methodology

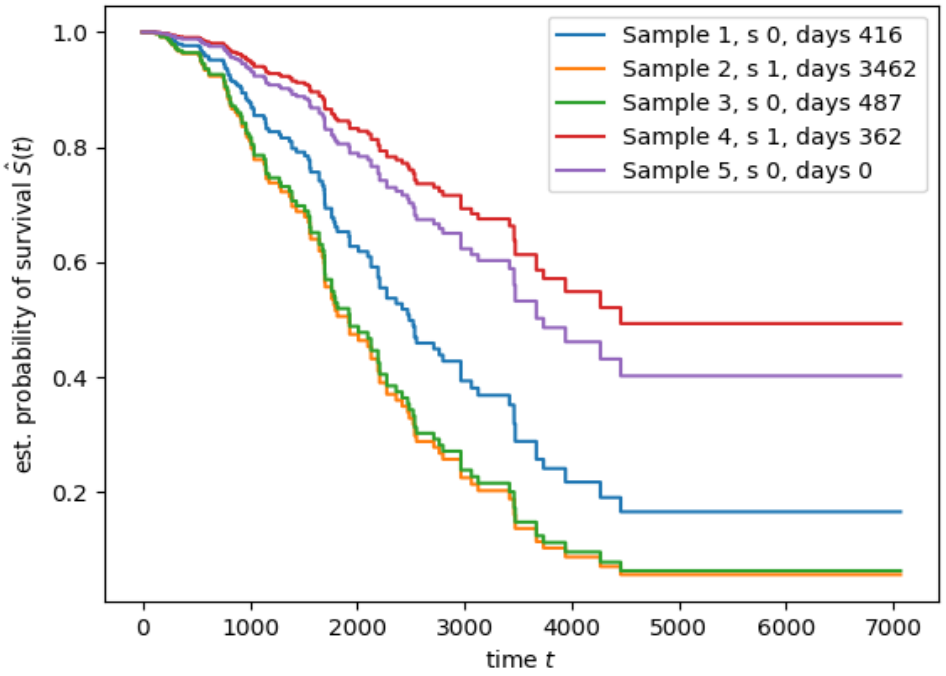
- Methylation data cleaning:
  - Remove feature if >20% is missing
  - Impute with mean feature, otherwise
- Normalize methylation per chip
- Extract and transform clinical and speciment data
- Using cross-validation find best Lasso model to fit for each cancer
- Model time-to-leave with Cox-Lasso analysis
- Find best model between “only Lasso” and “Lasso with different pre-processings”
- Interpret results

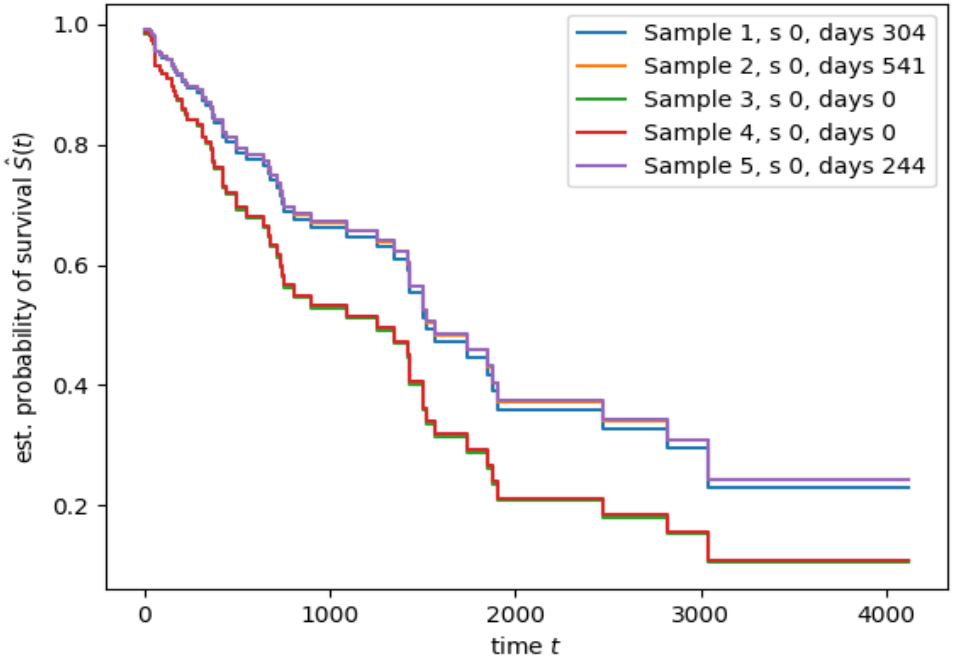
Breast	Colorectal	Kidney	Lung
387	1612	316	316
1612	1690	1964	2119
3856	1964	3301	3856
5286	5286	7384	5286
5361	7331	7394	5361
7383	7383	9412	7393
7628	8105	9911	8316
8523	8225	10057	9064
9411	8523	10958	9411
10441	11129	11072	9910
11825	11825	11826	10441
12983	13839	12080	10957
13220	16405	14000	11825
17741	17810	15822	13220
18844	18769	16407	15820
21007	20939	16461	17741
21707	21566	16911	18844
23208	23299	17812	21007
23300	23328	18771	
23329		21708	

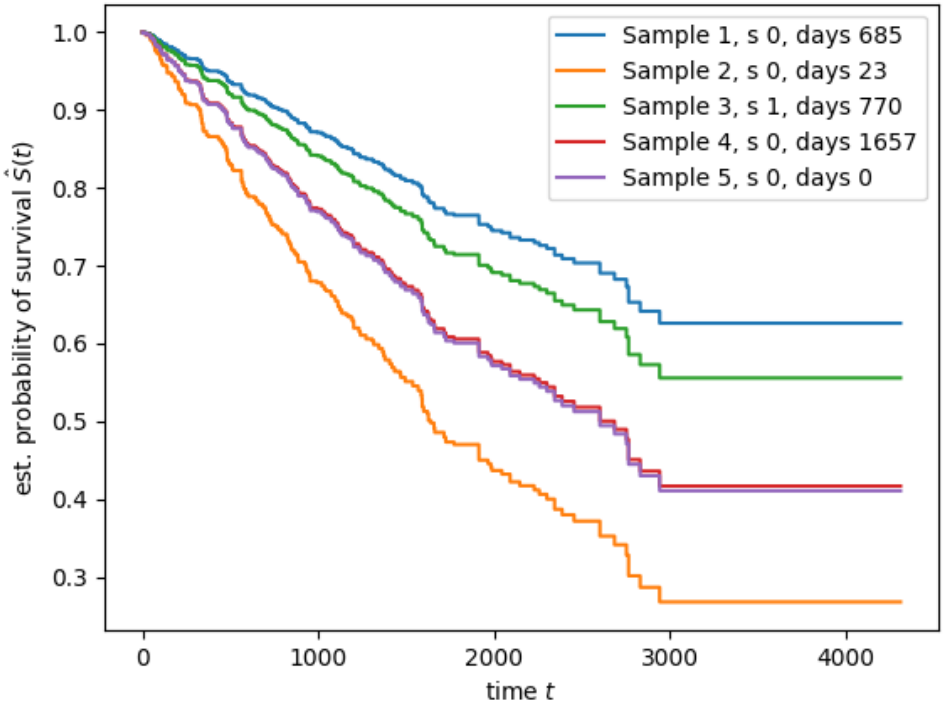
Results

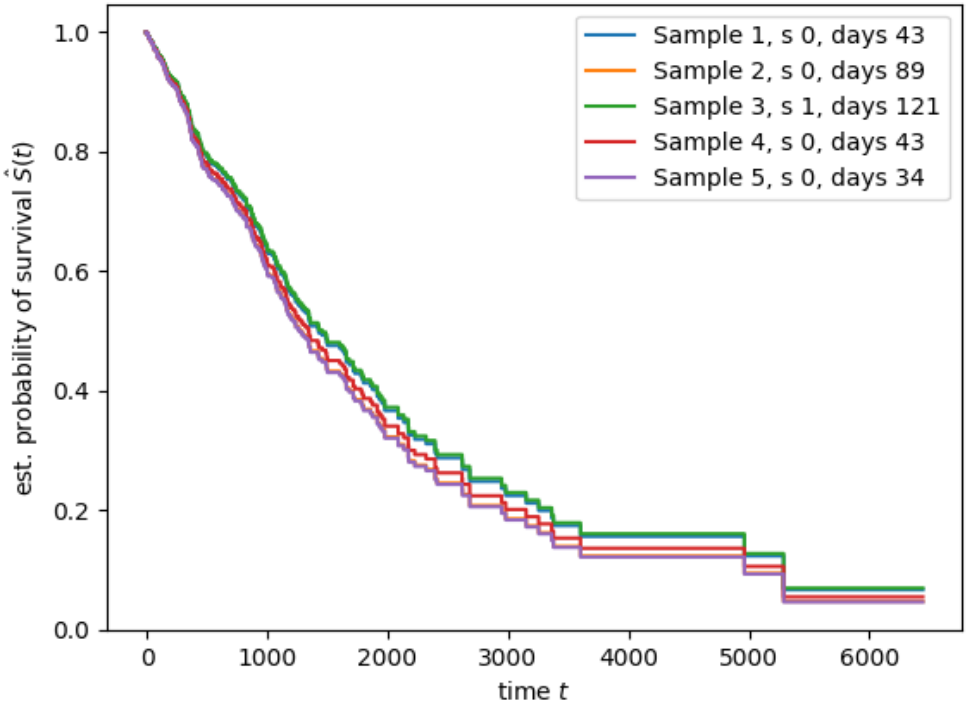
- Lasso experiment is repeated 5 times and methylation positions that appear often among the features with highest score are given on the table left
- Many features are repeating through different cancers
- Slight position change in some cases

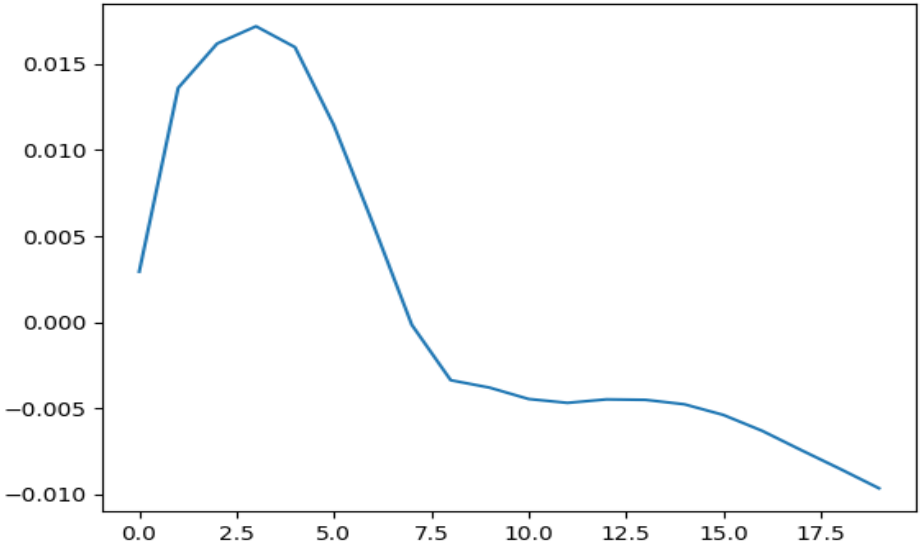
Average values for regularization coefficient			
Breast	Colorectal	Kidney	Lung
2.6	2.76	1.96	2.1

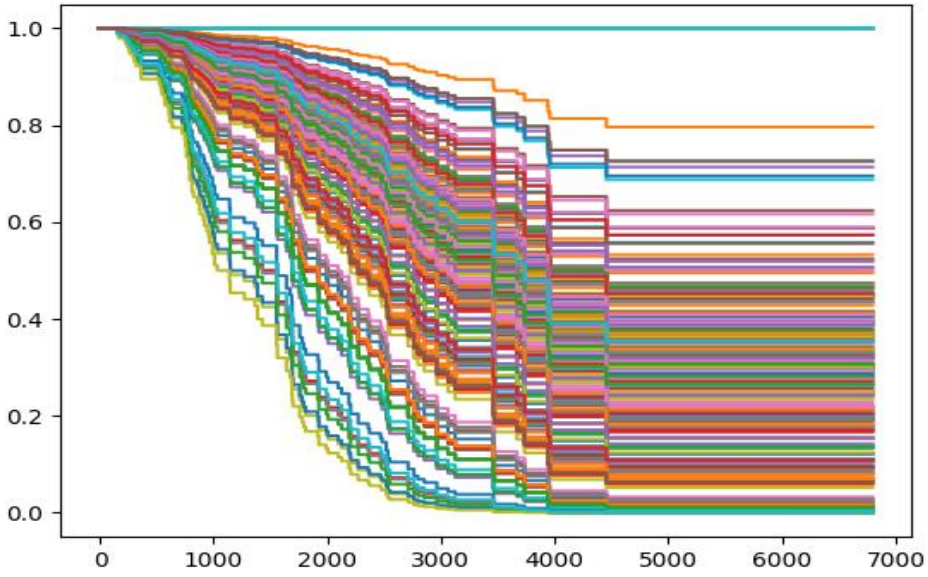

Sample 1, s 0, days 416
Sample 2, s 1, days 3462
Sample 3, s 0, days 487
Sample 4, s 1, days 362
Sample 5, s 0, days 0
est. probability of survival $\hat{S}(t)$
time t


Sample 1, s 0, days 304
Sample 2, s 0, days 541
Sample 3, s 0, days 0
Sample 4, s 0, days 0
Sample 5, s 0, days 244
est. probability of survival $\hat{S}(t)$
time t


Sample 1, s 0, days 685
Sample 2, s 0, days 23
Sample 3, s 1, days 770
Sample 4, s 0, days 1657
Sample 5, s 0, days 0
est. probability of survival $\hat{S}(t)$
time t


Sample 1, s 0, days 43
Sample 2, s 0, days 89
Sample 3, s 1, days 121
Sample 4, s 0, days 43
Sample 5, s 0, days 34
est. probability of survival $\hat{S}(t)$
time t





- Many common positions are selected for different cancers
- Lasso scores are similar. Cox-Lasso time-to-death predictions are close, but can be improved
- **Future work**: Create multi-view multinomial model + Use all available specimen and patient data

THANK YOU!