An Introduction to the 'PICS' Package, Version 1.0

Zequn Sun,Wei Wei and Dongjun Chung
Department of Public Health Sciences, Medical University of South Carolina (MUSC),
Charleston, SC, 29425.

February 15, 2017

1 Overview

This vignette provides basic information about the PICS'[1]' package. PICS stands for "Pathway-guided Identification of Cancer Subtypes". The PICS methodology for Pathway-Guided Identification of Cancer Subtypesis developed in PICS. The proposed approach improves identification of molecularly-defined subgroups of cancer patients by utilizing information from pathway databases in the following four aspects.

(1) integration of genomic data at the pathway-level improves robustness and stability in identification of cancer subgroups and driver molecular features; and

(2) summarizing multiple genes and genomic platforms at the pathway-level can potentially improve statistical power to identify important driver pathways because moderate signals in multiple genes can be aggregated; and

(3) in PICS, we consider this "operation" or "interaction" between pathways, instead assuming that each pathway operates independently during the cancer progression, which may be unrealistic; and

(4) PICS allows simultaneous inference in multiple biological layers (pathway clusters, pathways, and genes) within a statistically rigorous and unified framework without any additional laborious downstream analysis.

The package can be loaded with the command:

```
R> library("PICS")
```

2 Input Data

The package requires that the response consist of 4 components: (1) z-scores in the form of a either dataframe or matrix; and (2) time and status variable for Cox model of survival analysis in the form of vectors; and (3) pathway information is in the form of a list where gene names are included in pathway lists respectively.

The Cancer Genome Atlas (TCGA) is an example application for the 'PICS' package. The TCGA data was downloaded from the cBio Portal (http://www.cbioportal.org/) using the R package CGDS, and we used z scores for the mRNA expression data.

All the mRNA expression measures were centered and scaled to have unit variance according to standard practice. Only genes annotated in the KEGG database and that appeared in both datasets were considered for analysis.In the current stage, we focused only on gene expression measurements and aim at incorporating other data types into our model in the future.

Z scores for the mRNA expression data of 389 genes is given, also survival time and survival status of 50 observations are given in the form of numeric vector and binary vector respectively. A list of 4 pathways information can be found that consist of gene names.

The '`TCGA`' dataframe as z scores for the mRNA expression, the '`t`' vector as the survival time, the '`d`' vector as the survival status, and the '`path.list`' list of the pathways is loaded as follows:

```
R> data(TCGA)
R> TCGA[1:5,1:5]

        ACLY        ACO1       ACO2         CS       DLAT
1 -2.2410125 -0.48445531 -1.6346455  0.1378804 -3.5310321
2 -2.1301362  0.82116427 -0.9533701  0.6213512  0.6689948
3 -2.9122727 -0.08790649 -1.0975096 -0.2454025 -0.9433900
4 -1.1721514 -0.24249825 -0.7212639  0.1842386 -0.6188785
5  0.5383438  0.98012739 -0.7396043 -0.0699680  1.9573767

R> t[1:5]

[1] 43.89 40.97 49.12  2.00 46.59

R> d[1:5]

[1] 1 1 0 1 0

R> path.list[1]

$KEGG_CITRATE_CYCLE_TCA_CYCLE
 [1] "IDH3B"     "DLST"      "PCK2"      "CS"        "PDHB"      "PCK1"
 [7] "PDHA1"     "LOC642502" "PDHA2"     "LOC283398" "FH"        "SDHD"
[13] "OGDH"      "SDHB"      "IDH3A"     "SDHC"      "IDH2"      "IDH1"
[19] "ACO1"      "ACLY"      "MDH2"      "DLD"       "MDH1"      "DLAT"
[25] "OGDHL"     "PC"        "SDHA"      "SUCLG1"    "SUCLA2"    "SUCLG2"
[31] "IDH3G"     "ACO2"
```

3 Pre-filtering

We expected that only a parsimonious set of genes would be related to patient survival, i.e., the sparsity assumption. To eliminate the most unlikely predictors, we first conducted a supervised pre-filtering by fitting a Cox regression model of each mRNA expression measure on patient survival in the TCGA dataset. Only the gene expressions associated with patient survival at p-values smaller than a pre-specified cut-off were included in the subsequent analysis. Here we chose p = 0.5 as the default cut-off point.

```
R> train.list=prefilter(data=TCGA, time=t, status=d, p.cut=0.5, plist=path.list)
```

4 Gene Selection

This section is about Gene-level Analysis using a SPLS[**?**] Cox regression. In order to select key genes associated with patient survivals and effectively summarize them by taking into account

correlation among them, we fit a sparse partial least squares (SPLS) Cox regression model of patient survivals on gene expression measurements for each pathway.

PICS utilizes extention of SPLS Cox regression so that it can handle survival outcomes, where there are two main tuning parameters: 'eta' represents the sparsity tuning parameter and 'K' is the number of hidden (latent) components. Parameters can be chosen by (v-fold) cross-validation using the function 'selectGene'. The user specifies the range for these parameters and the cross-validation procedure searches within these ranges. 'eta' should have a value between 0 and 1. 'K' is integer valued and can range between 1 and $min\{p, (v-1)n/v\}$, where p is the number of predictors and n is the sample size. For example, if 10-fold cross-validation is used (default), 'K' should be smaller than $min\{p, 0.9n\}$. For the TCGA data, we search for 'K' between 3 and 5 and for 'eta' between 0.1 and 0.9 with the following command:

```
R> gene.results=selectGene( data=train.list, fold=5, K=c(3,5), etas=c(0.1,0.9),seed=123)
```

The function 'selectGene' returns the following output:

(1) Gene: cancer related genes selected by our approach.

(2) Gene beta: the SPLS regression coefficients of cancer related genes.

(3) W: the direction vectors for constructing scores, where score = x*w, where x is the original gene expression data.

(4) Score: the latent variables summarizing the gene expression data in each pathway, scores are linear combinations of the original gene expression data.

5 Pathway Selection

Next, in order to identify a parsimonious set of pathways associated with patient survivals, we fit a LASSO-penalized[4] Cox regression on latent components derived from all the pathways. First, the latent components generated from the SPLS step preserve pathway structure and also reflect correlation among genes and their association with survival outcomes. Second, this approach can potentially improve the stability of estimation in the subsequent analysis.

```
R> path.results=selectPath( gene.results, seed=123)
R> printPathResults(path.results)

$path.before
[1] "KEGG_CITRATE_CYCLE_TCA_CYCLE"    "KEGG_MAPK_SIGNALING_PATHWAY"
[3] "KEGG_TGF_BETA_SIGNALING_PATHWAY" "KEGG_THYROID_CANCER"

$path.after
[1] KEGG_CITRATE_CYCLE_TCA_CYCLE    KEGG_MAPK_SIGNALING_PATHWAY
[3] KEGG_TGF_BETA_SIGNALING_PATHWAY
4 Levels: KEGG_CITRATE_CYCLE_TCA_CYCLE ... KEGG_THYROID_CANCER

$path.beta
              rep.pathways..cols.  path.beta
1    KEGG_CITRATE_CYCLE_TCA_CYCLE 0.21115773
2    KEGG_CITRATE_CYCLE_TCA_CYCLE 0.00000000
3     KEGG_MAPK_SIGNALING_PATHWAY 0.17100370
4 KEGG_TGF_BETA_SIGNALING_PATHWAY 0.05360567
5             KEGG_THYROID_CANCER 0.00000000
```

The function 'printPathResults' returns the following output:

(1) path.before: pathways that were left from pre-filtering step.

(2) path.after: cancer related pathways selected by our approach.

(3) Gene beta: the SPLS regression coefficients of cancer related genes.

6 Pathway Effect Plot

In the LASSO step, we could identify the pathways associated with patient survivals using the latent components generated from the SPLS step.

Specifically, a pathway was selected if at least one of its latent components had non-zero LASSO coefficient estimate. Figure 1 shows the hazard ratio (HR) associated with each latent component in selected pathways.

Based on the TCGA data, pathways with the largest effect on survival (HR = 1.15) are CITRATE_CYCLE_TCA_CYCLE and MARK_SIGNALING pathways.

```
R> plotPathwayEffect(path.results=path.results)
```

7 Prediction

Survival predictions were made based on model parameters estimated from the TCGA data.

```
R> predicted=pred( path.results, newx=train.list, n=50)
R> predicted[1:3]

$risk.index
 [1] 1 3 0 3 1 2 3 0 0 1 3 0 1 2 1 3 0 0 1 1 1 0 0 1 2 2 1 0 1 1 1 0 1 0 0 0 0 2
[39] 3 2 1 3 2 3 3 3 2 2 3 3

$riskcat
 [1] "med"  "high" "low"  "high" "med"  "high" "high" "low"  "low"  "med"
[11] "high" "low"  "med"  "high" "med"  "high" "low"  "low"  "med"  "med"
[21] "med"  "low"  "low"  "med"  "high" "high" "med"  "low"  "med"  "med"
[31] "med"  "low"  "med"  "low"  "low"  "low"  "low"  "high" "high" "high"
[41] "med"  "high" "high" "high" "high" "high" "high" "high" "high" "high"

$cuts
[1] 0 2
```

8 Survival Curve

The predictive performance of PICS method was presented by Kaplan-Meier curves. 'survivalCurve' returns Kaplan-Meier curves of predicted patient subgroups based on the PICS approach.

```
R> summary( coxph(Surv(predicted$time, predicted$status) ~ predicted$riskcat) )
```

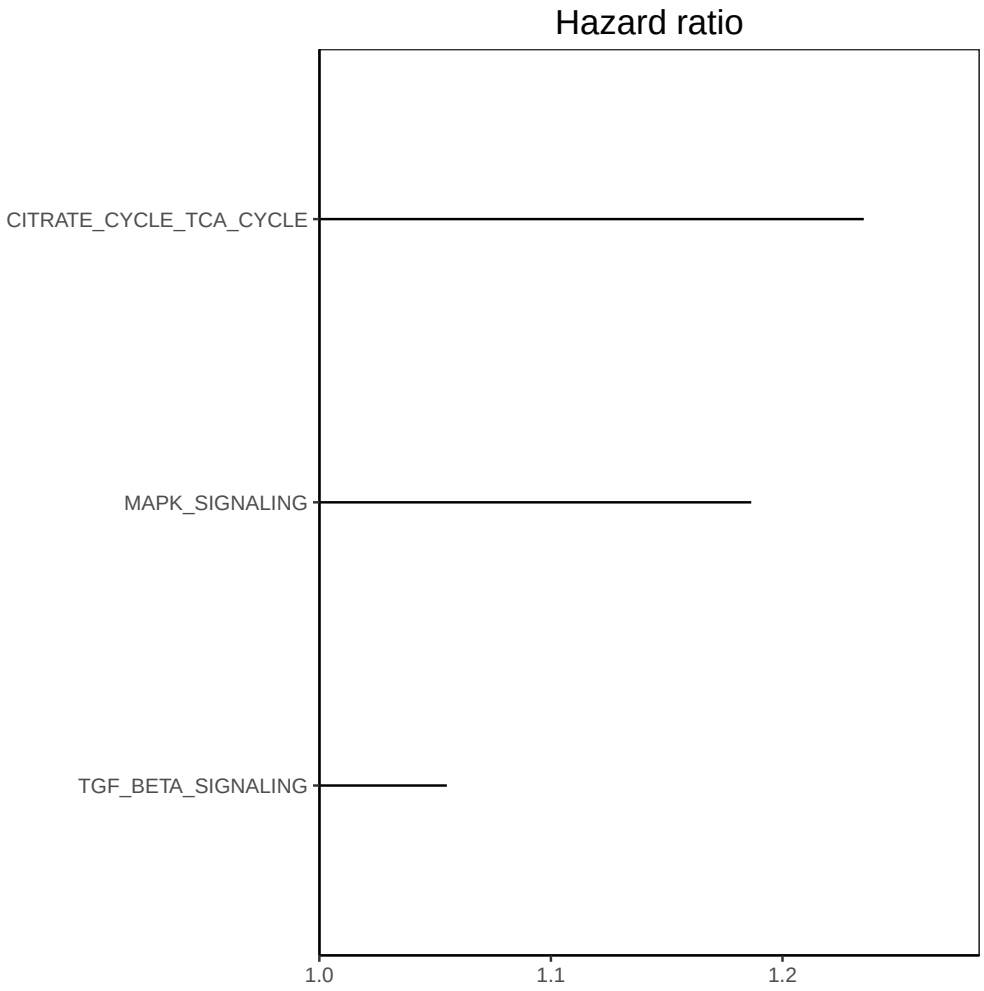

Figure 1: Hazard ratio (HR) associated with each latent component in selected pathways

```
Call:
coxph(formula = Surv(predicted$time, predicted$status) ~ predicted$riskcat)

  n= 50, number of events= 35

                          coef exp(coef) se(coef)      z Pr(>|z|)
predicted$riskcatlow -4.20595   0.01491  1.04289 -4.033 5.51e-05 ***
predicted$riskcatmed -1.07293   0.34201  0.40184 -2.670  0.00758 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                     exp(coef) exp(-coef) lower .95 upper .95
predicted$riskcatlow   0.01491     67.084  0.001931    0.1151
predicted$riskcatmed   0.34201      2.924  0.155591    0.7518

Concordance= 0.804  (se = 0.052 )
```

```
Rsquare= 0.582   (max possible= 0.987 )
Likelihood ratio test= 43.65  on 2 df,   p=3.326e-10
Wald test            = 20.5  on 2 df,   p=3.537e-05
Score (logrank) test = 41.06  on 2 df,   p=1.214e-09
```

```
R> survivalCurve(predicted)
```

Figure 2 shows the Kaplan-Meier curves of predicted patient subgroups based on the `PICS` approach. The `PICS` approach could distinguish the high risk group from the medium risk (log rank test, p=0.0075) and low risk (log rank test, p=0.00005) groups.

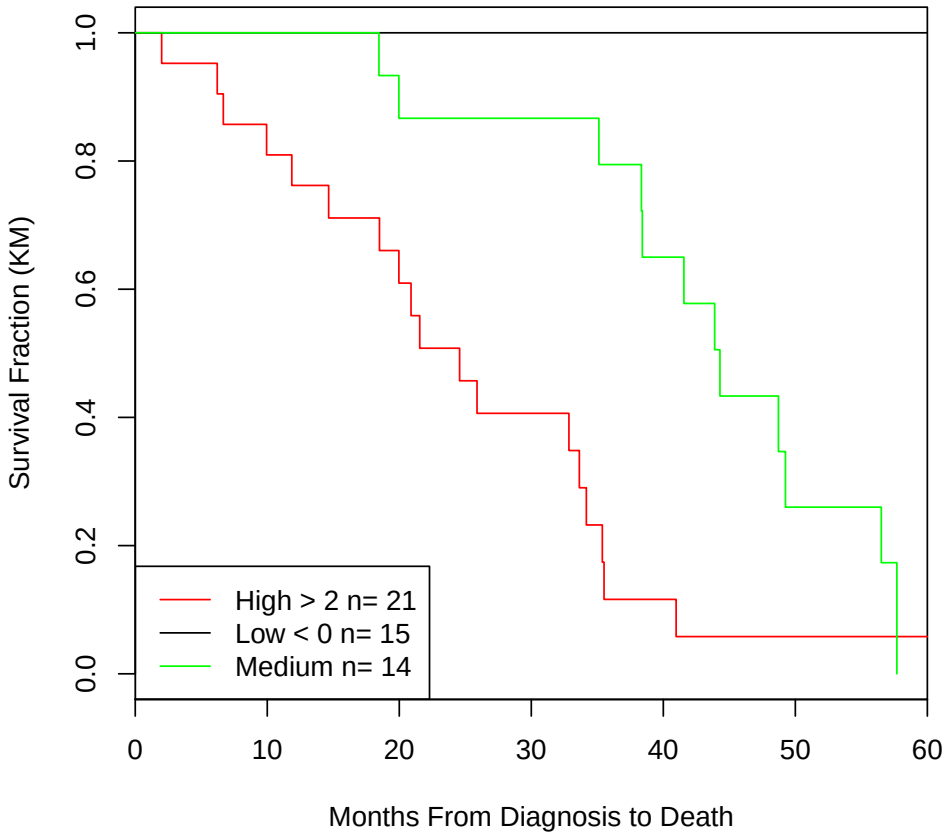

Figure 2: The observed survival curves for patient subgroups identified by the `PICS`

9 Survival ROC

The predictive performance of `PICS` method was further evaluated based on area under the time dependent receiver operating curve (ROC). '`survivalROC`' returns a ROC for survival plot.

```
R> survivalRoc(predicted)
```

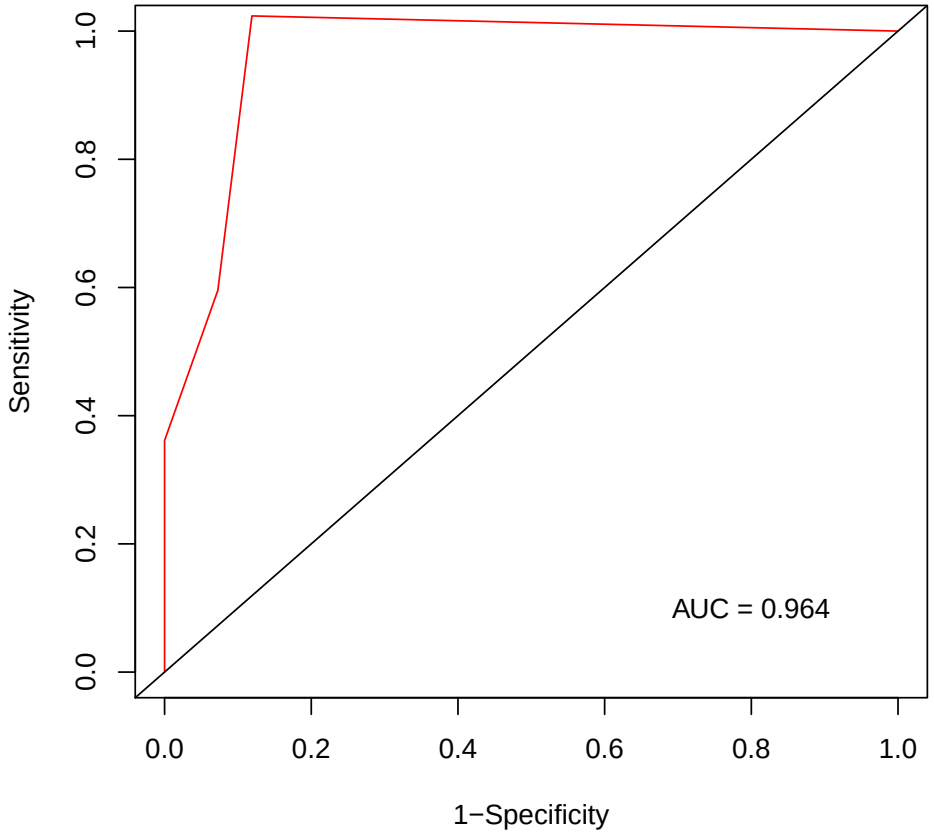

Figure 3: Time dependent Receiver Operating Curves

Figure 3 shows the ROC curves for survival and demonstrates an advantage of the `PICS` approachs. For the TCGA data, the area under curve (AUC) associated with the `PICS` approach was 0.964.

References

[1] Wei Wei, Zequn Sun, Willian da Silveira, Zhenning Yu, Andrew Lawson, Gary Hardiman, Linda Kelemen, Dongjun Chung (2017), "PICS: Pathway-guided identification of cancer subtypes". (submitted).

[2] Cancer Genome Atlas Research Network (2011), "Integrated genomic analyses of ovarian carcinoma", Nature, 474(7353):609-615.

[3] Chun, H. and Keleş, S. (2007) "Sparse partial least squares for simultaneous dimension reduction and variable selection", (`http://www.stat.wisc.edu/~keles/Papers/SPLS_Nov07.pdf`).

[4] Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso", Journal of the Royal Statistical Society (Series B), 58, 267-288.